Temporospatial Context-Aware Vehicular Crash Risk Prediction

by

Pouya Mehrannia

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Robert Dony Associate Professor, Dept. of Engineering, University of Guelph
Supervisor(s):	Otman Adam Al-Basir Professor, Dept. of ECE, University of Waterloo Behzad Moshiri Adjunct Professor, Dept. of ECE, University of Waterloo
Internal Member:	Kshirasagar Naik Professor, Dept. of ECE, University of Waterloo
Internal Member:	Mark Crowley Assistant Professor, Dept. of ECE, University of Waterloo
Internal-External Member:	Daniel Stashuk Professor, Dept. of Systems Design Eng., University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

With the demand for more vehicles increasing, road safety is becoming a growing concern. Traffic collisions take many lives and cost billions of dollars in losses. This explains the growing interest of governments, academic institutions and companies in road safety. The vastness and availability of road accident data has provided new opportunities for gaining a better understanding of accident risk factors and for developing more effective accident prediction and prevention regimes. Much of the empirical research on road safety and accident analysis utilizes statistical models which capture limited aspects of crashes. On the other hand, data mining has recently gained interest as a reliable approach for investigating road-accident data and for providing predictive insights.

While some risk factors contribute more frequently in the occurrence of a road accident, the importance of driver behavior, temporospatial factors, and real-time traffic dynamics have been underestimated. This study proposes a framework for predicting crash risk based on historical accident data. The proposed framework incorporates machine learning and data analytics techniques to identify driving patterns and other risk factors associated with potential vehicle crashes. These techniques include clustering, association rule mining, information fusion, and Bayesian networks.

Swarm intelligence based association rule mining is employed to uncover the underlying relationships and dependencies in collision databases. Data segmentation methods are employed to eliminate the effect of dependent variables. Extracted rules can be used along with real-time mobility to predict crashes and their severity in real-time. The national collision database of Canada (NCDB) is used in this research to generate association rules with crash risk oriented subsequents, and to compare the performance of the swarm intelligence based approach with that of other association rule miners.

Many industry-demanding datasets, including road-accident datasets, are deficient in descriptive factors. This is a significant barrier for uncovering meaningful risk factor relationships. To resolve this issue, this study proposes a knwoledgebase approximation framework to enhance the crash risk analysis by integrating pieces of evidence discovered from disparate datasets capturing different aspects of mobility. Dempster-Shafer theory is utilized as a key element of this knowledgebase approximation. This method can integrate association rules with acceptable accuracy under certain circumstances that are discussed in this thesis. The proposed framework is tested on the lymphography dataset and the road-accident database of the Great Britain.

The derived insights are then used as the basis for constructing a Bayesian network that can estimate crash likelihood and risk levels so as to warn drivers and prevent accidents

in real-time. This Bayesian network approach offers a way to implement a naturalistic driving analysis process for predicting traffic collision risk based on the findings from the data-driven model.

A traffic incident detection and localization method is also proposed as a component of the risk analysis model. Detecting and localizing traffic incidents enables timely response to accidents and facilitates effective and efficient traffic flow management. The results obtained from the experimental work conducted on this component is indicative of the capability of our Dempster-Shafer data-fusion-based incident detection method in overcoming the challenges arising from erroneous and noisy sensor readings.

Acknowledgements

The development of this thesis was an enjoyable and rewarding journey, but it would not have been possible without the support and encouragement of others. I would like to first express my sincerest gratitude to my advisors, Dr. Otman Basir and Dr. Behzad Moshiri, for their wisdom, guidance, support, and expertise. They provided me with an excellent environment, in which I was able to pursue a research area that interests me. I always received valuable guidance, endless encouragement, and constructive feedback from them. I am very grateful for everything that they have done.

I also would like to thank the members of my examining committee, Dr. Daniel Stashuk, Dr. Mark Crowley, and Dr. Sagar Naik for investing time in reviewing this work and for their valuable comments throughout this process to make the thesis come out in its current form. I also thank Dr. Bob Dony for his constructive feedback as my external examiner.

I have been fortunate to work with many wonderful people in the Pattern Analysis and Machine Intelligence (PAMI) lab, I am thankful to all of them for creating an enjoyable environment for creativity and innovation. Special thanks to my friends who made Waterloo a home for me. Thanks for the good times and the many laughs. I would like to thank the staff at the graduate office of the department of Electrical and Computer Engineering (ECE) who helped me a lot in processing administrative documents for my PhD program. They are treasures of the department.

I owe my warm thanks to my parents, Jalal and Mehri, for their patience and supports while I was immersed in this journey. They loved me unconditionally and inspired me from far away and I know how difficult this has been for them. I also would like to show my deepest appreciation to my brother and my sister, Nima and Nona, for always having my back and making me smile regardless how much I do not want to.

Last, but far from least; my better half, Sheyda, deserves special thanks for her support. Throughout my studies, she provided encouragement, sound advice, good criticism, good company, and lots of good ideas. Her patience, love, and encouragement throughout this journey was invaluable.

Dedication

To My parents, for their unconditional love and support

To my wife, for letting me on her magical mystery ride

To Canada, for being a home without discrimination

Table of Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Motivation

Traffic collisions are considered as the 9th leading cause of fatality and account for 2.2% of all deaths worldwide [1]. According to the World Health Organization (WHO), approximately 1.25 million people die annually in traffic collisions across the world. That means a vehicle crash takes one life every 25 seconds. Apart from the high number of fatalities, approximately 20 million people are injured or become disabled each year. In a similar report in 2004 by WHO, projections indicated that the death rate will rise by 65% over a 20-year period unless the commitment to prevention is increased. Even if the current rate remains constant, it still tears many families apart and imposes billions of dollars of financial loss on governments every year [2, 3]. By comparing the global leading causes of death in 2000 and 2012 in Figure 1.1 it is revealed that road injury is a growing cause of death. One of the main reasons for this growth is the increasing number of vehicles in use worldwide. The trend of worldwide vehicle registration in Figure 1.2 shows that this number had a 50% growth during the years from 2000 to 2012. To date, the number of vehicles in use has reached more than 1.2 billion and by extrapolation this number is likely to rise to 2 billion by the year 2035, illustrated in Figure 1.3.

Road accidents not only take lives but also damage vehicles and properties. Road-traffic injuries cost governments approximately 3% of their gross domestic product (GDP) [3]. Involved passengers can be severely injured, resulting in long-term recovery or permanent disability. Road accidents can also be life-changing for the families with members who sustain serious injuries. To that end, vehicle-crash prevention is an absolute necessity.

No	Causes of death, 2000	Deaths (million)	% of deaths	No	Causes of death, 2012	Deaths (million)	% of deaths
1	Ischaemic heart disease	6.0	11.3	1	Ischaemic heart disease	7.4	13.2
2	Stroke	5.7	10.7	2	Stroke	6.7	11.9
3	Lower respiratory infections	3.5	6.6	3	COPD	3.1	5.6
4	COPD	3.1	5.8	4	Lower respiratory infections	3.1	5.5
5	Diarrhoeal diseases	2.2	4.1	5	Trachea, bronchus, lung cancers	1.6	2.9
6	HIV/AIDS	1.7	3.2	6	HIV/AIDS	1.5	2.8
7	Tuberculosis	1.3	2.5	7	Diarrhoeal diseases	1.5	2.7
8	Prematurity	1.3	2.5	8	Diabetes mellitus	1.5	2.7
9	Trachea, bronchus, lung cancers	1.2	2.2	9	Road injury	1.3	2.3
10	Diabetes mellitus	1.0	2.0	10	Hypertensive heart disease	1.1	2.0
12	Road injury	1.0	1.9				

Figure 1.1: Comparison of global leading causes of death 2000 and 2012

The global average mortality rate caused by road accidents is 17.4 car-crash deaths per 100,000 people per year [3]. This number significantly varies across countries. In developing and underdeveloped countries the mortality rate is much higher than that of developed ones. In Canada, the number of fatalities per 100,000 population is 6, and Ontario has a rate of 3.52. Although statistics show that Canada is well-placed regarding the mortality rate, its rate is still high, so we should not leave the infrastructure unchanged. Considering the rapid growth of the population, the current crash-prevention technologies in transportation may not be able to cease or decrease the growth of the road-accident mortality rate in the future.

The aforementioned statistics indicate that road accidents are a growing concern around the world. Traffic-rule awareness, improved transportation systems, and effective precautionary systems are examples of approaches to prevent road accidents. These approaches have a considerable impact on traffic flow and safety. However, none of them is recognized as the predominant method for avoiding road accidents. One approach that can have a significant influence on traffic flow and safety is employing collision risk analysis models which are sometimes referred to as collision prediction models (CPMs). Risk analysis and prediction of vehicle accidents are becoming popular topics among researchers in road-safety analysis. Having the probability range or the risk level of potential accidents predicted, associated warnings and suggested precautionary actions can be made available to drivers. Therefore, drivers will have the chance to avoid the dangers or reduce the damage by responding quickly in advance.

Historical trend of worldwide vehicle registrations 1960-2012 (thousands)									
Type of vehicle	**1960**	**1970**	**1980**	**1990**	**2000**	**2005**	**2009**	**2010**	**2012**
Car registrations	98,305	193,479	320,390	444,900	548,558	617,914	684,570	723,567	773,323
Truck and bus registrations	28,583	52,899	90,592	138,082	203,272	245,798	295,115	309,395	341,235
World total	126,888	246,378	410,982	582,982	**751,830**	863,712	979,685	1,032,962	**1,114,558**

Figure 1.2: Growth in the number of vehicles in use worldwide

CPMs not only reduce road accidents but also are considered as an approach for safety-performance estimation. The prevailing approach to estimate safety performance is using collision rates. However, the non-linear relationship between collision frequency and the amount of traffic on a road segment makes collision rates inappropriate representatives of safety. Hence, CPMs are replacing collision rates as the primary tool for estimating road safety. These models also overcome the limitations of traditional road-safety measurement by facilitating an accurate and consistent quantification of safety performance [4].

1.2 Problem Statement

This thesis is concerned with the design of a crash prediction model that generates the likelihood of a particular vehicle crashing as a function of available crash contributing factors including the driver behavior, environmental considerations, location, and time. Three specific issues emerge by stating the problem with these elements: (1) the probabilistic nature of the model's output as a result of its individual-based prediction rather than a population-based one, (2) the awareness and adaptability of the model with regards to crash-associated factors, and (3) the model's ability to cope with temporal and spatial variations. In this section, a thorough description of the aforementioned issues is reported.

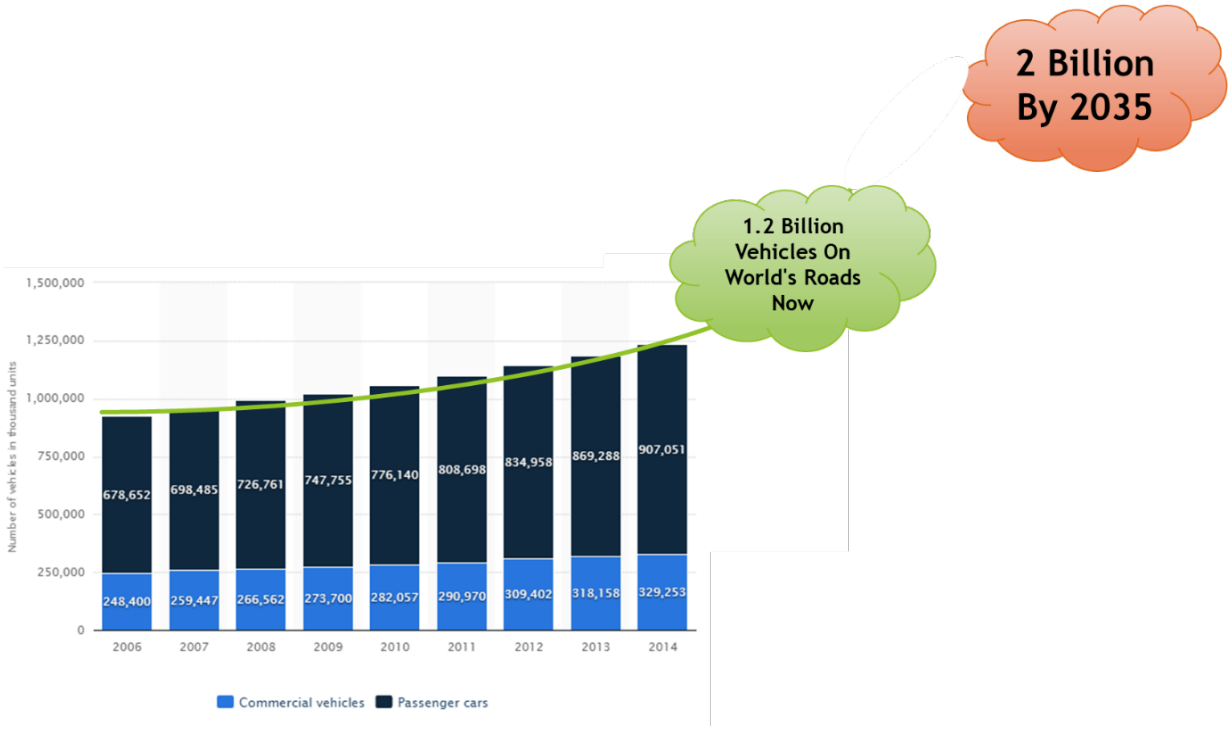

Figure 1.3: Number of vehicles in use worldwide

A motor-vehicle crash is defined as an accident resulted from the collision of two or more vehicles, the collision of a vehicle with a pedestrian or other moving or stationary obstructions, or a single vehicle run off the road. Motor-vehicle crashes may have inevitable consequences like injury, death, and property damage. There are a burgeoning number of factors contributing to motor-vehicle crashes. However, they all belong to 3 main categories, here referred to as the triangle of crash-contributing factors (Figure 1.4):

- Human factors
- Environmental factors
- Vehicle factors

Human factors are recognized to have the largest influence on the occurrence of motor-vehicle crashes. The age, experience, and skills of the road users, their attention and fatigue level, and use of intoxicating substances or cellphones are some examples of human factors

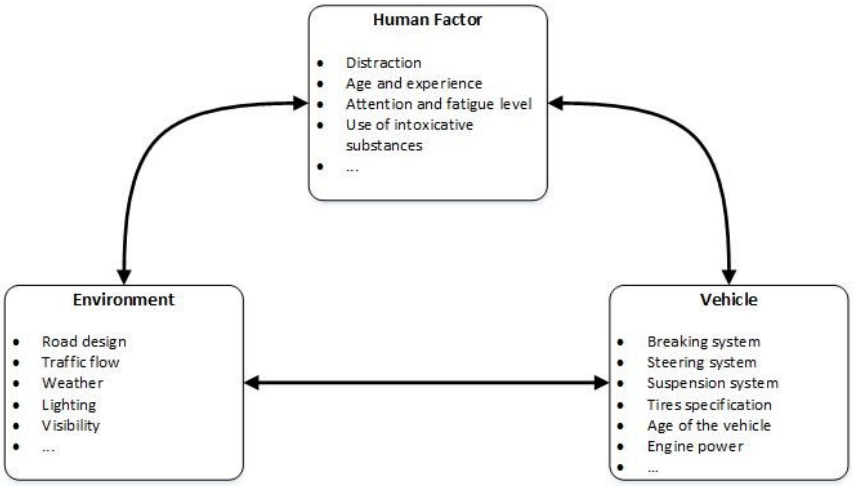

Figure 1.4: The triangle of road-accident contributing factors

[5, 6, 7, 8]. In spite of human factors' importance, not many research works investigate them for the road-accident prediction. This lack of consideration arises from the complexities that the existence of human factors brings to the model. The contributions of human factors are too complex to be discovered and measured by simple models. One of the goals of this study is to contribute to a deeper comprehension of the role of human aspects involved in the field of accident causation analysis.

The second group of factors responsible for crashes is environmental factors. This group includes the factors related to road design (e.g., the number and width of the driving lanes, the radius of curves, the speed limit, or the pavement quality), traffic flow (e.g., the daily traffic volume), weather, and lighting conditions of the roads [9, 10, 11, 12, 13, 14, 15]. Understanding the relationship between these factors and motor-vehicle crashes not only unveils their contribution in the occurrence probability of accidents but also helps in implementation of countermeasures to minimize the impact of particular factors in accidents. In other words, some accidents would be avoided if the design of a particular road was different, and therefore, identifying those aspects of the road design that have considerable potential in preventing future accidents is crucial.

The occurrence of motor-vehicle crashes can also be associated with vehicle factors. These factors include the braking, steering, and suspension systems, tire specifications, age of the vehicle, engine power and other vehicle specifications abetting in technical malfunctions or failure of the driver in avoiding collisions [17, 18]. Firgure 1.5 shows the number of fatal, serious and slight road accidents that were caused by vehicle defect fac-

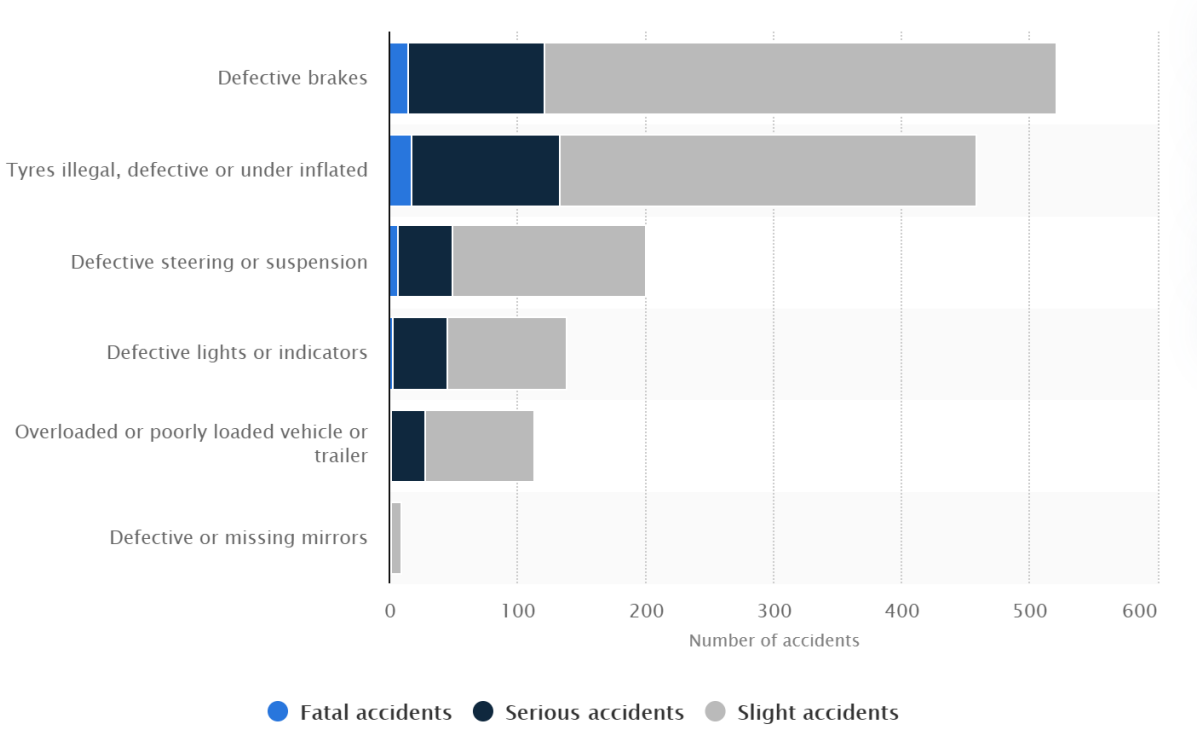

Figure 1.5: Vehicle defect factors [16]

tors in Great Britain (UK) in 2018. According to this figure, defective breaks, tires, and steering/suspension are the top three vehicle factors causing road accidents and are the most fatal ones. Vehicle factors are partly recorded in many crash datasets. Their importance may not be as human factor's (considering the number of crashes they cause), yet they are major causes of fatal accidents. Their connections to the type of crashes and their severity are still a problem of interest globally.

A crash risk prediction system is expected to identify hazardous situations. The main goal of such a system is to reduce the number of car accidents and improve a driver's decisions. One important feature of a CPM or a crash risk estimator that improves its ability to identify hazardous situations is context awareness. Context-aware systems are those with the ability to sense, reason, and react upon the current contextual information. Driver assistance systems can use the information coming from context-aware CPMs to link drivers with the physical environment surrounding them.

The total number of the variables contributing to car accidents is unknown, but numer-

ous variables are known to have correlation with the risk of car accidents. The dependencies between these variables cause road accident data to be heterogeneous. The heterogeneous nature of road accident data makes the crash risk analysis task difficult. As the number of variables increases, the number of inter-dependencies between them grows exponentially and overcoming the heterogeneity becomes more challenging.

Road-accident datastes contain a large amount of unprocessed information about the contribution and significance of crash-associated factors. An individual-based approach focuses on the current contextual information from the environment, vehicle, and driver and compares them against the historical data to obtain a crash risk that is associated to the vehicle of interest. On the contrary, a population-based prediction relies on the frequency of crashes happening in a predefined scenario. As a result, driver assistance systems developed based upon population-based approaches fail to provide crash risk estimations that are specific to the unique features of the vehicle and driver. Therefore, individual-based prediction is preferred and promises a strong foundation for developing driver assistance systems.

CPMs should also account for the temporal and spatial correlations in the road-accident databases. The non-random distribution of road-accidents in time and space demonstrates the importance of discovering spatial and temporal patterns in traffic accidents. Spatial and temporal analysis helps to identify these patterns and their characteristics. Although time and space variations have been recognized to be influential in the occurrence of crashes, limited research has been conducted to investigate the interaction between the location of crashes and the time they occur. In this thesis, time trends and demographic effects are investigated in addition to other dependencies.

1.3 Objectives

This work aspires to design a new preventive approach by which the number of road accidents and casualties dramatically decreases. The idea here is to find the probability range of crashes in advance by considering the data collected from the vehicle, the driver behavior, the road geometry and location, the weather, and other historical data on road accident track records. The proposed model is expected to make early prediction of potential road crashes possible by incorporating data analytics and machine learning algorithms and identifying crash-associated patterns. This model can be used as a basis for developing a pre-crash warning system that allows the drivers to be aware of hazards ahead of time. The resultant product of this project saves lives by alerting drivers ahead of time, enhances

the quality of driving experience, makes roads a safer place, and sets the foundation for development of intelligent safety-oriented self-driving cars.

The scope of this study is divided into four parts:

1. **Accident data clustering and insight induction:** Developing a tool by which road accident data repositories can be fully explored. The goal is to extract insights about the contribution of certain combinations of collision-associated factors in car crashes. Each combination results in a different crash configuration or severity level. This tool should serve as the core of a non-real-time collision prediction model. The model maps the risk factor combinations to certain probability ranges and risk levels. Fulfilling this objective requires identification of the frequent risk factor combinations in a crash dataset and estimating their severity levels.

2. **Knowledgebase approximation:** aggregation of insights from disparate datasets. Each country, region, or district owns a set of collision datasets containing elements that are selected for a specific purpose. Additionally, no single dataset presents a complete set of crash-contributing factors. Even if it does, investigating huge amount of data using the existing insight extraction methods would impose a high computational complexity. In order to equally investigate all the crash-contributing factors, an insight aggregation framework is needed. The ideal framework enables aggregation of insights from multiple datasets without considerable increase in the computational time.

3. **Context-aware collision analysis:** a naturalistic driving analysis that enables the model to process driving patterns in real-time. Context-aware colision prediction model examines the three main sources of contributing factors (the driver, vehicle, and environment) in different variations of time and location. It is desired that all the insights are aggregated before starting the context-aware collision analysis. Contextual information helps in recognition and classification of collision patterns and in adapting model's performance. Adaptive model allows the system to work under various circumstances. Finally, the output of the context-aware CPM can assist the driver in augmenting the probability of undertaking safe behavior.

4. **Traffic incident detection and localization:** a subsidiary system to identify and localize incidents using sensor networks. While the in-depth accident analysis and naturalistic driving studies set the foundation for the context-aware prediction model, there are plenty of subsidiaries that can be added to improve this compound. Traffic incident detection and localization system is one example of these subsidiaries.

Detecting and localizing traffic incidents enables timely response to accidents and facilitates effective and efficient traffic flow management. One of the frequent types of traffic collisions are head on accidents to the vehicles stalled in highways after occurrence of an incident. The automatic incident detection system can inform the drivers about the congestion in their path and the associated collision risks.

The outcome of this research, having achieved the above mentioned objectives, helps drivers take precautionary actions by recognizing and reducing the influence of crash-associated factors. Also, car insurance companies can benefit from the non-real-time model by using it as a basis for calculating the car insurance premiums.

1.4 Organization

This thesis is organized as follows: Chapter 2 provides a comprehensive background to crash prediction models from three perspectives of regression and statistical modeling, Artificial intelligence, and probabilistic modeling. The proposed crash risk analysis framework is formulated in Chapter 3, and an overview of the structure is presented. Chapter 4 proposes the accident data clustering and insight induction describing its design details and discussing the association rules derived from the National Collision Database of Canada. A knowledgebase aggregation and approximation framework using the disjunctive pooling rule variation of the Dempster-Shafer theory is proposed in Chapter 5. In Chapter 6, a context-aware collision analysis system is designed using a Bayesian network that learns the structure based on approximated knowledgebases. Chapter 7 introduces traffic incident detection and localization, as a component of the crash risk analysis framework, with the ability to automatically detect and localize incidents that happen on highways. Finally, conclusions and future work are presented in Chapter 8. The conclusion sections of chapters 4 to 7 provide a short summary of the achievements and the limitations pertinent to each of these chapters.

Chapter 2

Background and Literature Review

The statistical methodologies and their variants in univariate and multivariate regression frameworks have been successfully applied to the crash count estimation problem. These methods have enhanced our perception of the relationships between many risk factors and accident outcomes. AI-based and probabilistic techniques also capture these relationships from a different perspective and to a different extent. Exploring the literature of the crash prediction methods reveals how fast these methods are evolving and how promising they are in identifying the influence factors. At the same time, it shows there are gaps remaining in this field to accomplish a desired crash prediction model that can be integrated into a vehicle's alerting system.

This chapter reviews the background, the literature, and the state-of-the-art techniques for developing crash-prediction models. The techniques can be classified into three categories: Regression-based techniques, AI-based techniques, and Probabilistic techniques. Some examples of these methods are illustrated in Figure 2.1. The sections of this chapter introduce these categories along with notable approaches in them. The advantages and shortcomings of the methods as well as the gaps in the literature are also presented.

2.1 Regression-based CPMs

Regression-based collision-prediction models provide an estimation of the occurrence frequency of road-accidents. The developers of such models assume that crash frequency is an appropriate dependent variable for predicting road accidents. The estimation that such models produce is for a specific location and is highly correlated to the characteristics of

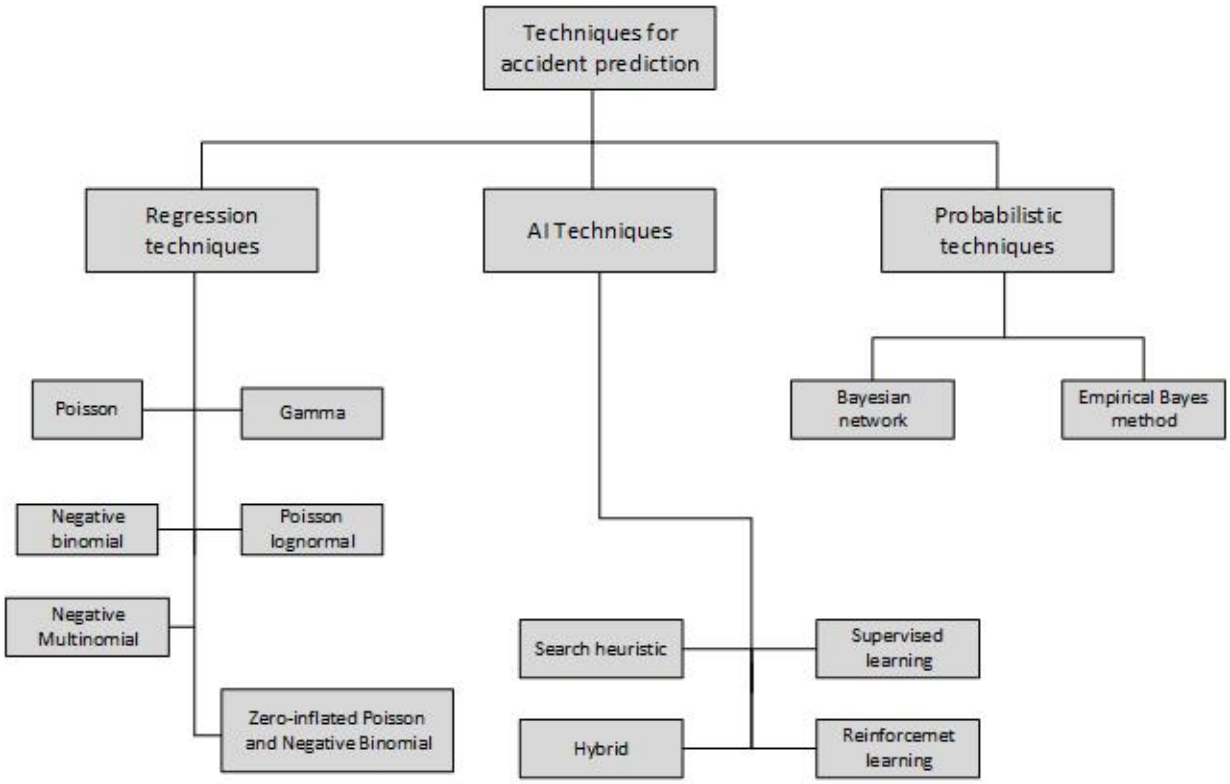

Figure 2.1: Crash prediction models

the site. Consequently, such models generally focus only on one of the risk factor classes (i.e., the environment) involved in accidents. Although the models in this domain have some issues, especially the lack of generalizability, they form a strong basis for integrating environmental variables and for finding the correlation of certain variables to crash likelihood.

Several regression-based modelling techniques have been used for crash prediction including multiple linear regression, Poisson distribution, negative binomial, random effect technique, and multiple logistic regression models. The multiple linear regressions modeling technique has been used in several fields to model the relationship between different explanatory variables and an outcome variable. This technique fits a linear equation to observed data to describe the relationship between these variables. Although multiple linear regression models are proposed and used widely in road accident studies, they have limitations to describe the random, non-negative, discrete, and typically sporadic events, which are all characteristics of road crashes.

One of the first regression-based CPMs developed for multi-lane roads was proposed by Persuad et al. [19] in 1993. From numerous independent variables which affect crash frequency, they chose average daily traffic (ADT) and hourly volume (VH) to express the relationship between crash data and traffic flow. Generalized linear models are the basis for their examination of this relationship. They show that the increase in traffic flow will result in an increase in crash rate. Although both ADT and VH can express the traffic flow, the authors believe the latter reflects a more appropriate expression. They draw this conclusion by observing that the accident risk on four-lane freeways is lower than on freeways with more than four lanes in the same traffic volume due to the greater freedom for drivers to maneuver and that VH as opposed to ADT takes the congestion conditions into account. Nevertheless, ADT is used more often in CPMs due to the difficulty in accurately measuring VH. In the same year, Knuiman et al. [20] employed negative binomial distribution to study the relationship between median width of four-lane roads and crash frequency. The results reveal that the crash rate decreases as the median width increases. Cross-over accidents, which involve a car traversing the median barrier on highways, are the main type of car crashes that wider medians can help to prevent.

The random and discrete nature of the crash events has made many researchers argue that Poisson regression models are more appropriate than multiple linear regression models. Greibe [21], for example, developed a Poisson distribution model for urbun intersections and road segments in Denmark and found significant explanatory variables for estimating the number of crashes in road segments. These variables include speed limit, road environment, parking facilities, number of minor side roads and number of exits per kilometer. For intersections, the most significant variables were found to be those describing the traffic flow. Abdel-aty et al.[22] introduced some limitations for Poisson model. One of these limitations is that the mean must equal the variance of road crash number (dependent variable). In most crash data, the variance value of the road crash number exceeds the mean value and, in such case, the data would be overdispersed.

The regression-based CPMs were conventionally developed by employing the Poisson-gamma hierarchy [23]. This would result in a negative binomial regression model as the tool for road safety performance estimation [24, 25, 26, 27, 28]. Negative Binomial (NB) regression method is considered as an alternative for the Poisson regression model since it does not require the equal mean and variance assumption. For some special purposes the Poisson distribution is still a fairly good representation, but more recent research suggests that a combination of models provide a better understanding of how safety is affected by explanatory variables.

Poisson-lognormal (PLN), Multivariate Poisson (MVP) and multivariate Poisson-lognormal (MVPLN) regression are also frequently used in literature to devise models for crash fre-

quency estimation. Such models perform better than those based on Gamma distribution in existence of outliers [29]. The MVPLN regression is adopted to model crash rate at different levels of severity [30], and is reported to be more appropriate than MVP approach for the analysis of multivariate collision count data [31].

Fridstrm et al. [32] also adopted negative binomial regression to examine the effect of traffic flow, speed limits, weather and lightening conditions on the crash rate. Their work also contributed by applying goodness-of-fit measurement. By considering Poisson regression besides negative binomial regression, Hadi et al. [33] related total crash rate and injury crash rate to annual average daily traffic (AADT) and environmental characteristics to develop a CPM for multilane and two-lane roads. Their findings show that crash rate increases by increasing AADT on the roads with high capacity. However, on the roads with low capacity, crash rate decreases as higher traffic flow in such roads will restrict freedom for drivers' maneuvers.

One of the earliest studies in which curves and tangents were separately analyzed was performed by Persaud et al. [34]. The crash frequency for two lane roads has been estimated using traffic flow and road geometry. Generalized linear modeling was employed for regression model calibration. An increase in AADT and Section length (L) were found to cause crash frequency to rise for both curves and tangents. Curvature (1/R) was found to have impact on crash frequency in tangents.

Some limitations are associated with traditional statistical regression models that has encouraged researchers to propose non-parametric methods and artificial intelligence (AI) models for predicting crashes. Recent studies on historical traffic data indicates that the applied statistical modeling fails when dealing with complex and highly nonlinear data [35], which could suggest that the relationship between the influence factors and traffic crash outcomes is more complicated than can be captured by a single statistical approach.

Regression-based models require assumptions about the distribution of data. Furthermore, they need a well-defined functional form, a linear functional form for example, between dependent variable and independent variables. These basic assumptions of the traditional statistical regression models play an important role in their accuracy and the quality of the outcome. If these assumptions are violated, it will result to inefficient estimations and incorrect inferences [27, 36, 37].

Table 2.1 summarises the advantages and disadvantages of different regression-based CPMS obtained from a few more detailed literature [38]. Based on the information in this table, the developer of the CPM model should consider many aspects of the available crash data and the form of the model's output to make a decision about the model type. If none of the regression-based models can address the requirements of the CPM, the AI-based

Table 2.1: Comparison of regression models for analysing crash-frequency [38]

Model Type	**Advantages**	**Disadvantages**
Multiple Logistic	Suitable to study the effect of one variable while controlling for other variables	used to analyze only crash binary outcomes
Multiple Linear	Easy to estimate crash number	Unable to describe adequately the random, non-negative, discrete, and typically sporadic events
Random Effects	Handle spatial correlation	The results from this technique may not be transferable to other data sets because the results are observation specific
Poisson	Handle with unavoidable discrete and more likely random events	Cannot handle over- and under dispersion(the mean must equal the variance of crash number)
Negative Binomial	Does not require the equal mean and variance assumption, able to describe adequately the random, non-negative, discrete, and typically sporadic events	Cannot handle with small sample sizes

and probabilistic models are other options to be considered.

2.2 AI-based CPMs

Artificial intelligence (AI) is widely used in real-life applications these days. The need for AI is amplified where the outcomes and the data change all the time. In this section, AI techniques are surveyed for road-accident prediction.

Previous studies show that neural network is among the most popular AI-based techniques used for crash prediction. Artificial neural networks (ANNs) are non-linear statistical data modeling tools capable of finding complex relationships between inputs and outputs in a system or patterns in data. In terms of performance, neural network analysis is equal to non-linear regression analysis. Consequently, neural network is an alternative of regression analysis for non-linear engineering problems. One advantage of ANN over regression analysis is that it does not need a pre-selected model to fit the data and sufficient hidden nodes guarantee the required accuracy. Neural network is known to exhibit a more realistic and accurate prediction [39]. its applications in transportation engineering can be found in the studies from early 1990s for traffic management and transportation engineering [40, 41].

ANN-based methods can be used for the prediction of accidents and their severity. A combination of decision tree with ANN using back-propagation was presented by chong et al. to train the network [42]. In this study, driver's seat belt usage, light condition of the roadway, and driver's alcohol usage were recognized as the most critical features in fatal injuries. Tambouratzis et al. also combined probabilistic neural networks and decision trees for the prediction of accident severity (light, serious, or dead). The implications in this study show that this combination enhances classification accuracy [43]. In both studies, road-accident historical data were used for developing a crash prediction model, and therefore, these models suffer from the inaccuracy attributed to the population-based nature of the prediction; the data-driven model will not be able to translate data gathered in real-time to a realistic crash probability.

Many other crash prediction models are developed based on AI techniques using a combination of neural networks, support vector machine, and decision tree [44, 45, 46, 47, 48]. These models, which are all created using observational data, reveal that traffic flow, road section length, infrastructure geometric characteristics, pavement surface conditions, lighting, weather conditions and driver behavior contribute to the occurrence of accidents. By considering these contributing factors in crash prediction models, the occurrence and severity of accidents can be reduced.

Evolutionary algorithms are also among successful AI techniques used for developing crash prediction models. Genetic algorithm (GA) and genetic programming (GP) are two popular examples of Evolutionary algorithms used in this context. GA is a heuristic based evolutionary search approach inspired by the process of natural selection [49]. GA applies genetic operations such as mutation, crossover, and selection to generate high-quality solutions to optimization and search problems. It starts with randomly generating a pool of candidates known as initial population in compliance with the rules of the problem domain. The next generation is a set of promoted candidates selected based on a defined fitness function. The fitness function ensures that well-fitted candidates have more chances to be selected for the reproduction process. In the reproduction process, crossover and mutation operations are used to alter the properties of the candidate solutions.

