2015-09-09

Vehicle Accident Severity Rule Mining Using Fuzzy Granular Decision Tree

Kiavarz Moghaddam, Hamid

UNIVERSITY OF CALGARY

Vehicle Accident Severity Rule Mining Using Fuzzy Granular Decision Tree

by

Hamid Kiavarz Moghaddam

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN GEOMATICS ENGINEERING

CALGARY, ALBERTA

AUGUST, 2015

Abstract

Road collisions are disasters that constitute a major cause of disability and untimely death. Therefore, the need for investigation of the conditions of road collisions and driver awareness on highways is critical.

A great deal of huge data, with regards to road collisions such as collision properties, road conditions, temporal information, environmental attributes, spatial measures and road geometry have been accumulated.

This thesis proposes a new fuzzy granular decision tree to generate road collision rules to apply to the discrete and continuous data stored in collision databases. To improve the efficiency of the algorithm, the fuzzy rough set feature selection is applied .The major highways in California are considered as a case study to examine the proposed approach. The experimental results demonstrate that the proposed method is more accurate and efficient than the traditional decision tree methods, with the less redundancy in constructing the decision tree.

Acknowledgements

I would like to take this opportunity to thank all the people who made this work possible. My deepest gratitude and appreciation goes to my supervisor, Dr. Xin Wang, for her remarkable guidance.

I would also like to thank the California Highway Patrol (CHP) for providing me with accreted collision spatial data. In particular, I wish to express my deep appreciation to CHP's department responsible for the State Integrated Traffic Records System, for helping me get the real dataset.

I am most grateful to the Faculty of Graduate Study and the Department of Geomatics Engineering at the University of Calgary.

Last but not least, I am indebted to my brother and friends, Majid Kiavarz and Majeed Pooyandeh, for their valuable comments and suggestions.

Dedication

I would like to dedicate this thesis to my parents, Bijan Kiavarz and Shahin Niknam, for their continuous love, trust and support.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Abbreviations

Abbreviations	Definition
ID3	Interactive Dichotomize 3
CART	Classification And Regression Tree
SLIQ	Supervised Learning in Quest
FRBS	Fuzzy Reasoning Based system
FGDT	Fuzzy Granular Decision Tree
PDO	Property Damage Only
SPRINT	Scalable PaRallelizable INduction
CDP	Criterium Decision Plus
CHAID	CHi-squared Automatic Interaction Detector
TAID	Theta Automatic Interaction Detector
FDT	Fuzzy Decision Tree
EDT	Entity Attribute Decision Tree
RDT	Reduce Attribute Decision Tree
FIS	Fuzzy Inference System
GrC	Granular Computing
IS	Information System
RST	Rough Set Theory
RFS	Rough set Feature Selection
FRFS	Fuzzy Rough Set Feature Selection

Chapter One: Introduction

1.1 Background Information

Road traffic collisions are a social and public health challenge, as they often result in injuries and fatalities (Organization 2013). The World Health Organization (WHO) reports that road collisions were the ninth leading cause of death in 2004 and are expected to be ranked as the fifth in 2030 (Organization 2013). WHO estimates that over 1 million people are killed each year in road collisions, which is equal to 2.1% of the annual global mortality, resulting in an estimated social cost of $518 billion (Peden, Scurfield et al. 2004). In North America, approximately 40,000 people are killed every year on the roads (Ameratunga, Hijar et al. 2006).

Previous traffic safety studies show that, in most cases, the occurrences of traffic collisions are seldom random in space and time; and, results indicate that the traffic collision concentration areas are dependent on geographic space (Anderson 2009). A concentration area is defined as an area or location where there is a higher probability for a collision to occur, based on historical data and spatial dependency. Thus, if the properties of high-risk locations on the roads can be identified, road safety managers can analyze the reasons for the high risk; and, vehicle drivers can be made aware of the danger and drive more carefully on the roads with high-risk properties.

In this regard, the identification of the risk parameters that significantly influence the severity of traffic collisions and determination of the relationships between factors and collision severity, which can be represented as vehicle collision severity rules, are important research topics.

A great amount of data attributes regarding road collisions, such as collision properties, road condition, temporal information, environmental attributes, spatial measures and road geometry, have been accumulated over time. However, it is still a challenge to analyze and extract rules from diverse historic collision data from large databases.

Data mining is a suitable solution to help decision-makers recognize the rules of vehicular collision severity of large databases (Shanthi and Ramani 2011). It is an approach that focuses on searching for new and interesting patterns rather than confirming the present ones. Therefore, it can be utilized for finding yet unrecognized and unsuspected relations between data. The main goal of inductive inference in this study is the analysis of a set of historical collision instances and discovery of collision severity rules, so that vehicle collision severities can be predicted by applying the discovered rules.

A decision tree is a data mining method for generating understandable rules (relations between attributes) (Peng and Flach 2001). It provides a hierarchical representation of the data and a decision path to create logical rules. The most frequent usage of decision tree algorithms are ID3 (Iterative Dichotomiser 3), C4.5 (extension of ID3), CART (classification and regression tree), SLIQ (Supervised Learning in Quest) and random tree, which are used as classic methods for the generation of decision trees (Lavanya and Rani 2011).

1.2 Problem Statement

The direct application of the traditional decision tree mining methods on collision datasets is not feasible. There are some issues in the traditional methods that reduce the data mining performance and the accuracy of the results. In this section, some of the issues are discussed.

Vehicle collision datasets include inconstant data. The inconsistency in the dataset confuses the decision tree methods. When a decision tree method is studied, wrong predictions can sometimes be observed when inconsistent data are present. For example, two instances in the dataset with same attribute values may fall into different collision severity classes, which may result in the traditional decision tree methods presenting the wrong decision for vehicle collision severity. Incorrect predictions in vehicle collision severity can cause erroneous decision-making for drivers.

Collision datasets usually have a large number of attributes, such as collision properties, road conditions, temporal information, environmental attributes and road geometry. Not every attribute contributes to collisions. Traditional decision tree methods are not good at handling datasets with large numbers of attributes, because many branches are generated with duplication and repetition of sub-trees.

Vehicle collision data contain both discrete and continuous attributes. For example, the weather condition at the time of collision is a discrete attribute; and, the spatial measures of the vehicle collisions, such as the distance between the collision location and intersections, are usually continuous attributes. The conditional entropy in traditional decision tree algorithms considers only discrete attributes. Therefore, discretization is generally applied to continuous variables before applying traditional decision tree methods. However, discretization of continuous values in a dataset increases the uncertainty of classification and reduces the accuracy of the final classification.

The last problem is related to the selection of the proper vehicle collision attributes. In traditional classification methods, an attribute is chosen solely based on information about a node, which creates redundancy in the decision tree.

1.3 Motivation

This research focuses on the design and implementation of a framework that provides a collision prediction model based on road locations. The historical data at collision locations with the properties of road and environmental status can be used to predict the probabilities of a collision and its severity. The motivation of this study focuses on the needs of two types of users:

- Drivers. The framework can provide a good sense of their risk in terms of specific conditions, such as different time intervals, weather or road surface conditions, on a road network based on the historical accident dataset. In this situation, the time consumption of the algorithm is important, since drivers can add optional attributes (conditions) to the application and allow the method make a new vehicle collision severity prediction.
- Traffic road decision-makers. Using the proposed method, they can investigate the collision severity on major roads and make predictions based on different scenarios. Traffic decision-makers can also observe the extracted rules and

determine which attributes have the most effect on the severity of collisions on specific roads. The information can help them make better decisions about lighting, road changes or designing a traffic alarm to increase the safety at critical road locations.

1.4 Research Objectives

The goals of this thesis are to address the challenges of inconsistent vehicle collision datasets and to mine the collision datasets to determine the collision severity rule. To achieve these objectives, the following steps needed to be taken:

1. Construction of a feature selection model to deal with inconsistent vehicle collision datasets and select those vehicle collision attributes that have minimum correlation and improve the performance of data mining methods without losing the accuracy of results. The feature selection model also had the capability of selecting the most information-rich attributes in a dataset without the loss of information needed for the classification and without transformation of the data.
2. Investigation of the role of granular computing in data mining and extraction of the optimal rules between vehicle collisions attributes by a decision tree method. In this step, the following questions needed to be answered:
    - What is a suitable learning approach using granular computing for employing vehicle collision attributes and generating efficient rules between them?

- How can the discrete and spatial data values be handled during the learning approach and rule generation process?
- What is the best solution using the concept of granularity to involve all attributes for constructing each level of decision tree rather than only considering the attribute of one node?

3. Application of a suitable reasoning method as a decision engine using the results of the learning method to determine the classes of vehicle collision severity. In the context of the vehicle collision classification task, answers were sought for the following research questions:
   - What is the best solution to involve all extracted rules to determine the final severity classification of a probable collision event?
   - How should the discrete and continuous data in dataset be incorporated into the process of vehicle collision severity classification?

1.5 Research Contribution

The research for this thesis makes three main contributions:

1. Application of a fuzzy rough set and approximation concept to approximate the inefficient inconstant vehicle collision dataset. A fuzzy rough set feature selection on vehicle collision data is employed to select the features that have minimum correlation and considered the inconstancy in the collision dataset.
2. Proposal of a fuzzy granular decision tree (FGDT) by introducing the fuzzy granular entropy to measure the degree of disorder or uncertainty of objects in

each granular dataset, with respect to both discrete and numerical data. In this regards, the spatial properties of collision events, which are considered continuous data, can be involved in constructing the decision tree. To ensure the accuracy of continuous data, the fuzzy granular entropy is calculated based on the exact value of the continuous value and fuzzy membership functions rather than using the categorized continuous attributes.

3. Proposal of the inference process of a fuzzy reasoning based system (FRBS), using fuzzy membership functions as input data and fuzzy granular decision tree rules, which are if-then linguistic rules whose antecedents and consequents are composed of fuzzy statements. The FRBS can be applied as a decision engine to use the extracted rules from FGDT to determine the final classes of vehicle collision severity with respect to all extracted rules.

1.6 Organization of the Thesis

This thesis is organized as follows: Chapter 2 provides a literature review on existing decision tree methods and their limitations and on fuzzy decision tree methods. Chapter 3 describes fuzzy rough set feature selection for inconsistent vehicle collision data and discusses how to deal with the inconstant data before constructing a decision tree. Chapter 4 presents the proposed method fuzzy granular decision tree for vehicle collision severity rule extraction and fuzzy reasoning as the decision engine. Chapter 5 describes the implementation of the proposed methods on twelve main highways as a case study, and Chapter 6 provides the conclusion of the thesis and discusses future research directions.

Chapter Two: Literature Review

2.1 Introduction

This chapter presents an overview of related research. Section 2.2 discusses rough set theory and feature selection, and Section 2.3 introduces conventional decision trees and their properties. Section 2.4 discusses fuzzy decision trees and their importance of them, and Section 2.5 presents a comprehensive overview of granular and rough set decision trees. Section 2.6 presents an overview of the fuzzy inference methods, and Section 2.7 describes some research on traffic collisions and road safety. Section 2.8 is a quick review on the overall view of the proposed methods in this research.

2.2 Rough Set Theory and Feature Selection

There are many feature selection methods in knowledge discovery and data mining. It may be expected that an increasing number of attributes (features) would increase the information to distinguish between decision classes; however, it is not true that increasing the size of features in the training dataset can increase the capability and accuracy of classification methods to determine the final decision efficiently (Beynon 2001, Jensen and Shen 2002).

A high-dimensional dataset increases the chances of a data-mining algorithm, like decision tree methods, and can find false patterns or rules that are not valid. Most data-mining techniques engage feature reduction or feature selection methods in order to cope with high dimensional datasets (Jensen and Shen 2002).

One successful approach in the feature selection process is using the rough set theory (RST) to obtain features (Ziarko 2015). Over the past twenty years, rough set theory has become a topic of interest for researchers who are employed in many domains, such as classifications (Jing 2015), clustering (Ho, Kawasaki et al. 2003), expert systems (Shen and Jensen 2004) and data mining. The success of using RST in feature selection is related to the lack of need for additional information, such as determining thresholds or expert knowledge, dealing with inconsistent data in datasets, and finding minimal knowledge in the data (Jensen and Shen 2009). The power of RST method is the finding of the most informative subset of the original features from a given dataset. One of the most important advantages of this method is that feature selection can be applied to a dataset with the removal of all other features from the dataset resulting in a minimal loss of minimal information.

2.2.1 Rough Set Feature Selection

The concept of rough set feature selection (RFS) is based on the indiscernibility of instances in the dataset. This method tries to select those features to make better discernibility between instances with respect to features (conditional attribute or feature) and the final decision classes (decision attributes of feature). For more clarification of the RFS concept, let I = (U, A) be an information system, where U is a non-empty set of finite instances (the universe) and A is a non-empty finite set of attributes such that Va is the value of each row for a specific attribute (a ∈ A). Suppose for any $P \subseteq$ A, there is an corresponding equivalence relation, which is called IND(P) (Jensen 2005):

$$IND(P) = \{(x, y) \in \mathrm{U}^2 \mid \forall a \in P, a(x) = a(y)\} \qquad (1)$$

In the above statement, x and y are two values of a decision table belong to attribute a ∈ A. *IND(P)* generates the partition of U, which is denoted U/IND(P), and can be calculated as follows (Jensen and Shen 2009):

$$U/IND(P) = \otimes\{a \in P : U/IND(\{a\})\} \quad (2)$$

$$A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y 6= \emptyset\} \quad (3)$$

The instances in a partition of U are indiscernible by a conditional attribute. Therefore, if (x, y) ∈ IND(*P*), then x and y are indiscernible by attributes from *P*. The equivalence classes of the *P*-indiscernibility relation are denoted as $[x]_P$. Let X ⊆ U.

In RST, the two concepts of the *P*-lower and *P*-upper are defined to approximate X using the information contained within them. The lower approximation (or positive region) is the union of all instances that can be classified certainly in one of the decision classes, whereas the upper approximation is a description of the instances that possibly belong to one of the decision classes. These two approximation set are defined as (Jensen 2005, Ziarko 2015):

$$PX = \{x \mid [x]_P \subseteq X\} \quad (4)$$

$$PX = \{x \mid [x]_P \cap X 6= \emptyset\} \quad (5)$$

Let *P* and *Q* be equivalence relations over U, then the positive, negative and boundary regions can be defined as:

$$POSP(Q) = \cup_{X \in U/Q} \underline{P}X \quad (6)$$

$$NEG_P(Q) = U - \cup_{X \in U/Q} \overline{P}X \quad (7)$$

$$BND_P(Q) = POS_P(Q) - NEG_P(Q \qquad (8)$$

The positive region contains all instances in the dataset of U that can be classified into the decision classes of U/Q using the information in the attributes of P. Boundary region *BNDP(Q)* is the set of instances that can possibly, but not certainly, be classified in decision classes. Negative region *NEGP(Q)* is the set of instances that cannot be classified to decision classes (Jensen 2005, Ziarko 2015*).*

2.2.2 Calculation of Dependency between Features

Determination of the lower and upper approximations help in the calculation of dependencies between attributes in feature selection methods, which is an important issue. A set of attributes (Q) depends totally on a set of attributes (P), if all attribute values from Q are solely specified by the values of attributes from P. In RST, dependency is defined in the following way: P, $Q \subset A$ and Q depend on P to degree k, where $0 \leq k \leq 1$.

$$K = \gamma_p(Q) = \frac{|POS_p(Q)|}{|U|} \qquad (9)$$

If $k = 1$, Q depends totally on P, if $0 < k < 1$, Q depends partially on P, and if $k = 0$ then Q does not depend on P.

2.2.3 Limitation of Rough Set Feature Selection

Datasets often include spatial data where the values of attributes may be both crisp and real valued: this is where many feature selection methods, such as those based on traditional rough set theory, face a problem. It is clear that there is a need for a method that provides the means of feature selection for crisp and real-value attributed datasets. Fuzzy sets (Zadeh 1965) prepare a mechanism by which real-valued features can be effectively managed.

The vagueness and uncertainty present in inconsistent datasets can be modelled by assigning values that belong to more than one label, using the membership function definition and lower and upper approximations of that dataset. This can be achieved with the use of fuzzy rough sets for feature selection (Sahatier 1992). Crisp RST can be extended with fuzzy RST, which allows all instances in upper and lower approximation sets to take values in the range of [0,1]. Therefore, in this research, the fuzzy RFS was applied to a case study dataset.

2.3 Decision Trees

Classification-rule learning involves finding rules or decision trees that partition the given data into predefined classes. For any realistic problem domain of the classification-rule learning, the set of possible decision trees is too large to be exhaustively searched. In fact, the computational complexity of finding an optimal classification decision tree is non-deterministic polynomial-time hard (NP-hard). The main existing decision tree algorithms, such as like Iterative Dichotomiser 3 (ID3) (Quinlan 1986), an extension of ID3 (C4.5) (Quinlan 1993), , SLIQ (Supervised Learning in Quest) (Mehta, Agrawal et al. 1996), and scalable parallel classifier for data mining (SPRINT) (Joshi, Karypis et al. 1998), use Hunt's method (Quinlan 1983) as the basic algorithm.