As opposed to GA, the candidate solutions in GP have a tree-like structure rather than a one dimensional array. The candidate solutions in GP, in fact, are computer programs encoded as a set of genes that the algorithm breeds them to find the best solution for the problem. Similar to GA, a fitness function is used for the selection process, and mutation and crossover are used as the reproduction operators. Compared to ANN, GP has a major advantage which is the transparency in the model; it makes the best approximation of the objective, found in the search space, visible by removing the black box effect. In addition, previous studies have shown that the models developed based on GP outperform the ones

developed based on traditional modeling techniques with regard to prediction performance [50, 51, 52].

A GP-based real-time crash-prediction model has been proposed by Xu et al. using traffic and weather situations in crash data as input [53]. The study has been performed for freeways and separate models developed in congested and uncongested traffic situations. The authors selected the most influential contributing factors in crashes by employing random forest technique. They used ramped half-and-half method [49] for initialization of the candidate solutions and setting the tree depth limit to six levels. The maximum tree depth were increased to 30 levels after the initialization phase. Xu et al. observed that traffic flow characteristics contribute in crash risk quite differently in congested and uncongested traffic conditions.

There are other works targeting accident prediction using evolutionary algorithms, but not all of them directly use GAs to solve the problem. Dixon et al. [54] used GA to reduce the distance between converged samples and ground truth labels for prediction. Damousis et al. used fuzzy expert system while GA is used to train the parameters with the purpose of increasing accuracy. Yang [55] also employed GA to train the parameters of support vector regression to predict highway traffic accidents. In these works, GA were utilized to train another learning method to enhance the performance of prediction.

Some of the most recent works in risk and severity prediction use Neural networks (NN). Zeng et al. [56], for example, proposed a traffic accident's severity prediction model using convolutional neural network for traffic accident's severity prediction. Another example by Yuan et al. [57] uses a deep learning approach on heterogeneous spatio-temporal data to predict traffic accidents. Deep learning has attracted a great attention from researchers in the past decade, but has been mostly used in the fields of text, image, and voice recognition. Other state-of-the-art examples of using deep learning for traffic accident risk and severity prediction can be found in [58, 59, 60], and [61].

2.3 Probabilistic CPMs

Human beings have the ability to estimate the threat level of planned actions in traffic. While driving, they evaluate different maneuvers like overtaking, lane changing, or intersection crossing according to a ratio of risk and time efficiency. When a certain maneuver starts, its consequences affect the future development of traffic situation. Probabilistic CPMs consider uncertainties originating from the possible behaviors of other traffic participants. Other uncertainties such as measurement inaccuracies, interaction of participants,

and limitation of driving maneuvers due to road geometry can also be considered in probabilitic CPMs. However, they may bring too much complexity depending on the method used in the model [62].

The outcome of probabilitic CPMs represents the probability of a crash for a specific trajectory based on predictions for other participants' behaviors and interactions. These models can be predictive or nonpredictive. The predictive approach works based on predicting a behavior and estimating the risk of crash in that situation. In contrast, nonpredictive methods rely on historical crash records and the evaluation of traffic arrangements that have resulted in dangerous situations.

An efficient and widely implemented class of methods for predicting traffic situations is simulation of single behaviors of traffic participants [63, 64]. These methods generate useful measures like time to collision or predicted minimum distance, but they do not consider actions of other traffic participants and measurements' uncertainties. For this reason, these methods may suffer from unsatisfying collision predictions [65]. A good and more sophisticated alternative is to use Monte Carlo methods that consider multiple simulations of other vehicles based on different initial states and changes in steering angle and acceleration. Examples of these methods can be found in [66, 67], and [68].

Another method for probabilistic investigation of crashes is stochastic reachability analysis. This method is an evolved version of reachability analysis by using stochastic information. Reachable sets contain all the possible future states of a system's trajectory for a given set of initial states and disturbance values. In this method, a particular path of a certain vehicle is evaluated as unsafe when the reachable sets of other vehicles cover all the possible positions that vehicle could move to [69, 70]. When the reachable sets are enhanced by stochastic information, the probability by which a vehicle's path may result in a crash is also reported. Stochastic reachable sets have also been studied in the fields of air traffic safety [71] and fault diagnosis [72, 73].

Bayesian inference can also be used as a probabilistic model as a whole, or can be combined with regression-based or AI-based methods to make the crash analysis probabilistic. Full Bayesian inference has been used widely for crash prediction; see, e.g., [74]. Five years of crash data is used in this study for aggregate investigation, and one year crash data along with real-time traffic and weather data were utilized for disaggregate models. Aggregate analysis uses historical data only and reveals contributing factors for each crash type. In contrast, disaggregate studies use surveillance systems and measurement devices to benefit from detailed traffic and weather data.

For a more realistic real-time crash prediction, CPMs have to consider the interactions between vehicles in addition to the measured data and information from surveillance sys-

Table 2.2: Descriptive comparison of regression-based, AI-based, and probabilistic CPMs

	Regression-based CPMs	**AI-based CPMs**	**Probabilistic CPMs**
Prediction type	Predicts the collision frequency and not the collision likelihood	Can predict the collision likelihood but most of the existing methods target collision frequency for simplicity	Can predict the collision likelihood but most of the existing methods target collision frequency for simplicity
Spatial considerations	More effective if applied to a specific location Highly correlated to the characteristics of the site	Can be Applied to any location depending on the learning method Is able to consider environment along with the driver and vehicle variables	Can be Applied to any location depending on the learning method Is able to consider environment along with the driver and vehicle variables
Temporal considerations	Needs the history of collisions in a period of time	Needs an information repository (collisions or conflicts) to learn the parameters	An information repository is not needed
constraints	The number of involved parameters should be limited	No restriction on the quantity of parameters Increasing the number of parameters arises the interdependency issue	No restriction on the quantity of parameters Increasing the number of parameters arises the interdependency issue

tems. The big issue for analysing traffic participants' interactions is the computational complexity. Even by taking only discrete actions (e.g., lane change) at discrete points of time into account, a fast-growing tree of possible situations is generated. The computational complexity in that case is in the order of $O(\mu^{\rho \cdot v})$, where μ is the number of possible actions, ρ is the number of time steps of the prediction, and v is the number of traffic participants [78].

Tables 2.2 and 2.3 summarize the descriptive and analytical comparison of regression-based, AI-based, and probabilistic CPMs. In general, AI-based and probabilistic CPMs are shown to outperform the regression-based CPMs. The reason lies in regression-based CPMs' need for a pre-selected model which is not guaranteed to be the best fit for the data in use. As opposed to the regression-based CPMs, AI-based and probabilistic CPMs are not limited to a specified model and hence can cope with a wider range of datasets and are more robust to noise and outliers.

Statistical models are not the best option for dealing with complex and highly nonlinear data. As a matter of fact, the relationship between the risk factors and crash outcomes is more complicated than can be captured by a single statistical approach. Statistical methods

Table 2.3: Analytical comparison of regression-based, AI-based, and probabilistic CPMs

	Regression-based CPMs	**AI-based CPMs**	**Probabilistic CPMs**
Xie et al. (2007) [75] Method: Bayesian neural network and negative binomial regression model Evalution: MAD [1] and MSPE [2]	MAD: 1.86 MSPE: 6.53	MAD: 1.42 MSPE: 2.48	
Xu et al. (2012) [53] Comparision: GP and binary logit models Evaluation: ROC curve [3]	Uncongested=59.24 Congested=55.58	Uncongested=67.42 Congested=60.5	
Yu et al. (2013) [74] Method: multi-level Bayesian analysis Evaluation: AUC			Maximum achieved: 0.78
Das et al. (2015) [76] Method: logistic regression Evaluation AUC	Maximum achieved: 0.7622		
Deublein et al. (2014) [77] Comparision: Regression(Reg), Emperical Bayes(EB), and Bayesian Network(BN) Evaluation: coefficients of correlation [4]	Reg: 0.688		EB: 0.711 BN: 0.731

mostly presume some strong assumptions, and hence, are not suitable for generalization. Multi-collinearity, which is the high degree of correlation between two or more independent variables, is another problem with statistical models in crash prediction. They also have poor performance when dealing with outliers and missing or noisy data [79]. This thesis presents an AI-based CPM combined with a Bayesian network configuration that mitigates many of the issues and the gaps addressed in this chapter. Next chapter formulates the problem of temporospatial context-aware vehicular crash risk prediction and shows the steps that should be taken to tackle this problem.

[1] Mean Absolute Deviance
[2] Mean Squared Predictive Error
[3] The reported values here are average of sensitivity values
[4] The reported values here are r-values.

Chapter 3

Road-accident Risk Analysis System: Machine Learning Approach

The purpose of this chapter is to introduce and formulate the problem of temporospatial context-aware vehicular crash risk prediction. An overview of the proposed crash risk prediction system is presented which introduces the overall system structure followed by the problem breakdown. The problem break down section explains how the different parts of the proposed framework address the objectives of this study. A brief introduction to the main operations and the relationships between them can be found in this chapter. Further details about the functionality of subsystems and their operations are discussed in chapters 4 to 7.

3.1 Problem Formulation

The problem being addressed in this thesis is the individual-based analysis of crash risk factors by considering information from the driver, the vehicle, and the environment in different variations of time and location. The aim is to propose a structure that is able to map the available information from these sources to a collision risk level. The risk level, as the output of this structure, can be used to inform drivers of their current behavior's collision risk influenced by other factors. A particular driving behavior may not be presumed dangerous in all circumstances; It is the presence of certain environmental and vehicle-related features in a certain situation, time, and location that characterizes some driving behaviors as not safe. The large number of influence factors and the multi-collinearity

of the problem demand for a collision risk assessment model to process the relationships between these influence factors.

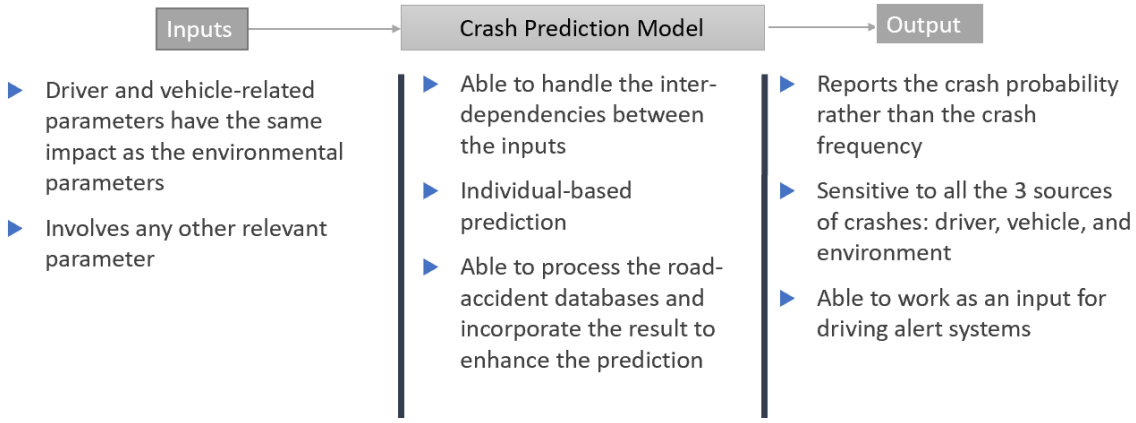

Figure 3.1: Problem definition

Figure 3.1 summarizes the problem definition by assuming the mapping model as a black box named crash prediction model (CPM). This model takes parameters from driver, vehicle, and environment as inputs and processes them with equal importance until certain patterns emerge that contribute to occurrence of accidents. In other words, this model does not focus on the influence of certain parameters individually, as opposed to the majority of works in the literature assigning weights to the parameters based on the degree of contribution they estimated for them. The way that this problem is formulated allows relevant inputs to be fed into the CPM, however, the degree of contribution of any input is not determined by a weight, but by their frequency of presence in collision patterns resulting into particular types of collisions.

This model is also formulated to handle the interdependencies between inputs. Different combinations of factors are assumed to be mapped into different risk levels. Although the proposed CPM is aimed to function in an individual-based manner, it does not ignore connections between collision-related factors. Moreover, the model should be able to process road-accident databases and incorporate the result to enhance the risk level prediction. Accident datasets are now populated with decades of accident data and are rich with patterns resulting in vehicle collisions. That is why part of the problem definition is dedicated to the model's ability to extract meaningful insights from data repositories and merging the information from various sources.

Another considerable gap in the literature is the way CPM outputs are reported. Most

of the CPM models predict crash frequencies which are not indicative of how dangerous driver behaviors are in certain circumstances. The output of the proposed model is aimed to report crash risk level rather than the crash frequency. It will be sensitive to all the 3 sources of crashes: driver, vehicle, and environment. The output has the potential to serve as an input to driver alert systems and as a source of information for insurance companies to compute premiums. The output is also expected to contribute to the safety requirements of autonomous vehicles.

3.2 Problem Breakdown Structure

Exploring historical collision data and in-depth accident analysis are proper first steps towards traffic collision analysis. CPMs that are developed based on historical data basically link a safety measure (e.g., collision risk) to a set of crash-associated variables. In contrast, in-depth accident analysis investigates contributing collision factors by detail reconstitutions of accidents with the aim of providing information on the chain of events that led to the collision [80]. Both methods share some shortcomings like providing limited amounts of data or requiring updated information repositories for a realistic prediction [81]. However, using them as a first step reveals interesting insights and general rules about contribution of crash-associated factors which can be helpful for triggering probabilistic models towards a low-cost real-time accident prediction.

Another approach in analyzing road-accidents is called naturalistic driving analysis. Naturalistic driving analysis is concerned with continuous collection of data from the sources that contain contributing factors. As mentioned in Chapter 1, these sources are mainly the road user's behavior, the vehicle, and the environment. In-depth accident analysis or naturalistic driving analysis cannot offer an accurate and robust crash prediction structure when used alone. However, a model based on the combination of both is likely to mitigate the shortcomings that each has individually. This procedure is illustrated in Figure 3.2 which shows how each method performs and how their combination can elevate the accuracy and robustness of the model.

Figure 3.3 is an overview of the problem formulation. This figure shows that the in-depth accident analysis and naturalistic driving analysis can work in harmony to perform collision risk analysis. This harmony takes advantage of prior collisions and the real-time driving conditions. It starts with extracting insights from a single collision dataset. Insights act as distilled information extracted from huge collision datasets. They are easier to store, simpler to access, and suitable to update. As shown in Figure 3.4, Chapter 4 of this thesis addresses the insight induction process from a single dataset. It is emphasized in the same

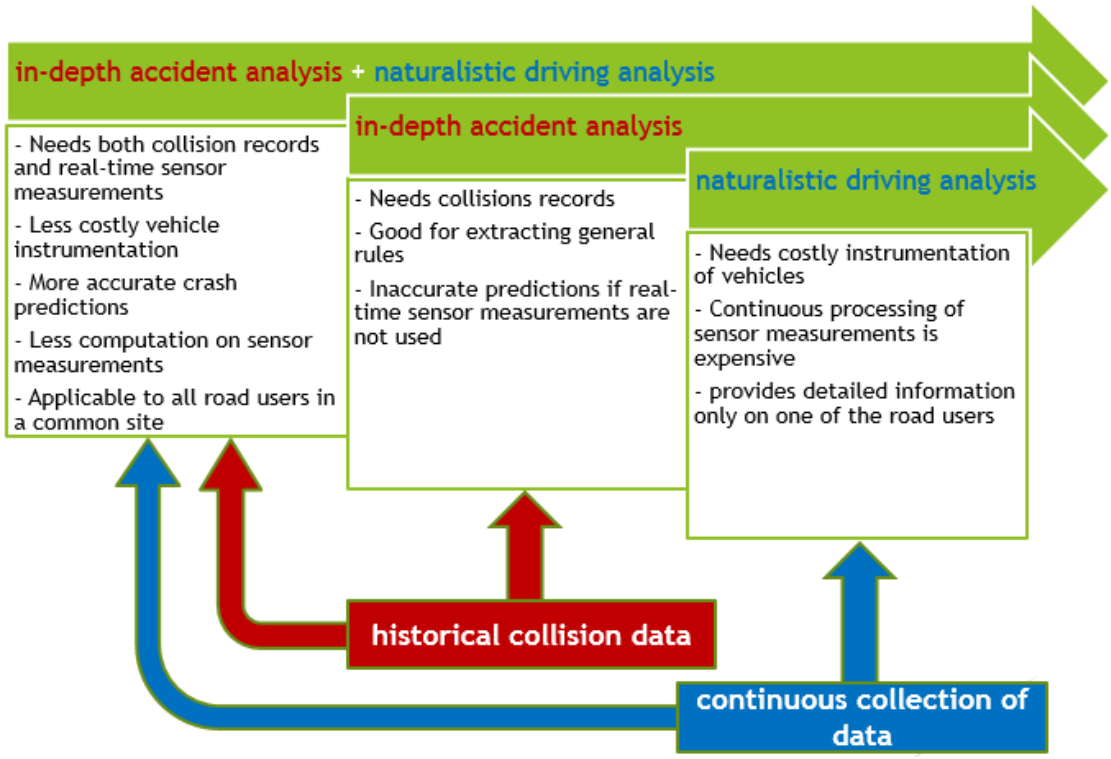

Figure 3.2: Comparison of in-depth accident analysis, naturalistic driving analysis, and their combination

chapter that data preprocessing is necessary for collision datasets, especially for decreasing heterogeneity, and hence, collision data clustering is also investigated.

After obtaining the framework to extract insights from a single dataset, integration of insights becomes prominent. There are many collision datasets available worldwide each of which points to different aspects of collision events based on the features they contain and the location in which the dataset is recorded. Some of them have been accumulating records for decades and hence are big in size. Insight integration is a significant contribution of this thesis which brings flexibility, adaptability, and context awareness to the CPM model. It also makes data processing, including road-accident data processing, less costly and less time consuming by eliminating the urge of huge datasets being involved directly in the model. Insight integration is investigated in Chapter 5, as shown in Figure 3.3.

The next sub-problem addresses the use of the integrated insights for training a collision

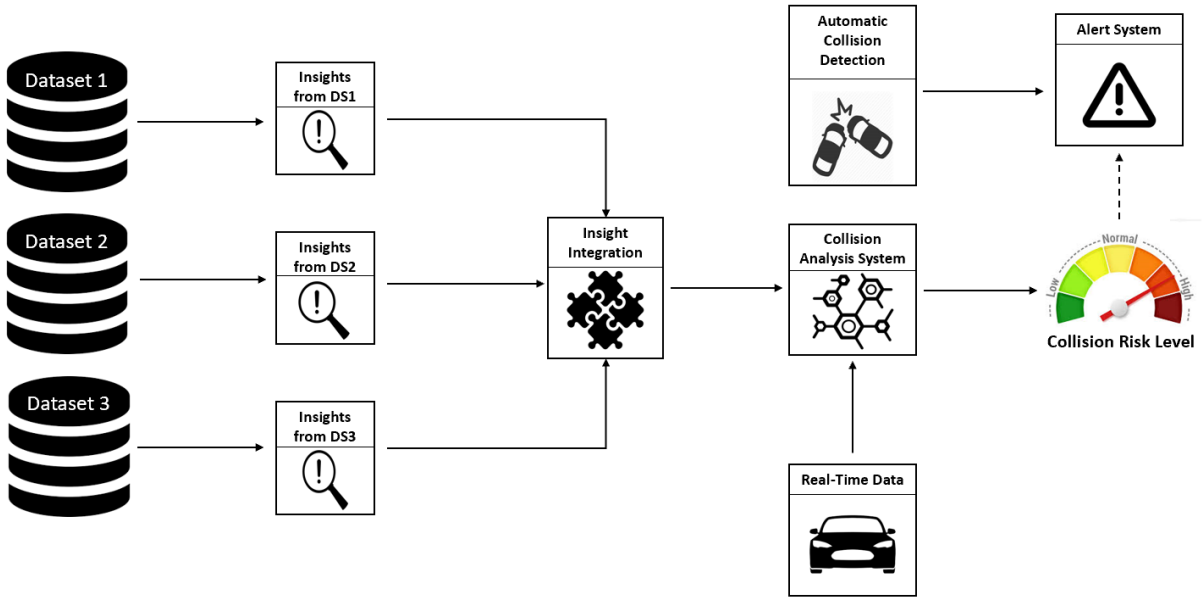

Figure 3.3: Problem formulation overview

analysis system. This system is the closest computational block to the user and generates the ultimate output this research is seeking for. All the real-time contributing measurements and information is passed through this system to be mapped into a collision risk level. This block is also where in-depth accident analysis and naturalistic driving studies meet. All discovered patterns resulting in occurrence of accidents collaborate in training a network that can classify different combinations of risk factors. At this stage, if any risk factor is recognized to have impact in the event of an accident, the overall impact will be analyzed by putting that factor alongside others to identify the accident probability region for them as a whole. This sub-problem is explored in Chapter 6.

There is another aspect of naturalistic driving analysis which is concerned with identification of traffic incidents in roads. Developing automatic traffic incident identification systems can be as important as collision risk analysis; Being connected to driving alert systems, they can inform drivers about blocked roads that should be avoided or to slow down before reaching an accident zone. Besides, any second can be crucial in informing the emergency personnel to be dispatched to the accident scene. Reducing the time gap between occurrence of accidents and presence of rescue team can save many lives. That is why an automatic collision identification system is investigated in Chapter 7 of this thesis.

Figure 3.3 shows how all the above-mentioned pieces are connected to create a well-

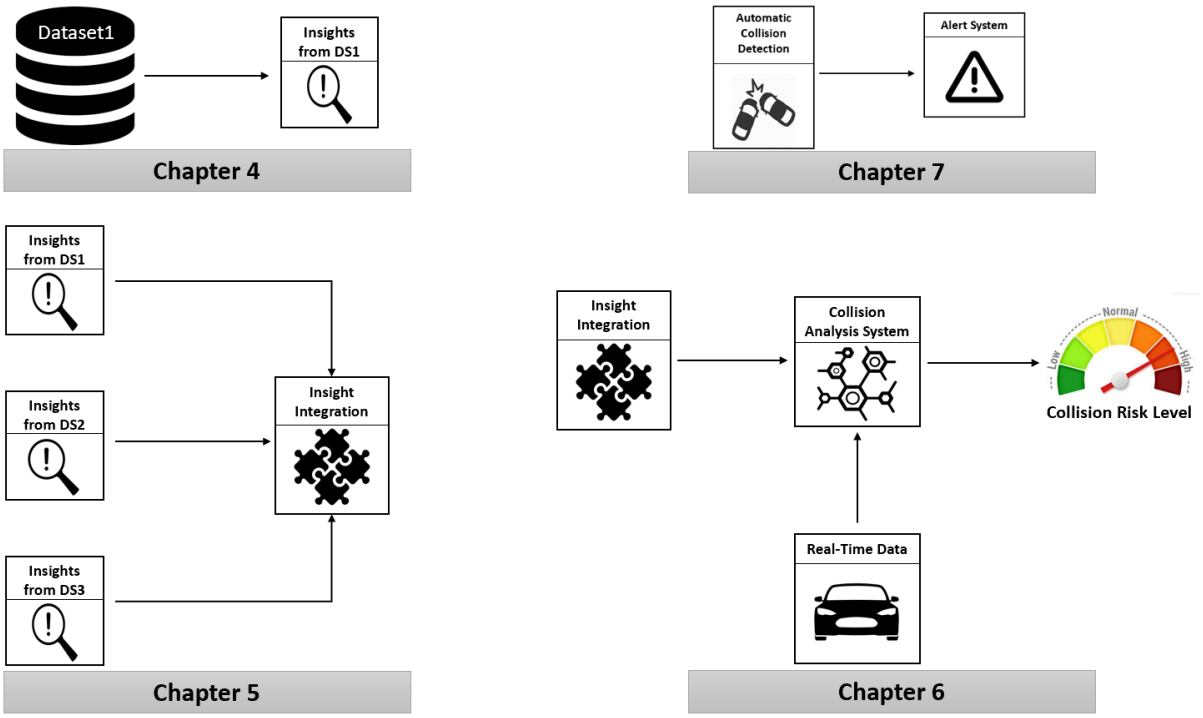

Figure 3.4: Problem breakdown

structured collision prediction and driver alert system. There are many other pieces that can be added to improve this complex but do not fit in the scope of this thesis. One of such pieces is the process of using the collision insights to generate suggestions to drivers in the event of high collision risk situations. However, for the scope of this thesis, I assume the risk level can be transmitted to a user interface for drivers as illustrated by a dashed line in the overview diagram in Figure 3.3.

Chapter 4

Accident Data Clustering and Insight Induction

4.1 Introduction

With the demand for more vehicles increasing, road safety is becoming a growing concern. Traffic collisions take many lives and cost billions of dollars in losses. This explains the growing interest of governments, academic institutions and companies in road safety. The vastness and availability of road accident data has provided new opportunities for gaining a better understanding of accident risk factors and for developing more effective accident prediction and prevention regimes. Much of the empirical research on road safety and accident analysis utilizes statistical models which capture limited aspects of crashes. On the other hand, data mining has recently gained interest as a reliable approach for investigating road-accident data and for providing predictive insights.

This chapter presents an approach, based on medoid clustering and association rule mining, for exploring and comprehensively understanding crash-contributing patterns in a given collision database. The Gower distance is used as a clustering criterion for handling the presence of categorical data. Furthermore, binary particle swarm optimization is employed to achieve insight discovery. The national collision database of Canada (NCDB) is used in this study to generate accident-related rules and evaluate the performance of the proposed approach. Although the number of attributes was limited with respect to human and vehicle-related factors, quite revealing insights were derived from the data pertinent to accident prevention and prediction. Experimental results are reported to demonstrate the efficiency of the proposed approach.

4.2 Accident Data Segmentation

Heterogeneity often exists in road-accident data. This heterogeneity may cause certain relationships between the variables to remain hidden. The heterogeneity in traffic data is the inconsistency in the values of risk-indicating variables in the same or almost the same circumstances [82]. It arises from collecting data from various scenarios in different circumstances while one or more critical variables are unobserved. Removing heterogeneity from road-accident data is a big challenge in road safety analysis [83]. Those certain relationships that may remain hidden can make the crash analysis task difficult. For example, certain risk factors related to specific vehicle types may not be significant in entire dataset, the influence of certain crash associated factors may be different for distinct conditions, or severity levels for collision contributing factors may be different for different accident types [84].

An example demonstrating the effect of heterogeneity in accident data can be seen in Figure 4.1. If the curvature and slope of the road (Road Alignment) are among the unobserved variables in the dataset, then there will be no feature in this road to differentiate its three segments. However, the different number of crashes per year in those segments shows that there are certain factors causing the vehicles in each segment to have different risks of accidents.

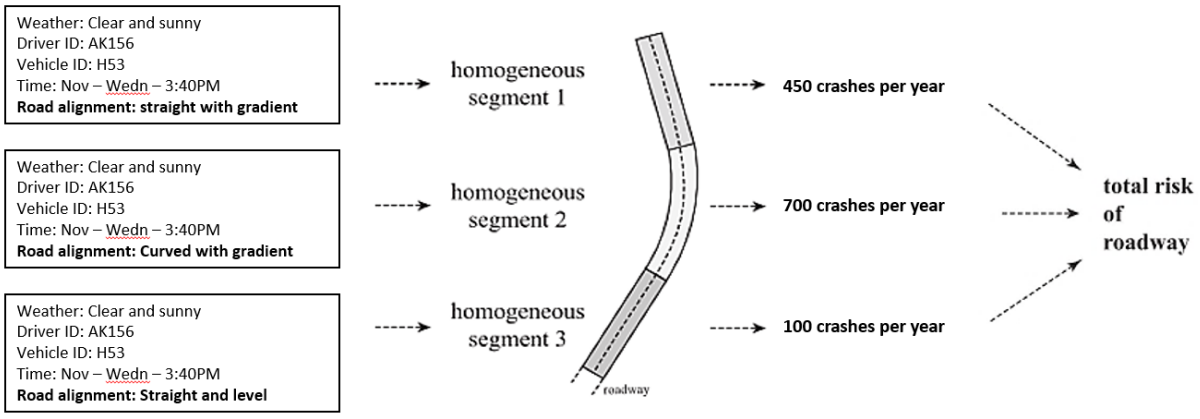

Figure 4.1: An example demonstrating the effect of heterogeneity in accident data

Data segmentation is one of the pragmatic solutions to the heterogeneity in the dataset. Data segmentation is concerned with the division of collision datasets into homogeneous segments of risk indicating variables. It will partition the data into relevant target groups to

minimize the impact of unobserved variables. This step will decrease the cost of computation by isolating the irrelevant data from each target group. Identification of homogeneous crash-data types allows subsequent road-accident analysis to deal with consistent crash configurations.

To apply data segmentation to the collision dataset, I opted for cluster analysis as a descriptive data-mining technique. In this unsupervised learning technique, the true number of clusters, as well as their form, are unknown. Therefore, I first utilized a cluster shape identification technique to reveal the true shape of the clusters.

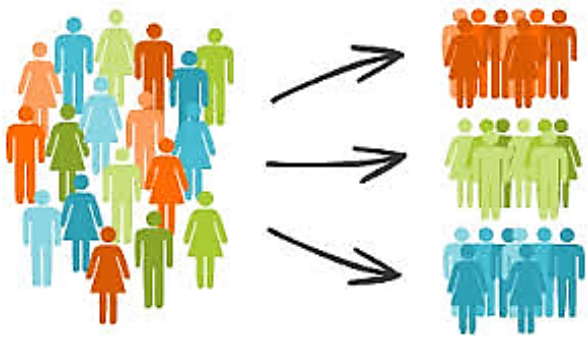

Figure 4.2: Data segmentation

Figure 4.3: Road-accident data segmentation

4.2.1 Cluster Shape Identification for Accident Data

Lack of prior information that defines the underlying data distributions generally happens in data clustering problems and makes it a challenging task. Using the wrong clustering algorithms for detecting clusters with unknown shapes and sizes may result in abnormal outcomes. In addition, different regions of the feature space may contain clusters of diverse shapes, and a single clustering algorithm can fail in finding all the clusters when its intrinsic

criterion does not fit well with the data distribution in the entire feature space. Road-accident data clustering is one of the key stages of in-depth accident analysis in this work which suggests using a proper technique to find the best algorithm for this purpose.

One of the solutions to uncover the underlying distribution of data in clusters is using multiple clustering algorithms with different clustering criteria. while a single clustering algorithm may not recover clusters of different shapes, I assume one proper clustering algorithm can be selected from a variety of clustering algorithms to provide the best approximation of the desired clusters. A cluster fitness function, however, is needed to identify the best algorithm that should be selected in the final clustering solution. If the final solution for clustering a dataset indicates poor performance with a single algorithm, it implies the existence of clusters with diverse types, otherwise that particular clustering algorithm deems appropriate for further examinations.

In this section, I introduce cluster shape identification which applies several clustering algorithms each representing a different objective function. These clustering algorithms are applied independently or in parallel to discover the shape of existing clusters. The big challenge here is that the objective function related to a clustering algorithm is not a good measure to indicate the quality of segments found by other clustering algorithms. Therefore, to identify the proper shape for a cluster, we need an external assessment criterion to judge the goodness of clusters found by diverse clustering algorithms.

One good option among the external assessment criteria is the stability of clusters when the dataset is resampled several times. Cluster stability has been used for other applications like multi-objective clustering and is found to be proper for this purpose [85, 86]. Stable clusters maintain their boundaries under resampling of the dataset. Formation of the same clusters regardless of minor changes in the dataset indicates their robustness and reliability. Let us assume that we have used clustering algorithm A_i to partition a dataset D in a way that it maximizes its corresponding objective function, f_i. In order to calculate the stability of obtained clusters, we resample D, M times and each time we get a new perturbed dataset D' and apply A_i to the perturbed dataset. Assuming that $P_j^i(D')$ is the partition obtained by applying A_i to the j^{th} perturbed dataset, we will have M partitions at the end of the run. The stability of A_i is then defined as the number of datapoints that never changed their clusters.

This definition requires identifying a single cluster in different partitions obtained through different runs of A_i. To do so, we must find the clusters that have more than 50 percent of their datapoints in common. In some partitions where a cluster is broken to smaller ones (and hence the number of clusters increases), we merge the smaller clusters to maintain the number of clusters as long as the merged cluster meets the minimum

Algorithm 4.1: Cluster Shape Identification

```
Data: Set of clustering algorithms A, Dataset D
Result: Selected clustering algorithm A*
for all A_i ∈ A do
    for j = 1 to M do
        Re-sample D to obtain perturbed dataset D';
        Apply A_i to D' and obtain P_j^i(D');
    end
    for all data points d_l ∈ D do
        if d_l never changed its cluster then
            Stability(A_i) = Stability(A_i) + 1;
        end
    end
end
A* = A_i with the highest Stability;
```

threshold of 50% similarity to its corresponding cluster in other partitions. If merging no combination of smaller clusters meets this criterion, the run is repeated, and the stability measure is penalized by a user-specified degree.

To ensure that different cluster shapes can be found using this algorithm, several independent clustering algorithms working based on different principals should be incorporated. Some common cluster shapes are spherical, hyper-ellipsoidal, and chained clusters. A suggested set of clusters to cover most common cluster shapes consists of:

1. Clustering around medoids (K-medoids) which minimizes the overall within-cluster distance and is suitable for finding spherical clusters. K-means clustering algorithm works based on the same concept and can be used as a substitute.

2. Expectation maximization (EM) algorithm which is a model-based clustering algorithm and can detect hyper-ellipsoidal clusters that may be overlapping.

3. Single link (SL) clustering which works based on minimum spanning tree capable of detecting chained clusters.

4. Spectral clustering which searches for clusters based on the spectral properties of the similarity graph obtained by using the inter-pattern distances.

4.2.2 Handling Mixed Data Types Using Gower Distance

Similarity-based clustering techniques use a specific distance function to measure the dissimilarity of the observations. Nonetheless, finding the pertinent distance function can be difficult. Common clustering algorithms are defined to deal with continuous variables. If the dataset consists of both continuous and categorical elements, the data is known to have a mixed type. A popular choice for clustering is Euclidean distance, but it is only valid for continuous variables. Since the type of road-accident data is mixed, Euclidean distance is not applicable to road-accident data clustering. In this study, a measure called Gower distance is used as the distance metric. Employing this measure allows the clustering algorithm to yield sensible results in the presence of mixed-type data.

Gower distance has a simple and comprehensive concept. For each variable type, we use a particular distance metric that works well for that type. The output is then scaled to fall between 0 and 1. At the end, the final distance matrix is created using a linear combination of all values with user-specified weights [87]. Let n be the size of dimension, and n_I , n_O, and n_N be the number of interval, ordinal, and nominal variables respectively. $\mathbf{X}$ is a mixture observation, characterized by n_I continuous variables and $n-n_I$ categorical variables.

$$\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_{n_I}}, \mathbf{x_{n_I+1}}, \ldots, \mathbf{x_{n_I+n_O}}, \mathbf{x_{n_I+n_O+1}}, \ldots, \mathbf{x_n}) \tag{4.1}$$

Thus, the vector x can be rewritten as follows:

$$\begin{aligned} \mathbf{X} &= (z_1, \ldots, z_{n_I}, q_1, \ldots, q_{n_O}, p_1, \ldots, p_{n_N}) \\ &= (\mathbf{Z}, \mathbf{Q}, \mathbf{P}) \end{aligned} \tag{4.2}$$

where $\mathbf{Z}$, $\mathbf{Q}$, and $\mathbf{P}$ represent subsets of $\mathbf{X}$ containing n_I interval, n_O ordinal, and n_N nominal variables. Gower's dissimilarity coefficient between the two mixture observations $x_i = (z_i, q_i, p_i)$ and $x_j = (z_j, q_j, p_j)$ can be calculated by the following equation:

$$\begin{aligned} D(x_i,x_j) = &\frac{\sum_{r=1}^{n_I} W_z^r D_{z_r}(x_i, x_j)}{\sum_{r=1}^{n_I} W_z^r} + \\ &\frac{\sum_{r=1}^{n_O} W_q^r D_{q_r}(x_i, x_j)}{\sum_{r=1}^{n_O} W_q^r} + \frac{\sum_{r=1}^{n_N} W_p^r D_{p_r}(x_i, x_j)}{\sum_{r=1}^{n_I} W_p^r} \end{aligned} \tag{4.3}$$

where W_z , W_q , and W_p are, respectively, the weights for interval variables, ordinal variables, and nominal variables. The metric used as the similarity measure for interval variables is a range-normalized Manhattan distance. $D_{z_r}(x_i, x_j)$ is the Manhattan distance along an interval variable z_r that can be computed as follows:

$$D_{z_r}(x_i, x_j) = \frac{|z_r^i - z_r^j|}{Max(z_r) - Min(z_r)} \tag{4.4}$$

For ordinal variables, they are first ranked, and then treated as interval variables (i.e., Manhattan distance is applied). Unlike interval and ordinal variables, nominal variables are first converted into binary columns by recoding each variable into a set of dummy binary variables. Then, k categories of nominal variables are separately converted into k sets of binary columns. Not all the measures for binary variables suit these dummy binary variables, considering that for a nominal variable, matching two individuals should have the same importance as when they do not match. In the case of this study, Dice similarity coefficient (DSC), also known as the Sorensen-Dice index or F_1 score, is adopted. DSC is used as a measure for comparing the similarity of two samples. Let:

- a be the number of times when the dummies 1 are aligned for both samples,
- b be the number of times when the dummies 1 for the first sample are aligned with the dummies 0 for the second sample,
- c be the number of times when the dummies 0 for the first sample are aligned with the dummies 1 for the second sample,
- And d be the number of times when the dummies 0 are aligned for both samples.

Then the DSC measure can be computed using the equation below:

$$DSC = \frac{2a}{2a + b + c} \tag{4.5}$$

The significant advantage of the Gower distance is that it is intuitive and straightforward to calculate. However, we should be careful about the presence of non-normality and outliers in the continuous variables since it is acutely sensitive to them. Hence, preprocessing of the data is required as an important step to remove the non-normality and outliers. The eventual outcome of measuring dissimilarity is a triangular matrix containing pairwise distances between data points.

4.2.3 Partitioning Around Medoids (PAM)

After calculating the distance matrix, a proper clustering algorithm should be selected. There are a few algorithms with the ability to handle a custom distance matrix. Among these algorithms, PAM was used in this study because it is relatively robust to noise and outliers [88]. The clustering steps for this algorithm are described in Algorithm 4.2:

Algorithm 4.2: PAM clustering

1. Pick k entities (data point) randomly to determine the initial medoids with k being the number of clusters.
2. Use the distance matrix to assign the entities to the cluster with the closest medoid to them.
3. For each cluster, find the observation that would minimize the average distance if it were chosen as the medoid for that cluster. Assign that observation as the medoid.
4. If none of the medoids has changed, end the algorithm. Otherwise, return to step 2.