2.3.1 Hunt's Method

Hunt's algorithm is an interactive method that grows a decision tree by partitioning the training objects into subsets. Let DT be the set of training objects that reach node t with the classes of decision. The general procedure is defined as below (Quinlan 1983):

1. If DT contains instances that belong to the same class C_j, then t is a leaf node labeled as C_j.
2. If DT is an empty set, then t is a leaf node labeled by the default class, C_j.
3. If DT contains instances $\{O_1, O_2 \ldots On\}$ that belong to more than one class.

This method uses an attribute test to split the data into smaller subsets. Suppose T is partitioned into subsets DT, where DT_i contains all the cases in DT that have outcome O_i of the chosen test. The decision tree for DT consists a decision node identifying the test and one branch for each possible outcome. The same tree building machinery is applied recursively to each subset of training cases.

Table 2.1 shows a training data set with four data attributes and three classes. Figure 2.1 shows how Hunt's method works with the training dataset. In case 3 of Hunt's method, a test based on a single attribute is chosen for expanding the current node. Attribute selection is normally based on the entropy gains of the attributes. The entropy of an attribute is calculated according to the information of the class distribution.

Table 2.2 shows the class distribution information of the weather data attribute of. For a continuous attribute, a binary test of all the distinct values of the attribute is considered. Table 2.3 shows the class distribution information of the collision time data attribute with three decision value of PDO (property damage traffic), injury and Fatal. Once the class distribution information of all the attributes is gathered, the entropy is calculated based on either information theory or the Gini index. One attribute with the most entropy gain is selected as a test for the node expansion.

2.3.2 ID3 Algorithm

The ID3 algorithm (Quinlan 1986) is a decision tree method in which the classification of instances in the dataset is specified by testing the values of their properties. It constructs the tree starting from a set of instances and a specification of properties in a top-down approach. The property of values is tested at each node of the tree, and the results are used to partition the instances in the dataset. This process is internationally done till the set in a given sub-tree belongs to one decision class. At each node, the node is chosen based on maximizing the information gain and minimizing entropy.

Table 2.1 Small Training Dataset of Vehicle Collisions

Granular	Weather	Surface	Lighting	Time	Severity
1	Clear	Dry	Day-Light	10:30	PDO
2	Clear	Dry	Dusky OR Dark	8:45	PDO
3	Clear	Dry	Dusky OR Dark	21:35	PDO
4	Clear	Not Dry	Day-Light	11:40	PDO
5	Clear	Not Dry	Dusky OR Dark	9:25	Injury
6	Clear	Not Dry	Dusky OR Dark	23:30	Injury
7	Rainy	Not Dry	Day-Light	10:00	Injury
8	Rainy	Not Dry	Dusky OR Dark	8:45	Injury
9	Rainy	Not Dry	Dusky OR Dark	19:25	Injury
10	Fog	Dry	Day-Light	19.:00	PDO
11	Fog	Dry	Dusky OR Dark	9:25	Injury
12	Fog	Dry	Dusky OR Dark	21:25	Fatal
13	Fog	Not Dry	Day-Light	10:44	Fatal
14	Fog	Not Dry	Dusky OR Dark	21:17	Fatal

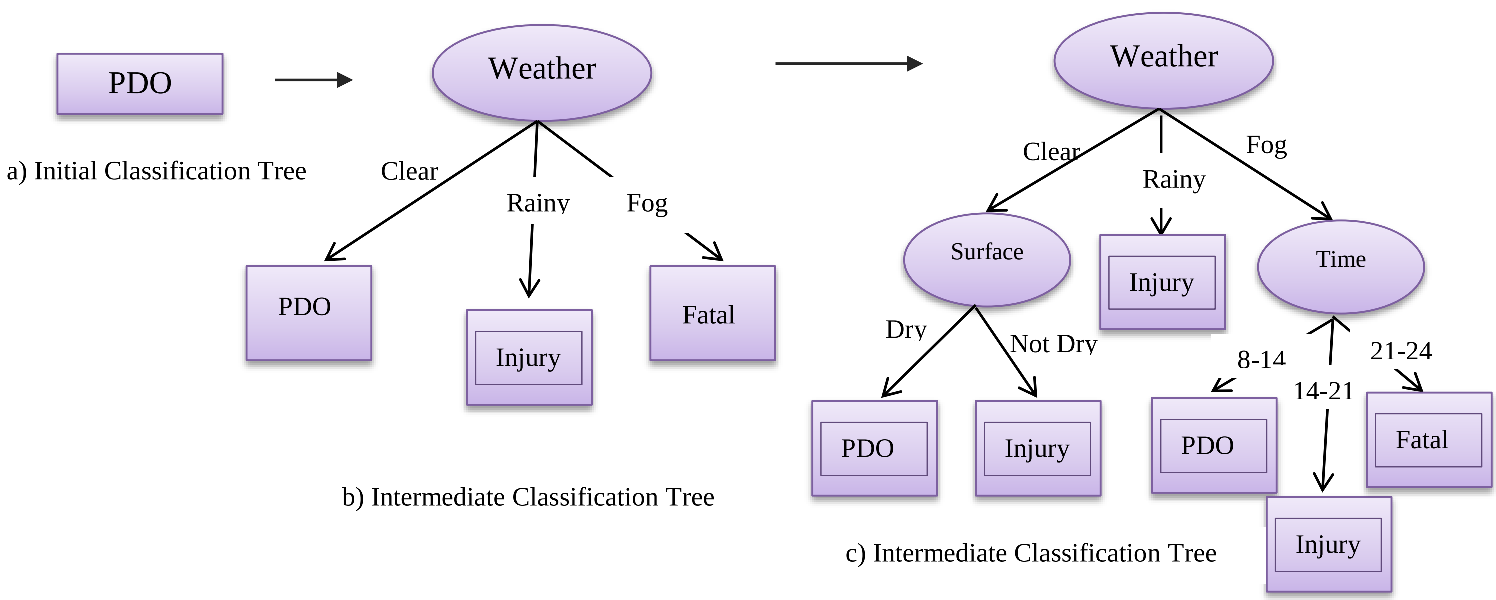


Figure 2.1 Demonstration of Hunt's Method

Table 2.2 Class Distribution Information of the Weather Attribute

Attribute Value	**Decision Classes**		
	PDO	Injury	Fatal
Clear	4	2	0
Rainy	0	3	0
Fog	1	1	3

Table 2.3 Class Distribution Information of Collision Attributes

Attribute Value	**Decision Classes**		
	PDO	Injury	Fatal
8 – 14	4	4	1
14 - 21	1	1	0
21 - 24	1	1	2

2.3.3 C4.5 Algorithm

The C4.5 algorithm was proposed by Quinlan (1993). The recursive partitioning of data is the main structure of the C4.5 algorithm. The algorithm considers all the possible tests and selects a test with the best information gain. For each discrete attribute, each distinct value belong to that specific attribute is tested. For each continuous attribute, the partitioning is applied to all attributes. In order to gather the entropy gain of all these tests efficiently, the training dataset belonging to the node in consideration is sorted by the values of the continuous attribute and the entropy gains of the binary divided based on each distinct value. This process is repeated for each continuous attribute.

2.3.4 C5.0/Sec 5 Algorithm

The C5.0 algorithm is an extension of the C4.5 and ID3 algorithms. It is a classification algorithm that can be applied to big datasets. It is better than the C4.5 algorithm with respect to the speed, memory and efficiency. The C5.0 method splits the dataset for those attributes that provide the maximum information gain. The process continues until the subset belongs to just one decision class. C5.0 easily handles multi-value attributes and missing attributes from dataset (Patil, Lathi et al. 2012).

2.3.5 Classification and Regression Tree

The classification and regression tree (CART) is a popular classification statistical method. CART splits the data by iteration until end nodes are achieved using the present criteria. CART operates by analyzing all explanatory variables and then determining which binary division of a single explanatory variable reduces deviance in the response variable (Breiman et al., 1984; Efron and Tibshirani, 1991; Venables and Ripley, 1997).

The process of CART is repeated for each part of the split data, continuing until the final nodes belong to one of the decision classes in a hierarchical tree. CART constructs a tree that describes the deviancy based on the original data. However, CART uses the dataset to fit it to tree. Therefore, to create a robust tree, a pruning method should be applied to the constructed tree. For example, the dataset is split into ten equal sets; and, to create the best size of tree, the tree is generated by nine parts and validated by the last remain part of the dataset.

It is obvious that the result of CART analysis is a decision tree, and a pruning method may select a smaller tree. A series of dichotomous splits is defined in each path of the tree, which specifies the conditions that lead to a most probable class. Therefore, the structure of the rules can be used for unknown observations to predict likely class membership.

2.3.6 CHAID

CHAID (CHi-squared Automatic Interaction Detector) is a basic decision tree learning algorithm. It was developed by Gordon V Kass (Kass 1980) in 1980. CHAID is easy to interpret, easy to handle and can be used for classification. CHAID is an extension of the AID (Automatic Interaction Detector) and THAID (Theta Automatic Interaction Detector) procedures. After detection of interaction between variables, it selects the best attribute for splitting the node which made a sub- node as a collection of equivalent values of the selected attribute. The method can handle missing values. It does not imply any pruning method (Patil, Lathi et al. 2012).

2.3.7 Some Limitations of Recent Decision Tree Algorithms

It is obviously all decision tree methods have their weakness and limitations. In this section some limitations of the mentioned decision tree algorithms are investigated respect to vehicle collision data base.

The fragmentation problem exists if data set gradually partitioned into smaller segments (Setiono and Liu 1998, Yao, Liu et al. 2005). Replication and repetition can cause fragmentation. Also, fragment can occur when many conditional features are involved in the process of construction diction tree. The algorithms like ID3 and C4.5 uses a top-down structure to recursively partition the training data. This strategy leads to the fragmentation problem because towards the end of the subdivision process, the size of the underlying data becomes quite small, even though a statistical test requires a data of significant size. The replication problem will be ignited if sub-trees are replicated in decision tree. It leads to the data set to be split into the smaller segments which indicate fragments problem (Setiono and Liu 1998, Yao, Liu et al. 2005). Partitioning in continuous data is the other issue in the decision trees (Jing-ti and Yu-jia 2009). Attributes having discrete values can be easily partitioned but continuous attributes like spatial properties of vehicle collision (ex. distance of collision to intersection locations or slope degree) have problem with partitioning in the process of creating decision tree . The algorithms like C4.5, C5.0 and CART apply the partitioning methods on data set and lose the accuracy of exact value of continuous attributes. Repetition problem is ignited if the features are repeatedly tested along a path in a decision tree(Setiono and Liu 1998, Yao, Liu et al. 2005). These repetitions explode data set into smaller and smaller segments, hence result in fragmentation.

ID3, 4.5 and C5.0 willing to take multi-valued attributes (Rui-Min and Miao 2010, Thakur, Markandaiah et al. 2010). They always select the attribute with many values which causes the wrong classification result. In some cases, some non-valuable attributes have the highest value of gain ratio (because of the formula of the information gain and entropy) and some valuable attribute's gain ratio becomes lower. So difficulty would come of making the root of the tree.

ID3 and C4.5 algorithm does not follow back in searching (Rui-Min and Miao 2010). Whenever certain layer of nodes in the tree chooses a property to test, it will not backtrack to reconsider this choice. In this way, the algorithm could easily solve the local optimal problems, but not global optimal issues.

The other issue is that the decision tree algorithm does not provide rowing learning. Indeed, ID3 algorithm cannot accept training sample incrementally, so the each increase of example requires abandoning original decision tree, to restructure new decision tree, and to lead to lots of overhead (Jing-ti and Yu-jia 2009).

Sometimes the dataset may have attributes with the range values. Current decision tree methods use the min-max, mean or medium value as representative value of that range which is not suitable. As an example the attributes such as slope has range value as inputs. We need to improve the membership grade and entropy calculation method to handle the attributes with the range value.

As vehicle collision events' databases contain both discrete and continuous attributes, the above-mentioned decision tree methods cannot apply the continuous values such as spatial measures based on their real values. So, it decreases the accuracy of decision tree. The fuzzy decision trees help to apply the real value of continuous values in vehicle collision data sets.

2.4 Fuzzy Decision Tree

Regarding decision trees, the ID3, CART, and C4.5 algorithms are among the most relevant ones. Fuzzy decision trees have also been proposed in the literature (Chang and Pavlidis

1977, Hori, Umano et al. 1999, Olaru and Wehenkel 2003, Janikow 2004, Tokumaru and MURANAKA 2010). Fuzzy decision trees combine the powerful models of mentioned decision trees with the ability to process uncertainty and imprecision of fuzzy systems. Moreover, fuzzy decision trees borrow the admirable properties of decision trees regarding their low computational induction cost, as well as the rules extraction properties in a low computational cost way.

2.4.1 Special Issues of Fuzzy Decision Tree Methods

Such as classical decision trees, fuzzy decision trees are constructed in a top-down manner by recursive partitioning of the training set into subnets. Here, we list some special issues of fuzzy decision trees (Chen, Wang et al. 2009):

- Attribute selection criteria in fuzzy decision trees

  A conventional method to select an attribute in classical decision tree is to choose the attributed with the highest information gain. But in fuzzy decision methods, the information gain is depended on fuzzy membership function values. So, the conventional information gain methods cannot be efficient. Some modifications and enhancements of the basic algorithm have been proposed and studied.

- Inference for decision assignment

  The inference procedure is an important part of FDT, and it needs to design the inference method so designing and analysis of inference methods increases the complexity of model.

- Stopping criteria

  In the conventional decision tree methods, the termination happens if all attributes belong to one decision class; or if all examples in the current node belong to the same class. In FDT, an example may occur in different classes with different membership values.

2.5 Granular Computing and Decision Trees

As mentioned in the decision trees issues section, decision trees suffer from getting overly complex and being easily affected by slight changes in the training set (Rokach 2008). To overcome these disadvantages, ideas from granular computing, in particular rough sets have been employed to investigate the relationships that exist between the complexity of the decision tree, data set consistency and the classification accuracy, as the dataset is transposed to a lower granular level.

In particular the following three forms of relationships will be explored:

1. The relation between accuracy, complexity and consistency of model based on the attribute binning.

2. The relation between accuracy, complexity and consistency of model based attribute reduction or selection.

3. The relation between accuracy, complexity and consistency of model based data modification.

2.5.1 Granular Computing

As discussed earlier, one of the problems with decision trees is that they get overly complex with large data set. Therefore, researchers have employed other techniques to construct decision trees that are small and accurate.

With combining the concept of simplicity and accuracy, the concept of granular computing will be emerged. Granular computing views the world by diving into entities called information granules that are grouped together due to their similarity, functional adjacency, in distinguishability or coherence (Bargiela and Pedrycz 2003). A highly detailed granular world can be abstracted into lower granulation using formal frameworks that approximates the original representation. This can be formally written as (Bargiela and Pedrycz 2003):

$$G =< X, Ģ, A\ldots >$$

where G is the granulation process, X is the object to be granulized, $Ģ$ is a family of reference and A refers to Attributes. In this method, knowledge is split into hierarchical steps. Regarding this hierarchical model, the information pyramid is built where the granules at the base are large in number and containing the most details and the granules at the apex have the smallest number of details, containing only the core information granules. This is illustrated in the figure below:

Moving from one step layer to another involves the process of abstraction. The issue of how to define the size of an information granule is one of fundamental problems in the field of granular computing In general, high information granularity levels are associated with a decrease in the usefulness of the concept.

Figure 2.2 Information Processing

In the framework of granular computing, the problem of lake comprehensibility can be modeled with the concept of high granulation. Decision trees also share the two primary motivations of all formal methods used in granular computing (Bargiela and Pedrycz 2003) :

1. A need to split a problem into smaller tasks.
2. A need to comprehend the problem without getting into unnecessary detail.

This sharing of principles between data mining and granular computing was also recognized by (Bargiela and Pedrycz 2003). Hence, granular computing serves as a powerful model for solving the problem of balancing accuracy and size of a decision tree.

2.5.2 Rough Sets and Decision Trees

Using the concept of Rough set in decision trees has mostly focused on either (a) employing the rough sets as an alternative of information gain to find an appropriate split attribute in decision tree construction, or (b) data pre-processing step using rough set theory (RST) to manipulate the dataset inputted in the decision tree. RST has also been used in comparison studies between the two approaches of decision making.

Han and Kim (2007) find out the disadvantages of traditional ID3 and C4.5 algorithms since the attributes are chosen without considering the correlations so that sometimes high frequency attributes are selected. The entity attribute decision tree (EDT) and the reduce attribute decision tree (RDT) methods are proposed by them. Also, they showed their advantages of accuracy and rule simplification (Han and Kim 2008).
RST has also been used to calculate the degree of dependence between the conditions and decision attribute. The condition attributes with the highest significance value is used as the splitting criteria in the construction of the decision tree (Han and Kim 2008, Wang and Ou 2008).

Within the same framework of finding appropriate selection of attribute and dealing with missing values, Li, Ruan et al. (2007) developed an algorithm that calculated the weighted mean roughness of every condition attribute and then selected the attribute that had the smallest mean roughness.

Rough set theory has also been used for feature selection in pre-processing step to eliminate redundant data. Zhou, Zhang et al. (2008) in comparisons with the C4.5 algorithm showed that the removal of redundant attributes can increase the prediction accuracy of a decision tree.

Minz and Jain (2003), (Minz and Jain 2005, Sikder and Munakata 2009) find out the importance of rough sets as a pre-processing step where the datasets is small and so no strong conclusions can be drawn. Their research shows the need to filter out redundant attributes.

Similar works were carried out by Yellasiri, Rao et al. (2007), which showed that rough sets pre-processed decision trees produced the highest accuracy compared with the ID3, C4.5 and CART algorithms.

Comparisons between rough sets and decision trees done by (Sikder and Munakata 2009) to identify important factors of earthquakes revealed that there were no statistically significant differences between the C4.5 method and rough sets. The fuzzy-rough sets approaches as a hybrid method have also been used extensively in rules induction step and compared with the C4.5 method produced better accuracy (Thangavel and Pethalakshmi 2009).

Decision tree has been used for rules generated by rough sets with the aim of visualization (Ilczuk and Wakulicz-Deja 2007). This research seeks the benefit of decision trees to provide comprehensibility in determining decisions.

Finally, after constructing an optimal decision tree, all rules are extracted from constructed decision tree. A decision engine should be applied to these rules to determine the final classes of severity for all vehicle event objects. This method can use the extracted rules to classify the vehicle collision severity.

2.6 Fuzzy Inference Systems

Vehicle collision events may be influenced by many environmental, human, spatial and road geometry factors. To use these influencing factors for vehicle collision severity prediction, the prediction accuracy largely depends on the quality of the factors, supplied with the efforts expanded in collecting, analyzing and processing. On the other hand, historical observations embedded with rich information can be used to infer the future collision severity. Therefore, this research focused on the prediction of vehicle collision severity based on historical recorded events. Road vehicle collision severity is, however, a complex, open and time variant system that often exhibits highly randomness and uncertainty. This challenge has bring up considerable research to be developed to this field. As a result, a wide range of classification and prediction method has been developed, such as the Kalman filter (Okutani and Stephanedes 1984) and it's extension (Gazis and Liu 2003), the support vector machine (Gazis and Liu 2003), Bayesian networks (Scutari, Howell et al. 2014), and hybrid approaches (Guo-jiang 2010). The fuzzy inference system (FIS) is a well-known method to deal with the incomplete dataset. This model is tolerant to noise, uncertainty. Also, it is an interpretable model and easy to incorporate expert and field knowledge (Wang and Liu 2008). Fuzzy approaches have the following properties with comparison to the other classification and prediction mentioned approaches: 1) the calculation of the dependency between inputs and out puts of a system 2) the fuzzy linguistic variables provides a natural way to deal with uncertainty; 3) the modeling of nonlinear systems can be handle by them; 4) the singular and linguistics output can be easily generated; 5) they are insensitive to random noise in the dataset.

A nonlinear mapping from its inputs space to output space will be constructed by fuzzy inference system with crisp inputs and outputs. This mapping is done by a number of fuzzy if-then rules which is extracted from decision trees. In particular, the antecedent of a rule defines a fuzzy region in the input space, while the consequent specifies the output in the fuzzy region. Basically a fuzzy inference system is composed of five steps as shown in Fig 2.3. The Structure of the Fuzzy Inference system is described as follows.

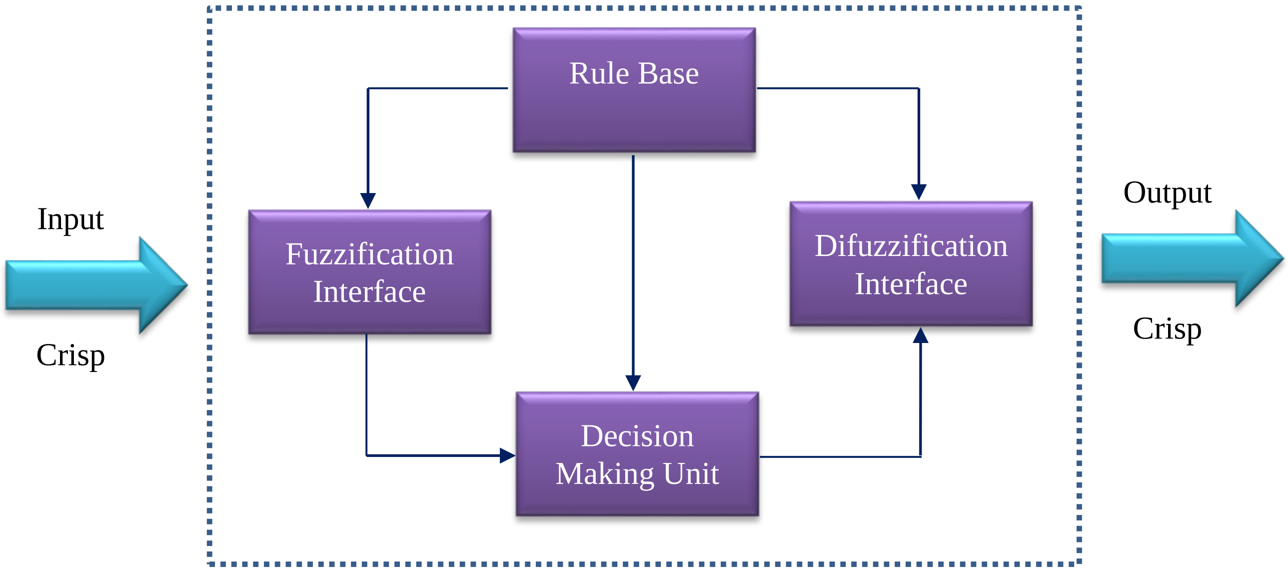


Figure 2.3 Structure of the fuzzy inference

Two types of FIS, namely Mamdani FIS (Mamdani and Assilian 1999) and Sugeno FIS (Sugeno 1985) are widely accepted and applied to solve many real-world problems.

In terms of use and applications, the Mamdani FIS is more widely used, mostly due to its reasonable results with a relatively simple structure, and its intuitive and interpretable nature of the rule base (Jassbi, Serra et al. 2006). Since the consequents of the rules in a Sugeno FIS are not fuzzy, this interpretability is lost; however, since the Sugeno FIS's rules can have as many parameters per rule as input values, this translates into more degrees of freedom in the design than a Mamdani FIS, thereby providing the system's designer with more flexibility in the design of the system (Mendel 2001). It should be noted that the Mamdani FIS can be used directly for either multiple input, single output (systems) or multiple input, multiple output (MIMO) systems, such as vehicle collision severity classification and prediction, whereas the Sugeno FIS can only be used in MISO systems.

Moreover, the classification and prediction of vehicle collision severity should be crisp to be useful for users and decision-makers. Since the Mamdani FIS has a defuzzification step to facilitate the crisp output, the Mamdani FIS was used in this research as the decision engine of determining the final class of events.

2.7 Traffic Collisions Road Safety Research

In general, the impact factors for collision risks and, thus, the reasons for variation can involve three interdependent realms: (1) risk exposure, (2) environment, and (3) social and psychological human factors. For instance,Thomson and Tolmie (2001) modeled risk as a function of exposure multiplied by hazard divided by traffic skill. Exposure, in their understanding, was mainly captured by time spent on the street, as their model refers to children. The degree of hazard was determined by the local traffic environment and traffic skills were considered as individual and household factors that affect a person's ability to deal appropriately with the hazard. Risk exposure is mainly an outcome of transport mode use, travel distances and traffic mix (Elvik, Vaa et al. 2009). For example, large proportions of cyclists may be associated with low risk levels, possibly because car users drive more carefully in regions where many people cycle (Vandenbulcke, Thomas et al. 2009).

In terms of Collison severity, driving speeds is a major factor. Environmental factors include the condition of the road network, road type and design, spatial context (e.g. density, land-use, distances to intersections, etc.), temporal context (e.g., darkness) and transport context (traffic density, speed and behaviour of other transport users). Social and psychological factors include socio-demographic and socio-economic structures, risk attitudes, lifestyles and associated behaviour. For instance, in a questionnaire survey in Norway of 900 young adults Elvik, Vaa et al. (2009) found that risk acceptance and risk-seeking were more common in rural location than in urban areas. Many status and other social variables are also associated with the socio-spatial attributes of the neighbourhoods where people live (Thomson and Tolmie (2001)). Various research works have differed in their focus and degree of detail with respect to the impact factors of collision risk. Differences refer to type of road, operational speed, traffic mix, road user mix, mode used, site etc.

Joly, Foggin et al. (1991) reported that "there has been little published on the geography of traffic accidents in two decades ago and has been changed over the last two decades. They discussed the difference between 'rate' and 'risk'. Indeed, accident rates referred to locations or areas, whereas accident risks refer to only to individuals or groups of road users related to some measure of exposure (Abdalla, Raeside et al. 1997). In reality, the location of accident shows the meaning of Collison rates, while accident risks point to people. Blatt and Furman (1998),Cummings, Koepsell et al. (1995) and Gooder and Charny (1993) investigated the differences between these two spatial approaches to road safety as place of accident (POA) and place of residence (POR). The data issues and possible deflection among POR, POA, and place of death (in case of fatal accidents) were considered on their study. We did a study overview of impact factors for accident risk which is followed by a review studies regarding to POA, studies based on POR. The POA and POR studies are discussed a briefs descriptions in below.

POA-based Approaches: The attributes of places where accidents are more likely to occur than elsewhere have been examined in many studies. Accident counts and control for areas and population size of the areas were considered in these studies. Overall, the results of these studies showed that high density and urbanity were associated with lower rates of severe injuries. The results were less conclusive when accident rates included all injuries.

A negative relation between the density and the fatalities per resident in a study of 448 counties in the USA was found by Ewing, Schieber et al. (2003). All fatalities were considered together including pedestrian fatalities.

Accident counts for San Antonio, USA were performed by Dumbaugh and Rae (2009) who focused on the neighbourhood level. Variables such as size, population density, socio-demographics, and traffic infrastructure attributes were associated with neighbourhoods. This study shows that large-scale retail outlets were associated with increased risk of accident and injury. On the other hands, the opposite was true for traditional designs with high density, walkability and small-scale neighbourhood shops. The result of this study showed a higher speed levels on arterials and less driving in traditional neighbourhoods.

A study of accidents in 80 cities in Germany was done by covering 60,000 residents(Holz-Rau and Scheiner 2013). A relation of dense cities and lower accident rates was an important result of this study. Moreover, it ascribed this to lower per capita travel volume, i.e. to risk exposure. Another finding of this study was an increase of accident rates with the extension of road networks, car ownership, and use. Other related studies reported similar results for towns and municipalities with less than 80,000 residents.

In contrast, the findings of Petch and Henson (2000) for Salford, UK, suggested higher rates of child pedestrian/cyclist casualties in inner-city areas. The findings included positive relationships between the fatality rates and over-crowding (persons per household), traffic volume and the proportion of no-car households. The latter relationship may suggest that children in areas with low levels of car ownership walk and bike more than average. Similarly, Joly et al. (1991) reported that child pedestrian and cyclist injury rates tended to be concentrated in Montreal, low-income areas

POA-based Approaches: The attributes of places where accidents are more likely to occur than elsewhere have been examined in many studies. Accident counts and control for areas and population size of the areas were considered in these studies. Overall, the results of these studies showed that high density and urbanity were associated with lower rates of severe injuries. The results were less conclusive when accident rates included all injuries.

A negative relation between the density and the fatalities per resident in a study of 448 counties in the USA was found by Ewing, Schieber et al. (2003). All fatalities were considered together including pedestrian fatalities.

Accident counts for San Antonio, USA were performed by Dumbaugh and Rae (2009) who focused on the neighbourhood level. Variables such as size, population density, socio-demographics, and traffic infrastructure attributes were associated with neighbourhoods. This study shows that large-scale retail outlets were associated with increased risk of accident and injury. On the other hands, the opposite was true for traditional designs with high density, walkability and small-scale neighbourhood shops. The result of this study showed a higher speed levels on arterials and less driving in traditional neighbourhoods.

In contrast, the findings of Petch and Henson (2000) for Salford, UK, suggested higher rates of child pedestrian/cyclist casualties in inner-city areas. The findings included positive relationships between the fatality rates and over-crowding (persons per household), traffic volume and the proportion of no-car households. The latter relationship may suggest that children in areas with low levels of car ownership walk and bike more than average. Similarly, Joly et al. (1991) reported that child pedestrian and cyclist injury rates tended to be concentrated in Montreal, low-income areas

POR-based approaches: The risk of injury for a population in POR-based approaches obtain better than studies based on POA.

Blatt and Furman (1998) studied collision severity in the USA. Rural areas and small towns are showing above average risk. This study is done for all drivers, male drivers, young drivers, and drivers involved in crashes with child fatalities. The risk figure represents twice the range of world's average in rural areas.

Scheiner and Holz-Rau (2007) reported case studies for two German regions and all age groups. The severity of injury and age were distributed but the travel mode was ignored in their study. Their study found that rural and suburban areas had the highest risk figures, with lower risk figures for city dwellers.

Research investigations of accident risks considering variables such as income, social status, or ethnic background are done by Abdalla, Raeside et al. (1997) and Edwards, Green et al. (2008) in the UK. The studies represented that risk figures are positively related to social deprivation. The studies on risks in UK focused on a small-scale level with 4,765 census areas in London distinguishing between age groups, severity of injury, and transport mode. This study reported that the most deprived areas to be approximately three times higher than those in the least deprived areas. The findings for wheelmen were similar, while the association between area deprivation and car occupants' injuries was less clear.

2.8 Overall View of Suggested Methodology

In this section, we present an overall view of the suggested methodology, as shown in Figure 2.4.

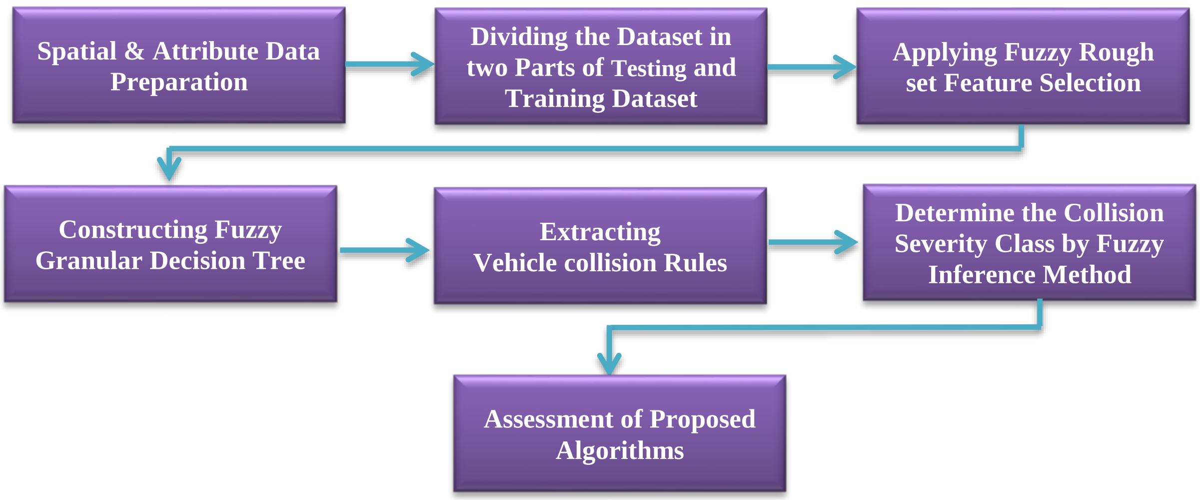


Figure 2.4 Overall view of suggested methodology

In the data preparation step, the dataset is investigated to remove those vehicle collision events with missing values and inconstant data. All spatial data, such as like vehicle collision events as a point layer, roads as polyline and districts as polygon layers, are then converted to the GCS_North_American_1983 datum coordinate system. In the next step, the vehicle collision events dataset is divided in two datasets – a training dataset and a testing dataset. Indeed, 70% of the original dataset is devoted to the training dataset, and 30% of the original dataset is assigned to the testing dataset.

A vehicle collision dataset contains a certain amount of redundancy between conditional attributes that will not discover knowledge and may, in fact, misinform the process. The fuzzy roughest feature selection is applied to find useful vehicle collision features to represent the data and remove redundancy. After determining the selected features, the fuzzy granular decision tree is constructed to extract the vehicle collision rules. The fuzzy inference model is used as the decision engine based on the extracted rules and is applied on the testing dataset to determine the severity class of vehicle collision events as a predictor model.

The accuracy, time consumption and discrepancies between the proposed method and some convention and common methods are investigated.

Chapter Three: Fuzzy Rough Set Feature Selection for Vehicle Collision Inconsistent Dataset

3.1 Introduction

Feature selection is a technique for selecting the attribute of a feature set as an important component of both supervised and unsupervised classification (Janecek, Gansterer et al. 2008). Most vehicle collision datasets contain a certain amount of redundancy that will not get knowledge discovery and may in fact misinform the process. The feature selection step finds useful vehicle collision features to represent the data and removes redundancy between attributes. Time is also saved during the decision tree process as a result of feature selection.

The main objective of feature selection in this study is threefold: preparation of the vehicle collision decision table with minimum correlation and redundancy, provision of faster efficient vehicle collision decision trees, and improved performance of the vehicle collision severity prediction via the predictor method. Vehicle collision datasets with the high dimensionality generate decision trees with a large number of nodes, redundancy and low performance. These problems restrict the usability of decision trees in the rule generation and classification for the collision data. This method prompted this research into the use of fuzzy rough sets for feature selection.