As can be seen, the steps are similar to the K-means algorithm, but K-means has cluster centers defined by Euclidean distance (called centroids) while cluster centers for PAM are restricted to be observations themselves (called medoids). Compared to the K-means algorithm, PAM is more robust to noise and outliers. It also benefits from having an observation serving as the exemplar for each cluster. However, similar to K-means, it still suffers from quadratic run-time and memory.

4.3 Association Rule Mining Using Binary Particle Swarm Optimization (BPSO)

Association rule mining [89] is an approach that uncovers hidden relationships between seemingly unrelated data in a rational database. It generates a set of rules that describes the underlying patterns in the dataset. The association of different features is discovered by determining how frequent they appear together in the dataset. The rules are in the form of if/then statements, which are used to extract certain facts usually from large information repositories. A rule $A \rightarrow B$ indicates that if A occurs then B will also occur. Extracting the rules might be a time-consuming task, but when the rules are extracted, the resultant model works fast enough to process the new generated data.

A preliminary review of the literature reveals that association rule mining has not been extensively employed to investigate road-accident data. Recent publications show

that association rule mining in cooperation with clustering methods can outperform old-fashioned approaches in extracting certain contributors associated with road accidents (e.g., [90]). An association-rule-based approach is also applicable to very large databases without ignoring less significant parameters in the database associated with vehicle crashes. This point is valuable in this study as we want to consider as many parameters as possible involved in car crashes. Those less significant parameters that are ignored in many crash-prediction models might be significant in specific scenarios and the superiority of this study is in unveiling these scenarios and situations that cannot be uncovered in most of other crash prediction methods.

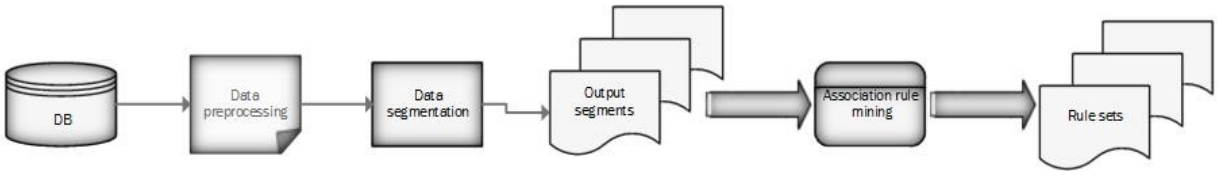

Figure 4.4: Association rule mining of collision databases

The association rule mining works mainly based on three interestingness measures of *Support*, *Confidence*, and *Lift*. These measures are the tools to find the significance of the rules. Although the common association rule mining methods are applied in advance and the results are used in developing the consequent model, their processing time increases exponentially with the size of the database. Evolutionary algorithms like Particle Swarm Optimization (PSO) can be used to expedite identifying significant rules.

4.3.1 Particle Swarm Optimization (PSO)

PSO is a population-based stochastic optimization approach introduced by Kennedy and Eberhart [91] that manipulates a number of candidate solutions at once. The main idea is taken from the collective behavior of social animals like bird flocking and fish schooling behaviors. A solution is referred to as a particle, and the whole population is called a swarm. The algorithm optimizes an objective function by iteratively improving the particles' positions in the search space. Each particle holds the essential information for its movement and decides about the next position by using its personal experience and that of the neighboring particles. Depending on how the topological neighbors are selected, the best neighboring particles' experiences can be defined as global best (gbest) or local best (lbest). Selecting a proper neighborhood affects the convergence and helps in avoiding getting stuck at a local minima.

Let us denote the number of particles by N and the dimension of the search space by D. The PSO algorithm in its basic form is given in Algorithm 4.3.

Algorithm 4.3: PSO algorithm

1. Initialize the position of each particle (X_i) randomly, restricted by the lower (L) and upper (U) bounds of the search space.
2. Apply the fitness function to all the particles to find the fitness value of them.
3. Use each Particle's current position to initialize its personal best $P_i \leftarrow X_i$.
4. Use the fitness function to initialize the global best as, $G \leftarrow Pi$, if $f(G) \geqslant f(Pi)$ for all values of i, where f is the objective function to be maximized.
5. Initialize each particle's velocity vector with 0.
6. Repeat the following steps until termination condition is met:
   for *each particle* $i = 1$ **to** N **do**
   - Pick random numbers r1, r2 from uniform distribution (0,1).
   - Update the velocity and position:
     $$V_i^{k+1} = WV_i^k + c_1 r_1 (P_i - X_i) + c_2 r_2 (G - X_i)$$
     $$X_i^{k+1} = X_i^k + V_i^{k+1}$$
   - **if** $(f(X_i) > f(P_i))$ **then**
     - $P_i \leftarrow X_i$
   - **if** $(f(G) < f(P_i))$ **then**
     - $G \leftarrow P_i$
7. The best position is stored in G which is reported as the optimal solution with $f(G)$ as the fitness value.

where V_i^k is the velocity of i$^{\text{th}}$ particle at k$^{\text{th}}$ iteration, W is the inertia factor, and c_1 and c_2 are positive constants.

4.3.2 Binary Particle Swarm Optimization (BPSO)

PSO was originally developed for continuous-valued spaces. However, many problems like traveling salesman problem (TSP) and assignment problems are defined for discrete-valued spaces. Binary version of PSO [92] is one of the variations of the PSO algorithm that makes it capable of dealing with discrete optimization problems. In this version, each particle represents a position in the binary space and their velocities are defined as probabilities of being in one state or the other. Since velocities represent probabilities, their values are restricted in the range of $[0, 1]$ by using the sigmoid function. The particles' position update equation in BPSO is given by:

$$x_{t+1}^{id} = \begin{cases} 1, & \text{if} \quad r < sig(V_{t+1}^{id}) \\ 0, & \text{otherwise} \end{cases} \tag{4.6}$$

where x_{t+1}^{id} and V_{t+1}^{id} are the position and velocity of particle id at iteration $t+1$, respectively, and r is a randomly generated number in $[0, 1]$.

4.3.3 Association Rule Mining Using BPSO

The BPSO-based association rule mining, by Sarath et al. [93], is a powerful tool for effective analysis of different information databases. Like the basic association rule mining algorithm, it finds the critical hidden information from the databases. What makes the BPSO-based one significant is its ability in improving computational efficiency and eliminating the need for finding the suitable threshold values for *Support* and *Confidence*. Unlike the commonly used association rule mining techniques like the apriory algorithm, the BPSO-based algorithm first searches for each particle's optimum fitness value and then the corresponding *Support* and *Confidence* values are reported to state the quality of rules.

As a preprocessing step for using BPSO-based association rule mining, the itemsets should be transformed and stored in a binary format. This helps accelerating the database scanning operation and calculation of *Support* and *Confidence* measures. Figure 4.5 explains the process of transformation and how they are stored for a straightforward application of the BPSO algorithm. This figure shows five records named T_1 to T_5 before the transformation process happen. These records are separately transformed into a binary format by considering a total number of 5 products I_1 to I_5. For each itemset, existence of any of these products in an itemset will make the corresponding cell to have a value of 1.

To represent an association rule as a particle position, each sequence is encoded as a separate rule. The consequents of the rules in this work are the ones describing the road-accident. Others can be placed in the antecedent part of the rule. Since the placement of the features in the rules is known, we do not need to add another dimension to distinguish the antecedents and consequents of the rules.

Support and *Confidence* can measure the strength of an association rule and can be calculated using Equations (4.7) and (4.8) respectively. *Support* of a rule $A \rightarrow B$ indicates the percentages of all itemsets, here accident records, in which A and B occurred together. *Confidence* of that rule indicates the percentage of itemsets in which when A occurs then B also occurs. Equation (4.9) is used to calculate another measure of interestingness for a

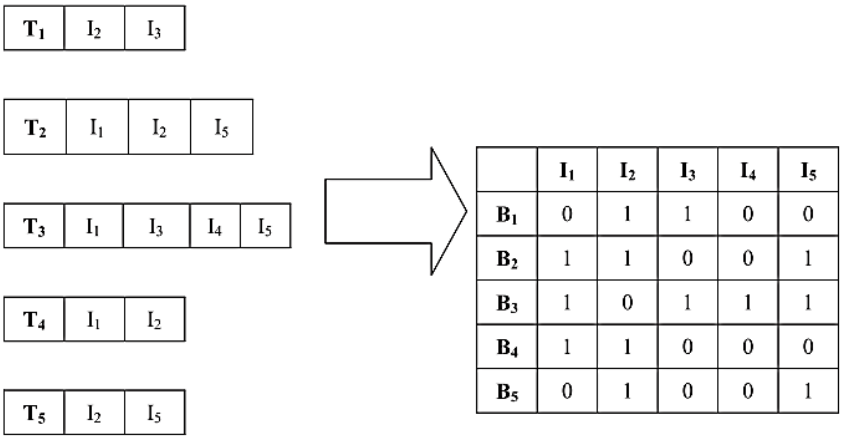

Figure 4.5: An example showing the transformation approach for BPSO [93]

rule called *Lift*. *Lift* is used to measure how dependant A and B are in the datasets. In case of a value greater than 1 for this measure, we can expect A and B appear together more frequently, whereas a value lower than 1 shows the opposite. A value of 1 for the *Lift* implies that A and B are independent of each other, which means no rule involving both can be drawn.

$$Support(A \rightarrow B) = P(A \cap B) \tag{4.7}$$

$$Confidence(A \rightarrow B) = P(A \mid B) = \frac{P(A \cap B)}{P(A)} \tag{4.8}$$

$$Lift(A \rightarrow B) = \frac{P(A \cap B)}{P(A)P(B)} \tag{4.9}$$

In order to define a fitness function in this study, I have used a weighted linear combination of *Support* and *Confidence*. The weights are used to control the effect of the database size: huge amount of road-accident data causes the values of *Support* to be scaled down significantly. Hence, many infrequent but valuable rules are pruned out. Consequently, a lower weight is assigned to *Support* to decrease the effect of the database size. The fitness function used in this BPSO-based association rule mining algorithm is given by:

$$f = W_s \times Support + W_c \times Confidnece \tag{4.10}$$

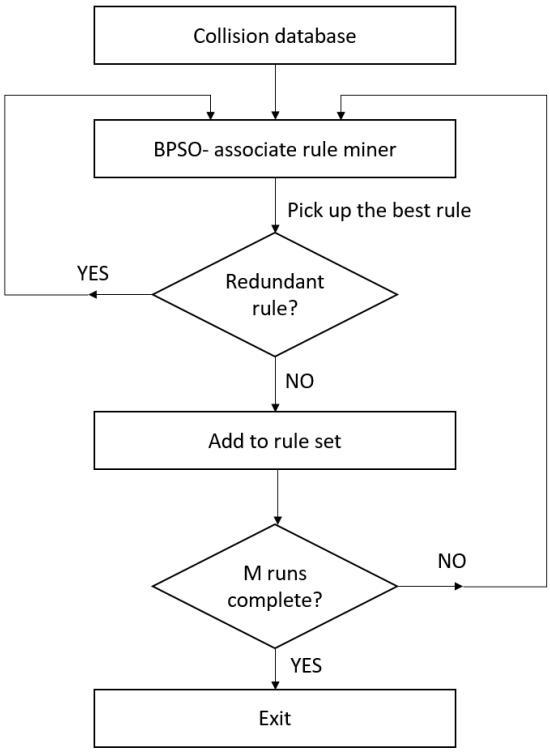

Figure 4.6: Block diagram of the BPSO-based association rule mining

The next step is to generate a population of particle swarms according to the calculated fitness value. The PSO search procedure usually continues until the maximum number of iterations is reached. Then, it outputs the top association rules for a given data. However, in this study, we continue the search until the algorithm starts to pick the rules which are already selected in the previous runs. The algorithm for this procedure is given in Algorithm 4.4:

Algorithm 4.4: BPSO algorithm

1 Initialize the cells of each particle randomly with either 0 or 1
2 Evaluate the particles using the fitness function
3 Update the personal and global best values
4 Update the particle positions
5 End the algorithm if the termination condition is met, otherwise return to step 2

A single run of this process will generate a single rule which is the best rule found in that run. To generate more rules, we have two options. The first is to fix the number of

rules to M upfront and run the algorithm M times. This approach will provide us with the top M rules from the data. If the rule found in any of the runs is previously found, that rule is discarded and new run is conducted to replace the discarded rule and maintain the population size of the rules. The second approach is to fix the number of runs of the algorithm and let it generate as many non-redundant rules as possible. In this approach, we only limit the upper bound of population size of the rules. In this study, I opted for the second approach to let the algorithm decide for the number of rules based on the amount of information residing in each cluster.

4.4 Experimental Settings

4.4.1 Dataset Description

To implement the data-driven accident analysis, we need to have a real and valid dataset to use. National collision database (NCDB) of Canada is the information repository that is used in this research and is obtained from Transport Canada [94]. This database contains all the police-reported motor vehicle collisions on public roads in Canada from 1999 to 2015. It consists of more than 3 million road accident records each representing occurrence of a collision. There are 22 different features available in this database which are divided into three categories. These categories are: (1) collision level data elements, (2) vehicle level data elements, and (3) person level data elements. The components of each category are represented in Table 4.1. Each component takes a certain number of unique values which is shown in the third column of this table. To simplify the appearance of these values in the association rules, I encoded each value with a number appended to the name of the component and its category. All possible values and their meaning for each variable in the database are available in appendix A and can be also accessed online in the NCDB dictionary[94]. The codes for component values are in the order of their appearance in the NCDB dictionary.

The category of collision level data elements is comprised of mostly environmental contributing factors. The collision severity, collision configuration, and number of involved vehicles are also among the data elements in this category which will be used to find the consequent of certain collision contributing factors happening at the same time. The vehicle and person level data elements contain a few elements from the vehicle and human contributing factors. This study shows that despite the limited number of attributes with respect to human and vehicle-related factors, quite revealing insights can be derived from the data pertinent to accident prevention and prediction.

Table 4.1: Description of the variables in use

Data element	Definition	unique values
Collision level data elements		
C_YEAR	Year	17
C_MNTH	Month	12
C_WDAY	Day of week	7
C_HOUR	Collision hour	24
C_SEV	Collision severity	3
C_VEHS	Number of vehicles involved	26
C_CONF	Collision configuration	20
C_RCFG	Roadway configuration	12
C_WTHR	Weather condition	19
C_RSUR	Road surface	11
C_RALN	Road alignment	8
C_TRAF	Traffic control	19
Vehicle level data elements		
V_ID	Vehicle sequence number	28
V_TYPE	Vehicle type	19
V_YEAR	Vehicle model year	55
Person level data elements		
P_ID	Person sequence number	37
P_SEX	Person sex	4
P_AGE	Person age	100
P_PSN	Person position	16
P_ISEV	Medical treatment required	5
P_SAFE	Safety device used	8
P_USER	Road user class	6

4.4.2 Collision Database Pre-processing

The process of data pre-processing is essential in this research to transform the raw data of accident records into an understandable format which can then be used to generate meaningful information through the further processing performed by the proposed algorithms. The collision database can be inconsistent or incomplete. It might also contain errors or lack certain behaviors or trends. Consequently, this stage is an inevitable part of the research and needs considerable attention.

Data pre-processing in this work consists of a series of steps. Data cleaning is the first step and is used to fill in the missing values, smoothing the noisy data, or resolving the inconsistencies in the data. During this process, the incomplete, incorrect, inaccurate, or irrelevant elements in the database are identified and resolved by replacing, modifying, or deleting actions. Data integration is another step which is taken during pre-processing. Since recorded observations for a road accident can be of different representations, data integration is used to put the data with different representations together and resolve the conflicts among them.

Data transformation and data reduction are other pre-processing steps used in this research. Since similarity measures will be used in this work as a basis for data segmentation through a clustering algorithm, normalization of data elements in the range of each feature is important. Here, the necessity of this step is highlighted due to the mixed nature of the data. I have also used sampling for the purpose of numerosity reduction. As mentioned in Section 4.4.1, the original database contains more than 3 million samples. In order to avoid big data analysis at this stage, I have used sampling to reduce the data and improve the performance in terms of computational time without losing much accuracy in the model.

4.5 Results and Discussions

4.5.1 Collision Data Segmentation

In the prediction problem that we investigate, many variables contribute, but NCDB database provides maybe a few of them. Other variables are either not given or cannot be determined directly, which causes heterogeneity in the dataset by adding hidden and unobserved variables. Road accident data segmentation is implemented in this study to mitigate the influence of these hidden and unobserved variables.

To remove the heterogeneity in road-accident data, I opted for cluster analysis as a descriptive data-mining technique. In this unsupervised learning technique, the true number of clusters, as well as their form, are unknown. Therefore, I first utilized a cluster shape identification technique by investigating the outputs of multiple clustering algorithms where cluster stability was chosen as the goodness function. For this purpose, 4 different clustering algorithms, each capable of finding different clustering shapes, were explored: clustering around medoids, expectation maximization (EM) algorithm, single link (SL) clustering, and spectral clustering. These four algorithms cover globular-shaped clusters, hyper-ellipsoidal-shaped clusters, chained clusters, and clusters that are based on the spectral properties of the similarity graph.

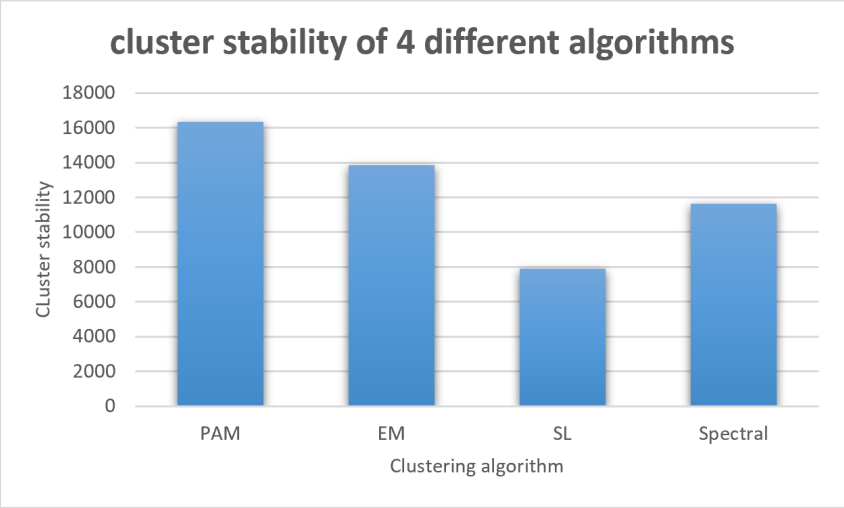

Figure 4.7: Cluster shape identification using cluster stability

This experiment on a 20000-sample set of the NCDB data revealed that the shape of the clusters for this particular data is globular. The stability of the four clustering algorithms used in this experiment are reported in Figure 4.7. As a result, I narrowed the clustering algorithm down to similarity-based clustering techniques that best conform to the clusters with globular shapes. In this case, similarity as the basis of internal clustering validation measures [95] can be used to discover closely-related patterns and PAM was adopted as the clustering algorithm to discover the clusters based on the similarity measure. The primary requirement of data segmentation using PAM clustering is to determine the number of clusters. Each cluster represents one segment which should be investigated individually to avoid the influence of hidden or unobserved variables.

In order to find the optimal number of clusters I utilized the silhouette score. This score is used to measure the similarity of an object to its own cluster compared to the other clusters. This score is in the range of -1 and 1 and is formulated in Equations (4.11) and (4.12). Higher values of silhouette score for an object indicates a better match to cluster it belongs. In these equations, $s(i)$ shows how similar object i is to the cluster that it has been assigned to considering its dissimilarity to the other clusters. For each datum i, $a(i)$ is the average dissimilarity of i with all other data within the same cluster, and $b(i)$ is the lowest average dissimilarity of i to any other cluster, of which i is not a member. The overall score for indicating the goodness of clustering can be calculated by averaging over all the $s(i)$s measured for each data point.

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{4.11}$$

$$S = \sum_{i=1}^{k} s(i) \tag{4.12}$$

where k is the total number of data points.

I implemented and applied the PAM clustering algorithm and iterated it over the number of clusters from 2 to 15 to record the silhouette scores corresponding to each number of clusters. I also implemented two other centroid-based clustering algorithms, K-means and fuzzy C-means, to compare them with the PAM algorithm. The silhouette scores obtained from these three algorithms are illustrated and compared in Figure 4.8. According to this figure, regardless of the clustering algorithm, the maximum silhouette score is achieved when the number of clusters equals to 3. The silhouette score of the models with more than 3 clusters declines as the number of clusters grows. In addition, PAM delivers a better model compared to K-means and fuzzy C-means at the point where the number of clusters is 3. These findings confirm that PAM is a suitable candidate among centroid-based clustering algorithms for road-accident data segmentation. It is also concluded that, for the specific part of the NCDB database investigated in this study, the optimal number of clusters is found to be 3.

Now that the number of clusters is set, the next step is to partition the data and describe how they are characterized. By exploring the segments, it was discovered that the road and weather conditions, driver's age and gender, and the severity of the accidents are the variables with the most contribution in the data segmentation for the current database.

The first cluster consists of collisions in which more than 90% of the involved vehicles are light duty (Passenger car, Passenger van, Light utility vehicles and light duty pick-up trucks) and the collisions have mostly happened while the traffic signal has been fully operational, the weather has been clear and sunny, the road surface dry and normal, the road straight and level, and the road configuration has been at an intersection of at least two public roadways. Cluster 2 consists of collisions in which more than 60% of the people involved were females. Regardless of the gender, people involved are mostly aged below 25 years. Moreover, more than 60% of the collisions did not cause injury and in cases where injury occurred, the injured person has been mostly the motor vehicle passenger and not the driver. Finally, cluster 3 consists of the collisions in which the driver has been injured.

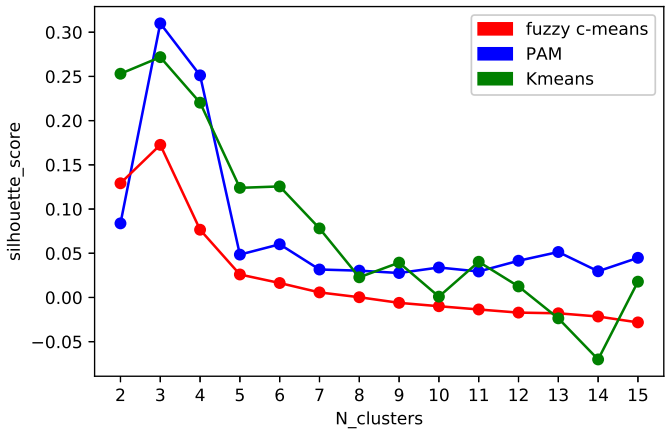

Figure 4.8: Silhouette score for different number of clusters

The drivers are mostly more than 25 years old and are men in 60% of the time. This cluster seems to be the complementary of cluster 2 and it is expected that the contributing factors have different influence on the drivers with the specified age and gender in this category.

4.5.2 BPSO Association Rule Mining

To generate association rules, BPSO-based association rule mining algorithm was applied to the data in each segment. At this stage, the consequent part of the rules is selected from the collision configuration and severity due to the nature of the database in use. The severity of collisions are obtained from the severity component of the NCDB dataset which can take low, injury, and fatality values. The crash likelihood is also defined to take low and high values based on frequency of rule occurrence. In this study, I set the crash likelihood threshold to 30 percent. Fitness value is the parameter based on which the rules are selected and is calculated using Equation (4.10). The fitness value threshold for selecting a rule was set to 60 percent based on empirical evidence.

The NCDB database is a collection of the collisions occurred in Canada, and therefore, non-collision situations are not covered in the current database. Table 4.2 contains the rules generated by the BPSO-based associate rule mining algorithm for cluster 1.

These rules show that November and December are the most probable months of the year that drivers show reckless driving behavior. Fridays are also the most frequent day of the week for seeing such behavior while Sundays and Mondays were observed to be the less

Table 4.2: Rules generated by the BPSO-based associate rule mining for cluster 1

Cluster 1 Traffic signal fully operational, weather clear and sunny, road surface dry and normal,road straight and level					
# Rule	Rule	Explanation	Fitness Value	Crash Likelihood	Severity
1	C_MNTH_{11-12} =>C_SEV_2	Crashes in November and December are both frequent and likely to have injuries	High	High	Injuries
2	C_MNTH_{1-3} =>P_ISEV_1 C_MNTH_{6-10} =>P_ISEV_1	Crashes in January to March and in June to October are frequent but not likely to have injuries	High	High	Low
	No rule for C_MNTH_{4-5}	Crashes in April and May are neither frequent nor likely to have injuries	Low	Low	Low
3	C_WDAY_5 =>C_SEV_2	Crashes on Fridays are both frequent and likely to have injuries	High	High	Injuries
4	C_WDAY_{2-4} =>P_ISEV_1	Crashes from Tuesday to Thursday are frequent but not likely to have injuries	High	High	Low
	No rule for C_WDAY_{1,6,7}	Crashes in weekends and on Mondays are neither frequent nor likely to have injuries	Low	Low	Low
5	C_HOUR_{15-17} =>C_SEV_2	Crashes from 3:00pm to 5:59pm are both frequent and likely to have injuries	High	High	Injuries
6	C_HOUR_{12-14,18} =>C_SEV_2	Crashes from 12:00pm to 2:59pm and from 6:00pm to 6:59pm are frequent but not likely to have injuries	High	High	Low
	No rule for C_HOUR_{00-11,19-23}	Crashes from 7:00pm to 11:59am are neither frequent nor likely to have injuries	Low	Low	Low

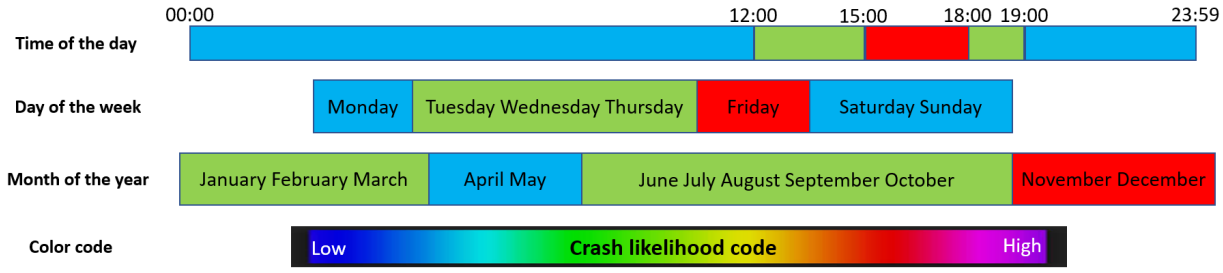

Figure 4.9: Summary of rules in cluster1

frequent ones. Another interesting fact in this cluster is that inconsiderate driving behavior mostly happens in the second rush hour of the day when people are going back home from work between 3PM and 5PM. It is evident that people have less degree of consciousness after work and can be easily distracted resulting into a non-fatal accident. Figure 4.9 summarizes the most probable time intervals in which drivers show careless behaviors.

Association rules in the second cluster, illustrated in Table 4.3, elucidate on the causes of accidents in the less experienced group of drivers. These rules show that light duty vehicle collisions at an intersection of at least two public roadways involved injuries when one vehicle conducts a left turn across opposing traffic. It is also observed that intersections with stop signs are a frequent crash site which results into right angle collisions and injuries. The third rule shows that drivers aged less than 20 years old are likely to have rear end collisions on the roads with no control. It is also found that vehicles manufactured before 2012 appear in non-fatal accidents more frequently than those manufactured later. The last rule in this cluster indicates that younger and less experienced drivers have more trouble in controlling the vehicle while the road surface is wet. These observations are perceptive for developing CPMs and encourage us to employ them accordingly.

The last cluster calls attention to the situations where even experienced drivers might have difficulty in preventing crash and controlling the vehicle. Populated in Table 4.4, the first rule highlights the curved parts of the roads, no matter leveled or gradient, where vehicles might run off the right shoulder of the road or roll over in the right ditch. According to the second and third rules, different sections of highways are also found to be dangerous when the weather condition is not suitable (rain, snow, freezing rain, and other conditions limiting visibility) or between 8PM and 2AM when the visibility range is restricted due to lack of light.

To evaluate the performance of BPSO-based association rule mining technique for road-accident insight induction, I measured the consistency of the extracted rules across different batches, each of which included 5000 observations. To that aim, the consistency check

Table 4.3: Rules generated by the BPSO-based associate rule mining for cluster 2

Cluster 2 Aged below 25 years, mostly females					
# Rule	Rule	Explanation	Fitness Value	Crash Likelihood	Severity
1	{C_RCFG_2 + V-Type_1 + C_Conf_33} =>{C_SEV_1 }	Light-duty vehicle Crashes due to a left turn across apposing traffic at an intersection of at least two public roadways are frequent and likely to have fatalities	High	High	Fatalities
2	{C_RALN_2 + C_TRAF_3} =>{C_CONF_35 + P_ISEV_2}	Stop sign locations along gradient roads are likely for right angle collisions resulting into injuries	High	High	Injuries
3	{C_RALN_4 + C_TRAF_18 + {P_AGE_<20}}=>C_CONF_21	If the road is curved and gradient and no control (traffic light or sign) is present, people aged below 20 are prone to rear-end collision	High	High	Low
4	{V_YEAR_<2012} =>C_SEV_2	People driving vehicles manufactured before 2012 are more likely to have injuries in accidents	High	High	injuries
5	C_RSUR_{2,5} =>C_SEV_2	Crashes are likely to have injuries when the road is wet or icy	High	Low	injuries

Table 4.4: Rules generated by the BPSO-based associate rule mining for cluster 3

Cluster 3 Aged above 25 years, mostly males, injury					
# Rule	Rule	Explanation	Fitness Value	Crash Likelihood	Severity
1	C_RALN_{3,4} =>C_CONF_4	If the road is curved (level or gradient), vehicles are likely to run off the right shoulder	High	Low	Low
2	{C_WTHR_{3-6} + C_RCFG_{8-12}} =>{ P_ISEV_{2 , 3}}	If it is raining, snowing or freezing rain is occurring or visibility is limited, crashes resulting into injuries and deaths are likely to happen in ramps; traffic circles; and express, collector, and transfer lanes of a freeway system	High	Low	Injuries and fatalities
3	{C_HOUR_{20-2} + C_RCFG_{8-12}} =>{P_ISEV_{2 , 3}}	After 8pm (when there is no sunlight), crashes resulting into injuries and deaths are likely to happen in ramps; traffic circles; and express, collector, and transfer lanes of a freeway system	High	Low	Injuries and fatalities

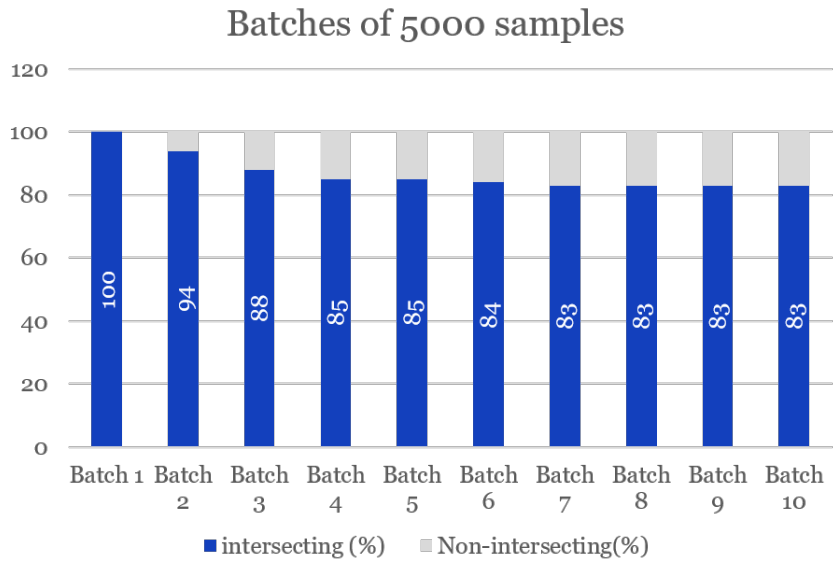

Figure 4.10: Evaluating the performance of BPSO-based association rule mining algorithm with NCDB database

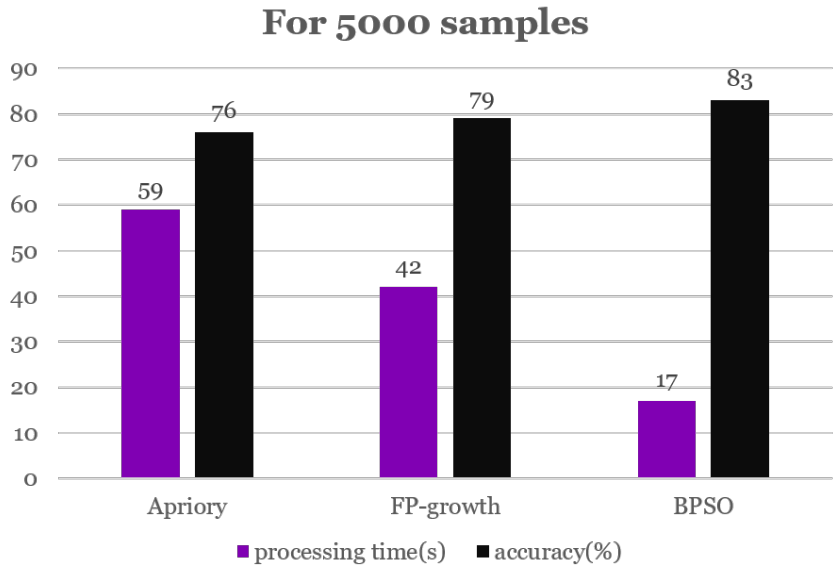

Figure 4.11: Comparison of BPSO-based association rule mining algorithm with Apriory and FP-growth algorithms

algorithm was developed with the following steps:

Algorithm 4.5: Consistency check algorithm

1. One batch is created by randomly selecting 5000 samples from the dataset.
2. The BPSO-based association rule mining is applied to the current batch.
3. If it is the first iteration go to step 1 otherwise go to step 4.
4. Calculate the percentage of rules' consistency in the current batch by counting the similar rules among all visited batches and dividing it by the total number of rules found in each iteration.
5. Terminate the program if the consistency remains constant for several consecutive batches. Otherwise, go to step 1.

Figure 4.10 shows the process of finding the consistency of the proposed insight induction algorithm applied to the NCDB database. This figure shows that 83 percent of the rules found by this algorithm are consistent across the whole database. One of the main reasons for why BPSO was selected in the first place was to reduce the processing time of rule induction in the database. Figure 4.11 compares the performance of road-accident insight-induction algorithm based on BPSO with apriory and FP-growth in terms of processing time and consistency. As illustrated in this figure, the BPSO-based algorithm has increased the degree of the consistency of extracted rules compared to commonly used association rule mining techniques. Moreover, the improvement of processing time achieved by this algorithm is significant which makes it a smart choice for large datasets like the road-accident dataset used in this study.

4.6 Conclusion

Understanding the relationship between the collision contributing variables starts with exploring the patterns of their appearance in a collection of collision samples. This chapter proposed medoid clustering and association rule mining for exploring and comprehensively understanding crash-contributing patterns in a given collision dataset. The clustering is adopted for the purpose of data segmentation which minimizes the effect of heterogeneity in data. The PAM algorithm was chosen based on a cluster shape identification method customized for the case study of road-accident data analysis. Cluster stability was the basis for the cluster shape identification algorithm that chooses the clustering algorithm. However, there is no certainty about the shape of the clusters based on the chosen clustering algorithm. In fact, clustering is an ill-defined problem and clustering shape and quality depends on how the clusters are used. The settings used here for the clustering shape

identification and the PAM clustering facilitate a solid segmentation of the dataset to tackle the issues arising from heterogeneity in accident data.

Handling the presence of categorical data is another challenge that emerges when dealing with a great majority of datasets. This challenge was addressed by introducing Gower distance as a criterion to measure dissimilarity of samples in the presence of categorical data. Insight induction is the core of this research and required a solid approach to reflect the relationship and contribution of variables in occurrence of traffic collisions. Association rule mining is a reliable approach for this aim. This thesis contributed to rule induction by enhancing its precision and computation time using binary particle swarm optimization. The rules generated the proposed method are accompanied with an estimate of the rule occurrence likelihood.

National collision database of Canada was investigated as a case study using the cluster identification, data segmentation, and association rule mining methods. In the numerical analysis, the chosen clustering method, PAM, divided the data into three segments based on existence of complex driving situations and driver characteristics. Association rule mining was then applied to these segments and insightful rules were successfully generated. The performance of the BPSO-based association rule mining proposed in this study was compared with that of other techniques in terms of consistency and processing time which showed significant improvement.

Chapter 5

Knowledgebase Aggregation and Approximation

5.1 Introduction

Knowledge discovery is becoming a central issue for industrial and government organizations. The ability of these organizations to conduct their business, effectively and efficiently, is heavily dependent on insights they derive from knowledgebases relevant to their businesses. This has led to the emergence of insights deduction systems as an important computing discipline.

It is typical that Big Data Knowledgebases tend to be disparate with high dimensionality. In the presence of high dimensional data, it may not be feasible to apply complicated functions directly to the dataset. As such, computational efficiency and knowledge fusion are major design concerns in insight induction systems.

There are quite similar issues in the case of road-accident data analysis. Deficiency of road-accident databases in containing all crash contributing factors is a significant barrier in discovering crash-associated patterns. Another issue is that researchers sometimes eliminate some of the available factors to avoid dealing with complicated relationships between variables. Although this approach simplifies the model implementation process, discarded information can be, in some cases, significantly correlated to the occurrence of accidents. Moreover, crash prediction models are often customized to incorporate distinct sets of contributing factors.

This chapter introduces knowledgebase approximation and fusion using association

rules aggregation as a means to facilitate accelerated insight induction from high dimensional disparate knowledgebases. There are two typical observations that make approximating knowledgebases of interest: (1) it is quite often that the insights induction can be derived based from a partial set of the samples, and not necessarily from all of them; and (2) generally speaking, it is rare that the knowledge of interest is contained in one knowledgebase, but rather distributed among a disparate set of unidentical knowledgebases. As a matter of fact, the insights derivable from knowledgebases tend to be uncertain, even if they were to be derived from a wholistic analysis of the knowledgebase. Thus, optimal knowledgebase approximation may yield the computational efficiency benefit without necessarily compromising insight accuracy.