3.2 The Reasons of Using Fuzzy Rough Set Method in Vehicle Collision Rules Mining

Many existing algorithms need expert knowledge and information for feature selection, which is a significant drawback. User knowledge is necessary to state the number of features or determine a threshold to terminate the algorithm. Therefore, the decision is dependent on user's judgments (Jensen 2005).

In our study, the vehicle collision dataset consists of twelve conditional features and one decision feature. The task of feature selection in a vehicle collision dataset is the determination of the smallest subset of conditional features, so that the resulting reduced dataset remains consistent with respect to the decision feature. A dataset is consistent if, for every set of instances whose attribute values are the same, the corresponding decision attributes are identical (Stefanowski and Tsoukias 2001).

The vehicle collision dataset used in this study was, however, not identical. This means that there are some instances with the same conditional feature values, but different decision feature values. The non-identical vehicle collision dataset creates vagueness in construction of the decision tree and in rule mining. One of main problems of existing feature selection methods is related to modeling of vagueness data.

With the utilization of rough set theory in the feature selection process, the vague concept can be modeled by the approximation of a vagueness set by a pair of precise concepts, called lower and upper approximations. The lower approximation, or positive region, is the union of all instances that can be classified certainly in one of the decision values; whereas the upper approximation is a description of the instances that possibly belong to one of the decision values (Jensen 2005, Wang and Ou 2008).

The values of attributes in vehicle collision datasets are both crisp and continuous (real valued), and this is where many feature selection methods encounter a problem. It is not possible to say whether two attribute values are similar and to what extent they are the same. The fuzzy rough set feature selection (FRFS) method employs degrees of membership of values in the rough set feature selection to solve the aforementioned problem; therefore, discretization of the values is not considered at all. For example, two values may both be mapped to the same label "Near to Road Intersection", but one may be much more near than the other: values 10 m and 55 m could both be mapped to this class, although they are significantly different. This is a source of information loss, which is contrary to the rough set ideology of retaining information content. Further discussion of this issue can be found in (Beynon 2004).

The combination of fuzzy sets and the process of fuzzification of a vehicle collision dataset provide a mechanism by which real-valued features can be effectively managed. The vagueness and uncertainty of vehicle collision datasets can be modeled by allowing values to belong to more than one label with various degrees of membership and involving the lower and upper approximation concept in the process of feature selection. This information may then be exploited by fuzzy methods to enable reasoning under uncertainty (Gupta, Saini et al. 2015).

In this section, we apply FRFS on vehicle collisions attributes. FRFS chooses the most information rich attributes in a dataset without transforming the data. The FRFS method is highly efficient, relying on a simple set of operations, which makes it suitable as a pre-processor for rules mining. Moreover, since it is often the case that vehicle collision data have values of attributes that may be both discrete and real-valued (continuous), FRFS should be better equipped to handle this uncertainty and vagueness (Jensen and Shen 2002).

3.3 Fuzzy Rough Set Feature Selection Process

To apply FRFS on a vehicle collision dataset, two datasets of collision vehicle events were considered. The first dataset contained real-valued (continuous) data and described the values of the conditional feature with nominal decisions, such as the values of morning, afternoon and evening for the time of collision. The second dataset was defined by fuzzy membership values, with corresponding fuzzy decision memberships, such as 10 am with fuzzy membership value between 0 and 1.

To simplify the following discussion, twelve common features were considered as conditional features attributes in this study: weather (CF_1), collision day of week (CF_2), road surface (CF_3), collision road type (CF_4), lighting (CF_5), collision time (CF_6), direction (CF_7), road condition (CF_8), road radius (CF_9), slope (CF_{10}), distance from intersection (CF_{11}), and distance from population center (CF_{12}). The decision feature was the collision severity. Among the twelve conditional features, CF_6, CF_9, CF_{10}, CF_{11} and CF_{12} were continuous variables, while CF_1, CF_2, CF_3, CF_4, CF_5, CF_7 and CF8 were discrete variables. The first step of FRFS is the creation of fuzzy equivalence classes, which are described in the next subsection.

3.3.1 Fuzzy Equivalence Classes

The formation of the fuzzy equivalence classes is the first step of the FRFS method. The fuzzy equivalence classes are defined based on the similarity of the instances in the decision table *S* when they have the similar membership function in *S*. For example, instances *x* and *y* are considered to be similar if their membership are same as each other. Using the definition of the fuzzy similarity relation, the family of normal fuzzy sets produced by a fuzzy partitioning of the universe *U* can play the role of fuzzy equivalence classes (Dubois and Prade 1992). For example, the equivalence classes of decision feature *DF* are those instances that are partitioned by their membership functions, such as *U/DF* = {MFPDO, MFInjury, MFFatal}. The fuzzy equivalence classes are calculated for all conditional features (CF) and the decision feature (DF) by the inclusion of a fuzzy similarity relation in the vehicle collision dataset.

3.3.2 Fuzzy Rough Set Lower and Upper Approximation

The next step in the process of the FRFS of the vehicle collision dataset is built on the notion of fuzzy lower and upper approximations to enable the reduction of dataset containing real-valued features. The fuzzy lower and upper approximations are derived from possibility and necessity theory (Dubois and Prade 1992). The lower fuzzy approximation evaluates the extent to which an instance is effected by the vehicle severity decision classes. The fuzzy upper approximation evaluates the extent to which an instance is consistent with the vehicle severity decision classes. Assume that *P* is a subset of universe *U*, the fuzzy *P*-lower approximation and *P*-upper approximations are defined in Equations 1 and 2 (Jensen 2005).

(1)

$$\mu \underline{p} x(F_i) = \sup_{F \in U/P} \min(\mu F(x), \inf_{y \in U} \max\{1 - \mu F(y), \mu x(y)\}) \quad \forall i \in U$$

$$\mu \bar{p} x(F_i) = \sup_{F \in U/P} \min(\mu F(x), \sup_{y \in U} \min\{\mu F(y), \mu x(y)\}) \quad \forall i \in U \qquad (2)$$

where Fi denotes a fuzzy equivalence class belonging to *U*/*P*; i is the number of membership functions; *x* and *y* are indicators of the instances in the dataset; and *sup* and *inf* are the supremum and infimum functions, respectively, to obtain the real values of fuzzy sets. The supremum or least upper bound of vehicle collision set *S* of real numbers of real membership value is defined to be the smallest real number that is greater than or equal to every number in *S*. The infimum or greatest lower bound of vehicle collision set *S* of real membership value is defined to be the largest real number that is smaller than or equal to every number in *S (Downarowicz, Frej et al. 2015)*. In other words, $\mu\underline{p}x$ and $\mu\bar{p}x$ determine the membership values of instances that belong to the fuzzy lower and upper approximation sets. These values are calculated for all instances belong to fuzzy equivalence classes.

The next step is the calculation of the positive region of each instance. Instance x does not belong to the positive region only if it can be classified with certainty to the vehicle severity classes. The membership of an instance x∈U, belonging to the fuzzy positive region can be defined by the fuzzy union of the fuzzy lower approximation of the fuzzy equivalence classes (Equation (3)).

$$\mu_{POS(DF)}(x) = \sup_{X \in U/\bar{D}F} \mu \underline{p} X(x), \qquad (3)$$

In the above equation the fuzzy positive region is calculated by the *sup* function of the fuzzy lower approximations.

An important step in FRFS is the identification of the dependencies between conditional attributes. Intuitively, a set of attributes DF depends totally on a set of conditional attributes $P \in$ CF, if all attribute values from DF are uniquely arranged by the values of attributes from P. If there exists a functional dependency between values of DF and P, then DF depends totally on P. In fuzzy rough set theory, dependency is defined in the following way:

$$\gamma_p(DF) = \sum_{X \in U} \mu_{POS_P(DF)}(X) \tag{4}$$

The dependency value is between 0 and 1, which is calculated for all composition of vehicle collision conditional features. The higher value of dependency for a conditional feature indicates a more significant feature. If the significance is 0, then the conditional feature is dispensable. The next section describes the method of feature selection based on dependency values.

3.4 Fuzzy Rough Set Quick Reduction

In this study, we have used the dependency value to select the conditional features with a high dependency. The higher dependency values of conditional features show the feature has the capability to separate the more instances in a specific vehicle severity decision class in the dataset. Therefore, the dependency values of all 12 vehicle collision conditional features were calculated, and the highest values were then selected and added to the selected features set. The process was iterated until the dependency value no longer increased. Hence, the selected set represents the chosen features. Figure 3.1 shows an example of the feature selection process for a dataset with the three conditional features.

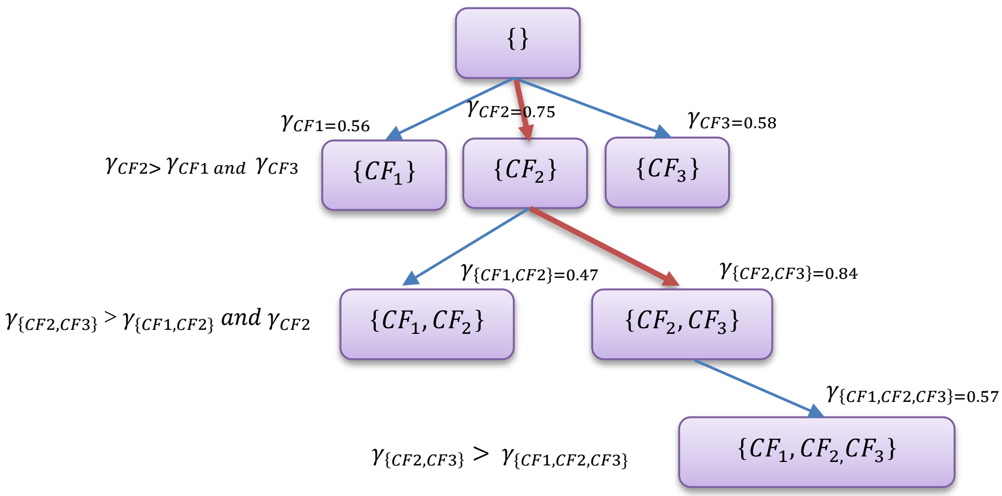


Figure 3.1 Path is taken by the fuzzy-rough quick reduction algorithm

At the first level of Figure 3.1, CF_2 had the highest value of dependency among the three conditional features; therefore, it was chosen as the first feature in the feature set. One feature from CF_1 and CF_3 was then added to the current selected feature set $\{CF_2\}$. The feature set $\{CF_2, CF_3\}$ increased the dependency value more than the feature set $\{CF_1, CF_2\}$; therefore, CF_3 was added to selected features set in the second level. After CF_1 was added to CF_2 and CF_3 in the third level, the value of dependency did not increase. Therefore, the algorithm stops, and the output of the selected feature set was $\{CF_2, CF_3\}$. The dataset can now be constructed based on the selected feature set, including all instances as rows and only those attributes appearing in the selected feature set as columns.

.

Chapter Four: Fuzzy Granular Decision Tree for Vehicle Collisions Severity Rules Extraction

4.1 Introduction

This section introduces the proposed method for the framework of fuzzy granular computing. As discussed previously, this research proposes a fuzzy granular decision tree (FGDT) for the construction of a decision tree from a vehicle collision database, in order to support discrete and continuous data by defining membership functions and fuzzification of data in the database. While decision tree methods, such as ID3, consider only discrete attributes, the fuzzy granular decision tree, which is an extension of the classic decision tree, perceives both discrete and continuous attributes. It applies the fuzzy set theory to represent the dataset and combines tree growing and granular computing to determine the structure of the tree.

To overcome the over-fitting problem in conventional decision tree methods, the FGDT chooses an attribute value in favor of all nodes at the same level when splitting a node. However, in conventional methods, the attribute is chosen solely based on the information about this node, not any other nodes at the same level. Thus, in the conventional decision tree, different nodes at the same level may use different attributes; and, the same attribute with all possible values that may be used at different levels causes the over-fitting issue.

Vehicle collision event rules mining was employed to demonstrate the potential of the fuzzy granular decision tree in solving the mentioned issues. Traffic collisions are usually caused by human, vehicle, environmental, roadway design and spatial factors

(Geurts, Wets et al. 2003). Due to the lack of sufficient human and vehicle historical damage data, this thesis utilized environmental, roadway design and spatial factors to test the proposed FGDT methodology, which is a generalization of the classical decision tree.

All the data in the training dataset were first fuzzified in the form of membership functions. The fuzzy granular entropy was then calculated for each record in the dataset. According to the calculated fuzzy granular entropy and generality and redundancy criteria, the fuzzy granular decision tree was constructed. The decision method of the final classification was done by training and testing data using a fuzzy rules based system, which is described in Chapter 5. The FGDT is discussed step by step in the following sections.

4.2 Granular Computing

In this section, the reasons for choosing granular computing as an appropriate mathematical model for rules mining and classification problems are discussed.

Granular computing (GrC) (Lin 1997) is a rapid development of granular computing, and a fast growing interest in this computation has been observed. Granular computation and granules as a subset of the universe are regarded as the primitive notion of GrC. The notion of a level consisting of a family of granules is referred to as a granulated view. Granules in different levels are joined by order relations into a hierarchy. A granule in a higher level can be decomposed into many granules in a lower level; and, conversely, many granules in a lower level can be combined into granules in a higher level. A granule in a lower level provides a detailed description of the granule in a higher level, and a granule in a higher level has a more abstract description than the granules in a lower level.

From the standpoint of GrC, a concept may be illustrated by a granule and be described or labeled by a formula. Once concepts are constructed and described, one can develop computational methods for the granule and the formula, such as the sub and super concepts and the disjoint and overlapped concepts (Yao 2004). These relationships can be conveniently expressed in the form of rules, with some associated quantitative measures indicating strength.

Knowledge discovery and data mining, especially rule mining, can be determined as a process of forming concepts and finding relationships between concepts in terms of granules and formulas by combining the results from granular computing and formal concept analysis. They are directly related to concept formation and concept relationship identification (Yao and Yao 2002). While concept formation involves the construction and description of classes, concept relationship identification involves the connections between classes. The rule mining and classification problem is then properly modeled by the GrC theory.

4.3 Information Tables

The information table is the basic concept in decision tree model and consists of the information about objects being compiled into an information table. Indeed, information tables are used in GrC models and provide a convenient way to describe a finite set of objects, called a universe, by a finite set of attributes. It represents all available information. The objects in the information table are only perceived, observed or measured by using a finite number of properties. It was defined by (Pawlak, Grzymala-Busse et al. 1995) as:

$$\mathrm{S} = (\mathrm{U}, \mathrm{At}, L, \{\mathrm{V_a} | \alpha \epsilon \mathrm{At}\}, \{\mathrm{I}_\alpha | \alpha \in \mathrm{At}\})$$

where U denotes a finite nonempty set of objects, At shows a finite nonempty set of attributes, L represents a language defined using attributes in At, Va is a nonempty set of values for $a \in At$, and $Ia{:}U \rightarrow Va$ shows an information function. In language L, an atomic formula is given by $a=v$, where $a \in At$ and $v \in Va$.

Formulas can be formed by logical negation, conjunction and disjunction. If formula φ is satisfied by object x, it can be written as $x|=S\varphi$ or in short $x|=\varphi$ (Yao 2001). If φ is a formula, set $m_S(\varphi)$ defined by $m_S(\varphi)=\{x \in U|\ x|=\varphi\}$ is called the meaning of φ in S. If S is understood, it can be simply written as $m(\varphi)$. The meaning of formula φ is the set of all objects having the property explained by formula φ. A connection between formulas of L and subsets of U is thus established.

With the introduction of language L, we have a formal description of concepts. A concept definable in an information table is a pair $(\varphi, m(\varphi))$, where $\varphi \in L$. Moreover, φ is a description of $m(\varphi)$ in S and the intension of concept $(\varphi, m(\varphi))$; and, and $m(\varphi)$ is the set of objects satisfying φ and the extension of concept $(\varphi, m(\varphi))$.

An example information table is given in Table 4.1. Based on the definition of the information table, we can find U = {O_1, O_2, $O_{3,}$ … O_{14}} and At = {A, B, C, D, class}. In addition, attribute A has the three possible values of V_A = {Clear, Rainy, Fog}. The other attributes have two possible values: V_B = {Dry, Not Dry}, V_C = {Day-Light, Dusky OR Dark}, V_D = {Day, Night} and V_{Class} = {PDO, Injury, Fatal}. (Note that PDO is the acronym for property damage only).

Table 4.1 Example of an Information Table

Granular	**A**	**B**	**C**	**D**	**Class**
O_1	Clear	Dry	Day-Light	Day	PDO
O_2	Clear	Dry	Dusky OR Dark	Day	PDO
O_3	Clear	Dry	Dusky OR Dark	Night	PDO
O_4	Clear	Not Dry	Day-Light	Night	PDO
O_5	Clear	Not Dry	Dusky OR Dark	Day	Injury
O_6	Clear	Not Dry	Dusky OR Dark	Night	Injury
O_7	Rainy	Not Dry	Day-Light	Day	Injury
O_8	Rainy	Not Dry	Dusky OR Dark	Day	Injury
O_9	Rainy	Not Dry	Dusky OR Dark	Night	Injury
O_{10}	Fog	Dry	Day-Light	Night	PDO
O_{11}	Fog	Dry	Dusky OR Dark	Day	Injury
O_{12}	Fog	Dry	Dusky OR Dark	Night	Fatal
O_{13}	Fog	Not Dry	Day-Light	Day	Fatal
O_{14}	Fog	Not Dry	Dusky OR Dark	Night	Fatal

For decision tree construction and rule mining tasks, it is assumed that information about objects is given by an information table, and each object is associated with a unique class label. Objects can be divided into classes that form a granulation of the universe. Without the loss of generality, it is assumed that there is a unique attribute class taking class labels as its value. The set of attributes is expressed as S= At ∪ {Class} and named information system (IS), where *At* is the set of attributes used to describe the objects, also called the set of descriptive or conditional attributes.

For developing a granule tree, the universe, which is a finite set of attributes in information table, should be split into grouping or partitions of the same class with the atomic formula of the class label. A set of granules associated with the atomic formula of the attribute value is then constructed. The selection of the most appropriate formula and its connected granule for each level of granular tree needs some quantitative measures to estimate the quality of a generated rule, as described in the following subsections.

Generality* and *Fuzzy Generality: Generality indicates the relative size of the granule. The generality is defined by the probability of each granule, which is defined by a formula. If a generality covers more instances of the universe, it is more general than the other granules (Yao and Yao 2002). Equation (4.2) shows the generality:

$$G(\varphi(a = v)) = \frac{|\mathrm{m}(\varphi)|}{|S|} \tag{1}$$

With the fuzzy granular decision tree (FGDT) proposed in this research, the fuzzy generality is introduced in Equation 2:

$$FG(\varphi(a = v)) = \sum_{i=1}^{c} \frac{\sum_{j}^{N} \mu_{ij}}{S} \tag{2}$$

where μ_{ij} is the fuzzy membership value of the j^{th} granule to the i^{th} class. Equation 2 is defined based on the concept of applying the value of fuzzy membership function of each attribute of collision events rather than using the number of granules. The summation of membership values of the granular set in the specific class is designed as a numerator and the summation of membership values of all granules in a specific formula in *S* is calculated as the denominator of this equation. The calculated generality presents the generality of the granular S of formula $a = v$ (Kiavarz and Wang 2014).

***Confidence*:** Given two formulas φ and ψ Zhao and Yao (2007) introduced symbol ⇒ to connect φ and ψ in the form of φ ⇒ ψ .We can illustrate φ (A = Clear) ⇒ ψ (Class = PDO) as an sample of the connection of the two formulas. The ratio of the number of granules in a granular set that are correctly classified by the generated rules to the number of granules in the granular set that are created by a formula that is termed as the confidence or absolute support. Thus, it is a measure of the correctness or precision of the inference. As Equation 3 specifies, if the value of confidence of a rule is kept high, fewer association rules will be mined, but their prediction accuracy will be quite high (Zhao and Yao 2007). To calculate the confidence of a rule that is constructed by fuzzy granular decision tree, we introduce the Fuzzy Confidence or Fuzzy Absolute Support by Equation 3.

$$FAS(\varphi \Rightarrow \psi) = \frac{\sum_{i=1}^{N} \mu_i(\varphi \wedge \psi)}{\sum_{i=1}^{N} \mu_i(\varphi)} \tag{3}$$

where μ_i is the membership value of the i^{th} granule to the decision class that is determined by formula ψ, and N is the number of granules in the granular set that is assigned by connection $\varphi \Rightarrow \psi$.

Coverage: Coverage is a measure that represents the suitability of the classification. It demonstrates the fraction of data in a class correctly classified by the rule (Kiavarz and Wang 2014). The value of coverage is estimated by the fraction of the number of granules in a granular set that are correctly classified by the number of granule in training data with the same class label (Zhao and Yao 2007). To calculate the coverage of a rule that is constructed with the FGDT, the fuzzy coverage (FCV) in introduced in Equation 4.

$$FCV(\varphi \Rightarrow \psi) = \frac{\sum_{i=1}^{N} \mu_i(\varphi \wedge \psi)}{\sum_{i=1}^{N} \mu_i(\psi)} \tag{4}$$

where μ_i is the membership value of the i^th^ granule to the decision class that is determined by formula ψ, and N is the number of granules in the granular set that is assigned by connection $\varphi \Rightarrow \psi$.

Conditional Entropy:

Conditional entropy is the most commonly used measure for selecting attribute values in the construction of the decision tree for classification. Many decision tree algorithms, such as ID3 and common granular decision tree, require data with discrete values. Discretization of a continuous variable is not easy, particularly the determination of the boundary of each interval. An example is the distance from a collision to an intersection. The fuzzy concept in the FGDT method is applied, with the dataset with the fuzzy expression forming the FGDT.

Based on this concept, fuzzy granular entropy is proposed to employ the continuous and discrete values in decision tree construction. Fuzzy granular conditional entropy is introduced based on defined membership values of each object in each granular set, due the data fuzzy expression (Kiavarz and Wang 2014). Equation (4.6) specifies the fuzzy granular conditional entropy (FGCE) with the given granular universe S:

$$Fuzzy\ Granular\ Conditional\ Entropy(a = v) = -\sum_{i=1}^{c} \frac{\sum_j^N \mu_{ij}}{S} log_2 \frac{\sum_j^N \mu_i}{S} \quad (5)$$

where μ_{ij} is the membership value of the j^th^ granule to the i^th^ class. This equation is defined based on the concept of applying the value of membership function of each factor of collision events rather than using the crisp values. The summation of membership values of granular in the specific class is designed as a numerator of Equation 1 and the summation of membership values of all granular in a specific formula is calculated as the denominator of this equation. The calculated entropy presents the entropy of the granular S of formula $a = v$ related to training data. In the fuzzy granular conditional entropy the membership function (which belongs to formula $a = v$ for each granule) are involved to calculate entropy. For example, the granular set of formula A(weather) = Clear is $\{O_1,O_2,O_3,O_4,O_5,O_6\}$ based on Table 3.1. Also the granular sets of decision classes of PDO, Injury and Fatal are $\{O_1,O_2,O_3,O_4,O_{10}\}$, $\{O_5,O_6,O_7,O_8,O_9,O_{11}\}$ and $\{O_{12},O_{13},O_{14}\}$, respectively. According to the defined membership function of A(weather), the membership values for objects are demonstrated in Table 4.2.

Table 4.2 Membership values of intances with
the connection A = weather ⇒ Class=Clear

Granules MF	MF Values
$\mu_{A(O1)}$	1
$\mu_{A(O2)}$	1
$\mu_{A(O3)}$	0.9
$\mu_{A(O4)}$	1
$\mu_{A(O5)}$	0.7
$\mu_{A(O6)}$	0.8

The value of FGCE is calculated based on Equation (1) which is equal to FGCE (weather = Clear) = 0.72*0.30 + 0.28 * 1.84 + 0 = 0.73.

4.4 Pre-processing of Inconsistent Data and Missing Attribute Values

Data preprocessing includes the handling of inconsistent data and missing attribute values in the dataset. The preprocessing is applied on the dataset before applying the feature selection algorithm. There are missing attribute values for some events as instances of the dataset, i.e. the values of one or more attributes are empty for some instances in the dataset. In of data preprocessing, the missing values are marked and then simply not used in fuzzy granular entropy calculations.

Inconsistent vehicle collision event data also emerge in the decision table during data collection. The instances with identical condition values and different decision values belonging to different severity classes create confusion in the construction process of the decision tree algorithm. The given vehicle collision database is considered as decision table $S = (U, CF, DF)$. When inconsistent data appear, a new decision table – $S' = (U', CF, DF)$ – can be defined according to those instances that have the maximum frequency decision value.

Table 4.3 presents a sample of vehicle collision decision table with inconsistent data. The three instances (O_1, O_3 and O_4) are inconsistent data, the individual values of the decision property Severity are PDO, PDO and Injury. The proportion of the decision value PDO is 66% of three instances, and the decision value of Injury is 33%. Therefore, the instances with the identical condition values and the decision value of PDO are selected, and the other instances are deleted. A new decision table is obtained as shown in Table 4.4.

Table 4.3 The Road Collision Decision Table with inconsistency Data

U	Weather	Surface	Lighting	Time	Severity
O_1	Clear	Dry	Day-Light	Morning	PDO
O_2	Clear	Not Dry	Dusky OR Dark	Morning	PDO
O_3	Clear	Dry	Day-Light	Morning	PDO
O_4	Clear	Dry	Day-Light	Morning	Injury
O_5	Raining	Not Dry	Day-Light	Night	Fatal

Table 4.4 The Road Collision Decision Table without Inconsistency Data

U	Weather	Surface	Lighting	Time	Severity
O_1	Clear	Dry	Day-Light	Morning	PDO
O_2	Clear	Not Dry	Dusky OR Dark	Morning	PDO
O_5	Raining	Not Dry	Day-Light	Night	Fatal

4.5 Constructing Road Vehicle Collision Fuzzy Granular Decision Tree

To create a road collision FGDT with minimum uncertainty, the subset of formula (attribute value) with the highest values of coverage, confidence, generality and minimum granular fuzzy entropy should be selected as a node of the tree. The road collision rules are then generated automatically based on the training data. Construction of the FGDT involves applying concepts, such as generality, which represents the presence of a granular set rather than the other granular set in the universe, fuzzy granular entropy, which measures the homogeneity of each granular set and decreases the redundancy by selecting those granular set that have the minimum redundancy rather than the other granular set to cover the universe. The redundancy in a FGDT means that an object in a granular set is repeatedly placed along a given branch of the tree. To reduce such redundancy, the proposed FGDT automatically selects more appropriate nodes based on measurement of the redundancy by counting the repetitive objects of a granular set and the universe in each step to select the node with minimum redundancy and maximum coverage of the universe granular set.

This process can recognize which granule is most appropriate at the end of each level to be broken down first, until it reaches to the granular set whose objects will be a subset of final classes. This granular set is called non-active granular set. The active granular set is a set of objects that belong to different classes. Thus, an active granule is further divided through efficient measures, and the construction of the tree is continued until all granules reach zero or near to zero fuzzy entropy, in which the union of all non-active granules is equal to the universe set. In other words, after the union of all non-active granules that cover the universe at the all levels, the FGDT construction is stopped.

The designed program extracts the rules from constructed tree automatically. In the constructed FGDT, each non-active granule is labeled by its decision class value. The union of all non-active granules in the two levels form a non-redundant covering solution of the consistent classification problem and the union of all inactive subsets in the two levels forms the universe. The steps of constructing FGDT shown below:

- **Set** *U* as the root node of a fuzzy granule tree at the initial stage.
- **Set** the status of *U* as active.
    - **While** the active granule(s) is available N
        - **Extract** all formula in the *U*
    - **While** the unprocessed formula is available M
        (1) Calculate fuzzy generality, fuzzy confidence, fuzzy coverage and fuzzy granular entropy condition for all granules.
        (2) Select the proper granule as a node with respect to below conditions:
            - Granule with the minimum entropy,
            - Granule with the minimum redundancy,
            - Granule with the maximum fuzzy generality, fuzzy confidence, fuzzy coverage.

(3) Determine the selected granule as an active or inactive granule.

(4) Update the activity status of N.

(5) Update *U* set based on the selected formula.

(6) Modify the fuzzy granule tree by adding the granule N∩m(a=v) as a new node, connecting it to N by an arc, and labelling it b ya=v.

End of Loop M

End of Loop N.

4.6 Reasoning with the Fuzzy Granular Decision Tree

Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made or patterns discerned. The rules that are extracted from the FGDT have a description structure based on the if-then phrase, which is called a linguistic rule. This research employs the fuzzy rules based system (FRBS) to make a final decision to specify the road vehicle collision events in the three classes of PDO, Injury and Fatal.

The process of FRBS is started from a given input to the output using a set of fuzzy if-then linguistic rules, which are generated from the FGDC, the antecedents and consequents of which are compounded fuzzy statements related by the concepts of fuzzy implication and the compositional rule of inference (Roisenberg, Schoeninger et al. 2009). An FRBS is a knowledge-based method, which implicates the information in the form of if-then fuzzy rules, i.e. if a set of conditions are satisfied, then set of consequents can be driven. The consequent is an output collision severity class. For example, it can be denoted as (Zadeh 1997):

$R_j: IF\ a_{j1}\ IS\ v_{j1}\ AND \ldots AND\ a_{jm}\ IS\ v_{jn}\ THEN\ b\ IS\ C_j$

where j = 1-L, which is the number of extracted rules, a_{j1} to a_{jm} and b are the input and output variables, and v_{j1} to v_{jn} and C_j are the involved antecedents and consequent labels, respectively.

The fuzzy reasoning is applied to determine the final classes of collision events in five steps: fuzzification of the input variables, application of the fuzzy operator (AND or OR) in the antecedent, implication from the antecedent to the consequent, aggregation of the consequents across the rules, and defuzzification . The following subsections present the descriptions of FRBS's steps (Straccia 2011).

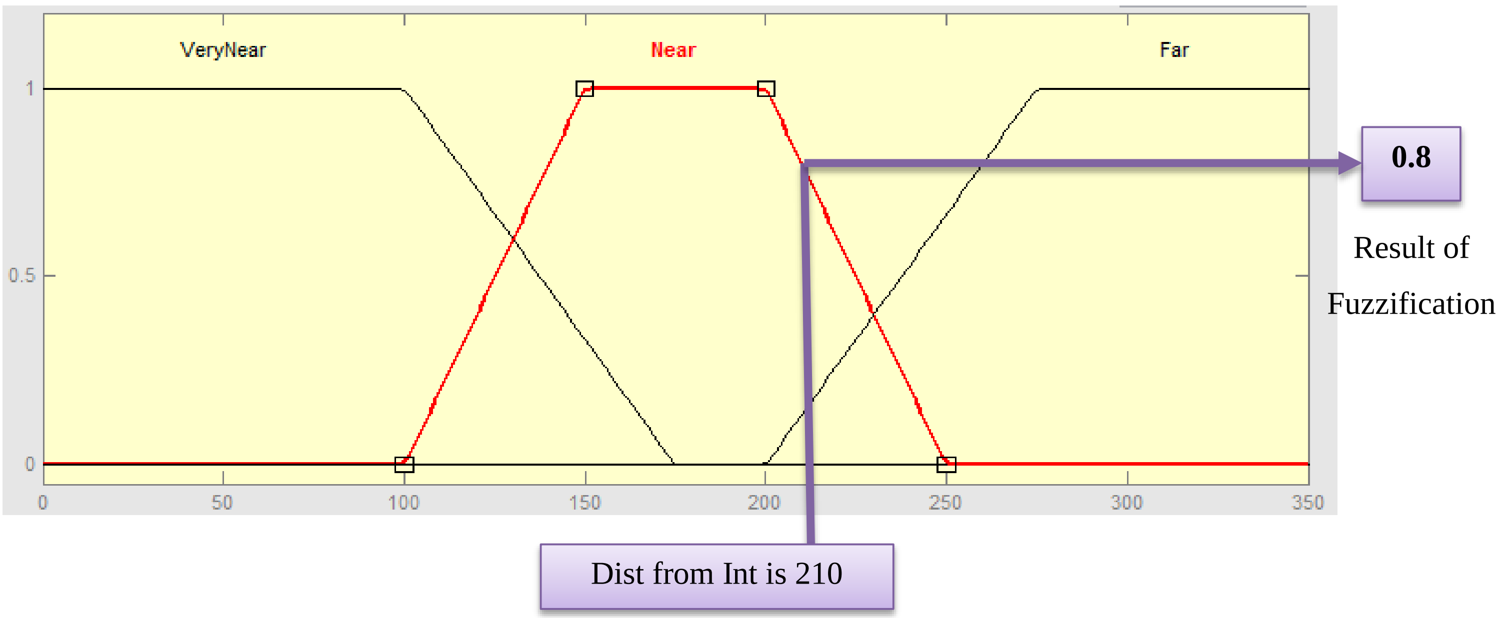
4.6.1 Fuzzification of Input Variables

As previously mentioned, there are two different types of data in a vehicle collision database – discrete and continuous. Many decision tree algorithms, such as ID3 and regular granular tree, require data with discrete values. This results in the loss of some information with continuous values in the database; therefore, this research strives to use the fuzzy concept to construct the decision tree, substituting the training data with the fuzzy expression and forming the FGDT.

The type of fuzzy membership function for each attribute is very significant in the creation of the FGDT. As such, various functions are tested, and an appropriate function for each factor is determined. Triangular and trapezoidal functions (with maximum equal to 1 and minimum equal to 0) are widely applied membership functions. This research uses triangular and trapezoidal membership functions, due to their simplicity, their learning capability, and the short amount of time required for designing the system. Based on the collision data in the fuzzy membership function, the fuzzification of collision data is applied to the database by the defined membership functions (Tahriri, Mousavi et al. 2014).

The first step is to take the inputs, including conditional attributes and objects, from the information table and determine the degree to which they belong to each of the appropriate fuzzy sets via membership functions. The input is a numerical value limited to the universe of discourse of the input variable (it has difference intervals for each attribute), and the output is a fuzzy degree of membership in the qualifying linguistic set (always the interval between 0 and 1).