I present a novel method to approximate a set of knowledgebases based on association rule aggregation using the disjunctive pooling rule. I show that this method can reduce insight discovery time while maintaining approximation accuracy within a desirable level. I initially devised the proposed method to enrich discovered insights from road-accident datasets. However, this method can be applied to other datasets to overcome various issues related to uncertainty of data and lack of information.

5.2 Motivation

Analyzing data becomes a challenging and expensive task when data starts to grow in volume, variety, and velocity. The accuracy and speed of many of the common predictive techniques degrade on high dimensional data. Abundance of data and high dimensionality may establish a more valuable resource, but entails incorporating more sophisticated predictive analysis. Moreover, the absence of accurate and well-organized data or the incapability of processing large datasets may result in false and spurious insights. Several projection pursuit and manifold methods like principal component analysis (PCA) and multidimensional scaling (MDS) are used for dimensionality reduction for high dimensional data. However, such methods typically rely on the assumptions such as the fact that variables are highly correlated or take only numeric values.

Typically, the underlying knowledge in a dataset is more important than the dataset itself in designing information systems. The knowledge extracted from a dataset is stored in knowledgebases which contain information at a higher level of abstraction. A knowledgebase stores general facts and rules which might be deduced from thousands of data samples. Therefore, the memory requirements for a knowledgebase is much lower compared to a conventional database.

Creating knowledgebases from datasets of reasonable size is simple, but the complexity of knowledgebase generation grows exponentially with the size of the feature space, especially when typical dimensionality reduction methods are not applicable. In the presence of high dimensional data, it may not be feasible to apply complicated functions directly to the dataset. As a result, many organizations refrain from using some fundamental features to avoid increased computational complexity while those features can enrich induced insights dramatically.

In the majority of cases, all the features that are needed to create a complete and comprehensive knowledgebase cannot be found in a single dataset. While data integration can be used to unify disparate datasets, it does not necessarily construct a reliable dataset containing all the features in one place. Even if it does, the emerging dataset would need a more intensive processing effort to output the desired knowledgebase. As a result, smaller datasets are processed and more and more partial knowledge is produced everyday. Generally speaking, it is rare that the knowledge of interest is contained in one knowledgebase, but rather distributed among a disparate set of unidentical knowledgebases. As such, computational efficiency and knowledge fusion are major design concerns in insight induction systems.

Limitations of memory and processing speed often require the knowledgebases to be approximated by aggregating the knowledge extracted from smaller datasets. In many cases knowledge approximation results in more accurate insights especially when the knowledge induction process relies on interestingness measures or in the presence of noisy and incomplete data. As a matter of fact, the insights derivable from knowledgebases tend to be uncertain, even if they were to be derived from a wholistic analysis of the knowledgebase. Thus, optimal knowledgebase approximation may yield the computational efficiency benefit without necessarily compromising insight accuracy. On the other hand, due to explosion of data, data mining methodologies and information retrieval mechanisms are being revolutionized. Therefore, it is essential to find faster mining approaches and gain deeper insight into recorded data to help make more effective decisions. Using approximation for reducing computational complexity is widely used for probability models and has been around for a long time. Examples of such approximation techniques can be found in [96] for discrete probability distributions and in [97] for probability models. For knowledgebases, the idea has been developed in the form of knowledge compilation or approximate knowledge fusion [98, 99, 100, 101]. Knowledgebases can also be aggregated to contain fused information. An example of such information fusion is introduced in [102] where ordered weighted averaging (OWA) is incorporated to fuse the decision lists of web search engines based on users' preferences.

Next sections present a novel approach based on disjunctive rule of combination to

approximate knowledgebases. The knowledge here is represented in the form of rules extracted using association rule mining (ARM) [103] techniques. To demonstrate the capacities of knowledgebase approximation, I apply this method to a well-known classification problem and show that it successfully generates rules that approximate correlations in the input dataset. This behavior is beneficial for knowledge fusion from multiple datasets and enhances computational efficiency when dealing with high dimensional data.

5.3 Combination of Evidence

Dempster-Shafer (DS) theory, often described as an extension of the probability theory or a generalization of the Bayesian inference method, offers an alternative for mathematical representation of epistemic uncertainty. As opposed to the traditional probability theory, where evidence is associated with single events, DS theory deals with evidence associated with sets of events and probability values assigned to sets of possibilities. DS theory works at higher levels of abstraction by adding a third aspect, called unknown, to the crisp logic. The basic idea is built upon obtaining degrees of belief from subjective probabilities and combining them using their independent items of evidence [104].

The three main functions used in the DS theory are the basic probability assignment function (BPA or m), the Belief function (Bel), and the Plausibility function (Pl). The BPA function assigns masses to all subsets of the entities in a system by mapping contents of the power set (P_Ω) to the interval between 0 and 1. The mass of subset p_i is commonly denoted by $m(p_i)$ and represents the amount of knowledge associated with that subset. In other words, $m(p_i)$ expresses the proportion of all available evidence that supports p_i but no particular subset of it. Each element $p_i \in P_\Omega$ is called a focal element of P_Ω if $m(p_i) > 0$, and the set of all focal elements is named a body of evidence (BOE). The following three equations represent the above description of m:

$$m : P_\Omega \rightarrow [0, 1] \tag{5.1}$$

$$m(\varnothing) = 0 \tag{5.2}$$

$$\sum_{p_i \in P_\Omega} m(p_i) = 1 \tag{5.3}$$

When multiple independent BOEs are available, which assumes existence of independent generic sources of information, we can use Dempster's rule of combination (DRC) to compute the aggregated BPA on p_i. Having two independent events p_a and p_b with their BPAs expressed by $m_1(p_a)$ and $m_2(p_b)$, DRC can be applied as follows:

$$m_1 \oplus m_2(p_i) = \begin{cases} 0, & \text{for } p_i = \varnothing, \\ \frac{1}{1-K} \sum_{p_a \cap p_b = p_i} m_1(p_a) m_2(p_b), & \text{otherwise.} \end{cases} \tag{5.4}$$

where

$$K = \sum_{p_a \cap p_b = \varnothing} m_1(p_a) m_2(p_b)$$

is a normalization constant called conflict degree and represents the amount of conflicting evidence between the two sources of information. DRC is purely a conjunctive operation which is AND-based and operates on set intersection. In the situation where not every source is reliable and at least one reliable source exists, a modified DRC, known as disjunctive pooling rule (DPR) [105], is more appropriate. As opposed to DRC, DPR is OR-based and operates on set union. DPR does not reject any of the information asserted by the sources and does not generate any conflict. It can be applied to two independent events p_a and p_b using Equation (5.5):

$$(m_1 \boxplus m_2)(p_i) = \sum_{p_a \cup p_b = p_i} m_1(p_a) m_2(p_b) \tag{5.5}$$

The other two key functions in the DS theory, the Belief and Plausibility functions, are two non-additive continuous measures perceived as the lower and upper bounds of the interval containing the exact probability at which p_i is supported [106]. Both functions are calculated based on basic probability assignment as indicated in Equations (5.7) and (5.8). The lower bound, Belief, is defined as the sum of all the masses of subsets of the set of interest, whereas the upper bound, Plausibility, is the sum of all the masses of the sets that intersect the set of interest.

$$Bel(p_i) \leq P(p_i) \leq Pl(p_i) \tag{5.6}$$

$$Bel(p_i) = \sum_{p_k | p_k \subseteq p_i} m(p_k) \tag{5.7}$$

$$Pl(p_i) = \sum_{p_k | p_k \cap p_i = \varnothing} m(p_k) \tag{5.8}$$

The two measures of Belief and Plausibility can be derived from each other by the following relations:

$$pl(p_i) = [1 - Bel(\overline{p_i})] \tag{5.9}$$

$$Bel(p_i) = [1 - Pl(\overline{p_i})] \tag{5.10}$$

where $\overline{p_i}$ denotes the complement of p_i.

DPR is more robust than DRC in the presence of conflicting evidence, and its use is appropriate when the conflict is due to poor reliability of some of the sources. In other words, DRC works based on the assumption that the belief functions to be combined are induced by reliable sources of information, whereas the DPR only assumes that at least one source of information is reliable, but we do not know which one. Both rules assume the sources of information to be independent. DPR is defined based on the union of the basic probability assignments (BPA) by extending the set-theory union and hence is an appropriate operator for insight aggregation. Some other characteristics of DPR that recommend it for this purpose are:

- Unlike conjunctive pooling, disjunctive pooling incorporates all the information asserted by the sources rather than selecting the part which is in consensus.
- The union does not generate any conflict
- No normalization procedure is required
- DPR is commutative and associative, but not idempotent
- The belief measure associated with aggregated BPAs is easily calculated via multiplication of individual BPA belief measures, i.e., $Bel(p_i) = Bel_1(p_i)Bel_2(p_i)$

5.4 Knowledgebase Approximation Framework

The focus of my approach for knowledgebase approximation is on integration of knowledge, which is drawn in the form of if/then rules using the ARM method, from smaller datasets with fewer features. These smaller datasets may be obtained from different data providers with their own objectives. In this case, approximating the knowledgebase corresponding to the integrated dataset would save the hassle of dataset integration and processing the bigger emerging dataset. Nonetheless, any large dataset can be broken into smaller ones by selecting only certain features to appear in each of them. As a result, we just need to deal with multiple lower-dimensional datasets requiring lower computation efforts.

Let us assume that N independent datasets are available for investigation indicated by DB_i and $i \in \{1, 2, \cdots, N\}$. These datasets are the main sources of information from which we aspire to obtain an approximated knowledgebase comprising all the attributes appeared in any of the datasets. Any pair of the datasets may share common features. Hence, the dimension of the corresponding integrated dataset is not necessarily equal to the sum of the number of features. At the end of this section I will show that common features help the DPR method to find the best approximation.

In order to induce knowledge from the smaller datasets, ARM is applied to each dataset which generates N independent rulesets. ARM explores and connects the attributes that contribute in the occurrence of a particular event or a set of events. Depending on the size and nature of the datasets, different ARM methods can be used. Given a minimum support threshold and a minimum confidence threshold, ARM finds all the strong association rules, that is, those whose confidence and support values are equal or greater than the thresholds. A rule that does not meet the thresholds is called a weak association rule.

Having mined all the association rules from the available datasets, there will be N independent rulesets available which are denoted by RS_i and $i \in \{1, 2, \cdots, N\}$. We assume each rule in the rulesets can be transformed into a piece of evidence in the form defined in the DS theory. Since ARM restricts the rules to those satisfying minimum support and confidence thresholds, masses corresponding to the emerging rules can be assumed to have non-zero BPAs and hence regarded as focal elements. Consequently, by defining a proper mapping, the rulesets can be transformed into independent bodies of evidence (BOEs) to which combination rules can be applied. The mass value assigned to a focal element is proportional to its generating rule's strength which is based on weighted linear combination of that rule's support and confidence.

My method incorporates DPR to combine the independent BOEs obtained from the lower dimensional datasets. DPR is a union-based operator and unlike DRC, which selects a

condensed part of evidence, DPR selects an extended piece of evidence based on the number and weights of the BOEs that can shape that extended piece of evidence in aggregate. In this study, pieces of evidence represent association rules and extending them will generate a rule with a larger number of antecedents. To merge the antecedents of multiple rules, their consequents should be the same. Therefore, the rules are filtered into groups in advance based on their consequents and the process of rule to BOE transformation and applying DPR is performed for each group separately.

As illustrated in Equation (5.5), DPR uses the values of BPAs in different domains to find a fused set of masses assigned to the higher dimensional domain. In a simplified version of the problem when the BPA values are disregarded, the strength of association rules are dismissed and all rules will have the same impact on generation of the elements in the fused set. Figure 5.1 elaborates on how DPR differentiates between strong and weak extended rules when BPA values are not considered. In this figure, $DB_q - R_k$ represents rule number k induced from DB_q and $DB_1 - R_{i1}$ is the first rule found in DB_1 that can generate a specified extended rule in aggregate with another rule $DB_2 - R_{i2}$ found in DB_2. In this figure, strong rule is an extended rule which is reproduced a good many times by the union of different pairs of rules, one from DB_1 and one from DB_2. In contrast, weak rule is reproduced relatively infrequent, which means not many rules from DB_1 and DB_2 could be augmented in their antecedents to form that extended association rule.

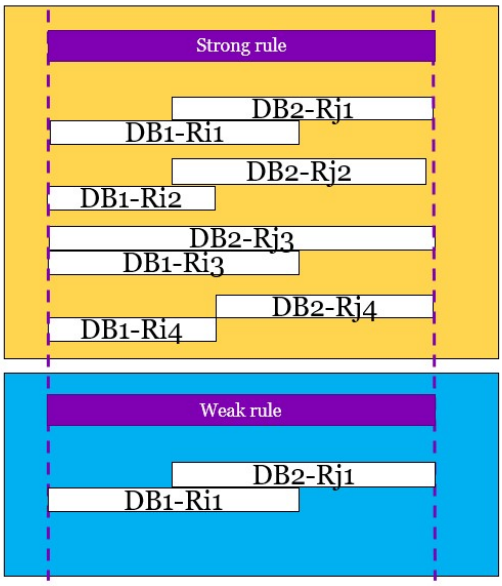

Figure 5.1: Strong vs. weak rules in DPR

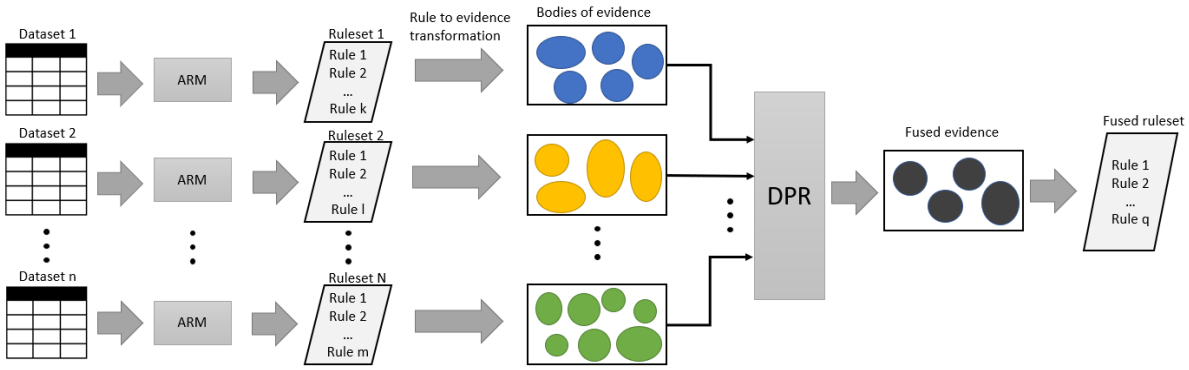

Figure 5.2: Application of DPR for knowledgebase approximation

In the real life scenario, the BPA values are not disregarded and each association rule in the rulesets is assigned a mass corresponding to the rule's strength. What changes in this case is that the strength of extended association rules is not measured only based on the count of reproduction times. Instead, the multiplications of the masses for any pairs of rules whose antecedents' union can reproduce the specified extended rule are accumulated. This procedure can also be applied to more than two datasets where the union should be conducted on a combination of rules from disparate datasets. Figure 5.2 outlines the proposed method in which N independent datasets are assumed available for investigation.

Mining transaction datasets for association rules typically generates a large number of rules. When ARM is used for subsequent prediction, most of the rules become unnecessary and can be eliminated using certain algorithms. In this method, however, I utilize every generated rule that satisfies the minimum support and confidence thresholds. Using the informative rulesets along with their dependant rules helps the DPR method to better sort the extended association rules based on their strength. When the fused rules and their corresponding BPAs are created then we can keep the informative ruleset and eliminate the dependant rules.

Rules' masses in this method are obtained by multiplication of their support and confidence, and normalizing them over the whole ruleset. Let us assume q rules are mined from a ruleset. If the rule r_i has the confidence $conf(r_i)$ and support $sup(r_i)$, then its mass is calculated using Equation (5.11).

$$m(r_i) = \frac{conf(r_i) \times sup(r_i)}{\sum_{j=1}^{q} conf(r_j) \times sup(r_j)} \tag{5.11}$$

When all the rules in N rulesets transformed into BOEs, DPR can be applied to integrate them into a single set of probability mass assignment indicated by fused evidence in Figure 5.2. This set is a combination of masses attributed to both strong and weak rules, but we can easily prune the masses and keep the stronger ones. To do so, we should consider the BPA in the fused set as a measure that is proportional to the multiplication of confidence and support of a rule that generated it. We refer to this measure as rule strength. Those BPAs in the fused set that can generate rules with a strength more than a minimum threshold will be selected as dominant BPAs and will be used to generate the integrated insights. The minimum strength threshold (MST) for this purpose is calculated using the minimum confidence and support of the rules in the original rulesets as indicated in Equation (5.12).

$$MST = Min(conf) \times Min(sup) \tag{5.12}$$

5.5 Application to Pattern Recognition

5.5.1 Evaluation Using Lymphography Case Study

Associative classification is an integration of association rule mining and classification which has been investigated widely in pattern recognition. Previous studies show that associative classification can achieve a high classification accuracy and is highly flexible at handling unstructured data. Among the algorithms proposed for classification based on multiple-class association rules CMAR and CPAR are shown to have competitive performance based on the experimental results in [107], [108]. In this section, I apply CMAR to the association ruleset in an approximated knowledgebase that is obtained by the association rule aggregation method and show that the accuracy of classification can be maintained when certain number of attributes are common between two datasets. In other words, a knowledgebase can be approximated by applying this method to its corresponding lower dimensional datasets when there is enough information shared among them.

I have used the lymphography dataset [109], obtained from UCI (university of California Irvine) machine learning database, to evaluate the DPR-based knowledge approximation framework. This dataset was recorded at the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. It contains 148 instances and 19 numerical valued attributes related to different aspects of lymphographic clinical data including the class attribute. Table 5.1 describes these attributes and the values used for them in the dataset. The class variable holds any of the four cases of normal, metastases, malign lymph, and fibrosis.

As described in the previous section, this study adopts ARM for extracting knowledge in the form of association rules. One of the criteria in ARM to select interesting rules is support which targets those rules that their components appear in the dataset adequately. Among the existing classes in the lymphography dataset two classes of normal and fibrosis contain very few samples and cannot satisfy the support measure. Therefore, our investigation is limited to the classes of metastases and malignant.

Table 5.1: Lyphography dataset features

Feature	**Feature description and values**
Lymphatic	A test for the overall lymphatic system: value 1 for normal; Value 2 for arched, value 3 for deformed; and value 4 for displaced
Block of afferent	value 1 for no; value 2 for yes;
Block of lymph c	value 1 for no; value 2 for yes;
Block of lymph s	value 1 for no; value 2 for yes;
By pass	value 1 for no; value 2 for yes;
Extravasates	expel from a vessel and is represented by 1 and 2;
Regeneration	value 1 for no; value 2 for yes;
Early uptake	value 1 for no; value 2 for yes;
Lymph nodes dimension	ranges from 0 to 3;
Lymph nodes enlarge	ranges from 1 to 4;
Changes in lymph	value 1 for bean, value 2 for oval and value 3 for round;
Defect in node	value 1 for no, value 2 for lacunars, value 3 for lacunars marginal and value 4 for lacunars central
Changes in node	value 1 for no, value 2 for lacunars, value 3 for lacunars marginal and value 4 for lacunars central;
Changes in structure	the structure of the lymphatic system; values 1 to 8, respectively, for: no, grainy, drop-like, coarse, diluted, reticular, stripped, and faint
Special forms	value 1 for no, value 2 for chalices and value 3 for vesicles;
Dislocation of node	value 1 for no and value 2 for yes;
Exclusion of node	value 1 for no and value 2 for yes;
Number of nodes	Values 1 to 7 for the number of nodes in the range of 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, and 60-69; and value 8 for equal or greater than 70

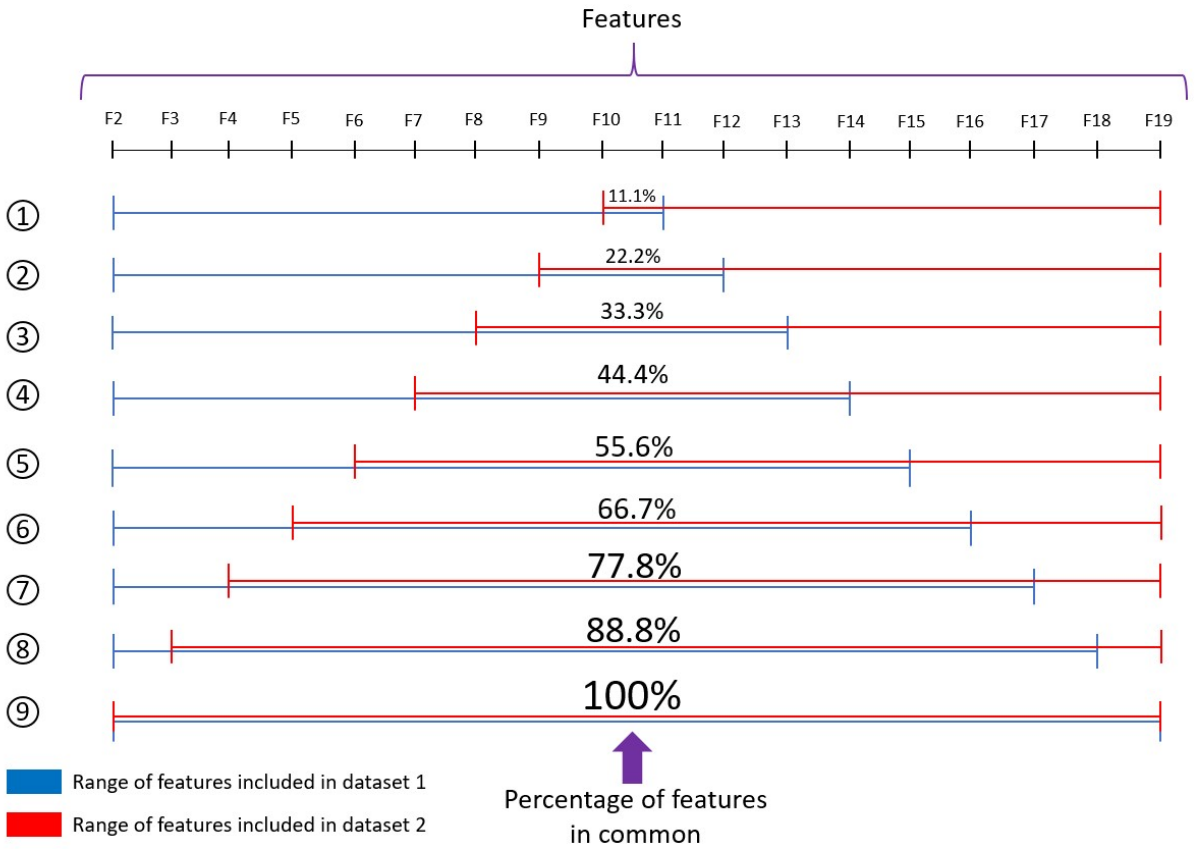

Figure 5.3: Lymphography dataset experiment

In order to approximate the knowledgebase for this dataset, I divided it into two smaller datasets in the feature space. This division was done 9 times and each time the percentage of common features was increased. Figure 5.3 shows this procedure where the percentage of common features is increased by 11.1 (2 features out of 18) in the range of 11% to 100%. The original dataset contains 18 features indicated by F2 to F19. As an example, when the percentage of features in common is 55.6%, the smaller datasets hold 14 features each, where F2 to F15 are in one and F6 to F19 are in the other. F6 to F15 are the features in common between the datasets in this example.

Once the original dataset is divided, ARM is applied to each smaller dataset to extract rulesets as illustrated in Figure 5.4. Before proceeding with integration, the rules were filtrated to separate the rules with consequences of metastases and malignant. The smaller rulesets with matched consequents are then independently aggregated base on the proposed

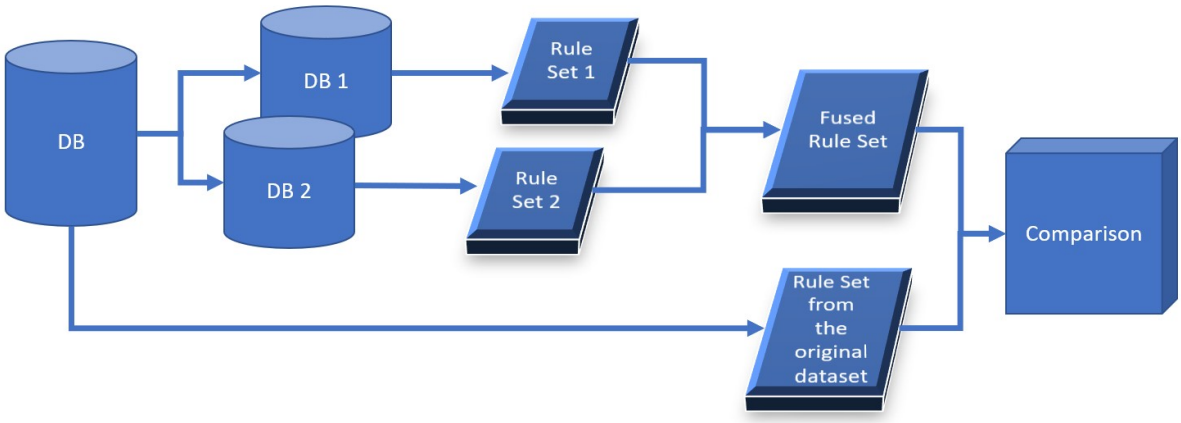

Figure 5.4: Evaluation framework

DPR-based integration framework to create two sets of extended rules with metastases and malignant as the consequents. These two sets can then be blended as a single set indicated by fused rules to represent the information in the approximated knowledgebase.

As shown in Figure 5.4, the ruleset from the original dataset was also induced as the ground truth for evaluating the proposed approximation method in this case study. As mentioned at the beginning of this section, the accuracy of associative classification is used to compare the knowledgebase and its approximation. CMAR classification method is applied to the fused ruleset and the ruleset induced directly from the original ruleset to compare the classification accuracy. For the classification purpose, 10-fold cross validation was used with a 70% percent of the original dataset's samples for training.

The DPR-based knowledgebase approximation can also be adopted to more than 2 datasets. In this case, there are two approaches available that have the same outcome. One is to apply the fusion operation to all datasets at the same time, an the other is to integrate rulesets two by two in a way that the resultant approximated ruleset of each two is fused with the next ruleset.

5.5.2 Results and discussion

In this study, the Apriori algorithm is used, which is the best-known ARM method and a simple approach for extracting association rules in a cohort dataset. However, other

Table 5.2: Approximation accuracy

Percentage of in-common features	**CMAR classification accuracy**	**Approximation accuracy**	**processing time (ms)**
11.1%	28.3%	34%	6
22.2%	41.8%	50.3%	12
33.3%	52%	62.5%	17
44.4%	61.2%	73.6%	26
55.6%	63.9%	76.9%	33
66.7%	69%	83%	48
77.8%	74.2%	89.2%	89
88.8%	79.3 %	95.4%	177
100%	83.1%	100%	1073

algorithms could be incorporated to generate insights for the proposed insight aggregation method. One important specification of the Apriory algorithm is that it uses a bottom up approach, i.e., one item is added to the frequent itemsets at a time and tested against the data. The breadth-first search nature of this algorithm makes it suitable for finding desired rules without considerable computation complexity from the small dataset in use here. Furthermore, it is easier to store the dependant rules of those in the smallest informative ruleset.

Based on empirical experiments, I set the value of support to 25%. This value provided enough frequent itemsets from which adequate useful rules were extracted. Moreover, the number of instances in the lymph dataset is 148 and a support of 25% guarantees that selected itemsets show up together in at least 37 instances. Although we can find so many rules with a confidence value of 100% if we decrease the support threshold, those rules are not necessarily useful since there are not enough occurrences of their itemsets. To better understand this fact, suppose there are only 2 instances of items A, B, C happening together. It is not infrequent that a rule with confidence of 100% is derived from it. If the number of instances is high (like 10K or more) then it makes sense to lower the support because even 1% in such case still has many instances.

As the goal of lymph classification was to relate predictor variables to the occurrence of two classes of metastasis and malignant, all predictor variables are limited to appear only in the antecedent (IF part), and lymph classes (outcome variable) to appear only in the consequent (THEN part). To generate all the strong association rules, the analysis is conducted by selecting any rule satisfying initial support threshold of 25% and confidence threshold of 40% for the generation of frequent item sets and rule induction. For each of the two smaller datasets, two types of rules were extracted for metastasis and malignant,

separately. I used orange3 in python 3.4 to apply the association rule algorithm.

After approximating the original dataset's knowledgebase by integrating the association rules from the smaller datasets, CMAR associative classification was applied to the aggregated rules. I applied CMAR to the variation of in common features. The accuracies are reported in the Table 5.2. The accuracy of CMAR in classifying the original dataset is 83.1% which is used to find the approximation accuracy in the third column of this table. An approximation accuracy of 100% indicates that the maximum possible classification accuracy is obtained, i.e., 83.1%.

One important achievement in the proposed method is decreasing the processing time. As discussed in previous sections, the processing time for finding association rules, or in general inducing insights, increases exponentially by the size of feature space. This approximation method decreases the processing time significantly while the approximated knowledgebase is highly accurate when there are enough features in common. The average processing time for the nine experiments reported in Table 5.2 using this method is 0.16s while running time of extracting association rules from the original dataset is about 1.1s. It is worthwhile to mention that the lymph dataset is a relatively small dataset containing only 143 samples, 18 features and 4 classes. The run time in this method is ten times less. It is trivial that using it for bigger datasets can save much more time.

Tables 5.3 and 5.4 report the number of correct and incorrect recovered rules in each experiment for the consequents of metastasis and malignant separately. The basis to distinguish correct and incorrect rules is the smallest informative ruleset induced from the original dataset. This basis is selected because the initial goal was to approximate the knowledgebase from the original dataset and choosing this basis conforms to that goal. In these tables, key features are those with higher relevancy to the consequents and can better distinguish the outcome. In the lymph dataset, 7 features exist that can be used as key features to differentiate between a metastasis and a malignant lymph.

To sum up, I introduced a knowledgebase approximation methodology to address two challenges in data analysis: (1) association rule mining efficiency at handling huge datasets, and (2) integration of induced rules from disparate datasets without the need for integration in data level. I proposed using disjunctive pooling rule along with basic probability assignment in Dempster-Shafer theory to combine the rulesets and assign a new measure of interestingness, i.e., rule strength, to the fused rules. The fused rulset is an approximate knowledgebase for the whole data available in disparate datasets. My experiments on the lymphography dataset in UCI machine learning database repository show that DPR-base approximation can achieve high accuracy when the number of features in common between two smaller datasets are above 60%.

Table 5.3: Metastasis

Percentage of in-common features	Number of key features in common	Number of correct recovered rules	percentage of correct recovered rules	Number of incorrect recovered rules	percentage of incorrect recovered rules
11.1%	2	11	36%	82	273%
22.2%	4	15	50%	65	217%
33.3%	5	16	54%	36	120%
44.4%	7	20	67%	22	74%
55.6%	8	23	77%	13	44%
66.7%	9	25	84%	8	27%
77.8%	11	26	87%	6	20%
88.8%	13	29	96%	2	6%
100%	14	30	100%	0	0%

Table 5.4: Malignant

Percentage of in-common features	Number of key features in common	Number of correct recovered rules	percentage of correct recovered rules	Number of incorrect recovered rules	percentage of incorrect recovered rules
11.1%	1	4	21%	16	84%
22.2%	3	6	32%	14	74%
33.3%	4	8	42%	11	58%
44.4%	4	8	42%	11	58%
55.6%	5	11	58%	7	37%
66.7%	7	15	79%	3	16%
77.8%	9	19	100%	1	5%
88.8%	11	19	100%	0	0%
100%	11	19	100%	0	0%

5.6 Application to Road-accident Datasets

In this section, I apply the proposed knowledgebase approximation method to road-accident datasets. In Chapter 4, I worked on the National collision database of Canada (NCDB) and mined top frequent patterns that result into accidents with certain configuration or severity levels. The NCDB dataset is organized and contains interesting features. Yet, it is lacking many important attributes specially in the vehicle and casualty categories which makes the derived insights unbalanced in those aspects. The proposed Knowledgebase approximation can resolve this issue by aggregating insights from NCDB and other road-accident datasets. The approximation experiment on the lymph dataset showed that we require at least 60 % feature similarity to obtain reasonable accuracy in approximation. Therefore, we need to act selective in choosing other datasets for this purpose.

After analyzing several open sourced road-accident datasets, I decided to use the road accidents and safety data of the Great Britain [110]. This Dataset, Which I call the GB dataset from now on, contains crash, vehicle, and casualty-related features that are reported in three different files named accidents, vehicles, and casualties respectively. The data in these files are linked through accident index and vehicle reference numbers. The main reason which makes this dataset appealing is the number of attributes that it has in common with NCDB dataset. Excluding the index and reference numbers and those features that are not counted as collision related factors, this dataset has 25 attributes, 16 of which are in common with the NCDB dataset. This means that 64% of its attributes are in common with NCDB. Since NCDB has 22 features in total, 72.7% of its features are in common with the second dataset. Consequently, we can be optimistic that the outcome of knowledgebase approximation provides us with insightful association rules.

The approximated knowledgebase will contain 31 unique features as opposed to the 22 in the NCDB dataset. The 9 new features added to the set are speed limit, vehicle manoeuvre, skidding and overturning, point of impact, vehicle leaving carriageway, junction control, number of casualties, light conditions, and pedestrian movement. knowing how any of these features, in conjunction with the currently available ones, can impact the risk of crash occurrences can be of great significance which is not attainable by processing these datasets apart.

The data in the three files in the GB dataset, Accidents, Vehicles, and Casualties, is separated based on timestamp. Since the knowledgebase approximation approach works based on association rules, it is useful to concatenate the data into a single dataset. having all the data in a single dataframe, we can run the same piece of code that we used for NSDB dataset to get the association rules out of the GB datset. Therefore most of the necessary data preprocessing, cleaning, and transformation functions needed to be applied on this dataset is previously established. However, a few preprocessing and transformation steps need to be done in advance. For example, we still need to transform many features to textual strings that is used in the initial dataset to represent the common features with the same names. When the GB dataset features were transformed into the same format as the ones used in NCDB, we proceed to feed it to the insight induction system that was introduced in Chapter 4 which prepares, clusters, and generates insights in the form of association rules.

A quick analysis of the insights derived from the GB dataset illustrate interesting road-accident patterns some of which are common with those derived from NCDB. for example, the rules with antecedents of day and hour show that morning and evening rush-hours are accounting for greater number of accidents which was also shown by the rules from NCDB. The evening rush-hour, however, accounts for a greater portion of the accidents

which can be due to the fact that drivers suffer from fatigue, anger, or other emotions that they carry after getting out of the workplace. The new insights also show that the severity of accidents In the Great Britain has been decreasing over the past years.

Without further analysis of the insights from the GB dataset, I integrated the insights from NCDB and GB datasets by using the knowledgebase approximation framework introduced in this chapter. As expected, there were quite a few association rules that were expanded after applying the integration function. The integrated rules are those rules that enough evidence in both individual datsets can support them. These new rules help us better understand the correlation of existing features and their collaboration in increasing the risk of accidents including fatal, serious, and slight ones. The top eight rules in the approximated ruleset are illustrated in Table 5.5.

The first rule in this ruleset is the same as one of those extracted from the NCDB dataset. The rule attests to the high crash risk of Light-duty vehicles when attempting a left turn across apposing traffic at an intersection of at least two public roadways. This rule views the left turn accidents as a general case which might be attributed to insufficient awareness of drivers on left turn when the lights turns yellow. Yellow lights, also called amber lights, can be confusing to drivers. According to section 44 of the highway traffic act, one must stop when approaching a yellow light if you can do so safely; otherwise, go with caution. Left-turning drivers, however, should only clear the intersection when no opposing traffic is passing through, which applies to the yellow light as well.When a collision happens during a left turn, the one who has made the left turn is at fault, even on yellow.

Left turn collisions can be even more out of control when the weather condition is severe. The second rule in Table 5.5 shows that left turning vehicles colliding with skidding vehicle in the opposing traffic are likely to have injuries and fatalities. Since this type of accident is not frequent, the risk is not too high, but the high severity of this type of accident places it among risky situations in roadways. When the car begins to skid, pushing the brakes or steering does not produce the normal result. Therefore, when the roads are wet, icy, or have snow, keeping greater distance with the car to the front and breaking early when approaching intersections is recommended.

The third rule is also concerned with left turn, but when pedestrians are involved. The rule says that when there are pedestrians involved in a left turn, vehicle is vulnerable to have right angle collisions or go off the road and hit stationary objects. In many instances, hitting a pedestrian or going off the road collisions happen at intersections with signaled and marked crosswalk. The reason can be attributed to the fact that both parties have the signal. Hence, the pedestrian steps into the crosswalk and the car has initiated the

Table 5.5: Rules generated by the knowledgebase approximation technique on NCDB and GB datasets

Knowledgebase approximation for NCDB and GB insight integration top 8 rules					
# Rule	Rule	Explanation	Fitness Value	Crash Likelihood	Severity
1	{C_RCFG_2 + V-Type_1 + C_Conf_33} =>{C_SEV_1 }	Light-duty vehicle Crashes due to a left turn across apposing traffic at an intersection of at least two public roadways are frequent and likely to have fatalities	High	High	Fatalities
2	{C_RALN_2 + V-Type_1 + C_Conf_33 + snow + skidded} =>{P_ISEV_{2,3} }	light-duty vehicles skidding in snowy weathers and colliding to left turning vehicles at an intersection of at least two public roadways is frequent and likely to have injuries and fatalities	High	Low	Injuries and fatalities
3	{C_RALN_2 + C_TRAF_3 + pedestrian} =>{{C_CONF_35 , C_CONF_02} + P_ISEV_2 }	When there are pedestrians involved in a left turn, vehicle is vulnerable to have right angle collisions or go off the road and hit stationary objects	High	Low	Injuries
4	{C_RALN_4 + C_TRAF_18 + {P_AGE_<20} + {speedlimit >80kmh}} =>C_CONF_21	If the road is curved and gradient and no control (traffic light or sign) is present, people aged below 20 are prone to rear-end collision when speed limit is above 80kmh	High	High	Slight
5	{{V_YEAR_<2012} + single carriageway + {speedlimit >80kmh} + overtaking} =>{C_SEV_{2,3}}	Overtaking in a single carriageway has injuries and fatalities when the vehicle is manufactured before 2012 and the speed limit is more than 80kmh	High	High	Injuries and fatalities
6	{{slip road , lane change}+{P_AGE_<20}} =>{P_ISEV_{2 , 3}}	Inexperienced drivers are likely to have severe accidents when attempting lane changing or merging in slip roads (ramps)	High	High	Injuries and fatalities
7	{C_WTHR_{3-6} + C_RCFG_{8-12}} =>{ P_ISEV_{2 , 3}}	If it is raining, snowing or freezing rain is occurring or visibility is limited, crashes resulting into injuries and deaths are likely to happen in ramps; traffic circles; and express, collector, and transfer lanes of a freeway system	High	Low	Injuries and fatalities
8	{{C_HOUR_{20-2} + C_RCFG_{2,3,8,9}} =>{P_ISEV_{2 , 3} +{pedestrian , motorcycle , pedal cycle}}	Pedestrian, motorcycle, and bicycle crashes resulting into injuries and fatalities are likely to happen during dark hours	High	High	Injuries and fatalities

left turning where resuming it results to hitting the pedestrian and stopping results to obstructing the opposing traffic. This problem arises mainly due to the reduced visibility of drivers when making a left turn. Drivers should be extra cautious during such situations where they may come to close contact with pedestrians.