Figure 4.1 represents as example of the distance of a collision event from an intersection, which was built on the rules. Each of the rules depends on resolving the inputs into a number of different fuzzy linguistic sets: severity is PDO, Injury or Fatal, distance is very near, near, far, and so on. Before the rules can be evaluated, the inputs must be fuzzified according to each of these linguistic sets. For example, to what extent is the distance of the collision from an intersection really near? Figure 4.1 shows how far the collision is (rated on a scale of 0 to 350, via its membership function) as the linguistic variable distance from the intersection. In this case, the distance was rated as 210 metres, which, given the graphical definition of near, corresponds to $\mu = 0.8$ for the distance from the intersection membership function (Zadeh 1965, Tahriri, Mousavi et al. 2014).



Input
Figure 4.1 Fuzzing the input value

At the fuzzification step, the crisp collision event data input is converted into fuzzy data by using defined membership functions. A value between 0 and 1 is assigned for each feature in the generated rules.

4.6.2 Application of Fuzzy Operator

After the inputs are fuzzified, the fuzzy degree of each attribute is determined as to which each part of the antecedent is satisfied for each rule. This subsection discusses about the multi-antecedent part. If the antecedent of a given rule has more than one part, the fuzzy operator is applied to obtain one number that represents the result of the antecedent for that rule. This number is then applied to the output function. The input to the fuzzy operator is two or more membership values from fuzzified input variables. The output is a single truth value (Mousakhani, Me'marzadeh Tehran et al. 2013).

Either the AND or OR operation can be used. They can be represented by two built-in AND methods: *min* (minimum) and *prod* (product). Two built-in OR methods are also supported: *max* (maximum) and *probor* (probabilistic OR method).

As the application of this research is vehicle collision severity classification, it is sensitive to each part of the rules. To consider each of the parts in the antecedent, the AND operator has been selected. Figure 4.2 presents an example of a multi-antecedent part and the use of the AND operator.

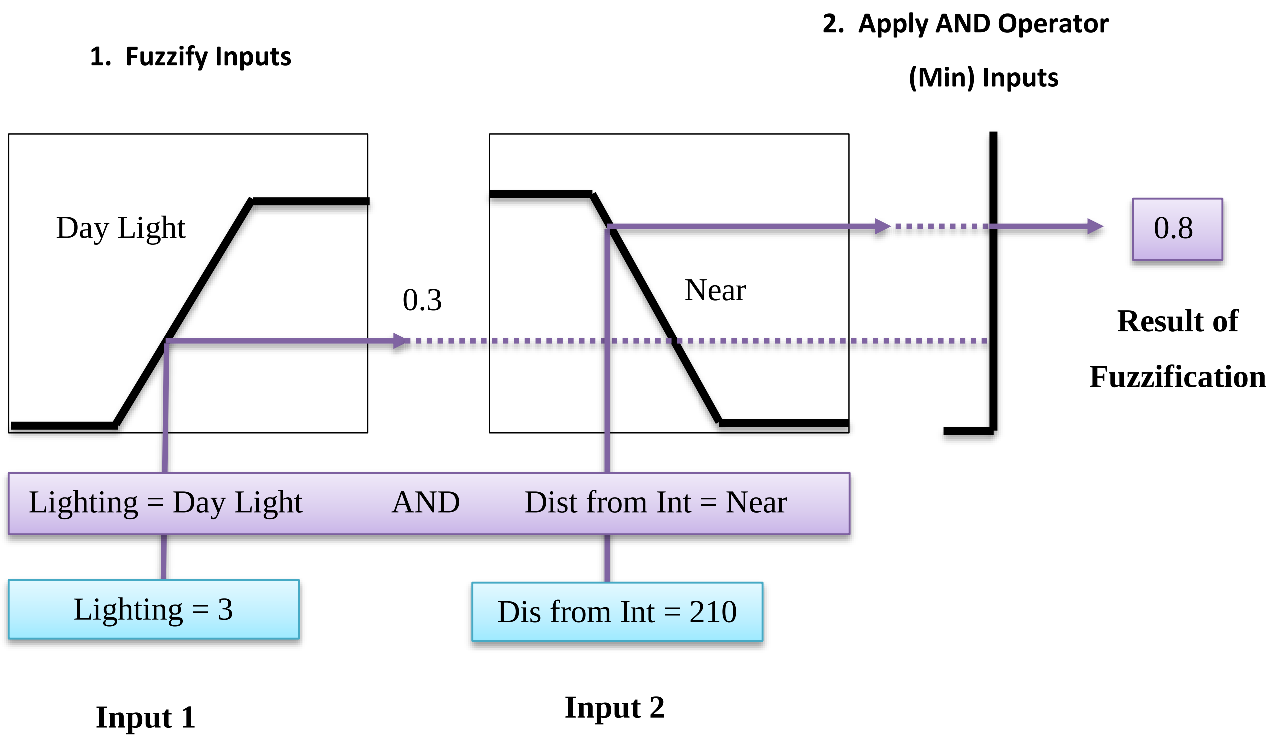


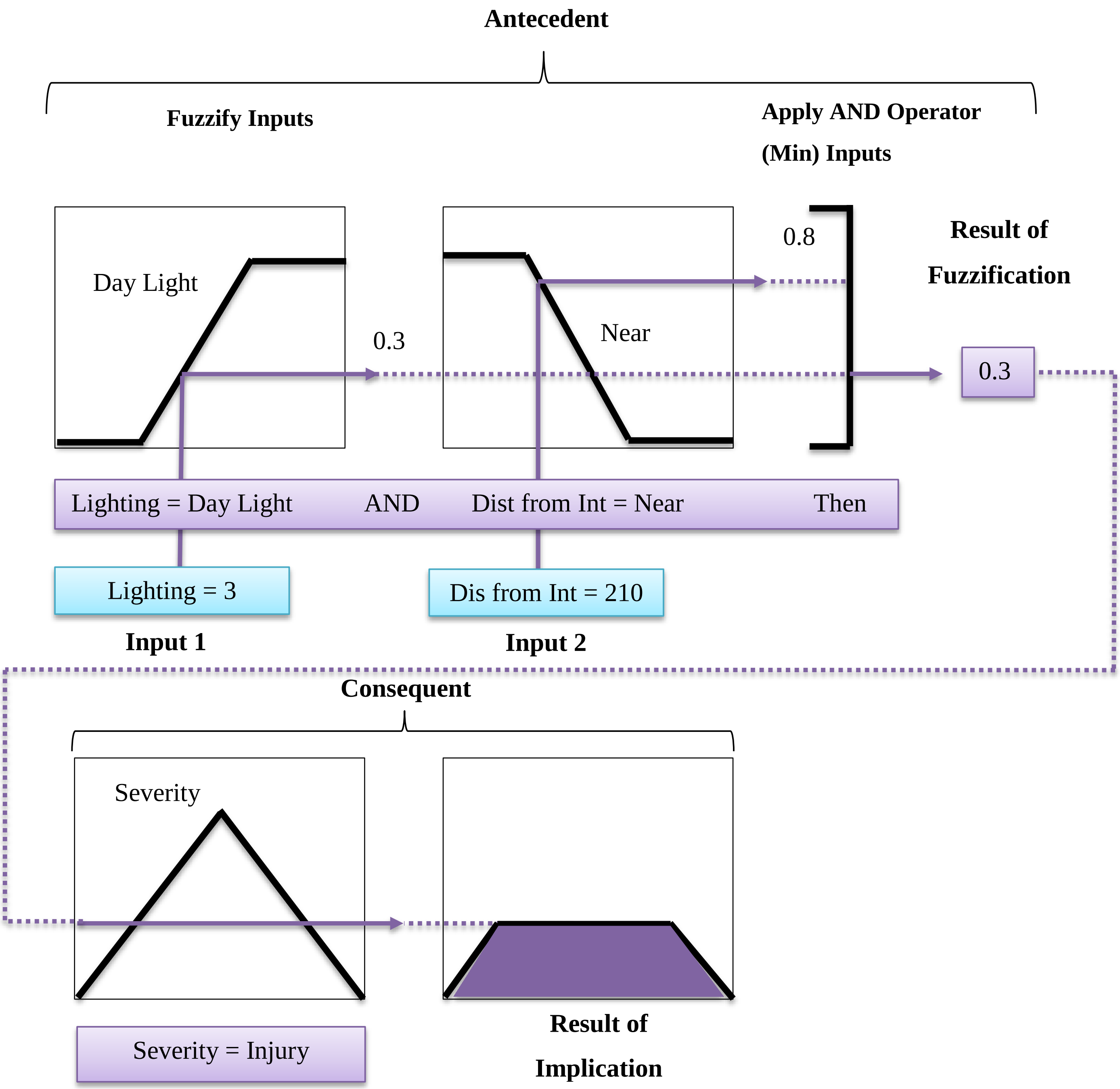
Figure 4.2 Application of Fuzzy Operator

4.6.3 Implication Step

Before applying the implication method, the rule's weight must be determined. Every rule has a weight (a number between 0 and 1), which is applied to the number given by the antecedent. Generally, this weight is 1 and, thus, has no effect at all on the implication process. In this research, the weight of each rule is considered as 1 (Zadeh 1965, Kulkarni 2001). After proper weighting is assigned to each rule, the implication method is applied.

A consequent is a fuzzy set represented by a membership function, with an appropriate weight attributed to the linguistic characteristics. The consequent is reshaped using a function associated with the antecedent (a single number). A single number is assigned as the input value for the implication process given by the antecedent, and the output is a fuzzy set.

Implication is implemented for each rule. Two built-in methods are supported, and they are the same functions used by the AND method: *min* (minimum), which truncates the output fuzzy set, and *prod* (product), which scales the output fuzzy set (Zhu, Wang et al. 2014). Figure 4.3 shows an example of the implication method applied to one rule of vehicle collision.



4.6.4 Aggregation of Outputs

Since the decisions in a fuzzy reasoning are based on the testing of all rules, the rules must be combined in some manner in order to make a decision. The fuzzy sets that represent the outputs of each rule are combined into a single fuzzy set by the aggregation process.

Aggregation only occurs once for each output variable, just prior to the fifth and final step, defuzzification. The list of truncated output functions returned by the implication process for each rule is the input of the aggregation process. The output of the aggregation process is one fuzzy set for each output variable (Liu, Jiao et al. 2013).

In the aggregation step, the order in which the rules are executed is unimportant so long as the aggregation method is applied to the outputs of all rules. Three built-in methods are available: *max* (maximum), *probor* (probabilistic OR) and *sum* (simply the sum of each rule's output set). As all outputs of rules can be important, the *max* function was selected as the aggregation function in this study (Liu, Jiao et al. 2013).

Figure 4.4 is an example of the use of the aggregation method on vehicle collision input data. All three rules have been placed together to show how the output of each rule is combined, or aggregated, into a single fuzzy set.

1. Fuzzy

3. Apply Fuzzy Operation AND (min)

2. Apply Implication Method (max)

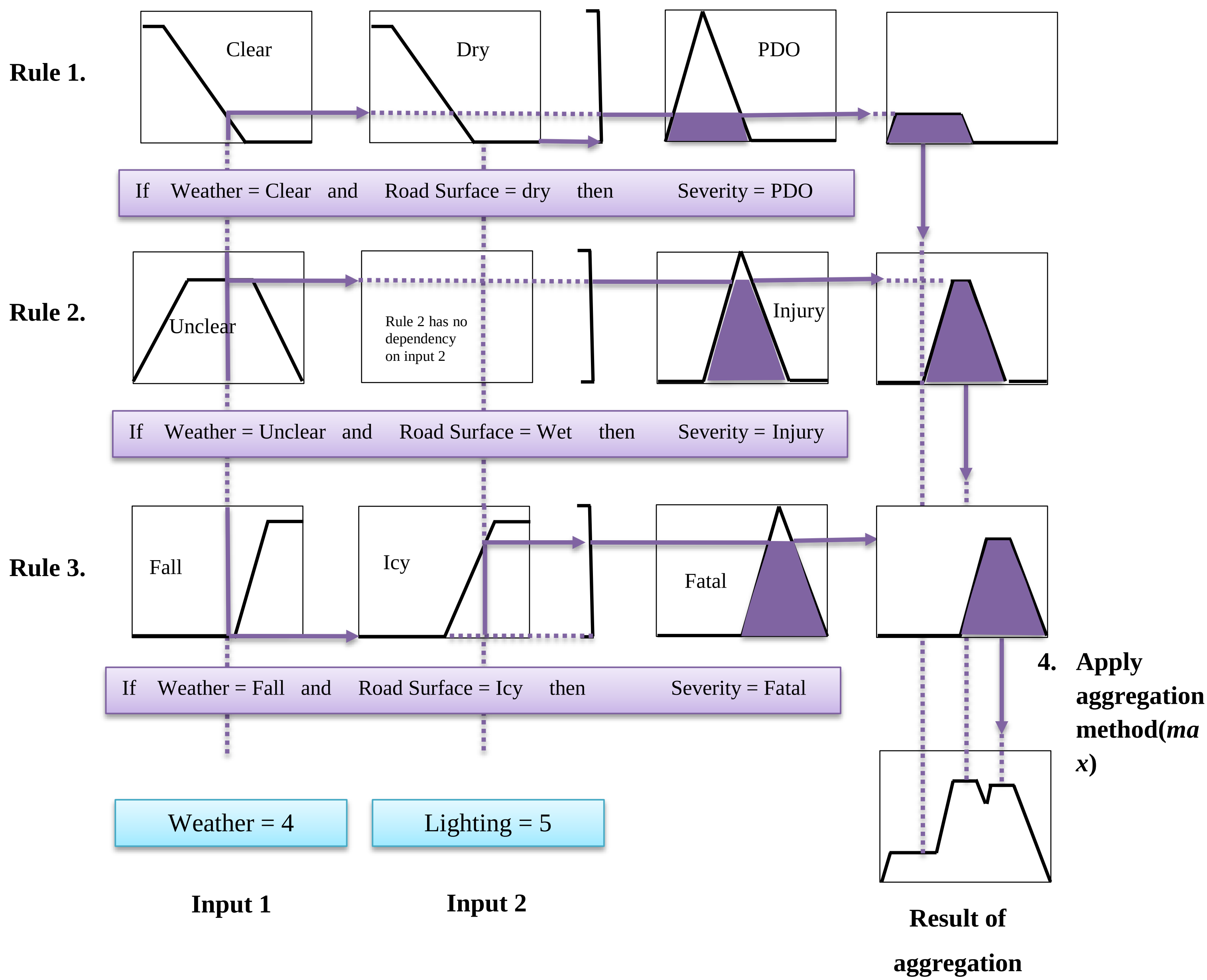


Figure 4.3 Aggregation step of Fuzzy Inference

4.6.5 Defuzzification

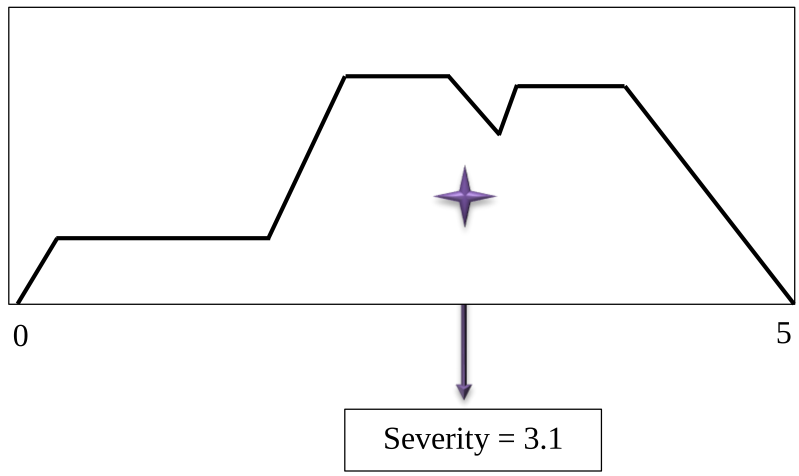
The last step is called defuzzification and is used to convert fuzzy values to the final crisp classes of PDO, Injury and Fatal value. In collision severity, a single crisp output is desired from an FRBS. The input for the defuzzification process in vehicle collision is a fuzzy set (as shown in the aggregate output fuzzy set), and the output should be a single number to determine the exact final severity class. As much as fuzziness helps the rule evaluation during the intermediate steps, the final desired output for each variable is generally a single number. However, the aggregate of a fuzzy set encompasses a range of output values, which must be defuzzified in order to obtain a single output value from the set (Pradhan 2013).

Perhaps the most popular defuzzification method is the centre of gravity, which returns the centre of the area under the curve. Five built-in methods are supported: centroid, bisector, middle of maximum (the average of the maximum value of the output set), largest of maximum, and smallest of maximum (Pradhan 2013).

In this research, the centroid method of defuzzification, which calculates the centre of gravity of the individual fuzzy sets aggregated with the maximum connective, was considered. The centroid defuzzification method finds a point representing the centre of gravity of the aggregated fuzzy set A on the interval [a,b], the calculation of which is (Erdik 2014):

$$z_{COG} = \frac{\int_a^b z.\mu_A(z)dz}{\int_a^b \mu_A(z)dz} \tag{6}$$

where z_{COG} is the crisp output, $\mu_A(Z)$ is the aggregated membership function and z is the output variable. In the case study section, there is an example of FRBS based on real data of collision events. Figure 4.5 is an example of deffuzzification method on vehicle collision input data.



Result of Aggregation

Figure 4.4 Defuzification Step of Fuzzy Inference

After calculating the single number with defuzzification, the decision attribute membership function can be performed to determine the final linguistic classes of severity (Kiavarz and Wang 2014). Figure 4.6 represents how the final classes of an object in the database can be determined. Indeed, it is an example of a membership function of severity in the class of PDO, Injury and Fatal. As this figure shows the final class was Injury.

After calculating the single number of defuzzification, the decision attributes membership function can be performed to determine the final linguistic classes of severity (Kiavarz and Wang 2014). The figure 4.6 represents how the final classes of an object in the database can be determined. Indeed, it is a sample of membership function of severity in the class of PDO, Injury and Fatal. As this figure shows the final class is Injury.

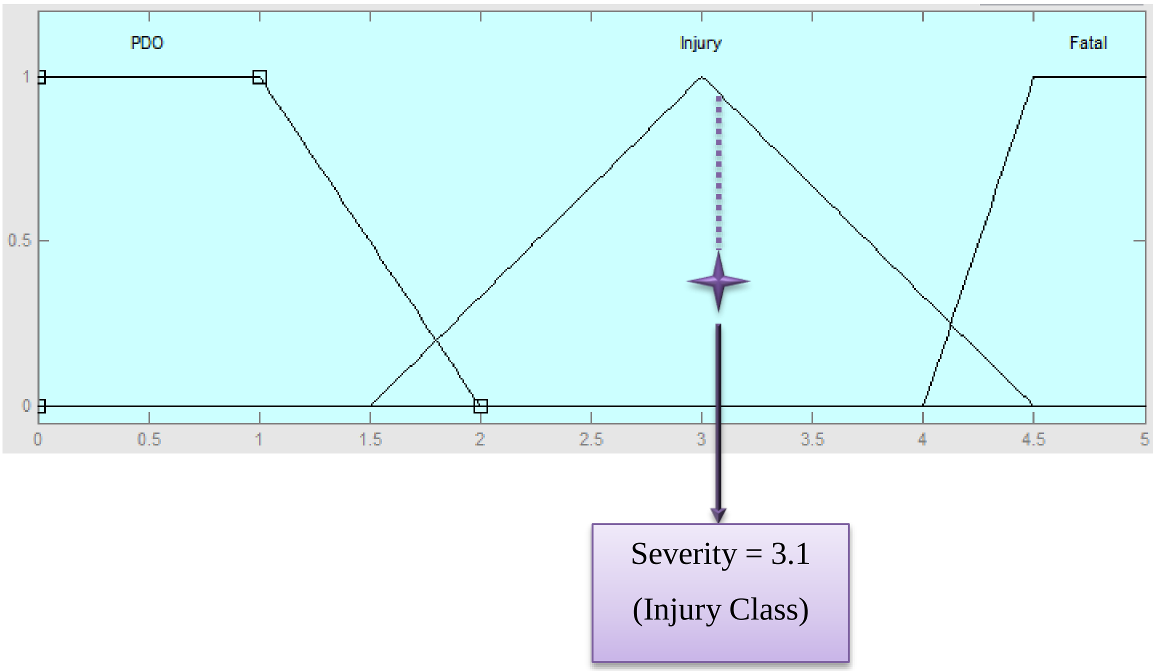


Figure 4.5 Determination of the Final Discrete Class

Chapter Five: Case Study and Results

5.1 Case Study Data Description

The data set for the case study contains traffic collision records of the twelve major highways in California. It holds the collision details from January 2011 to April 2013. The raw data were obtained from the California Collision Database & Synopses web site in a spreadsheet file. These data were stored in an excel file format with 16 attributes along with the coordinate system of collisions to describe each record. After the preprocessing step, the dataset had 45621 collisions instances in 12 counties of California (as shown in Figure. 5.1). All the instances in the data set were divided into 12 subsets based on the highways. The feature selection rule mining by decision tree and classification algorithms by fuzzy inference were applied to each subset. The 12 significant existing attributes were classified them into three groups: environmental, road geometry, and spatial attributes. The environment specific attributes were comprised of weather (CF1), collision day of week (CF2), road surface (CF3), collision road Type (CF4), lighting (CF5), collision time (CF6) and road surface condition (CF7). The road geometry related attributes are direction (CF8), road radius (CF9) and slope (CF10). The spatial attributes include distance from intersection (CF11) and distance from population center (CF12). The collision data set is separated into a training data set with 70% of the original dataset and a testing data set with 30% of original dataset.

All spatial data, including collision event points, of the California Roads and County's layer coordinate systems were converted to the GCS_North_American_1983 datum coordinate system. In the data preparation phase, the collision dataset was converted into two spatial datasets, using the latitude and longitude values in the two feature classes (CF_{11} and CF_{12}) separated into the training and testing datasets. The training and testing datasets were classified in three severity classes: fatal, injury and property damage only (PDO).

Traffic crashes occur due to the interactions of vehicle, driver, roadway and environmental factors. All these factors interact with each other and simultaneously influence the occurrence and severity of collisions.

Figure 5.1 The US Highway Road Vehicle Collisions in California State in the dataset

Although driver error is often a significant contributor to the occurrence of any particular collision event, the analysis of roadway and environmental factors help explain why collisions are more frequent in some locations than in the others. This research implemented spatial, road geometry and environmental factors for road collision severity rule generation and ignored driver error, due to the lack information in the dataset. The study helps to identify the potential of collisions in the other locations of the roads regarding the influential factors.

The fuzzy rough set feature selection (FRFS) method was applied to vehicle collision features. FRFS selected and reduced the attributes from the environmental, geometry and spatial collision features to improve the performance and accuracy of the decision tree and help handle a large number of features.

5.2 Applying the FRFS Method on the Dataset

To achieve minimum correlated features and redundancy, the FRFS was performed on the dataset. The selection process was conducted by calculating the dependency of features in each step. Features with maximum dependency were added to the selected feature set.

Table 5.1 presents the calculated dependency values for each step. Each column of the table shows the step of selecting the attributes with the highest dependency value. From the first column, it can be seen that conditional feature CF_1, with a dependency value of 0.557, had the greatest dependency degree in step 1. Therefore, CF_1 was chosen and added to the selected feature set. The process was then iterated, and the two-dependency degrees were calculated. The addition of CF_5 to the selected feature set caused the largest increase in dependency; therefore, the new candidate became $\{CF_1, CF_5\}$.

Table 5.1 The Fuzzy Rough Set Feature Dependency

$CF(\gamma'_1)$	$CF(\gamma'_2)$	$CF(\gamma'_3)$	$CF(\gamma'_4)$	$CF(\gamma'_5)$	$CF(\gamma'_6)$	$CF(\gamma'_7)$	$CF(\gamma'_8)$	$CF(\gamma'_9)$
CF_1 (0.557)	$\{CF_1,CF_2\}$ (0.516)	$\{CF_1,CF_5,CF_2\}$ (0.521)	$\{CF_1,CF_5,CF_9,CF_2\}$ (0.566)	$\{CF_1,CF_5,CF_9,CF_3,CF_2\}$ (0.572)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_2\}$ (0.576)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_2\}$ (0.576)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_{12},CF_2\}$ (0.579)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_{12},CF_{11},CF_2\}$ (0.571)
CF_2 (0.249)	$\{CF_1,CF_3\}$ (0.284)	$\{CF_1,CF_5,CF_3\}$ (0.286)	$\{CF_1,CF_5,CF_9,CF_3\}$ (0.606)	$\{CF_1,CF_5,CF_9,CF_3,CF_4\}$ (0.376)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_4\}$ (0.371)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_4\}$ (0.361)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_{12},CF_4\}$ (0.360)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_{12},CF_{11},CF_4\}$ (0.578)
CF_3 (0.357)	$\{CF_1,CF_4\}$ (0.286)	$\{CF_1,CF_5,CF_4\}$ (0.318)	$\{CF_1,CF_5,CF_9,CF_4\}$ (0.347)	$\{CF_1,CF_5,CF_9,CF_3,CF_6\}$ (0.452)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6\}$ (0.687)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_7\}$ (0.283)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_{12},CF_7\}$ (0.281)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_{12},CF_{11},CF_7\}$ (0.280)
CF_4 (0.279)	$\{CF_1,CF_5\}$ (0.573)	$\{CF_1,CF_5,CF_6\}$ (0.406)	$\{CF_1,CF_5,CF_9,CF_6\}$ (0.428)	$\{CF_1,CF_5,CF_9,CF_3,CF_7\}$ (0.287)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_7\}$ (0.285)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_8\}$ (0.406)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_{12},CF_8\}$ (0.399)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_{12},CF_{11},CF_8\}$ (0.376)
CF_5 (0.514)	$\{CF_1,CF_6\}$ (0.403)	$\{CF_1,CF_5,CF_7\}$ (0.231)	$\{CF_1,CF_5,CF_9,CF_7\}$ (0.258)	$\{CF_1,CF_5,CF_9,CF_3,CF_8\}$ (0.416)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_8\}$ (0.417)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_{11}\}$ (0.577)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_{12},CF_{11}\}$ (0.734)	
CF_6 (0.396)	$\{CF_1,CF_7\}$ (0.230)	$\{CF_1,CF_5,CF_8\}$ (0.406)	$\{CF_1,CF_5,CF_9,CF_8\}$ (0.419)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10}\}$ (0.654)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_{11}\}$ (0.577)	$\{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_6,CF_{12}\}$ (0.701)		

CF_7 (0.225)	{CF_1,CF_8} (0.4)	{CF_1,CF_5,CF_9} (0.598)	{CF_1,CF_5,CF_9,CF_{10}} (0.561)	{CF_1,CF_5,CF_9,CF_3,CF_{11}} (0.576)	{CF_1,CF_5,CF_9,CF_3,CF_{10},CF_{12}} (0.568)			
CF_8 (0.315)	{CF_1,CF_9} (0.510)	{CF_1,CF_5,CF_{10}} (0.521)	{CF_1,CF_5,CF_9,CF_{11}} (0.572)	{CF_1,CF_5,CF_9,CF_3,CF_{12}} (0.543)				
CF_9 (0.443)	{CF_1,CF_{10}} (0.510)	{CF_1,CF_5,CF_{11}} (0.558)	{CF_1,CF_5,CF_9,CF_{12}} (0.542)					
CF_{10} (0.460)	{CF_1,CF_{11}} (0.549)	{CF_1,CF_5,CF_{12}} (0.502)						
CF_{11} (0.5)	{CF_1,CF_{12}} (0.572)							
CF_{12} 0.413)								

The process stopped in iteration 9 (column $CF(\gamma'_9)$), because there was no increase in the degree of dependency in iteration 9. Ultimately, eight features {weather (CF_1), road surface (CF_3), lighting (CF_5), collision time (CF_6), road radius (CF_9), slope (CF_{10}), distance from intersection (CF_{11}) and distance from population centre (CF_{12})} were selected as the features to create the decision table.

5.3 Review of Selected Feature Definitions and Properties of Selected Highways in California

The eight selected features were determined based on the FRFS method. Table 5.2 presents a brief description of each selected conditional feature.

Table 5.2 The infused factors of vehicle collision severity

Type of Factors	Factors	Description
Road geometry	*Radius*	The lower road radius of curvature has more potential of accidence
	Slope	Locations on roads with higher slope have higher potential for accidence
Spatial Measures	*Distance from Intersection*	Locations on roads closer to intersections have higher collision potential
	Distance from Population Centers	Locations on roads closer to the population centers such as cities have higher collision potential
Environmental	*Weather*	Falling and Unclear condition increase the collision potential
	Surface	Dry and icy surface increase the collision potential
	Lighting	Dusky and dark road lighting have higher potential for accidence
	Time	Rush hours have high potential of road collisions

All collision events were divided into 12 subsets (feature classes) based on the major highways in California. The highways were selected based on the frequency of collisions and the diversity in influential factors of collision events. Rule generation and classification algorithms were applied to each and every subset. These highways were chosen based on the following types: U.S highways, interstate highways and state and county highways. The following subsections provide brief descriptions of the selected highways.

US Highway 50 run east and west from San Francisco to South Lake Tahoe and down through Carson City, Nevada. This is a major highway that travelers use to get from Sacramento to South Lake Tahoe. Weekday morning and evening rush hours in Sacramento and San Francisco expects traffic. Road closures are rare but can occur during the winter months. Carry chains and expect road closures during the winter between Placerville and Lake Tahoe. There were 1329 collision events considered to be related to this highway in the dataset.

US Highway 101 runs north and south along the coast from Los Angeles to beyond the Oregon border. This highway passes through most of California's major coastal communities. It is slow going through San Francisco and Los Angeles. Major wine producing regions like Napa, Sonoma, Santa Barbara and San Luis Obispo can be easily reached from this highway. The 'Pacific Coast Highway' (Hwy 1, PCH) often mirrors the 101 to the west but is slower as it winds, climbs and dips along the coast. There were 8023 collision events considered to be related to this highway in the dataset.

US Highway 395 runs north and south connecting Southern California with the Northern Sierras and Oregon. Traffic is heavier in winter with skiers heading to Mammoth

Mountain, June Lake, and Lake Tahoe. The southernmost portion from Hesperia to Ridgecrest is a dangerous stretch of two lane highway. Many travelers prefer to avoid this section by taking Highway 14 through Mojave. To the south, US highway 395 follows along the Eastern Sierras through the high desert communities of Lone Pine and Bishop. Mt. Whitney, the highest peak in the lower 48 states, and Death Valley, the lowest point in the U.S. are both accessible from Lone Pine. Ancient lava flows are evident in the southern regions. Road closures are common during the winter months between June Lake and the Nevada border. Carry chains orders occur between Bishop and Nevada during the winter. The 395 re-enters California north of Lake Tahoe and traverses the remote, less traveled, northeast corner of the Shasta Cascade. The stretch between Bishop and Inyokern has impressive volcanic topography. Totally 310 collision events are considered to be related to this highway in the dataset.

Interstate Highway 5 runs north and south and is the main nervure for travel within California. Interstate 5 is the only highway that traverses the entire state from the Oregon to Mexico borders and passes through major cities, like Sacramento, Los Angeles, Anaheim and San Diego. The stretch of highway between Redding and Bakersfield passes through Central Valley and the rural farm belt of the state referred to as the California Heartland. North of Redding, Interstate 5 climbs into the pine trees of the Shasta Cascade to the Oregon border. To the south, the entire stretch between Castaic and the Mexico border is urban sprawl. There were 46362 collision events considered to be related to this highway in the dataset.

Interstate Highway 8 runs east and west from San Diego paralleling the Mexico border, to Arizona. From San Diego, Interstate 8 climbs into the Cleveland National Forest, descends to El Centro in the desert and continues on through Yuma, Arizona. Off-roading is a popular activity in the desert sand dunes of Interstate 8. Totally 679 collision events are considered related to this highway in the dataset.

Interstate Highway 10 runs east and west connecting the beach city of Santa Monica with Los Angeles, the Inland Empire, Palm Springs and the Colorado River on the Arizona border. The popular mountain communities of Big Bear Lake and Idyllwild are also accessible from Interstate 10. Totally 3690 collision events are considered related to this highway in the dataset.

Interstate Highway 15 runs northeast from San Diego to Las Vegas, Nevada, and passes through Temecula wine country, the Inland Empire and the high desert communities of Barstow and Baker. Baker has the world's tallest thermometer. Interstate 15 is prone to Las Vegas traffic heading northbound on Friday night into Saturday and southbound on Sundays. Totally 2100 collision events are considered related to this highway in the dataset.

Interstate Highway 40 runs east and west from Interstate 15 in Barstow to Lake Havasu on the Arizona border. It has access to the Providence Mountains State Recreation Area from Interstate 40. Totally 184 collision events are considered related to this highway in the dataset.

Interstate Highway 80 runs east and west from San Francisco to North Lake Tahoe and down through Reno, Nevada. This is a major highway that travelers use to commute between San Francisco and Sacramento to North Lake Tahoe. Carry chain orders and road closures can be expected during the winter between Auburn and Lake Tahoe. There are 80 passes through the Gold Country town of Auburn and the nearby towns of Grass Valley and Nevada City. There were 1284 collision events considered related to this highway in the dataset.

Interstate Highway 580 runs southeast from San Rafael in the North Bay to Interstate 580 is a main highway for travelers coming from central and southern California to San Francisco and the North Coast. A toll may be required at the Richmond / San Rafael Bridge. There were 174 collision events considered related to this highway in the dataset.

State and County Highway 14 runs northeast from Interstate 5 in Santa Clarita to the U.S. highway 395 in Inyokern. Travelers from the south use this highway as a safer alternative to connect with U.S. highway 395. Although highway 14 becomes two lanes after the Mojave Desert, it is still a safer stretch of road than the southern portion of highway 395. Vasquez Rocks are located just north of the Interstate 5 connection. Highway 14 passes through the high-crime zones of Palmdale and Lancaster. The drivers try to have enough fuel to pass through these areas without stopping. There were 882 collision events considered related to this highway in the dataset.