Previously in Chapter 4, we extracted the rule from the NCDB dataset saying "If the road is curved and gradient and no control (traffic light or sign) is present, people aged below 20 are prone to rear-end collision". Now the integration of insights has added a

new element to this rule: speed limit. There is no speed limit reported in the NCDB dataset while it is reported in the GB dataset. Integration of insights has been able to further illuminate on this rule by stating that people aged below 20 are prone to have rear-end accidents in curved gradient uncontrolled roads when the speed limit is above 80 kilometers per hour. Yet, the severity of these types of accidents are often slight and injuries and fatalities are rarely reported. The speed limit in GB dataset is reported in miles per hour, hence, I transformed it to kilometers per hour to better reflect on the Canadian traffic accidents.

The fifth rule in the top eight integrated rules sheds light on severity of collisions for the vehicles manufactured before 2012 when overtaking in a single carriageway with the speed limit above 80 kilometers per hour. We previously figured out that vehicles older than 2012 have curtailed safety levels, but no other components of the high risk patterns for these vehicles were discovered. This new rule adds overtaking in single carriageways to the list of actions that drivers, specially those driving older vehicles, should be cautious about. Head-on collisions are one of the most dangerous ones that are likely to happen while performing this behaviour and speed limit of above 80 kilometers per hour can surely increase the intensity of the crash.

Exploratory data analysis of any of the two datasets show that inexperienced drivers, particularly drivers aged less than 20, have the highest rate of accidents among all the age bands. The sixth integrated rule shows that two manoeuvring behaviours are of high risk for inexperienced drivers: lane changing and merging in slip roads. Both situations needs good understanding of yielding and speed control. They both have a huge effect on surrounding traffic if done incorrectly and can result to fatal accidents in many occasions. safe performing of these maneuvers requires proper speed matching, gap spotting, lane changing indication, and performing the lane change when it is safe to do so.

Rule 7 in Table 5.5 is exactly the same as one of those extracted in previous chapter and emphasizes on the fact that bad weather conditions are directly associated with higher accident rates. The role of certain car's equipment is more felt while driving in harsh weather. Headlights, tail lights, and windshield wipers must be checked to be functional when they are needed. The tread of vehicle's tires should be also checked as balding tires can severely reduce traction on wet and icy roadways. Although the crash likelihood of these type of traffic accidents are low, they are often reported to have injuries and fatalities. Certain locations were high severity accidents of these types happen include ramps; traffic circles; and express, collector, and transfer lanes of a freeway system.

The eighth rule is concerned with the role of dark hours on vulnerable groups in traffic including pedestrians, motorcycles, and bicycles. The rules states that Pedestrian or (mo-

tor)bike crashes resulting into injuries and deaths are likely to happen during dark hours in ramps; traffic circles; and express, collector, and transfer lanes of a freeway system. It is clear that lack of visibility is the reason for most of these accidents. Since these vulnerable groups are more susceptible to take injuries or die in crashes, the risk and severity of this type of accident is high. hitting a pedestrian, bicycle, or motorcycle at a speed of over 50 kilometers per hour results in serious injuries and fatalities, and yet, a driver can seriously disable them in a crash where the driver is travelling only 15 kilometers per hour.

These integrated rules were selected and ranked based on their fitness value which are obtained using Equation (5.11). The whole list of rules are created based on the minimum strength threshold in Equation (5.12), but only the top eight are represented in Table 5.5. In this table, crash likelihood has been indicted by a binary value and reported as high or low. Crash likelihood here is calculated based on the ratio of number of particular crashes and the total crash quantities in a dataset. A high value for crash likelihood is indicative of frequent appearance of certain type of accidents compared to all crash-contributing patterns identified in the dataset. In this study, patterns with more than 5 % ratio of appearance are considered having a high crash likelihood. Crash frequency is often used by engineers as a base for calculating the probability of crash occurrence. However, in this study, a different fitness value is defined which is a better indicator of crash probability. The reason is that here all the patterns are compared with one particular outcome, or consequence, for calculating the chance of appearance of that particular outcome. This way, we understand the exact configuration of collisions that might occur when a certain pattern, or combination of features, is identified. Therefore, fitness value of a rule is what makes that rule important, which is based on measures of interestingness.

Outstanding statistics can be derived from the knowledgebase that is approximated from NCDB and GB datasets. Figures 5.5 and 5.6 show some of these interesting statistics that are calculated by considering the first 10000 rules in the approximated knowledgebase. There are 16 influence factors that appeared more frequent in the association rules with a consequent equal to certain type of crash and severity. Each of these 16 common influence factors appeared in more than 5% of the first $10k$ rules. Figure 5.5 shows that speeding has contributed in more than 28% of the first $10k$ rules and is the most contributing factor among the ones with more than 5% total contribution. However, speeding is not the factor with the most percentage of high-likelihood among the rules it has contributed to. The same figure shows that only 62% of the rules that speeding has contributed to have a high likelihood to happen. Impairment, on the other hand, is the 8^{th} most contributing factor in terms of the number of rules it has appeared in, but 95% of the rules containing impairment have a high likelihood to happen based on the specified measure of interestingness.

Figure 5.6 compares the severity of the accidents for each of the common influence

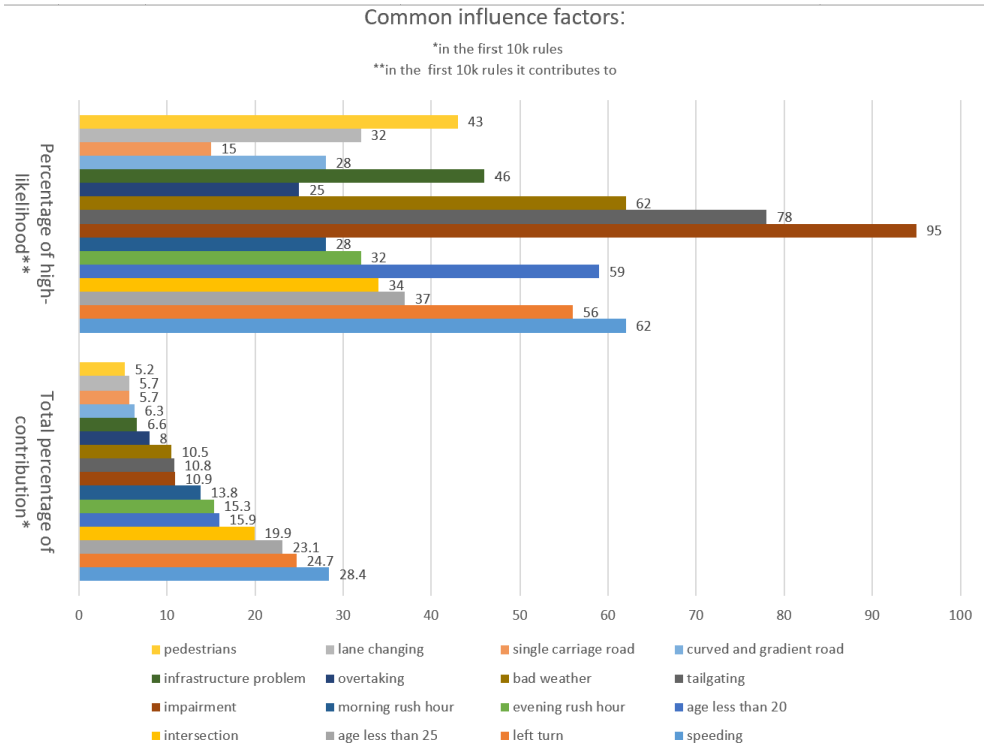

Figure 5.5: Statistics for common influence factors in the approximated knowledgebase

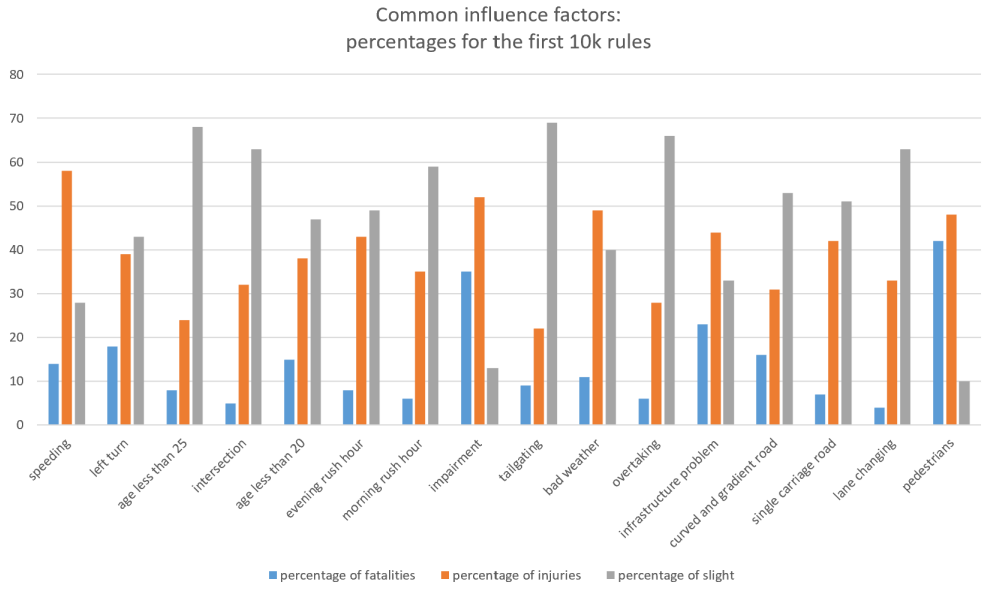

Figure 5.6: Severity level for common influence factors in the approximated knowledgebase

factors. The severity can belong to one of the three classes of fatalities, injuries, and slight. According to this figure, the accidents that contain pedestrians, impairment, and temporary infrastructure problems are among the deadliest types of accidents. The risk of an accident is not just dependant to a single influence factor. It is the combination of all contributing factors that determines the risk. This figure shows the influence of each factor separately on the severity of accidents by using statistics of the first $10k$ rules in the approximated knowledgebase. Next chapter explains how the risk level and likelihood of a certain accident can be estimated by considering the combination of contributing factors and by using a Bayesian network trained with fused association rules.

5.7 Conclusion

This chapter introduced knowledgebase approximation which aggregates association rules to facilitate insight induction from high dimensional disparate datasets. The method was tested on a lymphography dataset obtained from UCI machine learning database. The proposed method uses disjunctive pooling rule along with basic probablity assignment in Dempster-Shafer theory to combine the rulesets and assign a measure of interestingness. The evaluation of knowledgebase approximation using lymphography dataset showed that this method is more efficient for insight induction from huge datasets and decreases processing time significantly. It also showed that the association rules from disparate datasets can be fused without the need for integration at the data level. The findings from the application of knowledgebase approximation to the lymphography dataset are not necessarily indicative of knowledgebase approximation efficiency for other datasets. However, these findings helped us to set constrains on the characteristics of the datasets that can be combined with the national collision database of Canada using our proposed framework. At the end, knowledgebase approximation was applied to two different road accident datasets, one from Canada and one from the Great Britain, and fused insights were derived and recorded in a single approximated knowledgebase. The top eight rules, ranked based on their fitness values, are presented in Table 5.5, and show the traffic collision database of the Great Britain had significant contribution to the knowledgebase created in the previous chapter.

Chapter 6

Context-Aware Collision Analysis

6.1 Introduction

This thesis bases the development of context-aware collision analysis on performing in-depth accident analysis and naturalistic driving analysis in a consecutive order. Thus far, the in-depth accident analysis part is fulfilled by developing an insight induction structure followed by a knowledgebase approximation technique which generates association rules containing risk factors of disparate road-accident datasets. Now, it is time to accomplish the second phase, the naturalistic driving analysis, in a way that it can be built upon insights discovered from the previous phase. In other words, we are now concerned with creating a structure that can relate the measured risk factors of a moving vehicle to a comparative level of collision risk by using the information available from the induced insights which are in the form of integrated association rules.

This chapter explores Bayesian network as a potential solution for developing the naturalistic driving analysis structure. This study shows that Bayesian Network is a proper tool for predicting traffic collision risk based on the findings from the data-driven model. Bayesian Network is a relatively new method in the field of artificial intelligence with a variety of applications in reasoning under uncertainty and prediction of highly uncertain phenomena. it is a probabilistic graphical model representing random variables and their conditional dependencies in the form of a directed acyclic graph (DAG). It simplifies the application of Bayesian inference by acting as a comprehensive classifier.

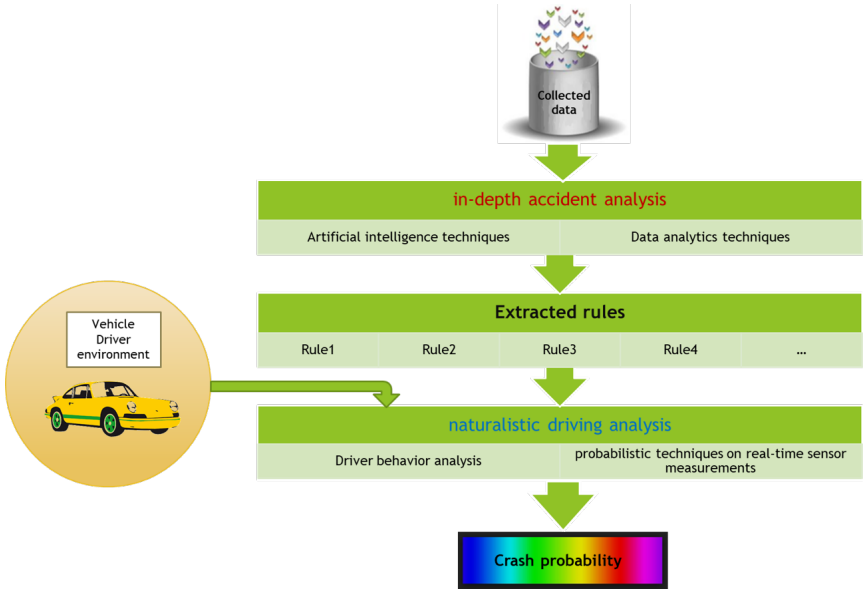

Figure 6.1: The overall stages of crash risk analysis

6.2 Motivation

The data-driven model introduced in chapters 4 and 5 consists of three different processes (segmentation, association rule mining, and insight fusion) and provides insightful patterns that are useful for investigation of collision probability while available databases are rich in collision contributing factors. However, the introduced data-driven model can not instantly generate the collision risk of a vehicle by sensing the situation in which the vehicle is positioned. The likelihood of a crash is significantly affected by the number of potential actions of the driver in a certain situation. Each of these potential actions might probabilistically lead to an accident. Therefore, the likelihood of an accident to occur at a time is the weighted sum of the crash probabilities of possible actions. Bayesian network is capable of structuring this probability model in a simplified directed acyclic graph (DAG).

To elucidate on this fact, consider a sample scenario with a vehicle on a highway. The vehicle in motion can perform a variety of actions at a time including line changing, acceleration, and deceleration. Since these possible actions are conducted by the driver, they lie in the driver behavior category of crash contributors. Knowing that the history of the driving behavior for each person is different, the probability of each action being conducted can be different from person to person. Moreover, the vehicle's surrounding environment can affect these probabilities. Therefore, the presence of the environmental

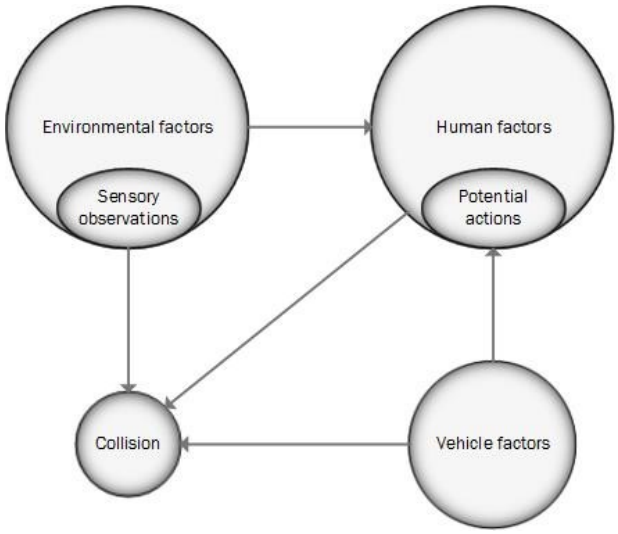

Figure 6.2: A basic DAG for predicting collision events

features can improve the estimation of the crash probabilities.

Crash prediction using static data has limited accuracy. Development of information and communication technologies as well as built-in vehicle sensory measurements can contribute to improving the design of real-time crash prediction models. Bayesian networks have the potential to find the likelihood of an accident given the updated values of all variables and measurements at all times and the relationships between them. The big challenge, however, is to find a way to use the approximated association rules from the previous chapter to learn the structure of the Bayesian network. the association rules obtained by integrating insights of distinct datasets contain the key information of the road-accident historical data, and hence, are plausible for constructing the structure of a Bayesian network that reveals the relationship between crash risk indicting variables in different variations of time and location.

Figure 6.2 illustrates a basic DAG for predicting collisions with respect to the three categories of crash contributing factors. All the three sets directly impact the probability of collision, and therefore, directed links are connecting them to the collision event. Moreover, people might react differently in diverse environmental situations or while driving different vehicles. This influence is represented by the arrows that links them and the collision node in Figure 6.2. Potential actions and maneuvers are included in the human factors category. However, I highlighted them in the model due to their importance in developing the theoretical model. The environmental and vehicle factors are not influenced by other

categories, hence there is no directed links ending on them.

Crash risk analysis can be implemented as a real-time driving alert system. For that purpose, sensory observations from the vehicle are required to report certain risk factors attributed to the environment or vehicle. Therefore, I have highlighted sensory observations in the environmental category in Figure 6.2 to show their significance. These observations have significant impact on the driver's actions and on the collision event itself. Updated weather and traffic data can be also incorporated to better reflect collision risk of the vehicle being driven. These measurements can enable the approximated probability model to combine prior knowledge with observed data and create a hybrid structure based on the rules generated in each segment.

6.3 Bayesian Network

Bayesian network (BN) is a family of probability distributions that admits a compact parameterization that can be naturally described using a directed graph. It can serve as a classification model to predict class membership probabilities for output variables. BN works based on Bayes' theorem and consists of two main components: a directed acyclic graph (DAG), and a set of conditional probability tables (CPTs). DAG's nodes represent random variables and its edges represent probabilistic dependences. Each random variable in the network possesses a CPT to specify the conditional distribution of nodes and their parents.

The power of BN is in the simplicity of calculating probabilities based on the connections in the network. if X is a BN, its joint probability density function can be written as the product of each node's individual density function conditioned by the occurrence of their parent variables. The probability of an event X=x is:

$$p(x) = \prod_{v \in V} p(x_v \mid x_{pa(v)}) \tag{6.1}$$

where v is a node in the directed acyclic graph $G = (V, E)$ with V and E being the set of nodes and links in G respectively, and $pa(v)$ is the set of parents of v. This equation then can be expanded and simplified using the chain rule as follows:

$$
\begin{aligned}
&P(X_1 = x_1, \ldots, X_n = x_n) \\
&= \prod_{v=1}^{n} P(X_v = x_v \mid X_j = x_j \text{ for each } X_j \text{ which is a parent of } X_v)
\end{aligned} \tag{6.2}
$$

BNs are normally represented in the form of A compact Bayesian network, which is a distribution in which each factor on the right hand side depends only on a small number of ancestor variables x_{A_i}:

$$
p(x_i \mid x_{i-1}, \ldots, x_1) = p(x_i \mid x_{A_i}). \tag{6.3}
$$

In a model with five variables, for example, we may choose to approximate the factor $p(x_5 \mid x_4, x_3, x_2, x_1)$ with $p(x_5 \mid x_4, x_3)$, and we can write $x_{A_5} = \{x_4, x_3\}$. Since the risk factor variables in these road accident datasets are discrete, we may think of the factors $p(x_i \mid x_{A_i})$ as CPTs, in which rows correspond to assignments to x_{A_i}, and columns correspond to values of x_i. The value of $p(x_v \mid x_{pa(v)})$ is an entry in X_i's CPT which can be incorporated into Equation (6.1) to calculate the probability of particular events.

6.4 Independencies in a BN

The directed acyclic graph in a BN captures dependency and independency relationships between variables. By using the principle of d-separation [111], the independencies can be recovered from the graph by identifying three types of structures. Any three nodes in an arbitrary BN can only have three possible structures: common parent, cascade, and V structure (see Figure 6.3). Each of these three structures leads to different independent assumptions.

If two variables are d-separated relative to an observed variable in a DAG, then they are conditionally independent on the observed variable in all probability distributions such a graph can represent. In other words, the two variables X and Y are conditionally independent relative to an observed variable Z when knowledge about X provides no information about Y once there is information about Z.

In the common parent structure, the three nodes are in the form of $A \leftarrow B \rightarrow C$. In this structure, if B is observed then A and C are independent, otherwise A and C are dependent. The reason is that having B observed all the information that determines the

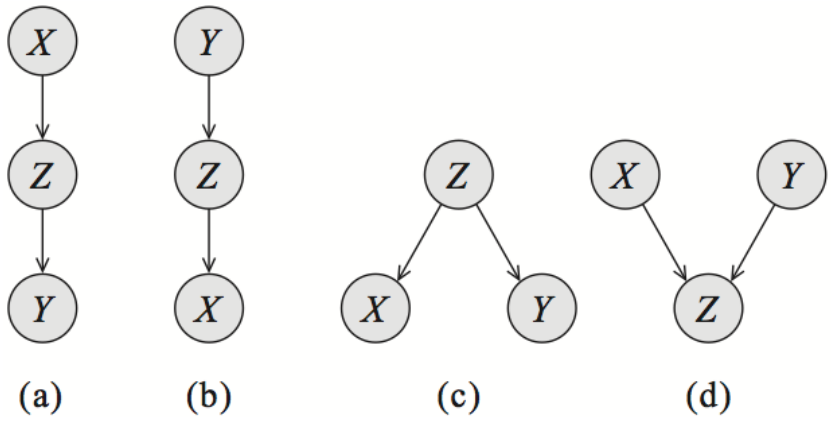

Figure 6.3: Possible structures for 3 nodes in a BN: cascade (a,b), common parent (c), and v-structure (d)

outcomes of A and C are available in B. One can simply refer to the CPTs that connect B to A and C and find the probability values attributed to these nodes. In this case, nothing else can affect A and C's outcomes.

The cascade structure represents the nodes in the form of $A \rightarrow B \rightarrow C$.In this structure, observing B makes nodes A and B independent. we can show this relationship in a brief format as $A \perp C \mid B$. The intuition for this independency is that B contains all the information that can determine C's outcome when B is observed. In this case, no matter what value A takes, it does not affect node C.

The third structure that any three nodes in a BN can have is the V-structure (also known as explaining away) which is in the form of $A \rightarrow C \leftarrow B$. In V-structure, unlike the previous structures, having the middle node, here C, observed makes the other two, A and B, coupled. Intuitively, when there is information about C, knowing the state of any of the parent nodes will affect the probability of the other as both are feeding node C. Therefore, $A \perp B$ if C is unobserved, but $A \not\perp B \mid C$ if C is observed.

The d-separation principle extends the notion of independency to general networks by applying the aforementioned three rules recursively over larger graphs. This principle states that two arbitrary nodes A and B are d-separated, given a set of observed nodes, if they are not connected by an active path. A path is called active given a set of observed variables O if for any connected triple of variables X, Y, and Z on the path one of the following holds:

- $X \leftarrow Y \leftarrow Z$, and Y is unobserved $Y \notin O$
- $X \rightarrow Y \leftarrow Z$, and Y is unobserved $Y \notin O$
- $X \leftarrow Y \rightarrow Z$, and Y is unobserved $Y \notin O$
- $X \rightarrow Y \leftarrow Z$, and Y or any of its descendants are observed.

6.5 BN Structure Learning Using Road-accident Knowledgebase

BNs can be constructed from the knowledge provided by subject-matter experts (also referred to as domain experts). When enough data from the domain is available, construction of BN becomes structure learning of the DAG from data. There exist a number of automatic structure learning algorithms in the literature most of which rely on heuristic search of structures that maximize some scoring criteria and are called score-based approaches. These approaches explore the structure space of the given variables to discover the best candidate that justifies the available data. The criterion defined in the score-based approaches evaluates how well the BN fits the data. A search algorithm is then adopted to find a structure with maximal score. Some examples of score-based approaches are those based on entropy [112, 113], Bayesian scoring [114, 115], and minimum description length [116].

Some research focus on the independence relationships for structure learning in the construction of BN. This approach, called constraint-based approach, employs the independence test to determine a set of constraints for the variables in the network. There are studies that combine both independence and scoring relationships to achieve better tailored Bayesian network structures [115, 117]. Similar to the score-based approaches, a search algorithm is needed to find the best DAG, here the one that conforms to the constraints. The structures described in Section 6.4 (V-structure, common parent, and cascade) can be identified in the network by doing independence test for the two nodes on the sides conditional on the node in the middle. The pitfall for this approach, however, is the amount of data samples needed to guarantee testing power. It works well with some other expert knowledge of structure, but it is not much reliable when not enough samples are available for the test.

Chapters 4 and 5 of this thesis explored and studied insights from the traffic accident datasets in the form of association rules. The approximated knowledgebase and integrated

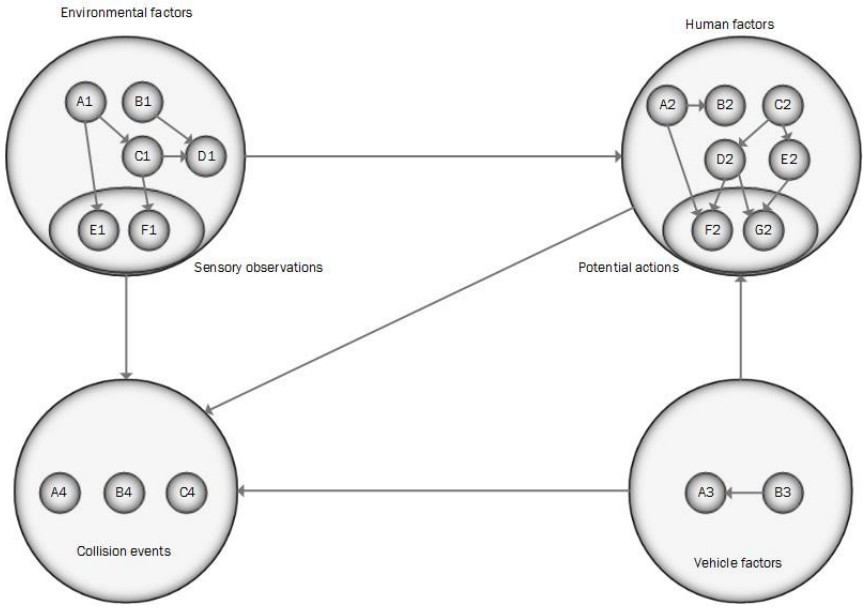

Figure 6.4: An example of a constructed BN indicating risk factor interconnections

association rules provide us valuable information about dependency of risk factor variables. Hence, it is convenient to use these association rules as a means to construct the consequent Bayesian network. Let us assume we have a certain number of rules for a specified segment of data after employing the clustering and insight induction techniques. The variables in the antecedent part of the rules have direct impact on the collision event and their level of contribution can be determined based on the measures of interestingness defined in Equations (4.7) to (4.9). These values can be used to create the CPTs for each node, which define the relationship between that particular node and their parent nodes.

Each of the categories in the triangle of contributing factors contain numerous attributes which might be inter-connected. Figure 6.4 is an example of a constructed BN which indicates these inter-connections and how all of them, as a group, impact the other groups of factors. In this example, three collision events exist and the probability of their occurrence depends on their parent nodes. The connections and relationships in such network can be identified by executing two phases once we have our association rules evaluated and sorted. These two phases are: 1) constructing the initial network by considering local independence knowledge, and 2) revisiting the network to prune the edges that are wrongly added to the network.

Evaluating association rules can be done using any of the interestingness measures

Algorithm 6.1: Constructing BN from association rules

Sort the selected association rules R_i according to the desired measure
for *each association rule* R_1 **to** R_l **do**
  Create the antecedent set $X = \{X_i = x_i\}_{i=1}^{n}$
  Create the consequent set $Y = \{Y_j = y_j\}_{j=1}^{m}$
  for $i = 1$ **to** n **do**
    for $j = 1$ **to** m **do**
      if $\{$*A directed edge from* X_i *to* Y_j *does not exist*$\}$
      AND $\{X_i$ *and* Y_i *are not d-separated*$\}$
      AND $\{$*adding a directed edge from* X_i *to* Y_i *does not create a cycle involving any of them*$\}$
      then
        Place an edge from node X_i to node Y_i

including support, confidence, and lift. Other measures obtained by combining these measures can also be used, as I did in Chapter 4 to restrict the number of rules. Combined measures are customized to reflect the contribution of each of the primary measures based on their importance in the context that they are used for. Sorting the rules are also of great importance in the proposed BN constructing approach as the algorithm for phase two adds connecting edges in a iterative manner. Adding further edges in the later iterations of the algorithm is highly dependent to those already added to the network, and hence, it is crucial to feed the rules to the function in the correct order.

After identifying and sorting the association rules, the initial network topology construction phase can be initiated. This phase uses the local independence information to add new edges to the network as indicated in Algorithm 6.1. One edge can be a potential candidate if it is part of the association rule that is under examination. The nodes of that edge should first go through the Independence test to make sure they are not d-separated. If they are, then they can not be directly connected as it would contradict their d-separated relationship. If they are not d-separated, they should go through another test to examine whether the corresponding edge creates a cycle in the graph or not. Since DAG is an acyclic graph, the edge can not be added if its presence creates any cycles. This is the step where ill-sorted rules can impact the structure of the network as the incorrect edges added earlier may prevent formation of other edges if it creates a cycle with them.

It is possible that unnecessary or wrong edges are placed in the process of network construction. Unnecessary edges can increase the computation complexity of the BN and

Algorithm 6.2: Pruning edges from initial BN

Data: A directed acyclic graph $G = (V, E)$
for *each edge* $e \in E$ *connecting nodes* v_i *and* v_j **do**
    Remove edge e temporarily
    if v_i *and* v_j *are d-separated* **then**
        insert e back to the network
    else
        permanently remove edge e

wrongly placed edges can impact the outcome of the nodes that we are interested in their values. Since we have constantly checked directly connected nodes for independency and cycles, we do not need an extensive pruning procedure. Hence, we use the d-separation test again to remove the edges that do not impact the independence relationships of the nodes in the network. The pruning algorithm is shown in Algorithm 6.2. In this algorithm, each edges is temporarily removed from the network and its two nodes are tested to be d-separated without that edge. If the two nodes become d-separated, that edge has significant importance in the determining the dependency of nodes in the constructed graph. In the case, the edge is inserted back to the structure of the network to maintain the flow of information. In the case where the nodes become not d-separated, the edge is permanently removed.

Figure 6.5 is an example of the BN constructed from the top four association rules that were mined in Section 5.6 as part of the approximated knowledgebase. Three different variables are observed as the consequent parts of these four rules: C_SEV, P_ISEV, and C_CONF. These variables indicate the road-accident configuration and its severity. The road configuration at the collision location is also playing a role in the antecedent of some rules and affects the severity of accidents. Descriptions of the variables, and the number of unique values they can take, are given in Table 4.1. There are three new variables that are fused into these rules and are originated from the GB dataset. These variables are the speed limit, presence of pedestrian, and vehicle's maneuvering. The description and values of the GB dataset's variables are presented in appendix B.

This figure is just an example of how the network is produced from association rules and it is trivial that we need to input more rules to the algorithms 6.1 and 6.2 in order to obtain a solid context-aware prediction system. Each rule that is added to the selected set may grow the network depending on how much new information it presents that connects the variables in the antecedent and consequent sides. We assume that the network expansion speed

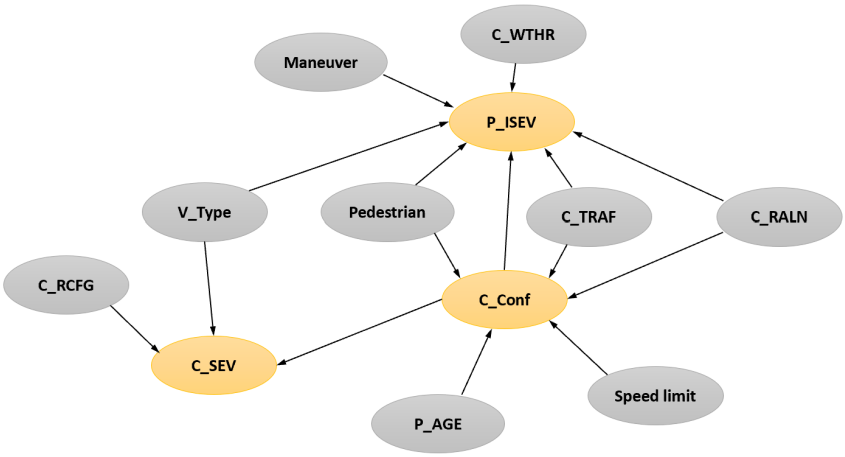

Figure 6.5: The Bayesian network constructed from Section 5.6's top four rules

will decease after a certain number of rules as the variable connections have considerable overlapping between different rules. Moreover, the less interesting rules may introduce new connections that may impose cycles with those of the more interesting ones, and therefore, can not be added to the network.

The beauty of this method is that it inherently has a knowledge integration procedure with it. Each new rule brings more variables and more connections to the network layout and the new rules can be extracted from distinct datasets. We just have to be mindful of doing proper pre-processing on the datasets in use to match the name of the identical variables and their identical values. When the structure of the BN is shaped, finding the CPT tables corresponding to each node of the graph is what remains to complete the BN. The probability values contained in CPTs can be calculated through running typical learning algorithms on the dataset.

In the next section, I evaluate the performance of the proposed BN construction method. It is worthwhile to mention that the main reason for using association rules in the proposed BN construction method is that the information from different resources can be merged into a single collision risk analysis system. To the best of my knowledge, only a few structure learning algorithms work on multiple datasets, and others can only be used if those multiple datasets are integrated in the data level. I have previously discussed the challenges and drawbacks of merging datasets at the data level. I proposed knowledgebase approximation to overcome those drawbacks and created a union of collision patterns in the form of association rules. The proposed BN network construction method in this chapter is intended to build the network from those reinforced association rules. This method can

also be applied to the association rules extracted from a single dataset and may be used as a substitute for other structure learning approaches.

6.6 Performance Evaluation and Interpretation of Results

Computation, analysis, and inferences of a single dataset are rarely extended by inferences obtained from other datasets. The reason is that multiple scenarios can happen when disparate datasets are recorded to represent a similar concept. In some scenarios, structure learning can not be applied to more than one dataset because of the contrasting joint distributions imposed by selecting different controlled or observed variables. For example, assume that we are observing a set of variables $X = \{A, B, C\}$ to learn a predictive or diagnostic model for one of those variables, say A, based on the remaining variables. The joint distribution in this case is different from when variable B is controlled to study its effect on another variable A. The associations between A and C are different in these two scenarios and hence the data can not be merged.

The situation can get more extreme in the case study of this thesis's interest. Dealing with road-accident datsets, we encounter a scenario in which each dataset is measuring different variables. Even for those variables that are the same, they may be semantically similar but not necessarily identical. One example is when the same quantity is measured using different scales and methods with no apparent mapping from one to the other.

The big advantage of the proposed approach is that we do not need to be concerned about applying the structure learning method directly to the dataset. The connections are concentrated in the association rules and we can use our selection of rules to build the BN structure on their basis. The datasets are only used later for generating the CPT tables. The challenges about integration of knowledge from multilpe datasets are addressed in previous chapter and here I just focus on how accurate the consequent BN functions.

In this section, I evaluate the performance of the proposed structure learning method on the NCDB and GB datasets. The approximated knowledgebase ruleset created from the NCDB and GB datasets in previous chapter are used for this purpose. I also compare the performance with the method introduced by Tsamardinos et. al. [118]. Tsamardinos states that unlike pairwise correlations, pairwise causal relations are transitive, meaning that if A is causing B, and B is causing C, then A is causing C. They use this inference along with some other inferences to induce causal knowledge from multiple datasets.

Table 6.1: Evaluation of BN construction from association rules (ARs) and comparison with multi-source causal analysis[118]

predicted variable	**reported-value type**	**structure learning from ARs**	**Multi-source causal analysis**
Collision severity	#correct predictions	15962	13080
	%correct predictions	79.81	65.4
Casualty	#correct predictions	14650	12720
	%correct predictions	73.25	63.6
Collision configuration	#correct predictions	13686	12212
	%correct predictions	68.43	61.06
Collision likelihood	#correct predictions	16260	15138
	%correct predictions	81.3	75.69

Table 6.1 summarises the prediction accuracy of the proposed structure learning method and compares it with multi-source causal analysis. In this experiment, 50000 entries were used to train probability values of CPTs and 20000 entries to test the model. I selected four variables to be predicted from other variables: collision severity, casualties, collision configuration, and collision likelihood. These four variables are chosen since they can help calculate the risk of collisions in a given driving situation while enough information about the observed variables are available.

In order to better reflect the accuracy of these methods, I used 10 rounds of cross-validation using different partitions and combined the validation results over the rounds. This way, a more accurate estimate of model prediction performance is reported. Considering the number of train and test samples that is used, 71.5% of the data are for training the network and 28.5% for validation. By using cross validation on standard-sized partitions of train and test data, problems like overfitting or selection biases are mitigated and better insights on the model's generalization power is provided.

The number and percentage of correct predictions for the chosen predicted variables are reported in Table 6.1. The results show that structure learning from ARs performs significantly better than multi-source causal analysis when two road-accident datasets are used to build the network. The reason is that the proposed method uses more sophisticated relationships between the variables. Indeed, multi-source causal analysis does not consider the patterns including a mixture of variables from disparate sources. Instead, it tries to splice the causal inferences from those nodes that appear in both sides. They decide about the points of splice by analysing direct causes, common latent causes, and existence of consistent causal graphs.

Among the four predicted variables, collision likelihood is a latent variable and was manually added to the experiment's dataset as this variable is not reported in the road-accident datasets. I added the likelihood of collisions to each sample by measuring how frequent that specific collision configuration happens. The other three variables have measured values in the road-accident datasets and can be predicted as a node of interest in the BN. Latent variables are not directly observed but make understanding the data easier. We can refer to collision likelihood as a hypothetical variable as we never know all the data points in the accident analysis that are exactly the same to discover the ratio of collision occurrence. Even if we could do that, the ratio is so small that may not seem significant. Since this research cares about the contribution of variables in occurrence of traffic accidents, I use the count of appearance of certain patterns in a collision dataset to estimate the likelihood of crashes. The values obtained from this approach represent the percentage of accidents in which those patterns are present.

Since integrated association rules are used to construct the Bayesian network, using samples from individual datasets will lack information about those variables that are not observed in that specific dataset. I addressed this issue by neutralizing the effect of those variables in the outcome of the predicted variables. In the BN's flow of inferences, if we encounter a node that is not observed, we assume equal probability for occurrence of the descendants of that node. In other words, we assume a full degree of ignorance for that variable instead of supplying prior probabilities.

Constructing causal models by using large-scale integration of data is not an easy task. The automated, or semi-automated, integrative analysis of multiple data sources should be able to handle data obtained over different experimental conditions, different variable sets, and semantically similar but not identical sets. The empirical results are indicative of how structure learning using ARs can increase learning performance of causal relations compared to a typical multi-source causal analysis. [118] provided evidence that multi-source causal analysis increases learning performance compared to learning from individual datasets. Therefore, It can be said that the proposed structure learning method outperforms both individual dataset analysis and learning from combined multiple sources.

Risk of a traffic collision, in this context, is the possibility of loss, injury or fatality. I have assumed fatality as the highest type of loss. So, if a driver is in a situation that matches the characteristics of a frequent type of accident whose severity is high and mostly contains fatality, that driver is experiencing the highest level of risk. If we show the level of risk by the colors illustrated in the output of Figure 6.1, then the mentioned driver would be notified as being in a red or purple zone. Collision likelihood is Representative of the chance of being involved in an accident but does not delineate the hardships imposed to the people involved. This variable is reported along with collision severity and number

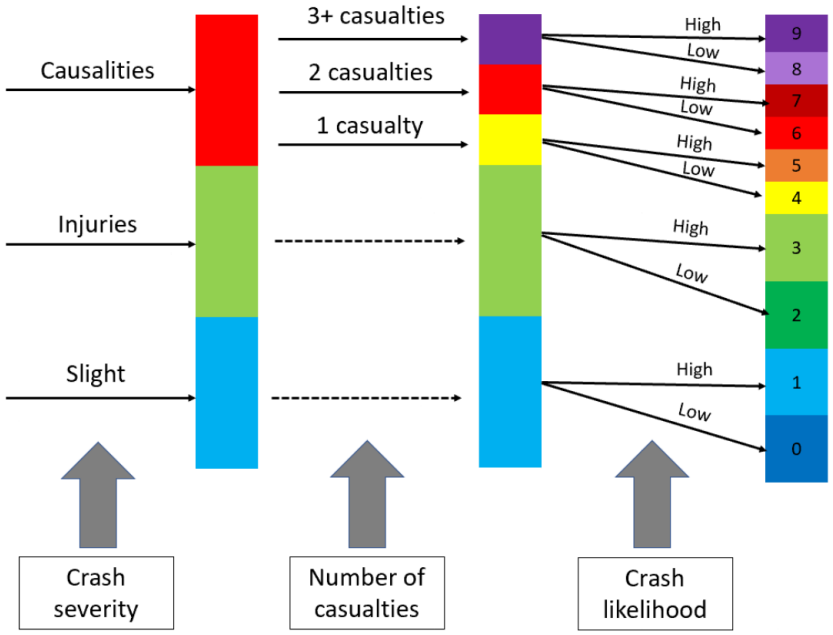

Figure 6.6: Transforming linguistic values of crash severity, number of casualties, and crash likelihood to risk level

of casualties to better picture the risk of the potential accidents that drivers are prone to experience.

The outputs of the Bayesian network are used to generate risk levels. Having the risk level, drivers are notified if the risk is too high to be alert and prepared. Figure 6.6 shows one way to use the output of the Bayesian network and transforming them into a risk level spectrum. Collision severity and number of casualties are used along with the collision likelihood to form levels of collision risk. Collision severity includes three classes of fatalities, injuries, and slight. The number of casualties shows the number of people died as a result of a crash. The risk level, as crisp value, is considered to be from 0 ro 1, but here the linguistic values of collision likelihood, collision severity, and injury level are used to generate a risk level. Based on the definition of risk, collision likelihood should act as an scalar that can change the risk degree obtained from the total severity. The total severity itself is first partitioned into three degrees(High, Medium, Low) based on whether the accident contains fatalities, injuries, or none of them. Then the number of casualties divides the High level into three new levels based on whether the accident had one, two, or more casualties. Eventually, the crash likelihood, which can be indicated as high or low,

doubles the number of risk levels by splitting each of them into two. This way ten different risk levels are obtained which are numbered from 0 to 9, as seen in Figure 6.6.

It is trivial that a particular pattern of variables in the NCDB and GB datasets may not always result into a single severity level. This is because the severity and casualties of an accident depends on numerous variables, many of which are absent from these datasets. The effect of hidden variables are minimized by applying a segmentation method in Chapter 4, however, we can not completely isolate them as they are, to a great extent, influencing accidents' severity. The exact make and model of the vehicle, for example, is highly correlated to the severity as different vehicles have specific safety factors. The datasets in this experiment have addressed this correlation partly by including the model year. Car manufacturers intensify their products' safety every year and it is not a bad assumption to associate the model year to safety and the damage that those cars can avert.

6.7 Conclusion

This chapter introduced a Bayesian-network-based approach for implementation of naturalistic driving analysis on association rules. The directed acyclic graph of a Bayesian network has the potential to work with real-time measurements fed into a collision prediction model which brings a great flexibility to the whole system. It is worthwhile to mention that we did not test the Bayesian network with actual real-time measurements. However, the structured and trained network can be provided with such measurements to both test and improve the model's efficiency in real driving applications. The chapter introduced a methodology to learn the structure of the Bayesian network from association rules, which is one of the contributions of this thesis. The great advantage of using association rules in the proposed Bayesian network construction method is the ability of the consequent model in merging the information from different resources as a single collision risk analysis system. There is still extensive research going on about structure learning algorithms that can work on multiple datasets. Previous studies mostly rely on integration of datasets in the data level which entails numerous challenges and drawbacks. The proposed method here overcomes those drawbacks by constructing a Bayesian network using the union of collision patterns in the form of association rules. The whole process brings combined knowledge to a risk analysis system that accepts real-time measurements from the vehicle, the environment, and the driver. Spatial and temporal aspects can also be integrated into this system to create a compact temporospatial context-aware collision analysis system. This method can also be applied to the association rules extracted from a single dataset and may be used as a substitute for other structure learning approaches.

Chapter 7

Traffic Incident Detection and Localization

7.1 Introduction

Traffic incident detection and localization is an important application in traffic management systems. The ability to detect and localize traffic incidents enables a timely response to accidents and facilitates effective and efficient traffic flow management. This chapter presents a sensor-network based approach for tackling the problem of incident localization. Traffic count sensors, which tend to be an element of the road infrastructure, are used as the source of traffic sensory data. Such sensors come in a variety of types and capabilities, providing the potential for complementary and redundant information gathering. Thus, it is conceivable to fuse such sensory information to achieve insightful and accurate incident detection and localization. In this context, the Dempster-Shafer (DS) theory of evidence is used as the foundation for fusing traffic sensory data. In this study, a traffic model generator and two traffic-counting sensory systems are employed for acquiring traffic data pertinent to the distribution of cars on a given road segment. The incident localization performance of the DS fusion technique is compared to the ordered weighted averaging (OWA) operator [119] and the sole sensory systems. Experimental analysis on the performance of the proposed approach is provided.

7.2 Motivation

An important indicator of survival rates after occurrence of an accident is the time between the accident and when emergency medical personnel are dispatched to the scene. Reducing this time decreases mortality rates by 6%[120]. One approach for decreasing delay uses automatic incident detection and notification systems that sense the time that traffic accidents occur and immediately notify emergency personnel. Moreover, drivers approaching the site can be notified and possibly change their route or at least slow down for safety. Consequently, these systems save time, improve road safety, and reduce mortality. One important piece of information that should immediately be reported to rescue services is the exact location of the incident, without which dispatching a rescue team becomes useless. Hence, incident localization is an essential part of automatic incident detection and notification systems. This study describes automatic incident localization achieved through simple steps and inexpensive practices utilizing sensor fusion in a sensor network.

Sensor networks are widely used to obtain data about vehicle surroundings, traffic flow, weather conditions, driver behavior and other factors affecting the collision-prediction process. However, the complexity of such networks poses major challenges, including noisy and erroneous measurements, incomplete and low-quality sensor data, interference of network nodes, and big volumes of data, to name a few. Based on the structure of the sensor network, it seems promising to exploit sensor data fusion methods to overcome the above-mentioned obstructions. In this study, two sensory systems obtain information from the environment. One of these systems is assumed to be noisy. The goal is to employ a data-fusion method to exploit the information from both systems so as to obtain more-accurate estimation of the collision location.

The classical inference for data fusion is based on Bayes criteria. The Bayesian inference is used to estimate the degree of certainty of multiple sensors providing information about measured data. It uses an a priori probability of a hypothesis to produce the a posteriori probability of this hypothesis. Some limitations for Bayesian criteria are [106]:

- no representation of ignorance is possible
- the prior probability may be difficult to define
- the result depends on the choice of prior probability
- the inference assumes coherent sources of information
- it is complex with a large number of hypotheses

- it has poor performance with non-informative prior probability

Another method, Dempster-Shafer (DS) theory, overcomes many of the classical inferences for data fusion. DS is often described as an extension of probability theory or a generalization of the Bayesian inference method. It is based on obtaining degrees of belief for one question from subjective probabilities for a related question. Dempster's rule also can combine such degrees of belief when they are based on independent items of evidence[121]. The method assigns masses (weights) to the subsets of the entities that constitute a system and calculates the confidence measure of each possible state, based on data from new and old evidence. Some highlighted features of this method are:

- it can be used without prior probability distributions
- slight changes in the input influence the output
- it is highly efficient with bodies of evidence in pseudo-agreement
- ability to deal with ignorance and missing information

This study employs DS data-fusion technique and evaluates its performance with regard to the quality of transmitted data and the estimated location of a collision. The performance of this method will also be compared with that of the ordered weighted averaging (OWA) fusion technique. OWA, first introduced by Ronald R. Yager in 1988 [119], is another operator in applied mathematics used to aggregate data. A mapping is called an OWA Operator of dimension n if it has an associated weighting vector W in the range $[0, 1]$ subject to a summation of 1 for all the weights.

I use MATLAB to generate the proposed traffic model and data fusion techniques, and the MATLAB fuzzy toolbox to get the output of each individual sensory system. Model performance, evaluated by several measures, confirms the accuracy of the proposed collision location estimator.

7.3 Related Work

Numerous methods have been proposed in the literature on Intelligent Transportation Systems(ITS) used to detect and predict road events. Intelligent systems require sensing technologies to get a perception of the surrounding environment. In [120] and [122], the

authors used acoustic signals as the input for their automatic traffic accident detection and notification systems. Acoustic signals can simply be collected by smartphones, especially now that everyone owns at least one smart device. These devices can be triggered to detect high-decibel acoustic events and so pinpoint the occurrence of accidents. However, triggering the built-in microphone to avoid false-positives is challenging.

Zhang et al. explained in [122] that a secondary sensory system is incorporated that detects variations in acceleration to lower the probability of false-positives. They used a client/server application that relays accident information to the server via HTTP and provides an interface that allows third-party observers to access reported data. For feature extraction from the acoustic signals, digital-signal-processing algorithms can be quite useful. The authors compared the results of Discrete Wavelet Transform (DWT), Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), Real Cepstral Transform (RCT) and Mel Frequency Cepstral Transform (MFCT), and incorporated statistical classifiers such as nearest mean, maximum likelihood, and nearest neighbors. They concluded that the maximum likelihood classifier in conjunction with RCT gives the best performance in low signal-to-noise ratios (SNR). However, for high SNRs, DWT and RCT are comparable in accuracy but DWT is computationally more efficient. The downside of using digital signal processing algorithms is the high computational load, which makes them hard and expensive to implement in real-time scenarios.

A real-time accident detection scheme, introduced by Sherif et al. in [123], uses WSN and RFID technology to inform authorities about accidents through a wireless interface. In this case, an embedded hardware board is required in the vehicle to check the status of airbags, gather all the essential information, prepare it as a packet, and send it to the information center. The issue with this technique is that all the vehicles need to be equipped with external hardware, which is quite expensive to buy and install.

A less costly idea is to use what already exists and is being operated. The contents of social media contain a lot of information which can be used for different purposes. [124] and [125] propose a social media-based traffic status monitoring. In [124], Fu et al initiated their system by generating related keywords and then applied iterative query expansion algorithm as an association rule to extract real-time transportation related tweets. They also incorporated a summarization algorithm to eliminate redundant tweets. In [125], the authors investigated the potential of twitter in real-time incident detection in the United Kingdom(UK). They utilized the Support Vector Machine (SVM) algorithm as a classifier an achieved the overall accuracy of 88.27%. When we consider social media as a data center for event detection, we should also note the countries or cities of implementation. Not every place can employ this method, like those countries whose access to twitter is banned by their government.

The other standpoint is to study the behaviour of drivers and find the likelihood of an incident in a specific intersection. In [126], Popsecu et. al studied the distance and the time for changing lanes to be used as extra information to the already in use systems for ITS. They asserted that there is no need to use the information reported by every single vehicle because the traffic data for the vehicles in the same vicinity are highly correlated. [127] introduces an approach to track each vehicle from the images of an intersection to identify the events resulted from a chain of behaviours. The authors proposed a spatio-temporal Markov random field to tackle the issue of tracking vehicles with occlusion effect. It determines the state of each pixel in an image and its transits along time and x-y axes. In fact, the algorithm acts upon each pixel whether it is assigned to vehicle A or B. They could demonstrate a success rate of 93%-96% to track multiple vehicles at intersections with occlusion effect. In another study, Kinoshita et al. [128] addressed the issue of distinguishing traffic congestions caused by reasons other than incidents. They used a probabilistic topic model to describe traffic states, and analyzed the differences between congestions caused by incidents and usual congestions based on the probe-car data.

Not many algorithms for the ITS has exploited data fusion methods. Fusion techniques can be a powerful asset to increase the accuracy and efficiency when the infrastructure of the sensory systems and the detection/prediction algorithms has been established. Although not all the sensor fusion methods in the literature are employed for collision detection, there are some good examples of their usage in similar works which help to find the ideal model for collision detection. Faouzi et. al [129] conducted a general survey on data fusion techniques in different areas of ITS and the challenges that still needs to be addressed. In [130] Otto first addresses the problem of environment perception and situation evaluation in Advanced Driver Assistance Systems (ADAS) in commercial vehicles. Five kinds of sensors had been utilized as a sensor network. A monocular camera is exploited as optical detection and classification of pedestrians. It spots the pedestrians just behind the windshield and a few centimeters above. A short-range radar sensor and a long range one are combined and positioned below the license plate for detection modes of long range and short range. A blind spot radar monitors the vehicle and is mounted on the lowest right step tread utilizing one antenna for transmission and four antennas for receiving the echoes. Additionally, a laser scanner plays the role of reference for pedestrian tracking. At the end, an Extended Kalman Filter with Joint Integrated Probabilistic Data Association (EKF-JIPDA) is developed to fuse the data obtained from the sensors and track pedestrians from a truck in real-time. The accuracy of the results is evaluated and compared to another data fusion technique: Extended Kalman Filter with Global Nearest Neighbor (EFK-GNN). The results demonstrated that EKF-JIPDA outperforms the EKF-GNN significantly, especially in crowded places. The author adds that the system might act

faulty in some situations such as distinguishing various objects, like pedestrians and other non-relevant objects, and identification of object with relatively little distances.

Using data fusion techniques for tracking multiple targets in a cluttered environment is the challenge addressed by Zahir et al. [131]. The suggested algorithm for data fusion is to use Cheap Joint Probability Data Association (CJPDA) and multiple model particle filter (MMPF). The MMPF is used to perform nonlinear filtering with switching dynamic models and the CJDAF is used to estimate the joint measurement-target probability association. Additionally, two fusion schemes of Federated Kalman Filtering (FKF) and Centralized Kalman Filtering (CKF) are compared to standard sequential Kalman filtering. The comparison of the three fusion schemes showed that CKF works better than FKF, which was expectable as FKF simplifies the dependency of inputs. However, the rationale behind using these specific filters and method has not been explained.

Although this study focuses on data fusion for traffic related application, data fusion techniques are proved to be useful in other fields as well. Ramirez et al. [132] proposed two algorithms for forest fire detection. The first algorithm used a threshold method when nodes are equipped with temperature, humidity, and light sensors. The second algorithm uses the DS theory for data fusion and assumes the data are from two sensors measuring temperature and humidity. They showed that both methods can detect fires in their initial stages. However, both algorithms reported false positives when the sensors were exposed to direct sunlight.

To address the problem of counterintuitive results when pieces of evidence highly conflict in DS theory Yuan et al. [133] took both statistic and dynamic sensor reliability into consideration. They suggested combining the evidence distance function and the belief entropy to obtain the dynamic reliability of each sensor report. Then, a weighted averaging method is adopted to modify the conflict evidence by assigning different weights to the evidence according to sensor reliability. The proposed method has better performance in conflict management and fault diagnosis since the information volume of each sensor report is taken into consideration.

In certain scenarios we deal with qualitive data, for example, when we want to consider the data reported by witnesses of an incident. A data fusion model is proposed by Golestan et al. [134] that is capable of combining the data generated from human-based sources with those generated by physical sensors. In their proposed model, the unstructured soft data is presented by undergoing a novel soft-data-association process through which the data is semantically analyzed and accurately structured in a fuzzy random variable. They have shown that their model is capable of handling both soft and hard data.

7.4 Methodology

In this section, I present the details of the techniques and algorithms that are used for accurate localization of motorway incidents. The block diagram in Figure 7.1 illustrates an overview of the system model proposed in this thesis. I developed an extendable traffic model generator to generate the required input data for the purpose of evaluating the proposed model. The sensory systems are a sequence of traffic counter sensors that complement each other for full coverage of vehicle counts in a particular section of the road. They collect data from the traffic model, measure the locations of vehicles, and produce a sequence of vehicle count numbers. This section also introduces a Fuzzy-based sensory system that enables individual sensory systems to predict the location of an incident. The main contribution lies in Figure 7.1's fusion block. This block employs Dempster-Shafer evidence theory for sensor data fusion and exploits the processed information from all of the sensory systems to obtain a more accurate prediction of the incident's location. The performance of the proposed method is evaluated and compared with individual sensory systems and another fusion method (i.e., OWA method) through evaluation and comparison blocks. Each of the blocks in Figure 7.1 will be further elucidated in this section.

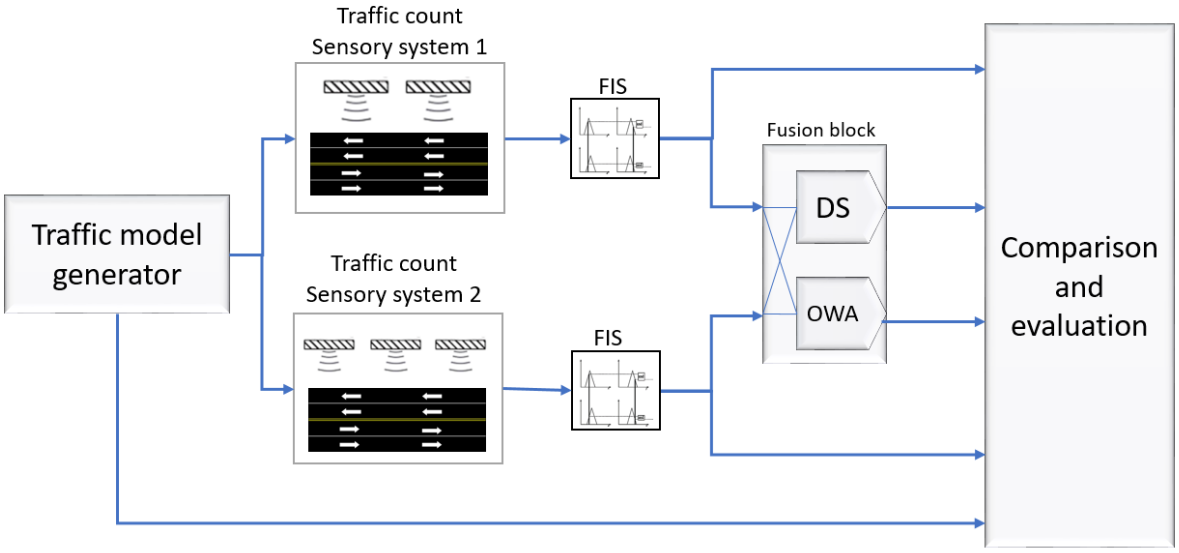

Figure 7.1: Road incident localization model

7.4.1 Traffic Modeling

A traffic or transportation network is a spatial network structure that allows vehicular movements. A complete mathematical model for the traffic flow represents the interaction between vehicles and the network infrastructure such as highways, traffic signs, traffic signals and control and information systems. Three variables in particular represent the traffic: flow, density, and speed. The traffic flow (q) is the number of vehicles per unit time; the traffic density (ρ) is the number of vehicles per unit length; and the speed (v) is the distance covered per unit time, which is dependent on flow and density.

Traffic models may vary based on the level of details being simulated. Depending on the purpose of the simulation, different models can be generated. For example, traffic models can be run either in a continuous approach or a discretized approach [135]. The primary difference between these implementation platforms is the method used to represent traffic flow. At a high level, platoons of cars are modeled rather than individual cars, while low-level models are concerned with the behavior of each vehicle on the network.

This chapter deals with the accuracy of predicted incident locations using sensory-system measurements. Hence, a low-level traffic model satisfies the requirements of the system. The model arises from a traffic model generator that uses two distribution functions to provide different scenarios. These scenarios are later used to evaluate the performance of the proposed technique under various configurations.

A four-lane motorway is assumed here, but the method is applicable to roads with any number of lanes. I represent the road with a grid in which each row represents one lane of the road and each cell is a location that may be occupied by a car. When a collision happens in a motorway, congestion increases around the location of occurrence. Figure 7.2 illustrates the street model when there is no collision. Suppose a collision occurs at a random longitudinal location. Then, the congestion increases in the vicinity of the collision point, and the further we go from that point, the lighter the traffic flows. To model this behavior, two probability distributions -Normal and Lognormal- are considered for the location of cars. In probability theory, Normal, namely Gaussian distribution, is a widely-used continuous probability distribution. Gaussian distribution is used in natural and social sciences to represent random variables with unknown distribution. Physical quantities that are expected to be the sum of the number of independent processes also tend to have a Gaussian distribution. The probability density of the distribution is:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{7.1}$$

where:

- μ is the mean or expected value of this distribution. Because in normal distribution expected value is equal to the mode and the median, μ can be also regarded as the median and the mode of this distribution.
- σ is the standard deviation, which is a measure quantifying the amount of variation or dispersion of a set of data values. A low standard deviation means the points tend to be close to the expected value. In contrast, a high standard deviation is an indication of a wider range that data points are spread over.
- σ^2 is the variance.

Normal distribution is usually denoted by $N(\mu, \sigma)$. When defining a normally distributed random variable X, it is indicated as follows:

$$X \sim N(\mu, \sigma) \tag{7.2}$$

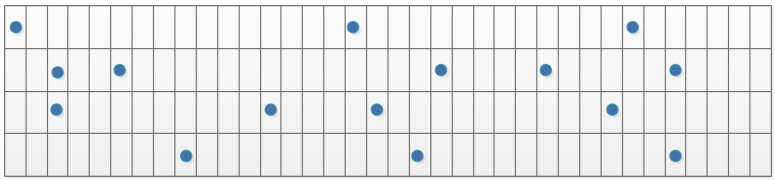

Figure 7.2: No Congestion

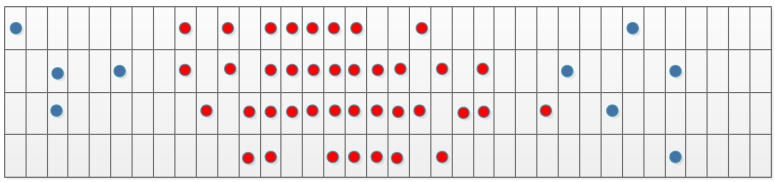

Figure 7.3: Congestion (Collision at location 16)

A lognormal distribution is also a continuous probability distribution of a random variable whose logarithm is normally distributed, meaning that if a random variable X is

lognormally distributed, then $Y = ln(X)$ has a normal distribution. In the same way, if Y has a normal distribution, then $X = e^Y$ has a lognormal distribution. The distribution is also referred to as the Galton distribution. A lognormal process is the statistical realization of the multiplicative product of a large number of independent positive random variables. Additionally, the lognormal distribution is the maximum entropy probability distribution for a random variable X, for which the mean and variance of $ln(X)$ are specified. With μ and σ being respectively the mean and standard deviation of a variable's natural logarithm, a lognormally distributed random variable X can be defined as:

$$X = e^{(\mu+\sigma Z)} \tag{7.3}$$

where Z refers to a standard normal variable.

The probability density function of this distribution can be written as:

$$\mathcal{N}(ln(x); \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(ln(x)-\mu)^2}{2\sigma^2}\right) \tag{7.4}$$

Based on empirical evidence, any collision increases the density of cars in both directions, those approaching the site slowing and those at it or leaving resuming normal speeds only gradually. Both normal and lognormal distribution can model this behavior. As opposed to normal distribution, lognormal distribution considers the fact that vehicles leaving the site move faster than those approaching it, and so proposes asymmetric distribution. Moreover, based on central limit theorem, when independent random variables are added together, their normalized sum tends toward a normal distribution even if the original variables are not normally distributed themselves, and lognormal processes are the statistical realization of multiplicative product of many independent random variables. By opting for these distributions, we are incorporating the effect of other random variables.

Figure 7.2 demonstrates the output of traffic generator when a collision occurs at location 16. When the model generates more than one car in a single location, it means the cars are in multiple lanes. However, this number should not exceed the number of lanes, which is four in this study. To sum up, the traffic model generator provides the locations of a specific number of cars when a collision happens at a certain random point. Similar to any other mathematical models, this model entails some discrepancies between the generated traffic model and the real world. For example, when an accident occurs, the number of blocked lanes may vary and the distribution of the vehicles changes accordingly. However, I disregard these scenarios in the scope of this thesis.

7.4.2 Fuzzy Inference System (FIS) Based Sensory Systems

Sensor modules are required for perception of the environment. A sensor module is defined as a component consisting of a sensor and a data processor enabled for algorithmic calculation. The latter provides interpretation or perception services based on the sensor data. This is often also referred to as an intelligent sensor. The sensor modules used in traffic and driver assistance systems are diverse. In this study, any vehicle sensing technologies like inductive loops can be responsible for obtaining the count of the vehicles in the sensor range. Two sensory systems are used here: one a combination of two sensor modules and the other a combination of three. Both sensory systems cover a specified length of the road with the capacity of L vehicles in each lane. The configuration of these sensory systems is illustrated in Figure 7.4, which is based on the traffic model. The sensory system with three sensor modules is expected to have a better accuracy in getting the location of the collision. However, sensors of this system are assumed to be noisy. To do so, the output of each sensor in this system is altered by summation of the actual count and a random integer between -20 and 20. The sensory systems used in this study are integrated with a fuzzy inferencing system which enables them to predict the location of an incident based on their own perceived information.

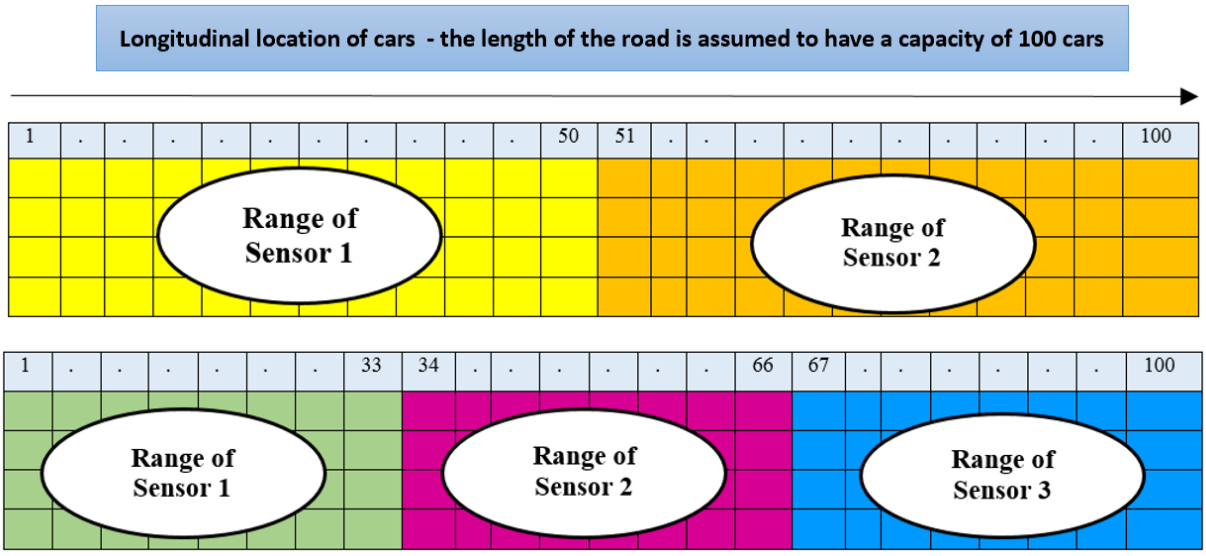

Figure 7.4: Configuration of sensors in sensory systems

FIS is a system that uses fuzzy set theory for mapping. This study employs the FIS proposed by Mamdani[136] for the sensory system decision makings. An example of Mam-

dani's FIS is shown in Figure 7.5. To compute the output of this FIS, one must go through six steps:

1. determining a set of fuzzy rules,
2. fuzzifying the inputs using the input membership functions,
3. combining the fuzzified inputs according to the fuzzy rules to establish a rule strength,
4. finding the consequence of the rule by combining the rule strength and the output membership function,
5. combining the consequences to get an output distribution, and
6. defuzzifying the output distribution.

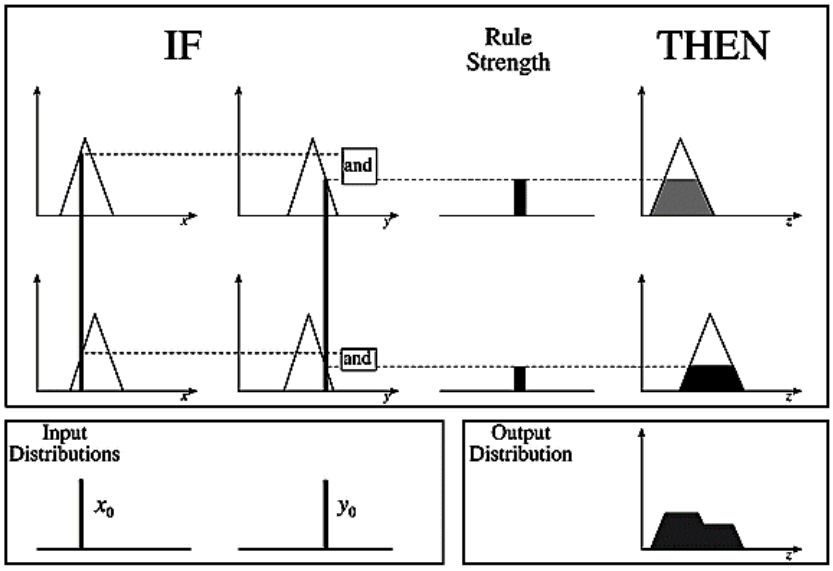

Figure 7.5: Mamdani fuzzy inference system

Fuzzy rules are a collection of linguistic statements that describe how the FIS should make a decision regarding classifying an input or controlling an output. Fuzzy rules are always written in the form of if-then rules. For incident detection using the information gathered by introduced sensors, I introduce membership functions to define the meaning of high, medium and low. The process of taking an input such as quantity of cars and

processing it through a membership function to determine if it belongs to "high", medium, or low is called fuzzification. Sensor measurements and the incident locations are fuzzified using triangular membership functions. The maximum vehicle count from the first sensory system is 200 (i.e., the capacity of the section of road observed by each of its sensors), in other words, 4 lanes times 50 (50 being the capacity of one lane in the range of each sensor). The maximum vehicle count for the second sensory system is 132, or 4 lanes times 33 (33 being the capacity of one lane in the range of each sensor).

After the fuzzification step, the fuzzy systems in Figure 7.6 and Figure 7.7 are used as the core of each sensory system to predict the collision point.

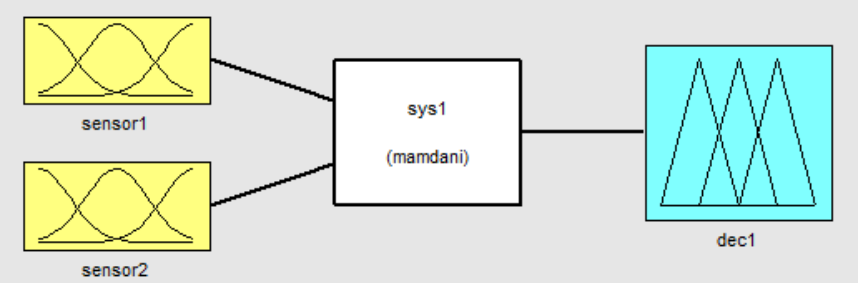

Figure 7.6: Fuzzy model of system 1

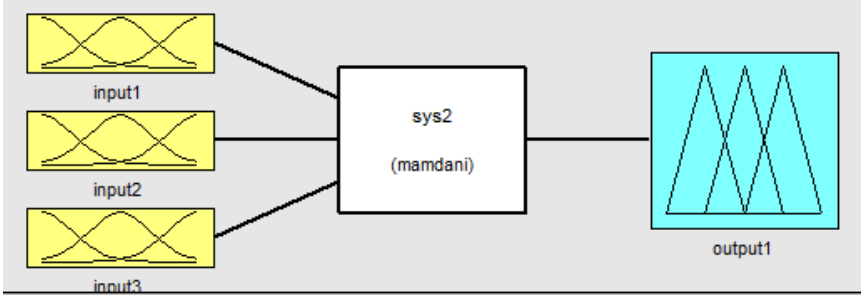

Figure 7.7: Fuzzy model of system 2

At this point, the outputs of all the fuzzy rules are combined to obtain a single fuzzy distribution in the output. This study desires to come up with a single crisp output from the FIS. This crisp number is obtained through the process of defuzzification. The result of defuzzification in the proposed system model is the location of the incident within the observed section of the road. Figure 7.8 is an example of how the sensory systems calculate the incident location. In this example, the first sensory system comprises two sensors. The first sensor has counted 38 vehicles in the upstream of the road and the second sensor has

counted 143 in the downstream. The yellow color shows the activation of each membership function involved in the rules and the blue color shows the strength of each rule in general. The last element in the rule strength column shows the combination of all activated rules and the red line is the defuzzified crisp output which is 67.1 and represents the predicted point of collision by this sensory system.

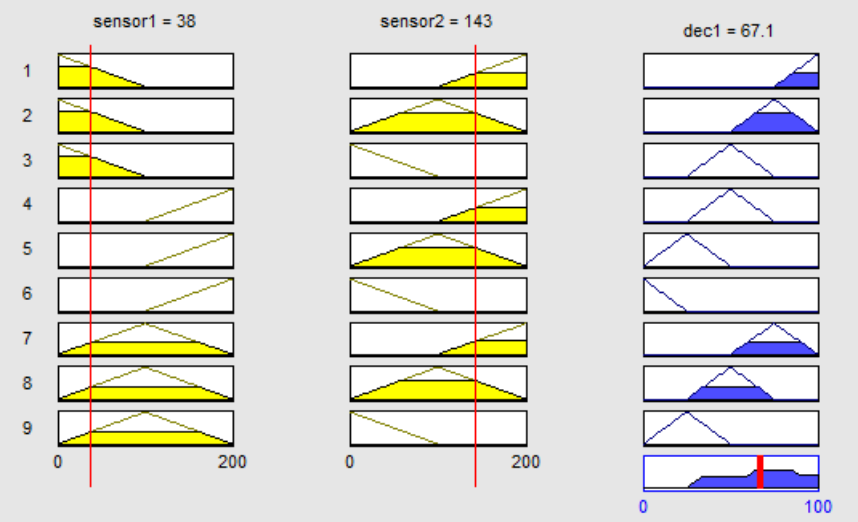

Figure 7.8: Example of rules' activation

7.4.3 Sensor Fusion Model

The probability theory is based on crisp logic, comprising zero or one. It does not consider any third possibility because the probability definition is based on set theory and crisp logic. It considers the probability of occurring or not occurring of an event. In contrast, the DS theory incorporates a third aspect which is the unknown. DS theory deals with assigned measures of belief in terms of mass, as opposed to the probabilities [106]. It allows statements of ignorance about likelihood of events. Consequently, a belief-based decision is made on the basis of two numbers: the degree to which an event is supported by the evidence (belief), and the degree to which there is a lack of evidence to the contrary (plausibility).

The two described sensory systems perceive information in the form of evidence. The two sensors in the first sensory system are able to measure the exact count of cars but the whole system suffers from lack of evidence perceived from the part of road it covers. In

addition, the three sensors in the second sensory system are noisy and do not have the desired precision but yet reveal more evidence compared to the other sensory system. By defining several allocations of belief to the location of an incident, DS theory offers a natural way of combining evidence to find a fused allocation of belief that deals both with ignorance and with conflict between the original beliefs. Having the belief and the plausibility for the ensemble, a decision can be made based on more comprehensive information.

In the proposed method, the masses are generated for 8 sections of the road covered by the sensory systems. These masses in total comprise the length of the road twice by each system. In this way, we ensure that the content includes various aspects of the information. The masses are combined using Desmpster's rule of combination illustrated by Equation (5.4). Using the definitions for the belief and plausibility functions (Equations (5.6), (5.7), and (5.8)) and also the combination rule of DS, the values of Bel and Pl can be calculated for five equal-length areas in the road. As the exact probability of a collision in each area is between the Bel and Pl for that area, these values are used along with the location of the area to estimate the collision location.

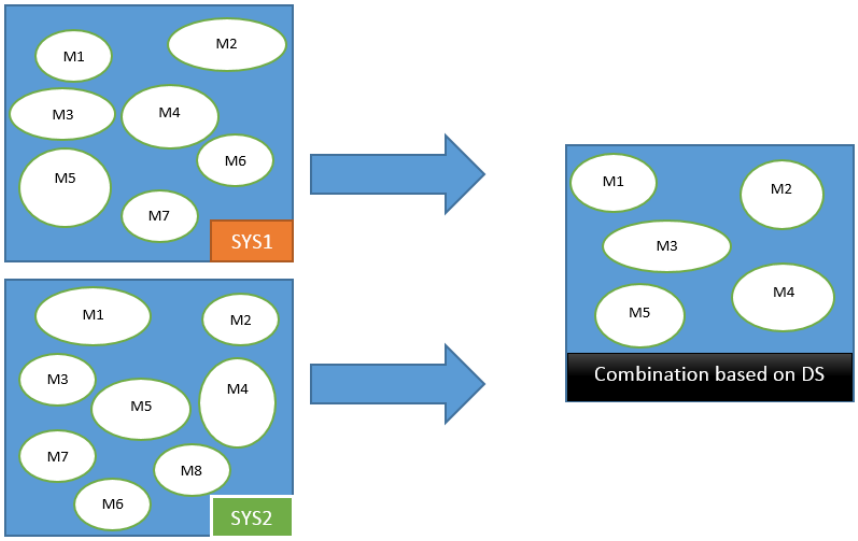

Figure 7.9: Combination rule of DS

7.5 Evaluation Criteria

The proposed model incorporates different sensory system models with varying detection capabilities. The model can be extended to fit the purpose of evaluation of different fusion architectures that can be designed for specific event detection like a car collision and data association techniques which are the core of any fusion paradigms. The aim here is to generate random car collisions in the designed traffic model which can be sensed by the sensors employed and then the data is fused with different fusion paradigms. The outcome then needs to be evaluated in order to find the quality of the sensory systems.

7.5.1 Normalized Mean Squared Error (NMSE)

The main metric used to evaluate the proposed multi-sensor fusion systems considering the ground truth values from the traffic model generator is NMSE. Mean squared error (MSE) assesses the quality of a predictor. If $\hat{Y}$ is a vector of n predictions, and Y is the vector of observed values corresponding to the inputs to the function which generated the predictions, then the MSE of the predictor can be estimated by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{7.5}$$

The prediction here is the car collision location from the sensory systems and the observation is the real collision location derived from the model generator. MSE is normalized by the length of the road to have a tangible value for the presented error. In order to have a complete evaluation scheme, different scenarios are generated by changing the traffic distribution function, number of cars and σ. Number of cars in the observing area which corresponds to the density of traffic and σ corresponds to the distribution of the cars which shows the severity of the accident. NMSE is employed to present the performance of all these scenarios in the tables and figures of the evaluation section.

7.5.2 Best Performance in Scenario (BPS)

I define BPS as a measure that provides the information about the best possible performance in each scenario. This measure is useful when the operator requires that a method reaches to a specified accuracy and precision. Although it does not provide any information about the conditions that the best performance has been achieved, it is important to

know the best performance that a method can achieve in a certain scenario. The formula to calculate this measure is:

$$BPS(i) = \min_{j} NMSE(T(i,j)) \tag{7.6}$$

Where i is the scenario to be investigated, j is the parameter that is being altered to obtain the scenario, and T is the sensory system or fusion operator.

7.5.3 Interval Percentage of Performance Improvement (IPPI)

From an economic standpoint, adding an extra cost for fusion techniques might not be efficient in some scenarios. Meaning that, the enhancement of the quality of data might be either minuscule or limited in the sense of interval of effectiveness. To quantify this effectiveness, IPPI is proposed. This criterion considers the actual performances of both sensory systems and also the intervals in which fusion techniques outperform them. Hence, IPPI could be defined as:

$$IPPI(i) = \frac{1}{2} * \frac{I_{DFS1}(i)}{I_{Total}(i)} + \frac{1}{2} * \frac{I_{DFS2}(i)}{I_{Total}(i)} \tag{7.7}$$

Where i is the scenario to be investigated, I_{DFS1} and I_{DFS2} are the lengths of the interval of the varying variable in which data fusion technique outperforms the sensory system 1 and 2 respectively. In addition, I_{Total} is equal to the total length of the varying variable.

7.5.4 Average Performance Improvement Ratio (APIR)

This evaluation criterion measures the performance improvement ratio for the whole range of the altered parameter in the scenario. In order to find a single value for this ratio, the average of performance improvement ratio is used over the range in which the altered parameter covers that scenario. This criterion is informative in the sense that it allows the operator to compare each fusion method with the best performance achieved while fusion is not applied. APIR provides this comparison for the whole scenario and gives a general view about the value of the fusion method by generating the properness ratio. This measure is calculated as below:

$$APIR(i) = \frac{\sum_{j=1}^{N} \frac{NMSE(T_f(i,j))}{NMSE(T_{BS}(i,j))}}{N} \tag{7.8}$$

In this equation, i is the scenario to be investigated, j is the parameter that is being altered to obtain the scenario, T_f is the fusion operator, T_{BS} is the best sensory system operator considering NMSE as the performance, and N is the number of experiments in the scenario.

7.6 Results

7.6.1 Evaluation Based on NMSE

For evaluation of the proposed method, several scenarios have been studied. First, the number of cars is fixed and the performance of the DS and OWA are compared to the performance of individual sensory systems for a range of σ for both normal and lognormal distributions. Then, the effect of number of cars has been studied with a fixed value of σ for both normal and lognormal distributions.

Normal Traffic Distribution with Fixed Number of Cars

Figure 7.10 compares the performance of the DS-based approach with that of the sensory systems and OWA. As σ increases, DS tends to outperform the individual sensory systems and OWA. OWA is an averaging technique and locates the point of collision somewhere in between of the sensory systems' estimations. Taking into consideration the similarities in these systems, it can be inferred that the actual point of collision can reside out of the region between estimations of individual sensory systems which precludes OWA from getting the higher quality information.

Although The DS fusion technique performs well in most of the scenarios, it fails to enhance the quality of the information for $\sigma \leq 17$ using the normal distribution and for 350 number of cars in the model. The reason is that low values of σ makes the model generator to create a model which all the cars are concentrated in the collision point symmetrically. This fact is the worst scenario in the sense of providing information to the sensor setups. What happens is that only one sensor have cars in its range and no matter where in that range the collision has happened, the fuzzy inference system will estimate the collision point in the center of the range to minimize the error. DS also can not generate enough evidence to decide based on and consequently it can not perform any better than the best sensor setup.

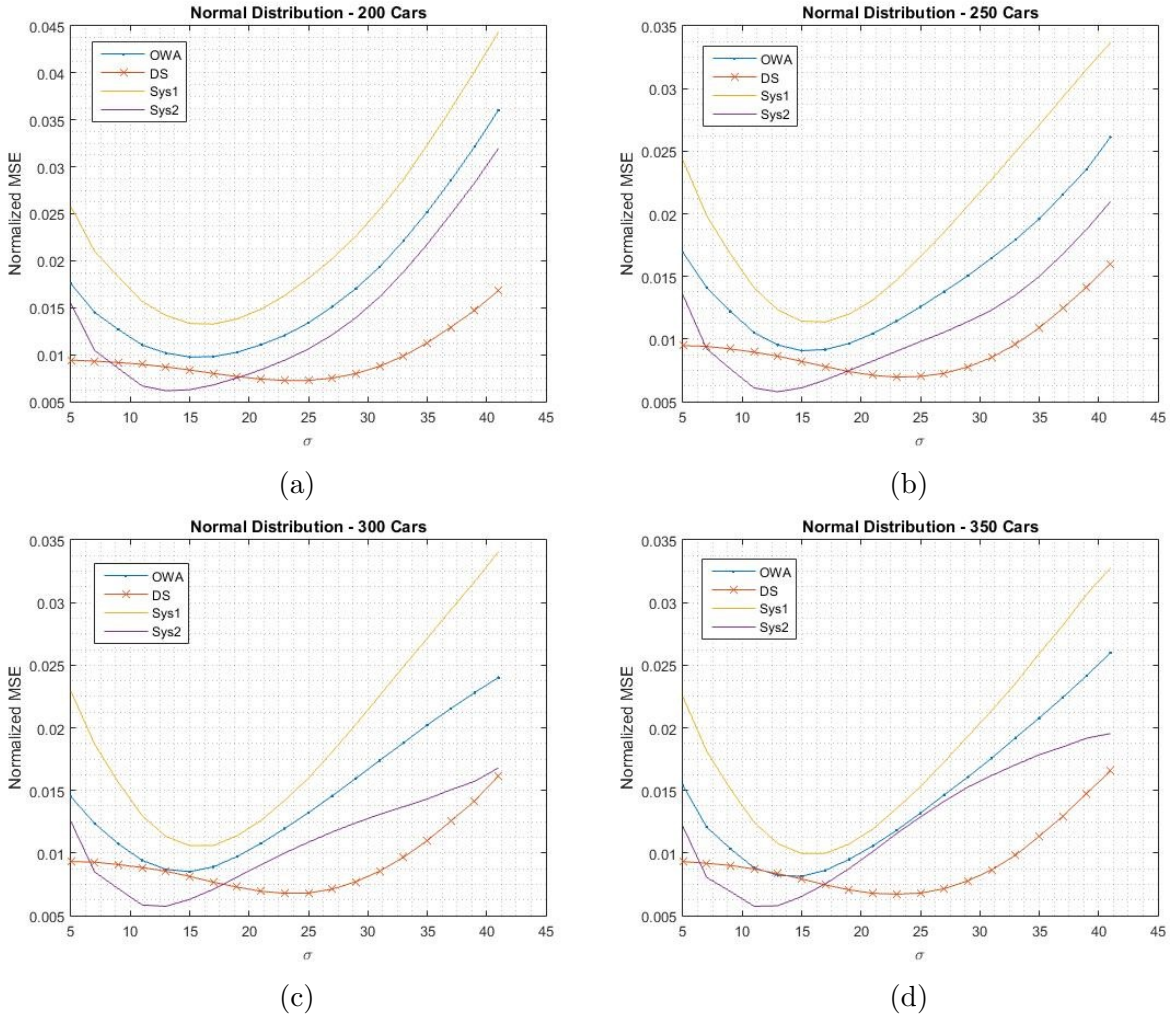

Figure 7.10: Effect of σ for fixed number of cars in normal distribution

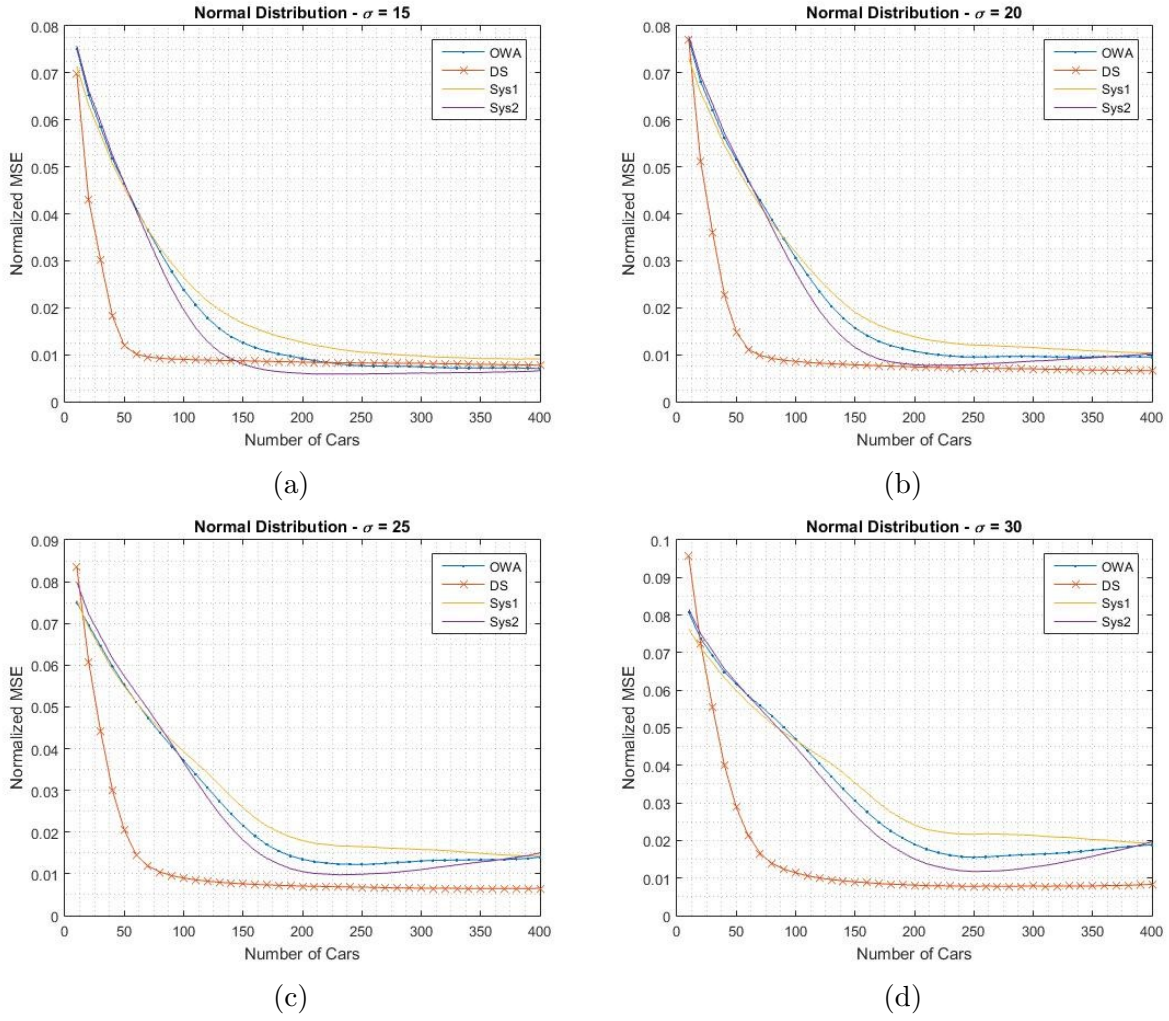

Figure 7.11: Effect of number of cars for fixed value of σ in Normal distribution

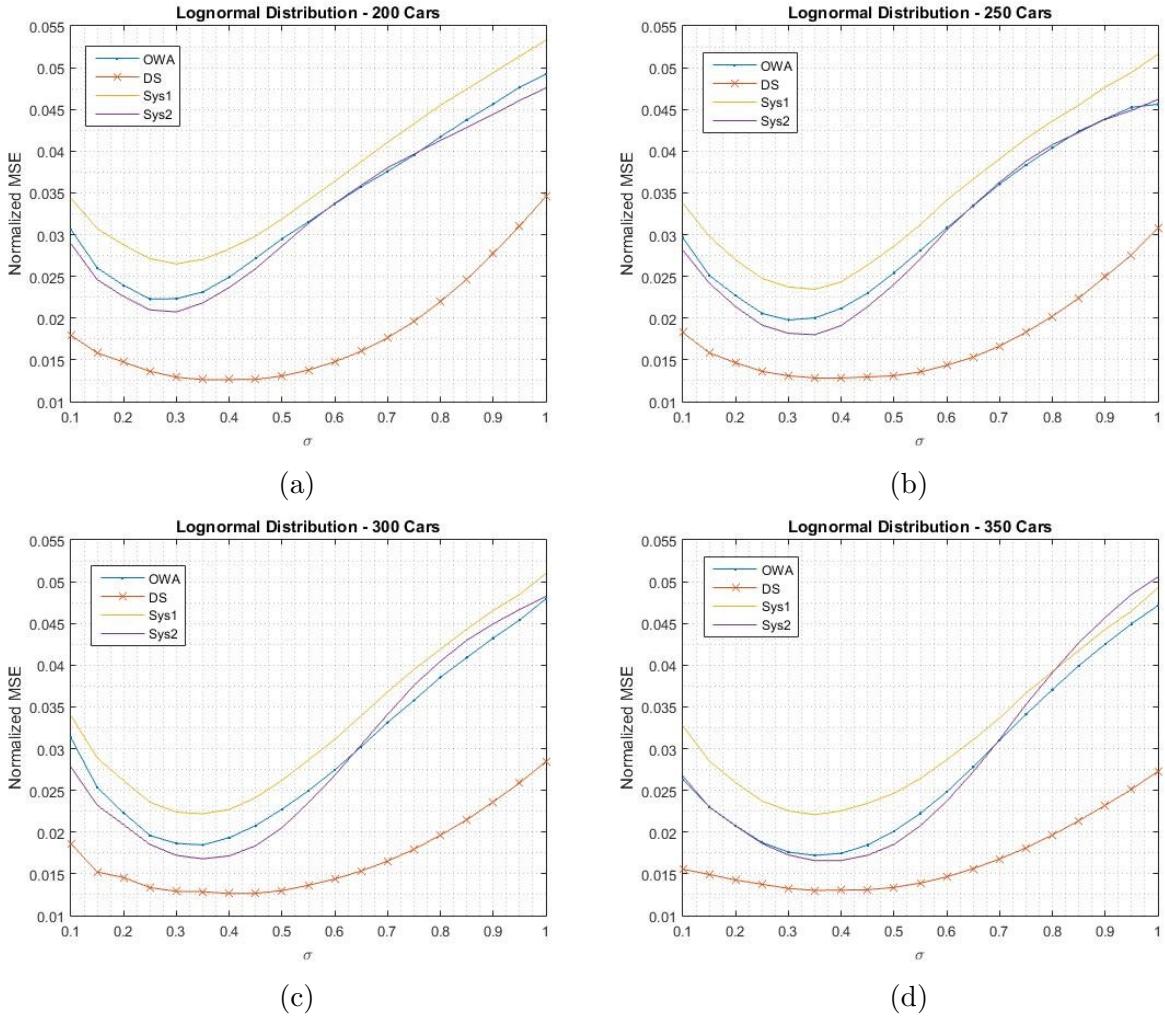

Figure 7.12: Effect of σ for fixed number of cars in lognormal distribution

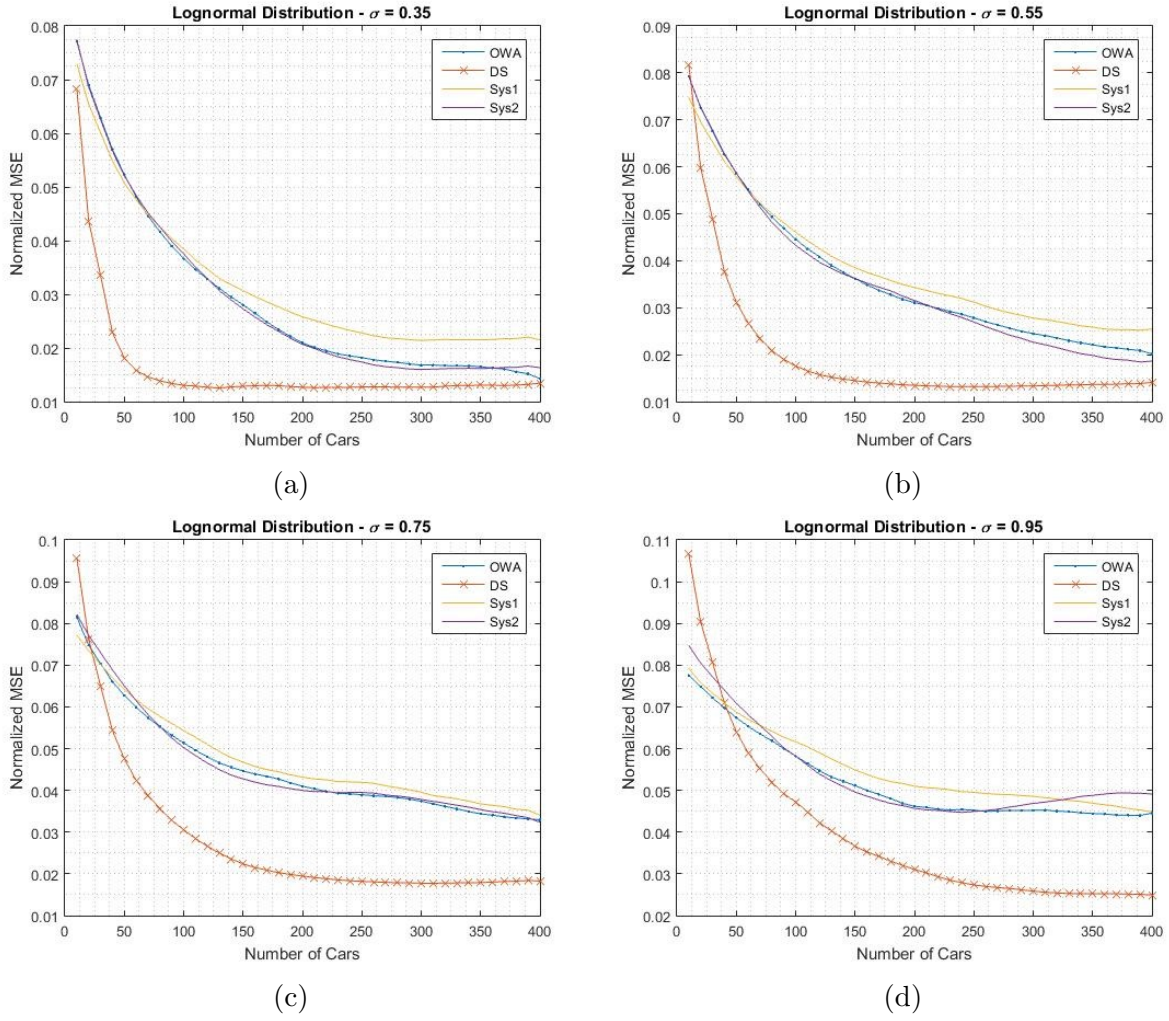

Figure 7.13: Effect of number of cars for fixed value of σ in lognormal distribution

Normal Traffic Distribution with Fixed σ

For the second scenario, σ was fixed and the performances are evaluated when number of cars vary. From Figure 7.11 it could be deduced that, similar to argument of the previous figure, DS method outperforms the sensory systems and OWA. In addition, in Figure 7.11, we see that the performances tend to converge to a single value as the number of car increases. This is because adding more cars while so many of them already exist provides sensory systems with little information compared to when the number of cars is small.

Lognormal Traffic Distribution with Fixed Number of Cars

In lognormal distribution, DS is shown to be a promising data fusion technique that significantly enhances the quality of the estimation as illustrated in Figure 7.12. As opposed to a normal distribution, lognormal distribution has asymmetrical nature. Hence, even for small values of σ, sufficient information is provided for sensors. The lognormal distribution in most cases let the other sensors to capture a non-zero value as the count of cars. Therefore, this asymmetric nature of lognormal distribution will provide DS method with some evidence to judge about the belief and plausibility of decisions.

Normal Traffic Distribution with Fixed σ

It is shown in Figure 7.13 that significant performance of DS is irrelevant to the number of cars in lognormal distribution which is similar to the conclusion derived from the normal distribution.

7.6.2 Evaluation Based on BPS

Table 7.1 demonstrates the calculated values of BPS for sensory systems alone, and also, for the data fusion techniques. This table provides the information about the best possible performance in each scenario. Although it is highly dependent on the exact operating point, but is useful when the operator requires that a method reaches to a specified accuracy and precision. According to this table, in all the scenarios except normal distribution with fixed number of vehicles and normal distribution with $\sigma = 15$, DS is superior in BPS compared to OWA and sole sensory systems. This fact shows that DS is a more reliable candidate for collision detection purposes. It can also be inferred that normal distribution is not as informative as lognormal distribution for DS method.

Table 7.1: Comparison of performance based on BPS criterion

	Scenarios	Sensor1	Sensor 2	OWA	DS
Normal Distribution	$\sigma = 15$	0.0089	0.0057	0.0069	0.0078
	$\sigma = 20$	0.0102	0.0076	0.0091	0.0065
	$\sigma = 25$	0.0142	0.0097	0.0117	0.0064
	$\sigma = 30$	0.0191	0.0116	0.0143	0.0077
	nCar = 200	0.0125	0.0056	0.0087	0.0071
	nCar = 250	0.0107	0.005	0.0076	0.0067
	nCar = 300	0.0097	0.005	0.0067	0.0065
	nCar = 350	0.0091	0.005	0.0065	0.0065
Lognormal Distribution	$\sigma = 0.35$	0.0211	0.0158	0.0138	0.0122
	$\sigma = 0.55$	0.0249	0.0183	0.0202	0.0131
	$\sigma = 0.75$	0.0341	0.324	0.328	0.0176
	$\sigma = 0.95$	0.0449	0.0445	0.0413	0.025
	nCar = 200	0.0258	0.0196	0.0196	0.012
	nCar = 250	0.0217	0.0161	0.0163	0.0126
	nCar = 300	0.021	0.0159	0.0157	0.012
	nCar = 350	0.0217	0.0157	0.0155	0.0127

7.6.3 Evaluation Based on IPPI

According to Table 7.2, the calculated IPPIs for DS and OWA show that DS is more effective in all of the developed scenarios when compared to OWA. Moreover, this table provides further information about the optimal point of operation. For DS, it is observed that it is most cost effective when $\sigma = 20$ for normal distribution and when $\sigma = 0.35$ for lognormal distribution. Besides, it can be inferred from this table that when the number of cars is sufficient, i.e., the sufficient amount of information is fed to DS, regardless of distribution, DS outperforms sensory systems and OWA technique.

In Contrast, the performance of OWA using IPPI is dependent to both the number of cars and the distribution of vehicles in the road. It is observed that OWA performs better for lognormal distribution due to its asymmetrical nature. The optimum operating point for this technique is when $\sigma = 25$ and $nCar = 350$ for normal distribution and when $\sigma = 0.95$ and $nCar = 300$ for lognormal distribution.

Table 7.2: Comparison of performance based on IPPI criterion

	Scenarios	OWA	DS
Normal Distribution	$\sigma = 15$	66.25%	52.50%
	$\sigma = 20$	98.75%	53.75%
	$\sigma = 25$	97.50%	60.00%
	$\sigma = 30$	96.25%	50.00%
	nCar = 200	81.58%	50.00%
	nCar = 250	84.21%	50.00%
	nCar = 300	84.21%	50.00%
	nCar = 350	84.21%	55.26%
Lognormal Distribution	$\sigma = 0.35$	100.00%	73.75%
	$\sigma = 0.55$	97.50%	56.25%
	$\sigma = 0.75$	97.50%	73.75%
	$\sigma = 0.95$	92.50%	77.50%
	nCar = 200	100.00%	65.79%
	nCar = 250	100.00%	60.53%
	nCar = 300	100.00%	76.32%
	nCar = 350	100.00%	76.31%

Table 7.3: Comparison of performance based on APIR criterion

	Scenarios	OWA	DS
Normal Distribution	$\sigma = 15$	1.3676	0.8169
	$\sigma = 20$	1.8227	0.8903
	$\sigma = 25$	2.2351	0.9094
	$\sigma = 30$	2.2591	0.8846
	nCar = 200	1.367	0.7827
	nCar = 250	1.1542	0.7454
	nCar = 300	1.17779	0.7569
	nCar = 350	1.3467	0.8576
Lognormal Distribution	$\sigma = 0.35$	1.8328	0.9913
	$\sigma = 0.55$	1.9569	0.9632
	$\sigma = 0.75$	1.8392	1.0013
	$\sigma = 0.95$	1.4836	1.0135
	nCar = 200	1.8416	0.9866
	nCar = 250	1.7503	0.9829
	nCar = 300	1.692	0.9939
	nCar = 350	1.6081	1.0096

7.6.4 Evaluation Based on APIR

According to Table 7.3, it is observed that in all scenarios DS has always shown performance improvement in average for the experiments performed for each scenario. This is inferred from APIR being greater than 1. Considering Equation (7.8), APIR value of greater than 1 shows the superiority of the fusion method. Furthermore, the greater the value of APIR for a fusion method, the more superior is that fusion method compared to single sensory systems. OWA, as opposed to DS, shows superiority in only 3 scenarios and with a slightly greater than 1 value which is not an acceptable performance for a fusion method in this study.

7.7 Conclusion

This chapter developed a traffic incident localization approach comprised of FIS-based sensory systems and a DS-based fusion module. For performance evaluation, a traffic model generator is developed and incorporated to the model. The traffic model generator is based on two distribution functions: normal and lognormal. The sensory setups are combinations of two and three simple vehicle-count-based sensors extracting the number of vehicles in their range. Based on the density of vehicles in each section, a fuzzy inference system is used to find the single sensory setups decision about the location of the collision. The two sensory systems' information is then aggregated in data level using the DS method and in the decision level using the OWA. The OWA performance is mostly in between of each of the sensory setups and occasionally better than both of them. This is observed by NMSE, BPS, IPPI, and APIR evaluation criteria. The reason is that OWA tends to merge the decisions and minimize the overall decision error. Therefore it outperforms the single sensor setups when the collision point lays in between their estimations. The DS method, on the other hand, is performing satisfactory compared to the single sensory setups and OWA considering all evaluation crieria. DS theory uses masses of evidence and generates the belief and plausibility of decisions. By aggregating all the belief and plausibility functions of both sensory setups, a significant performance is reached for collision detection in the proposed platform.

Chapter 8

Conclusion and Future Directions

This chapter provides a summary of the contributions of this thesis towards context-aware analysis of road-accident datasets and presents suggestions for future work. Section 8.1 summarizes the main ideas used in this thesis and highlights the achievements that make this study different from those of its kind. Each chapter of this thesis applies different state-of-the-art methods and algorithms to advance the quality of insights and the platform that produce them. Section 8.2 contains the research opportunities to further extend these methods and algorithms. I introduce the directions that this work can be driven to in the future.

8.1 Conclusion

Discovering potential risks of accidents and communicating timely warnings to the drivers remains a major problem for the society. Driving has become a habit for many people as they do it almost every day to commute to their workplace, school, gymnasiums, and other places and facilities. Being behind the wheel almost everyday drives us to our comfort-zone where we are unaware of the potential dangers we may confront. With the abundance of data about traffic accidents and the advances made in data analytics and information technologies, there is a great opportunity to discover patterns of collision-contributing factors that may lead driving individuals to severe accidents. Driving alert and assistance systems, which are widely used in modern cars these days, can provide valuable cautionary advice to drivers if simply integrated with these patterns and being updated every often.

Such technology can be also used in autonomous vehicles within all the the ranges from partially automated to fully automated. Doing so will enable them to foresee a

wider range of conflicting actions and help them better plan their journey to avoid those conflicts. It is only a matter of time before automated vehicles pervade the streets in urban and rural areas, making us feel the urge for the vehicles to have long-term journey planing capabilities.

The research presented in this thesis focused on developing methods to turn raw accident data into insightful association rules that can be used to construct a Bayesian network with the ability to predict the risk of potential accidents. This work can be considered in the category of collision prediction models which have a significant influence on traffic flow and transportation safety. Predicting the probability range and the risk level of potential accidents are regarded as valuable assets for car manufacturers and insurance companies. Insurance companies are tending to use more sophisticated solutions to offer more realistic premiums to their customers. Any improvement in creation of premiums, even though small, can make a huge difference in the revenue of insurance companies. Car owners and drivers are also willing to have a more accurate premium tailored to their driving experience, their vehicles safety, and the locations and times that they frequently drive in. All these factors are, in some way, associated to the risk level of a particular driver driving a certain vehicle. This study has offered a solution to reform classic collision analysis approaches by replacing individual analysis of risk factors by a wholistic analysis of crash contributing factors in the form of crash contributing patterns.

Understanding the relationship between the collision contributing variables starts with exploring the patterns of their appearance in a collection of collision samples. Chapter 4 of this thesis presents an idea based on medoid clustering and association rule mining for exploring and comprehensively understanding crash-contributing patterns in a given collision dataset. The clustering is adopted for the purpose of data segmentation which minimizes the effect of heterogeneity in data. The clustering method was chosen based on the output of a cluster shape identification method that I customized for the case study of road-accident data analysis. Handling the presence of categorical data is another challenge that emerges when dealing with a great majority of datasets. This challenge was addressed by introducing Gower distance as a criterion to measure dissimilarity of samples in the presence of categorical data. Insight induction is the core of this research and required a solid approach to reflect the relationship and contribution of variables in occurrence of traffic collisions. I found association rule mining a reliable approach for this aim and enhanced its precision and computation time by incorporating binary particle swarm optimization.

I used the national collision database of Canada and applied the cluster identification, data segmentation, and association rule mining methods on 20000 randomly selected samples. In the numerical analysis, the chosen clustering method, medoid clustering, divided

the data into three segments based on existence of complex driving situations and driver characteristics. Association rule mining was then applied to these segments and insightful rules were successfully generated. The performance of the BPSO-based association rule mining proposed in this study was compared with that of other techniques in terms of consistency and processing time which showed significant outperformance.

In the course of experiments, I noticed that many of the datasets that industry demands, including road-accident databases, are deficient in descriptive factors. This is a significant barrier for obtaining meaningful relationships from those datasets. For this reason, I presented the concept of knowledgebase approximation in Chapter 4 to facilitate and accelerate insight induction from high-dimensional disparate knowledgebases. This chapter introduced Dempster-Shafer theory as a means to elevate the amount of information obtained from disparate datasets by fusing their association rules. I tested this method on the lymphography dataset and applied it to the road-accident database of the Great Britain to enrich the insights induced from the national collision database of Canada.

Investigations on the road-accident historical data and in-depth collision analysis are powerful tools to explore contributing collision risk factors. In-depth collision analysis provides information on the chain of events leading to an accident by detailed reconstitution of accidents, and historical data play an important role in it. However, these techniques can not fulfill the requirements of a context-aware collision analysis model as they entail some shortcomings that does not allow generation of realistic predictions tailored to upcoming specific situations. As a result, I decided to tackle these shortcomings by conducting naturalistic driving analysis on the reinforced insights. Naturalistic driving analysis allows continuous, and sometimes real-time, processing of collected data from the sources that contain contributing factors, and hence, is a good complement for the previous steps to further expand the risk analysis model.

Using a Bayesian network is one of the ways to implement a solid naturalistic driving analysis framework to predict traffic collision risk based on the findings from the data-driven model. The directed acyclic graph of a Bayesian network can work with the real-time measurements fed into a collision prediction model which brings a great flexibility to the whole system. The big challenge, however, was to find a way to build the network based on the approximated knowledgebase containing association rules. Chapter 6 of this thesis elaborates on a methodology to learn the structure of the Bayesian network from association rules. When the structure is decided, the conditional probability tables can be simply generated from the data.

The great advantage of using association rules in the proposed Bayesian network construction method is the ability of the consequent model in merging the information from

different resources as a single collision risk analysis system. There is still extensive research going on about structure learning algorithms that can work on multiple datasets. Previous studies mostly rely on integration of datasets in the data level which entails numerous challenges and drawbacks. Chapter 5 proposed knowledgebase approximation to overcome those drawbacks and created a union of collision patterns in the form of association rules, and Chapter 6 used those integrated association rules to construct a Bayesian network. These whole process brings combined knowledge to a risk analysis system that accepts real-time measurements from the vehicle, the environment, and the driver. Spatial and temporal aspects can also be integrated into this system to create a compact temporospatial context-aware collision analysis system. This method can also be applied to the association rules extracted from a single dataset and may be used as a substitute for other structure learning approaches.

The in-depth accident analysis and naturalistic driving studies organize the foundation of the context-aware prediction model when combined. At this point there are plenty of subsidiaries that can be added to this compound. One of these subsidiaries is a traffic incident detection and localization system. Detecting and localizing traffic incidents enables timely response to accidents and facilitates effective and efficient traffic flow management. One of the frequent types of traffic collisions are head on accidents to the vehicles stalled in highways after occurrence of an incident. The automatic incident detection system can inform the drivers about the congestions in their path and their associated collisions risks if integrated with the crash prediction model.

Sensor networks are widely used to obtain data about vehicle surroundings, traffic flow, weather conditions, driver behavior and other factors affecting the collision prediction process. They are powerful source of information for a crash prediction model and can leverage the prediction accuracy if used as part of the model. These networks, however, may communicate erroneous and noisy sensor readings. Chapter 7 exploits sensor data fusion methods to overcome inaccuracies in a sensor network used for the purpose of automatic incident detection.

This study designed a crash risk analysis ensemble has the potential to perform individual-based investigation of the crash risk factors by considering the information from the driver, the vehicle, and the environment in different variations of time and location. The outcome showed the effectiveness of this ensemble with the available road-accident datasets. It is more focused on presenting admissible crash analysis techniques rather than presenting the induced insights. With the advances in telematics and the abundance of traffic data, crash risk prediction models are great complements for advanced driving assistance systems to provide drivers with life saving information.

8.2 Future Directions

Collision risk analysis and development of collision prediction models are active areas of research. New efficiency issues arise as the volume of traffic data and their generation speed is increased. Moreover, with the growing number of in-use vehicles every year, the need for enhanced safety equipment and technologies is being sensed more than ever. In this section, I show possible future directions of collision prediction and risk analysis by presenting a number of suggestions for further expanding the research in this thesis.

- **real-time collision risk analysis:** Chapter 6 of this thesis developed a Bayesian network based approach which has the potential to investigate real-time information. This Bayesian network uses the outcome of data-driven model to learn the network structure and calculate conditional probabilities. Real-time sensor measurements and environmental updates can be combined by the historical data enabling the model to calculate the crash risk and likelihood for a single vehicle in real-time.

- **Human factor:** The databases in this study is less concentrated on human factors compared to the vehicle and environmental factors. More driver behavior attributes, specifically, can add valuable information to the resultant insights. One of the future directions is to incorporate datasets more concentrated on human factors and augment their information using the knowlegebase approximation framework. Those developed parts of the model also have room for modification to generate more realistic segments and rules about driver behaviour.

- **Clustering methods:** Rule mining is an important part of the model that not only forms the foundation for discovering underlying rules in accident databases, but also contributes to learning the structure for the Bayesian network and other steps used in this research. For each of the clusters shaped during the data segmentation process, running the BPSO-based association rule mining algorithm produces several rules, which are the best ones selected with respect to their fitness values. The quality of clusters has a significant influence on the rules since a proper clustering technique prevents the model from generating redundant rules in different clusters. This fact is an encouragement for utilizing more powerful clustering algorithms in the future to empower the model in generating more informative rules.

- **Information fusion:** Identification of road-accident risk factors is essential in developing an effective crash prediction model. Some of these factors may be influential all the time, while some other impact in specific situations. No matter which ones

are more effective in the occurrence of an accident, a comprehensive model should be able to explore all the attributes connected to a collision event. There are numerous powerful data fusion techniques in category of data association, state estimation, and decision fusion that can complement the data-driven model by more accurately integrating the knwoledgebase constructed by different databases.

- **Other vehicles' information:** This thesis introduced a context-aware accident risk prediction approach which is based on factors from three categories: the driver, the evironment, and the vehicle. One other set of factors that can elaborate on the probability of a particular vehicle to crash is the other vehicles' information in the vicinity. Context-aware architectures based on VANET's on board unit (OBU) are suggested to be used along with the framework in this thesis to improve the results in the future work.

References

[1] Road traffic injuries. http://www.who.int, accessed 2017-08-17.

[2] Margie Peden, Richard Scurfield, David Sleet, Dinesh Mohan, Adnan A Hyder, Eva Jarawan, Colin D Mathers, et al. World report on road traffic injury prevention, 2004.

[3] World Health Organization. *Global status report on road safety 2015*. World Health Organization, 2015.

[4] Tarek Sayed and Paul De Leur. *Collision prediction models for British Columbia*. Ministry of Transportation and Infrastructure Victoria, BC, Canada, 2008.

[5] Kristian LL Movig, MPM Mathijssen, PHA Nagel, T Van Egmond, Johan J De Gier, HGM Leufkens, and Antoine CG Egberts. Psychoactive substance use and the risk of motor vehicle accidents. *Accident Analysis & Prevention*, 36(4):631–636, 2004.

[6] Donald A Redelmeier and Robert J Tibshirani. Association between cellular-telephone calls and motor vehicle collisions. *New England Journal of Medicine*, 336(7):453–458, 1997.

[7] Kenneth Wade Ogden. *Safer roads: a guide to road safety engineering*. 1996.

[8] Eleni Petridou and Maria Moustaki. Human factors in the causation of road traffic crashes. *European journal of epidemiology*, 16(9):819–826, 2000.

[9] Shaw-Pin Miaou. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, 26(4):471–482, 1994.

[10] Letty Aarts and Ingrid Van Schagen. Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention*, 38(2):215–224, 2006.

[11] Matthew G Karlaftis and Ioannis Golias. Effects of road geometry and traffic volumes on rural roadway accident rates. *Accident Analysis & Prevention*, 34(3):357–365, 2002.

[12] Min Zhou and Virginia P Sisiopiku. Relationship between volume-to-capacity ratios and accident rates. *Transportation research record*, 1581(1):47–52, 1997.

[13] Jean-Louis Martin. Relationship between crash rate and hourly traffic flow on interurban motorways. *Accident Analysis & Prevention*, 34(5):619–629, 2002.

[14] Xiao Qin, John N Ivan, and Nalini Ravishanker. Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accident Analysis & Prevention*, 36(2):183–191, 2004.

[15] Xin Pei, SC Wong, and Nang-Ngai Sze. The roles of exposure and speed in road safety analysis. *Accident Analysis & Prevention*, 48:464–471, 2012.

[16] Statista. *GOV.UK. (September 26, 2019). Number of road accidents caused by vehicle defect factors in Great Britain (UK) in 2018, by severity*, 2019. https://www.statista.com/statistics/323086/road-accidents-caused-by-vehicle-defect-factors-severity-in-great-britain-uk/ [Accessed 2020-01-26].

[17] Michel Bedard, Gordon H Guyatt, Michael J Stones, and John P Hirdes. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accident Analysis & Prevention*, 34(6):717–727, 2002.

[18] John Langley, Bernadette Mullin, Rodney Jackson, and Robyn Norton. Motorcycle engine size and risk of moderate to fatal injury from a motorcycle crash. *Accident Analysis & Prevention*, 32(5):659–663, 2000.

[19] Bhagwant Persaud and Leszek Dzbik. Accident prediction models for freeways. *Transportation Research Record*, 1401:55–60, 1992.

[20] Matthew W Knuiman, Forrest M Council, and Donald W Reinfurt. Association of median width and highway accident rates (with discussion and closure). *Transportation Research Record*, 1401:70–82, 1993.

[21] Poul Greibe. Accident prediction models for urban roads. *Accident Analysis & Prevention*, 35(2):273–285, 2003.

[22] Mohamed A Abdel-Aty and A Essam Radwan. Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5):633–642, 2000.

[23] Karim El-Basyouny and Tarek Sayed. Accident prediction models with random corridor parameters. *Accident Analysis & Prevention*, 41(5):1118–1123, 2009.

[24] Mark Poch and Fred Mannering. Negative binomial analysis of intersection-accident frequencies. *Journal of transportation engineering*, 122(2):105–113, 1996.

[25] Ezra Hauer. *Observational before/after studies in road safety. Estimating the effect of highway and traffic engineering measures on road safety.* 1997.

[26] John Hinde and Clarice GB Demétrio. Overdispersion: models and estimation. *Computational statistics & data analysis*, 27(2):151–170, 1998.

[27] Shaw-Pin Miaou and Dominique Lord. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and bayes versus empirical bayes methods. *Transportation Research Record*, 1840(1):31–40, 2003.

[28] Ziad Sawalha and Tarek Sayed. Traffic accident modeling: some statistical issues. *Canadian Journal of Civil Engineering*, 33(9):1115–1124, 2006.

[29] Hoon Kim, Dongchu Sun, and Robert K Tsutakawa. Lognormal vs. gamma: extra variations. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 44(3):305–323, 2002.

[30] Eun Sug Park and Dominique Lord. Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record*, 2019(1):1–6, 2007.

[31] Karim El-Basyouny and Tarek Sayed. Collision prediction models using multivariate poisson-lognormal regression. *Accident Analysis & Prevention*, 41(4):820–828, 2009.

[32] Lasse Fridstrøm, Jan Ifver, Siv Ingebrigtsen, Risto Kulmala, and Lars Krogsgård Thomsen. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. *Accident Analysis & Prevention*, 27(1):1–20, 1995.

[33] Mohammed A Hadi, Jacob Aruldhas, Lee-Fang Chow, and Joseph A Wattleworth. Estimating safety effects of cross-section design for various highway types using negative binomial regression. *Transportation Research Record*, 1500:169, 1995.

[34] Bhagwant Persaud, Richard A Retting, and Craig Lyon. Guidelines for identification of hazardous highway curves. *Transportation Research Record*, 1717(1):14–18, 2000.

[35] Matthew G Karlaftis and Eleni I Vlahogianni. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387–399, 2011.

[36] Lorenzo Mussone, Andrea Ferrari, and Marcello Oneta. An analysis of urban collisions using an artificial intelligence model. *Accident Analysis & Prevention*, 31(6):705–718, 1999.

[37] Dursun Delen, Ramesh Sharda, and Max Bessonov. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis & Prevention*, 38(3):434–444, 2006.

[38] Mohammad Al-Marafi and Kathirgamalingam Somasundaraswaran. Review of crash prediction models and their applicability in black spot identification to improve road safety. *Indian Journal of Science and Technology*, 11(5):1–7, 2018.

[39] Darçin Akin and Bülent Akba. A neural network (nn) model to predict intersection crashes based upon driver, vehicle and roadway surface characteristics. *Scientific Research and Essays*, 5(19):2837–2847, 2010.

[40] John F Gilmore, Khalid J Elibiary, and Naohiko Abe. Traffic management applications of neural networks. In *Working Notes, AAAI-93 Workshop on AI in Intelligent Vehicle Highway Systems*, pages 85–95, 1993.

[41] Ardeshir Faghri and Jiuyi Hua. Evaluation of artificial neural network applications in transportation engineering. *Transportation Research Record*, 1358:71, 1992.

[42] Miao M Chong, Ajith Abraham, and Marcin Paprzycki. Traffic accident analysis using decision trees and neural networks. *arXiv preprint cs/0405050*, 2004.

[43] Tatiana Tambouratzis, Dora Souliou, Miltiadis Chalikias, and Andreas Gregoriades. Combining probabilistic neural networks and decision trees for maximally accurate and efficient accident prediction. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2010.

[44] Abdul Wahab, Chai Quek, Chin Keong Tan, and Kazuya Takeda. Driving profile modeling and recognition based on soft computing approach. *IEEE transactions on neural networks*, 20(4):563–582, 2009.

[45] Xu Qu, Wei Wang, Wenfu Wang, Pan Liu, and David A Noyce. Real-time prediction of freeway rear-end crash potential by support vector machine. Technical report, 2012.

[46] F Rezaie Moghaddam, Sh Afandizadeh, and M Ziyadi. Prediction of accident severity using artificial neural networks. *International Journal of Civil Engineering*, 9(1):41, 2011.

[47] Yisheng Lv, Shuming Tang, Hongxia Zhao, and Shuang Li. Real-time highway accident prediction based on support vector machines. In *2009 Chinese Control and Decision Conference*, pages 4403–4407. IEEE, 2009.

[48] Xiugang Li, Dominique Lord, Yunlong Zhang, and Yuanchang Xie. Predicting motor vehicle crashes using support vector machine models. *Accident Analysis & Prevention*, 40(4):1611–1618, 2008.

[49] John R Koza. *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press, 1992.

[50] Hossein Etemadi, Ali Asghar Anvary Rostamy, and Hassan Farajzadeh Dehkordi. A genetic programming model for bankruptcy prediction: Empirical evidence from iran. *Expert Systems with Applications*, 36(2):3199–3207, 2009.

[51] Terje Lensberg, Aasmund Eilifsen, and Thomas E McKee. Bankruptcy theory development and classification via genetic programming. *European Journal of Operational Research*, 169(2):677–697, 2006.

[52] Thomas E McKee and Terje Lensberg. Genetic programming and rough sets: A hybrid approach to bankruptcy classification. *European Journal of Operational Research*, 138(2):436–451, 2002.

[53] Chengcheng Xu, Wei Wang, and Pan Liu. A genetic programming model for real-time crash prediction on freeways. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):574–586, 2012.

[54] Kevin R Dixon, Carl E Lippitt, and J Chris Forsythe. Supervised machine learning for modeling human recognition of vehicle-driving situations. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 604–609. IEEE, 2005.

[55] Zhen Qi Yang. Highway traffic accident prediction based on svr trained by genetic algorithm. In *Advanced materials research*, volume 433, pages 5886–5889. Trans Tech Publ, 2012.

[56] Ming Zheng, Tong Li, Rui Zhu, Jing Chen, Zifei Ma, Mingjing Tang, Zhongqiang Cui, and Zhan Wang. Traffic accidents severity prediction: A deep-learning approach-based cnn network. *IEEE Access*, 7:39897–39910, 2019.

[57] Zhuoning Yuan, Xun Zhou, and Tianbao Yang. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 984–992, 2018.

[58] Jie Bao, Pan Liu, and Satish V Ukkusuri. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. *Accident Analysis & Prevention*, 122:239–254, 2019.

[59] Athanasios Theofilatos, Cong Chen, and Constantinos Antoniou. Comparing machine learning and deep learning methods for real-time crash prediction. *Transportation research record*, 2673(8):169–178, 2019.

[60] Muhammad Aqib, Rashid Mehmood, Ahmed Alzahrani, and Iyad Katib. In-memory deep learning computations on gpus for prediction of road traffic incidents using big data fusion. In *Smart Infrastructure and Applications*, pages 79–114. Springer, 2020.

[61] Junhua Wang, Yumeng Kong, and Ting Fu. Expressway crash risk prediction using back propagation neural network: A brief investigation on safety resilience. *Accident Analysis & Prevention*, 124:180–192, 2019.

[62] Matthias Althoff, Olaf Stursberg, and Martin Buss. Model-based probabilistic collision detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 10(2):299–310, 2009.

[63] Alexander Barth and Uwe Franke. Where will the oncoming vehicle be the next second? In *2008 IEEE Intelligent Vehicles Symposium*, pages 1068–1073. IEEE, 2008.

[64] Jrg Hillenbrand, Andreas M Spieker, and Kristian Kroschel. A multilevel collision mitigation approachits situation assessment, decision making, and performance tradeoffs. *IEEE Transactions on intelligent transportation systems*, 7(4):528–540, 2006.

[65] K Lee and H Peng. Evaluation of automotive forward collision warning and collision avoidance algorithms. *Vehicle system dynamics*, 43(10):735–751, 2005.

[66] Andreas Eidehall and Lars Petersson. Statistical threat assessment for general road scenes using monte carlo sampling. *IEEE Transactions on intelligent transportation systems*, 9(1):137–147, 2008.

[67] Adrian Broadhurst, Simon Baker, and Takeo Kanade. Monte carlo road safety reasoning. In *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, pages 319–324. IEEE, 2005.

[68] Adrian Broadhurst, Simon Baker, and Takeo Kanade. *A prediction and planning framework for road safety analysis, obstacle avoidance and driver information*. Carnegie Mellon University, the Robotics Institute, 2004.

[69] Christian Schmidt, Fred Oechsle, and Wolfgang Branz. Research on trajectory planning in emergency situations with multiple objects. In *2006 IEEE Intelligent Transportation Systems Conference*, pages 988–992. IEEE, 2006.

[70] Jur Pieter van den Berg. *Path planning in dynamic environments*. PhD thesis, Utrecht University, 2007.

[71] Jianghai Hu, Maria Prandini, and Shankar Sastry. Aircraft conflict detection in presence of spatially correlated wind perturbations. In *AIAA Guidance, Navigation, and Control Conference and Exhibit*, page 5339, 2003.

[72] Jan Lunze and J Schroder. Sensor and actuator fault diagnosis of systems with discrete inputs and outputs. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(2):1096–1107, 2004.

[73] Jochen Schröder. *Modelling, state observation and diagnosis of quantised systems*, volume 282. Springer Science & Business Media, 2002.

[74] Rongjie Yu and Mohamed Abdel-Aty. Multi-level bayesian analyses for single-and multi-vehicle freeway crashes. *Accident Analysis & Prevention*, 58:97–105, 2013.

[75] Yuanchang Xie, Dominique Lord, and Yunlong Zhang. Predicting motor vehicle collisions using bayesian neural network models: An empirical analysis. *Accident Analysis & Prevention*, 39(5):922–933, 2007.

[76] Subasish Das, Xiaoduan Sun, Fan Wang, and Charles Leboeuf. Estimating likelihood of future crashes for crash-prone drivers. *Journal of traffic and transportation engineering (English edition)*, 2(3):145–157, 2015.

[77] Markus Deublein, Matthias Schubert, and Bryan T Adey. Prediction of road accidents: comparison of two bayesian methods. *Structure and Infrastructure Engineering*, 10(11):1394–1416, 2014.

[78] Ismail Dagli and Dirk Reichardt. Motivation-based approach to behavior prediction. In *Intelligent Vehicle Symposium, 2002. IEEE*, volume 1, pages 227–233. IEEE, 2002.

[79] Jose C Principe, Neil R Euliano, and W Curt Lefebvre. *Neural and adaptive systems: fundamentals through simulations*, volume 672. Wiley New York, 2000.

[80] Dominique Fleury and Thierry Brenac. Accident prototypical scenarios, a tool for road safety research and diagnostic studies. *Accident Analysis & Prevention*, 33(2):267–276, 2001.

[81] Nicolas Saunier, N Mourji, and B Agard. *Investigating collison factors by mining microscopic data of vehicle conflicts and collisions*. PhD thesis, École Polytechnique de Montréal, 2010.

[82] Markus Deublein, Matthias Schubert, Bryan T Adey, Jochen Köhler, and Michael H Faber. Prediction of road accidents: A bayesian hierarchical approach. *Accident Analysis & Prevention*, 51:274–291, 2013.

[83] Matthew G Karlaftis and Andrzej P Tarko. Heterogeneity considerations in accident modeling. *Accident Analysis & Prevention*, 30(4):425–433, 1998.

[84] Sachin Kumar, Durga Toshniwal, and Manoranjan Parida. A comparative analysis of heterogeneity in road accident data using data mining techniques. *Evolving systems*, 8(2):147–155, 2017.

[85] Martin HC Law, Alexander P Topchy, and Anil K Jain. Multiobjective data clustering. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

[86] Tilman Lange, Mikio L Braun, Volker Roth, and Joachim M Buhmann. Stability-based model selection. In *Advances in neural information processing systems*, pages 633–642, 2003.

[87] Brian S Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. Miscellaneous clustering methods. *Cluster Analysis, 5th Edition*, pages 215–255, 2011.

[88] Jacob Kogan, Charles Nicholas, Marc Teboulle, et al. *Grouping multidimensional data*. Springer, 2006.

[89] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[90] Sachin Kumar and Durga Toshniwal. A data mining framework to analyze road accident data. *Journal of Big Data*, 2(1):26, 2015.

[91] Russell Eberhart and James Kennedy. A new optimizer using particle swarm theory. In *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, pages 39–43. Ieee, 1995.

[92] James Kennedy and Russell C Eberhart. A discrete binary version of the particle swarm algorithm. In *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*, volume 5, pages 4104–4108. IEEE, 1997.

[93] KNVD Sarath and Vadlamani Ravi. Association rule mining using binary particle swarm optimization. *Engineering Applications of Artificial Intelligence*, 26(8):1832–1840, 2013.

[94] Statistics and Data - Transport Canada. http://www.tc.gc.ca, accessed 2017-12-17.

[95] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu, and Sen Wu. Understanding and enhancement of internal clustering validation measures. *IEEE transactions on cybernetics*, 43(3):982–994, 2013.

[96] C Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

[97] Hilbert J Kappen and Wim Wiegerinck. Second order approximations for probability models. In *Advances in Neural Information Processing Systems*, pages 238–244, 2001.

[98] Bart Selman and Henry A Kautz. Knowledge compilation using horn approximations. In *AAAI*, pages 904–909. Citeseer, 1991.

[99] Pedro Zuidberg Dos Martires, Anton Dries, and Luc De Raedt. Knowledge compilation with continuous random variables and its application in hybrid probabilistic logic programming. *arXiv preprint arXiv:1807.00614*, 2018.

[100] Barbara Dunin-Ke, Linh Anh Nguyen, Andrzej Szałas, et al. Tractable approximate knowledge fusion using the horn fragment of serial propositional dynamic logic. *International Journal of Approximate Reasoning*, 51(3):346–362, 2010.

[101] NIU Dangdang, LIU Lei, and LYU Shuai. Knowledge compilation methods based on the clausal relevance and extension rule. *Chinese Journal of Electronics*, 27(5):1037–1042, 2018.

[102] Amir Hosein Keyhanipour, Behzad Moshiri, Majid Kazemian, Maryam Piroozmand, and Caro Lucas. Aggregation of web search engines based on users preferences in webfusion. *Knowledge-Based Systems*, 20(4):321–328, 2007.

[103] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[104] Arthur P Dempster. A generalization of bayesian inference. In *Classic works of the dempster-shafer theory of belief functions*, pages 73–104. Springer, 2008.

[105] Didler Dubois and Henri Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Computational intelligence*, 4(3):244–264, 1988.

[106] Xavier Gros. *NDT data fusion*. Elsevier, 1996.

[107] Wenmin Li, Jiawei Han, and Jian Pei. Cmar: Accurate and efficient classification based on multiple class-association rules. In *icdm*, page 369. IEEE, 2001.

[108] Xiaoxin Yin and Jiawei Han. Cpar: Classification based on predictive association rules. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, pages 331–335. SIAM, 2003.

[109] *Lymphography dataset*, 1988. https://archive.ics.uci.edu/ml/datasets/Lymphography [Accessed: 2018-09-30].

[110] *Road Safety Data - Great Britain*, 2019. https://data.gov.uk/dataset/road-accidents-safety-data [Accessed 2019-04-05].

[111] Dan Geiger, Thomas Verma, and Judea Pearl. d-separation: From theorems to algorithms. In *Machine Intelligence and Pattern Recognition*, volume 10, pages 139–148. Elsevier, 1990.

[112] Luis M de Campos. Independency relationships and learning algorithms for singly connected networks. *Journal of Experimental & Theoretical Artificial Intelligence*, 10(4):511–549, 1998.

[113] Edward Herskovits. *Computer-based probabilistic-network construction*. PhD thesis, Stanford University USA, 1991.

[114] Wray Buntine. Theory refinement on bayesian networks. In *Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*, pages 52–60. Morgan Kaufmann Publishers Inc., 1991.

[115] Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

[116] Joe Suzuki. A construction of bayesian networks from databases based on an mdl principle. In *Uncertainty in Artificial Intelligence*, pages 266–273. Elsevier, 1993.

[117] Moninder Singh and Marco Valtorta. An algorithm for the construction of bayesian network structures from data. In *Uncertainty in Artificial Intelligence*, pages 259–265. Elsevier, 1993.

[118] Ioannis Tsamardinos and Asimakis P Mariglis. Multi-source causal analysis: Learning bayesian networks from multiple datasets. In *IFIP international conference on artificial intelligence applications and innovations*, pages 479–490. Springer, 2009.

[119] Ronald R Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on systems, Man, and Cybernetics*, 18(1):183–190, 1988.

[120] Jules White, Chris Thompson, Hamilton Turner, Brian Dougherty, and Douglas C Schmidt. Wreckwatch: Automatic traffic accident detection and notification with smartphones. *Mobile Networks and Applications*, 16(3):285–303, 2011.

[121] Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

[122] Yunlong Zhang, Lori M Bruce, et al. Automated accident detection at intersections. Technical report, Mississippi State University, 2004.

[123] Hossam M Sherif, M Amer Shedid, and Samah A Senbel. Real time traffic accident detection system using wireless sensor network. In *2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pages 59–64. IEEE, 2014.

[124] Kaiqun Fu, Chang-Tien Lu, Rakesh Nune, and Jason Xianding Tao. Steds: Social media based transportation event detection with text summarization. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 1952–1957. IEEE, 2015.

[125] Angelica Salas, Panagiotis Georgakis, and Yannis Petalas. Incident detection using data from social media. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 751–755. IEEE, 2017.

[126] Otilia Popescu, Sarwar Sha-Mohammad, Hussein Abdel-Wahab, Dimitrie C Popescu, and Samy El-Tawab. Automatic incident detection in intelligent transportation systems using aggregation of traffic parameters collected through v2i communications. *IEEE Intelligent Transportation Systems Magazine*, 9(2):64–75, 2017.

[127] Shunsuke Kamijo, Yasuyuki Matsushita, Katsushi Ikeuchi, and Masao Sakauchi. Traffic monitoring and accident detection at intersections. *IEEE transactions on Intelligent transportation systems*, 1(2):108–118, 2000.

[128] Akira Kinoshita, Atsuhiro Takasu, and Jun Adachi. Real-time traffic incident detection using a probabilistic topic model. *Information Systems*, 54:169–188, 2015.

[129] Nour-Eddin El Faouzi, Henry Leung, and Ajeesh Kurian. Data fusion in intelligent transportation systems: Progress and challenges–a survey. *Information Fusion*, 12(1):4–10, 2011.

[130] Carola Otto. *Fusion of data from heterogeneous sensors with distributed fields of view and situation evaluation for advanced driver assistance systems*, volume 8. KIT Scientific Publishing, 2013.

[131] Messaoudi Zahir, Oussalah Mourad, and Ouldali Abdelaziz. Multiple target tracking using cheap joint probabilistic data association multiple model particle filter in sensors array. *International Journal of Artificial Intelligence & Applications*, 3(4):1, 2012.

[132] Arnoldo Díaz-Ramírez, Luis A Tafoya, Jorge A Atempa, and Pedro Mejía-Alvarez. Wireless sensor networks and fusion information methods for forest fire detection. *Procedia Technology*, 3:69–79, 2012.

[133] Kaijuan Yuan, Fuyuan Xiao, Liguo Fei, Bingyi Kang, and Yong Deng. Modeling sensor reliability in fault diagnosis based on evidence theory. *Sensors*, 16(1):113, 2016.

[134] Keyvan Golestan, Fakhri Karray, and Mohamed S Kamel. An integrated approach for fuzzy multi-entity bayesian networks and semantic analysis for soft and hard data fusion. In *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE, 2015.

[135] Yassmin Shalaby. *An integrated framework for coupling traffic and wireless network simulations*. PhD thesis, 2010.

[136] Ebrahim H Mamdani and Sedrak Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of man-machine studies*, 7(1):1–13, 1975.

APPENDICES

Appendix A

NCDB Dataset Variables' Description and Their Values

Possible values and their meaning for each variable used in the national collision database (NCDB) of Canada is summarized in this appendix.

Table A.1: Possible values for C_YEAR

Code	**Description**
19yy-20yy	yy = last two digits of the calendar year. (e.g., 90, 91, 92)

Table A.2: Possible values for C_MNTH

Code	Description
01	January
02	February
03	March
04	April
05	May
06	June
07	July
08	August
09	September
10	October
11	November
12	December
UU	Unknown
XX	Jurisdiction does not provide this data element

Table A.3: Possible values for C_WDAY

Code	Description
1	Monday
2	Tuesday
3	Wednesday
4	Thursday
5	Friday
6	Saturday
7	Sunday
U	Unknown
X	Jurisdiction does not provide this data element

Table A.4: Possible values for C_HOUR

Code	Description
00	Midnight to 0:59
01	1:00 to 1:59
02	2:00 to 2:59
03	3:00 to 3:59
04	4:00 to 4:59
05	5:00 to 5:59
06	6:00 to 6:59
07	7:00 to 7:59
08	8:00 to 8:59
09	9:00 to 9:59
10	10:00 to 10:59
11	11:00 to 11:59
12	12:00 to 12:59
13	13:00 to 13:59
14	14:00 to 14:59
15	15:00 to 15:59
16	16:00 to 16:59
17	17:00 to 17:59
18	18:00 to 18:59
19	19:00 to 19:59
20	20:00 to 20:59
21	21:00 to 21:59
22	22:00 to 22:59
23	23:00 to 23:59
UU	Unknown
XX	Jurisdiction does not provide this data element

Table A.5: Possible values for C_SEV

Code	Description
1	Collision producing at least one fatality
2	Collision producing non-fatal injury
U	Unknown
X	Jurisdiction does not provide this data element

Table A.6: Possible values for C_VEHS

Code	Description
01-98	01 - 98 vehicles involved.
99	99 or more vehicles involved.
UU	Unknown
XX	Jurisdiction does not provide this data element

Table A.7: Possible values for C_CONF

Code	Description	
Single Vehicle in Motion:		
01	Hit a moving object	e.g., a person or an animal
02	Hit a stationary object	e.g., a tree
03	Ran off left shoulder	Including rollover in the left ditch
04	Ran off right shoulder	Including rollover in the right ditch
05	Rollover on roadway	
06	Any other single vehicle collision configuration	
Two Vehicles in Motion - Same Direction of Travel:		
21	Rear-end collision	
22	Side swipe	
23	One vehicle passing to the left of the other, or left turn conflict	
24	One vehicle passing to the right of the other, or right turn conflict	
25	Any other two vehicle - same direction of travel configuration	
Two Vehicles in Motion - Different Direction of Travel:		
31	Head-on collision	
32	Approaching side-swipe	
33	Left turn across opposing traffic	
34	Right turn, including turning conflicts	
35	Right angle collision	
36	Any other two-vehicle - different direction of travel configuration	
Two Vehicles - Hit a Parked Motor Vehicle:		
41	Hit a parked motor vehicle	
Other		
QQ	Choice is other than the preceding values	
UU	Unknown	
XX	Jurisdiction does not provide this data element	

Table A.8: Possible values for C_RCFG

Code	Description
01	Non-intersection e.g., 'mid-block'
02	At an intersection of at least two public roadways
03	Intersection with parking lot entrance/exit, private driveway or laneway
04	Railroad level crossing
05	Bridge, overpass, viaduct
06	Tunnel or underpass
07	Passing or climbing lane
08	Ramp
09	Traffic circle
10	Express lane of a freeway system
11	Collector lane of a freeway system
12	Transfer lane of a freeway system
QQ	Choice is other than the preceding values
UU	Unknown
XX	Jurisdiction does not provide this data element

Table A.9: Possible values for C_WTHR

Code	Description
1	Clear and sunny
2	Overcast, cloudy but no precipitation
3	Raining
4	Snowing, not including drifting snow
5	Freezing rain, sleet, hail
6	Visibility limitation e.g., drifting snow, fog, smog, dust, smoke, mist
7	Strong wind
Q	Choice is other than the preceding values
U	Unknown
X	Jurisdiction does not provide this data element

Table A.10: Possible values for C_RSUR

Code	Description	
1	Dry, normal	
2	Wet	
3	Snow (fresh, loose snow)	
4	Slush ,wet snow	
5	Icy	Includes packed snow
6	Sand/gravel/dirt	Refers to the debris on the road, not the material used to construct the road
7	Muddy	
8	Oil	Includes spilled liquid or road application.
9	Flooded	
Q	Choice is other than the preceding values	
U	Unknown	
X	Jurisdiction does not provide data element	

Table A.11: Possible values for C_RALN

Code	Description
1	Straight and level
2	Straight with gradient
3	Curved and level
4	Curved with gradient
5	Top of hill or gradient
6	Bottom of hill or gradient ("Sag")
Q	Choice is other than the preceding values
U	Unknown
X	Jurisdiction does not provide this data element

Table A.12: Possible values for C_TRAF

Code	Description
01	Traffic signals fully operational
02	Traffic signals in flashing mode
03	Stop sign
04	Yield sign
05	Warning sign (Yellow diamond shape sign)
06	Pedestrian crosswalk
07	Police officer
08	School guard, flagman
09	School crossing
10	Reduced speed zone
11	No passing zone sign
12	Markings on the road (e.g., no passing)
13	School bus stopped with school bus signal lights flashing
14	School bus stopped with school bus signal lights not flashing
15	Railway crossing with signals, or signals and gates
16	Railway crossing with signs only
17	Control device not specified
18	No control present
QQ	Choice is other than the preceding values
UU	Unknown
XX	Jurisdiction does not provide this data element

Table A.13: Possible values for V_ID

Code	Description
01 - 98	01 - 98
99	Vehicle sequence number assigned to pedestrians
UU	Unknown. In cases where a person segment cannot be correctly matched with the vehicle that he/she was riding in, the Vehicle Sequence Number is set to UU.

Table A.14: Possible values for V_TYPE

Code	Description	
01	Light Duty Vehicle (Passenger car, Passenger van, Light utility vehicles and light duty pick up trucks)	
05	Panel/cargo van $\leq$ 4536 KG GVWR	Panel or window type of van designed primarily for carrying goods.
06	Other trucks and vans $\leq$ 4536 KG GVWR	Unspecified, or any other types of LTVs that do not fit into the above categories(e.g., delivery or service vehicles, chip wagons, small tow trucks etc.)
07	Unit trucks $>$ 4536 KG GVWR	All heavy unit trucks, with or without a trailer
08	Road tractor	With or without a semi-trailer
09	School bus	Standard large type
10	Smaller school bus	Smaller type, seats $<$ 25 passengers
11	Urban and Intercity Bus	
14	Motorcycle and moped	Motorcycle and limited-speed motorcycle
16	Off road vehicles	Off road motorcycles (e.g., dirt bikes) and all terrain vehicles
17	Bicycle	
18	Purpose-built motorhome	Exclude pickup campers
19	Farm equipment	
20	Construction equipment	
21	Fire engine	
22	Snowmobile	
23	Street car	
NN	Data element is not applicable	e.g., "dummy" vehicle record created for the pedestrian
QQ	Choice is other than the preceding values	
UU	Unknown	
XX	Jurisdiction does not provide this data element	

Table A.15: Possible values for V_YEAR

Code	Description
19yy-20yy	Model Year $19YY$ to $20YY$ where $00 \leq YY \leq CurrentYear + 1$
NNNN	Data element is not applicable (e.g., "dummy" vehicle record created for the pedestrian)
UUUU	Unknown
XXXX	Jurisdiction does not provide this data element

Table A.16: Possible values for P_ID

Code	Description
01-99	01-99
NN	Data element is not applicable (e.g., dummy" person record created for parked cars)
UU	Unknown (e.g., applies to runaway cars)

Table A.17: Possible values for P_SEX

Code	Description
F	Female
M	Male
N	Data element is not applicable (e.g., dummy" person record created for parked cars)
U	Unknown (e.g., applies to runaway cars)
X	Jurisdiction does not provide this data element

Table A.18: Possible values for P_AGE

Code	Description
00	Less than 1 Year old
01 - 98	1 to 98 Years old
99	99 Years or older
NN	Data element is not applicable (e.g., dummy" person record created for parked cars)
UU	Unknown (e.g., applies to runaway cars)
XX	Jurisdiction does not provide this data element

Table A.19: Possible values for P_PSN

Code	Description
11	Driver
12	Front row, center
13	Front row, right outboard, including motorcycle passenger in sidecar
21	Second row, left outboard, including motorcycle passenger
22	Second row, center
23	Second row, right outboard
31	Third row, left outboard
32	Third row, center
33	Third row, right outboard
etc.	
96	Position unknown, but the person was definitely an occupant
97	Sitting on someone's lap
98	Outside passenger compartment (e.g., riding in the back of a pick-up truck)
99	Pedestrian
NN	Data element is not applicable (e.g., dummy person record created for parked cars)
QQ	Choice is other than the preceding value
UU	Unknown (e.g., applies to runaway cars)
XX	Jurisdiction does not provide this data element

Table A.20: Possible values for P_ISEV

Code	Description
1	No Injury
2	Injury
3	Fatality (Died immediately or within the time limit.)
N	Data element is not applicable (e.g., dummy" person record created for parked cars)
U	Unknown (e.g., applies to runaway cars)
X	Jurisdiction does not provide this data element

Table A.21: Possible values for P_SAFE

Code	Description
01	No safety device used or No child restraint used
02	Safety device used or child restraint used
09	Helmet worn (For motorcyclists, bicyclists, snowmobilers, all-terrain vehicle riders)
10	Reflective clothing worn (For motorcyclists, bicyclists, snowmobilers, all-terrain vehicle riders and pedestrians)
11	Both helmet and reflective clothing used (For motorcyclists, bicyclists, snowmobilers, all-terrain vehicle riders and pedestrians)
12	Other safety device used
13	No safety device equipped (e.g., buses)
NN	Data element is not applicable (e.g., dummy" person record created for parked cars)
QQ	Choice is other than the preceding values
UU	Unknown (e.g., applies to runaway cars)
XX	Jurisdiction does not provide this data element

Table A.22: Possible values for P_USER

Code	Description
1	Motor Vehicle Driver
2	Motor Vehicle Passenger
3	Pedestrian
4	Bicyclist
5	Motorcyclist
U	Not stated / Other / Unknown

Appendix B

GB Dataset Variables' Description and Their Values

This appendix describes some of the variables of the GB dataset used in this thesis as a secondary dataset used for reinforcing the association rules obtained from the NCDB dataset. In order to keep the balance of the the number of variables in common, I did not use all the variables in this dataset. This appendix describes the variables that are not in common with those of the NCDB dataset.

Table B.1: Possible values for number of casualties

Code	**Description**
1	one casualty
2	two casualties
3	three casualties
4	four casualties
5	five casualties
6	more than five casualties

Table B.2: Possible values for speed limit (in miles per hour)

Code	Description
1	20
2	30
3	40
4	50
5	60
6	70

Table B.3: Possible values for light conditions

Code	Description
1	Daylight
4	Darkness - lights lit
5	Darkness - lights unlit
6	Darkness - no lighting
7	Darkness - lighting unknown
-1	Data missing or out of range

Table B.4: Possible values for vehicle manoeuvre

Code	Description
1	Reversing
2	Parked
3	Waiting to go - held up
4	Slowing or stopping
5	Moving off
6	U-turn
7	Turning left
8	Waiting to turn left
9	Turning right
10	Waiting to turn right
11	Changing lane to left
12	Changing lane to right
13	Overtaking moving vehicle - offside
14	Overtaking static vehicle - offside
15	Overtaking - nearside
16	Going ahead left-hand bend
17	Going ahead right-hand bend
18	Going ahead other
-1	Data missing or out of range

Table B.5: Possible values for junction location

Code	Description
0	Not at or within 20 metres of junction
1	Approaching junction or waiting/parked at junction approach
2	Cleared junction or waiting/parked at junction exit
3	Leaving roundabout
4	Entering roundabout
5	Leaving main road
6	Entering main road
7	Entering from slip road
8	Mid Junction - on roundabout or on main road
-1	Data missing or out of range

Table B.6: Possible values for skidding and overturning

Code	Description
0	None
1	Skidded
2	Skidded and overturned
3	Jackknifed
4	Jackknifed and overturned
5	Overturned
-1	Data missing or out of range

Table B.7: Possible values for vehicle leaving carriageway

Code	Description
0	Did not leave carriageway
1	Nearside
2	Nearside and rebounded
3	Straight ahead at junction
4	Offside on to central reservation
5	Offside on to centrl res + rebounded
6	Offside - crossed central reservation
7	Offside
8	Offside and rebounded
-1	Data missing or out of range

Table B.8: Possible values for first point of impact

Code	Description
0	Did not impact
1	Front
2	Back
3	Offside
4	Nearside
-1	Data missing or out of range

Table B.9: Possible values for pedestrian movement

code	Description
0	Not a Pedestrian
1	Crossing from driver's nearside
2	Crossing from nearside - masked by parked or stationary vehicle
3	Crossing from driver's offside
4	Crossing from offside - masked by parked or stationary vehicle
5	In carriageway, stationary - not crossing (standing or playing)
6	Same as code 5 but masked by parked or stationary vehicle
7	Walking along in carriageway, facing traffic
8	Walking along in carriageway, back to traffic
9	Unknown or other
-1	Data missing or out of range