State and County Highway 99 runs north and south from Red Bluff to Bakersfield. Highway 99 parallels interstate 5 and is a more safe drive. Highway 99 travels through the major cities of Fresno, Modesto and Sacramento. There were 5097 collision events considered related to this highway in the dataset.

5.4 Transformation and Fuzzification of Input Data

In this step, the discrete values in the information table are converted to digital representation, which can be assigned to the fuzzy membership function. The information table used in our study contained integer and float values for all attributes. Therefore, the categorical variable were identified and coded by converting text into integers. For example, weather was derived to include the input discrete values of A (clear), B (cloudy), C (raining), D (snowing), E (fog) and F (wind). The text values were converted to digital values such as 1 (clear), 2 (wind), 3 (fog), 4 (raining) and 5 (snowing), based on the three membership functions of clear, unclear and fall. The descriptions in the following subsections represent the membership functions and their equation of each mention of influential factors of vehicle collisions used in the dataset.

Fuzzy membership functions were designed using MATLAB 2012, which is a high-performance language for technical computing. MATLAB was selected for several reasons. It integrates computation, visualization and programming in an easy-to-use environment, where problems and solutions are expressed in familiar mathematical notations. MATLAB also has a special toolbox to handle fuzzy programming, i.e. the Fuzzy Logic Toolbox.

- ***Weather***
    - This factor shows the weather condition at the time of the collision. Falling and unclear conditions increase the collision potential. Table 5.3 shows the values of the weather conditions in the information table and their corresponding digital values assigned based on the fuzzy membership function, which is represented in Figure 5.2.

Table 5.3 Conversion of Text Value Format to Digit Representation of Weather Attribute

Text Value in the Dataset	Assigned Digital Values
Clear	1
Cloudy	2
Wind	3
Fog	4
Raining	5
Snowing	6

Prior to calculating the fuzzy granular conditional entropy (FGCE) and constructing the fuzzy granular decision trees, the membership functions of each factor had to be specified using expert knowledge. Based on the collision data in the information table of the fuzzy membership function, the fuzzification of collision data in the information table was applied by the membership functions. The membership functions of the weather attributes are represented in Figure 5.2.

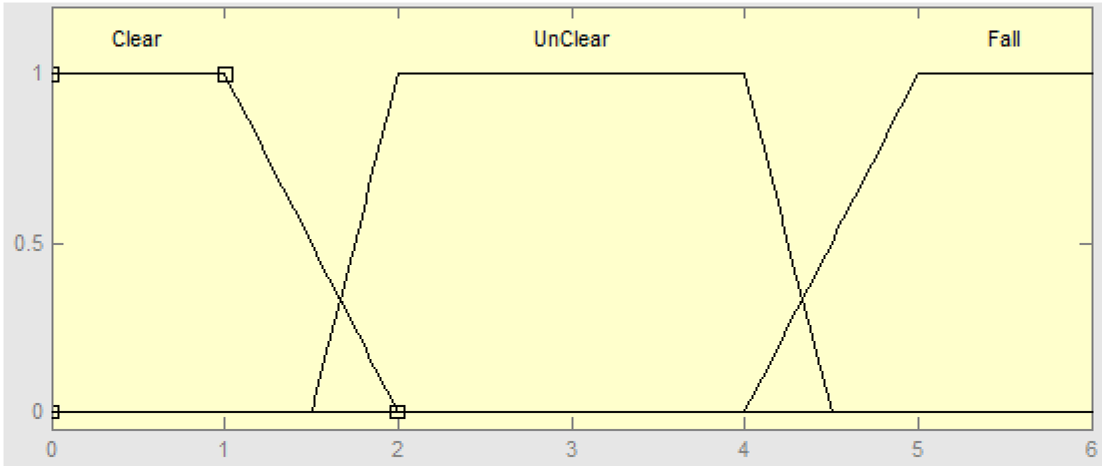


Figure 5.2 Membership Functions of Weather Attributes

The following equations express the membership functions of the weather attribute:

$$\mu_{Clear} = \begin{cases} 1 & x < 2 \\ 2 - x & 1 \leq x \leq 2 \end{cases}$$

$$\mu_{Unclear} = \begin{cases} \frac{x - 1.5}{0.5} & 1.5 \leq x \leq 2 \\ 1 & 2 \leq x \leq 4 \\ \frac{4.5 - x}{0.5} & 4 \leq x \leq 4.5 \end{cases}$$

$$\mu_{Fall} = \begin{cases} x - 4 & 4 \leq x \leq 5 \\ 1 & x > 5 \end{cases}$$

- ***Road Surface***

  This factor shows the road surface condition at the time of the collision. Wet and icy surfaces increase the collision potential. Table 5.4 shows the values of the road surface conditions in the information table and their corresponding digital values assigned based on the fuzzy membership functions, which are represented in Figure 5.3.

Table 5.4 Conversion of Text Value Format to Digital Representation of Road Surface Attributes

Text Value in the Dataset	Assigned Digital Values
Dry	1
Wet	2
Slippery (Muddy, Oily, etc.)	3
Snowy or Icy	4

The membership function of Weather attribute represents in figure 5.3:

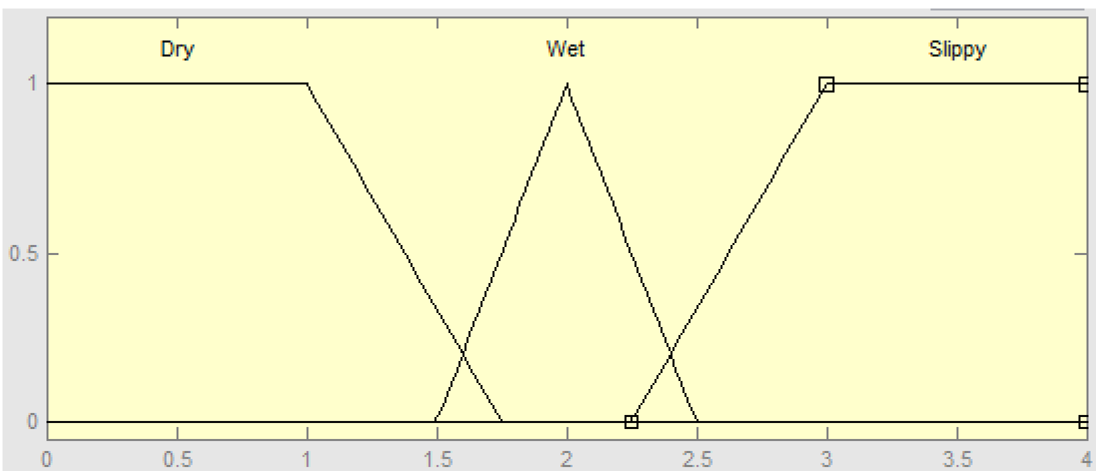


Figure 5.3 Membership Functions of Road Surface Attributes

The following equations express the membership functions of road surface attribute:

$$\mu_{Dry} = \begin{cases} 1 & x < 1.75 \\ \dfrac{1.75 - x}{0.25} & 1 \leq x \leq 1.75 \end{cases}$$

$$\mu_{Wet} = \begin{cases} \dfrac{x - 1.5}{0.5} & 1.5 \leq x \leq 2 \\ 1 & x = 2 \\ \dfrac{2.5 - x}{0.5} & 2 \leq x \leq 2.5 \end{cases}$$

$$\mu_{Slippy} = \begin{cases} \dfrac{x-2.75}{0.75} & 2.75 \leq x \leq 3 \\ 1 & x > 3 \end{cases}$$

- ***Road Lighting***

  This factor shows the road lighting condition at the time of the collision. Dusky and dark road lighting present higher accident potential. Table 5.5 shows the values of road lighting conditions in the information table and their corresponding digital values assigned based on the fuzzy membership functions, which are represented in Figure 5.4.

Table 5.5 Conversion of Text Value Format to Digital Representation of Road Lighting Attributes

Text Value in the Dataset	Assigned Digital Values
Daylight	1
Dusk – Dawn	2
Dark - Street Lights	3
Dark - No Street Lights	4
Dark - Street Lights Not Functioning	5

The membership function of Weather attribute represents in figure 5.4**:**

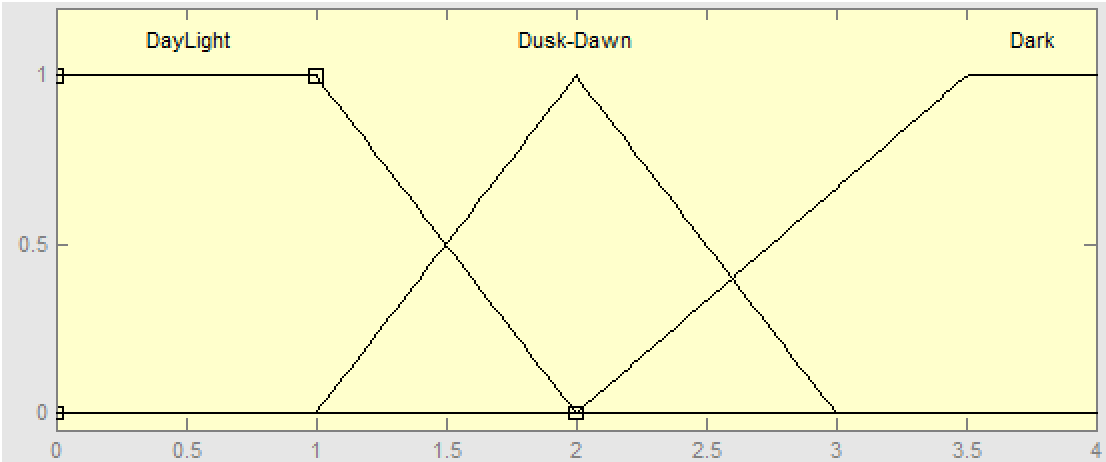


Figure 5.4 Membership Functions of Road Lighting Attributes

The following equations express the membership functions of road lighting attribute:

$$\mu_{Daylight} = \begin{cases} 1 & x < 2 \\ 2 - x & 1 \leq x \leq 2 \end{cases}$$

$$\mu_{Dusk-Dawn} = \begin{cases} x - 1 & 1 \leq x \leq 2 \\ 1 & x = 2 \\ 3 - x & 2 \leq x \leq 3 \end{cases}$$

$$\mu_{Dark} = \begin{cases} \dfrac{x - 2}{1.5} & 2 \leq x \leq 3.5 \\ 1 & x > 3.5 \end{cases}$$

- ***Time of Collisions***

This factor shows the time of day of collisions. Rush hours have a high potential for collisions. Time attributes have continuous values; therefore, there was no need for conversion to digital values. In this study, the exact values of the time attribute in the information table were applied in the construction of the fuzzy granular decision tree (FGDT). The designed membership functions are illustrated in Figure 5.5.

Figure 5.5 Membership Functions of the Time of Collisions

The following equations express the membership functions of time attribute:

$$\mu_{Night} = \begin{cases} 1 & x < 4 \\ \dfrac{8-x}{4} & 4 \leq x < 8 \end{cases}$$

$$\mu_{Morning} = \begin{cases} \dfrac{x-6}{2} & 6 \leq x < 8 \\ 1 & 8 \leq x \leq 12 \\ \dfrac{15-x}{3} & 12 < x \leq 3 \end{cases}$$

$$\mu_{Afternoon} = \begin{cases} \dfrac{x-12}{4} & 12 \leq x < 16 \\ 1 & 16 \leq x \leq 18 \\ \dfrac{22-x}{4} & 18 < x \leq 22 \end{cases}$$

$$\mu_{Night} = \begin{cases} \dfrac{x-18}{2} & 18 \leq x < 20 \\ 1 & x \geq 20 \end{cases}$$

- ***Slope***

This factor shows the situation of the vertical curve of the road. Locations on roads with higher slope have higher potential for collisions. The slope attribute has continuous values; therefore, there was no need for conversion to digital values. In this study, the exact values of the slope attribute in the information table were applied in the construction of the FGDT. The designed membership functions are depicted in Figure 5.6.

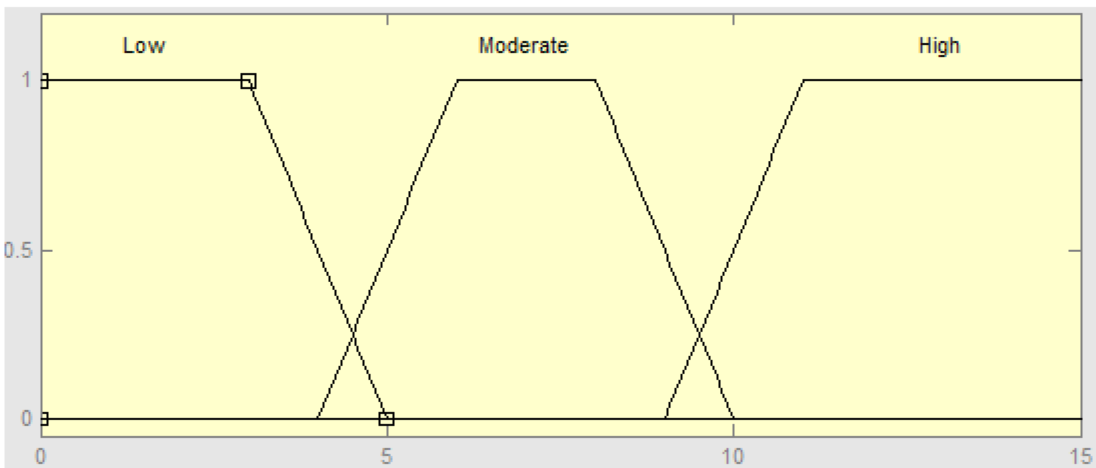


Figure 5.6 Membership Functions of the Road Slope

The following equations express the membership functions of slope attribute:

$$\mu_{Low} = \begin{cases} 1 & x < 3 \\ \dfrac{5-x}{2} & 3 \leq x < 5 \end{cases}$$

$$\mu_{Moderate} = \begin{cases} \dfrac{x-4}{2} & 4 \leq x < 6 \\ 1 & 6 \leq x \leq 8 \\ \dfrac{10-x}{2} & 8 < x \leq 10 \end{cases}$$

$$\mu_{High} = \begin{cases} \dfrac{x-9}{2} & 9 \leq x < 11 \\ 1 & x \geq 11 \end{cases}$$

$$\mu_{Far} = \begin{cases} \dfrac{x-200}{50} & 200 \le x < 275 \\ 1 & x \ge 275 \end{cases}$$

- ***Radius***

This factor shows the vertical radius of road at the location of a collision. Road locations where the vertical radius is small have greater collision potential. The vertical radius of road attribute has continuous values; therefore, there was no need for conversion to digital values. In this study, the exact values of the radius attribute in the information table were applied in the construction of the FGDT. The designed membership functions are figured out in the figure 5.7:

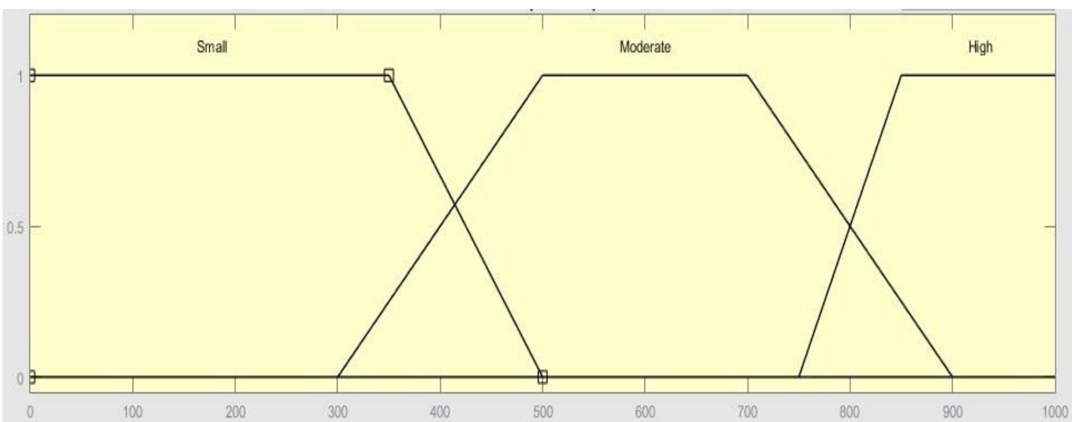


Figure 5.7 Membership Functions of the Road Radius

The following equations express the membership functions of collision distance from intersection attribute.

$$\mu_{small} = \begin{cases} 1 & x < 350 \\ \dfrac{500-x}{150} & 350 \le x < 500 \end{cases}$$

$$\mu_{Moderate} = \begin{cases} \dfrac{x-300}{200} & 300 \le x < 500 \\ 1 & 500 \le x \le 750 \\ \dfrac{1000-x}{250} & 750 < x \le 1000 \end{cases}$$

$$\mu_{High} = \begin{cases} \frac{x-900}{250} & 750 \leq x < 900 \\ 1 & x \geq 900 \end{cases}$$

- ***Distance from Intersection***

This factor shows the distance of the crash location to the nearest intersection. Locations closer to intersections have higher collision potential, especially close to cities. The distance from the intersection attribute has continuous values; therefore, there was no need for conversion to digital values. In this study, the exact values of the distance from the intersection attribute in the information table were applied in the construction of the FGDT. The membership functions are illustrated in Figure 5.8.

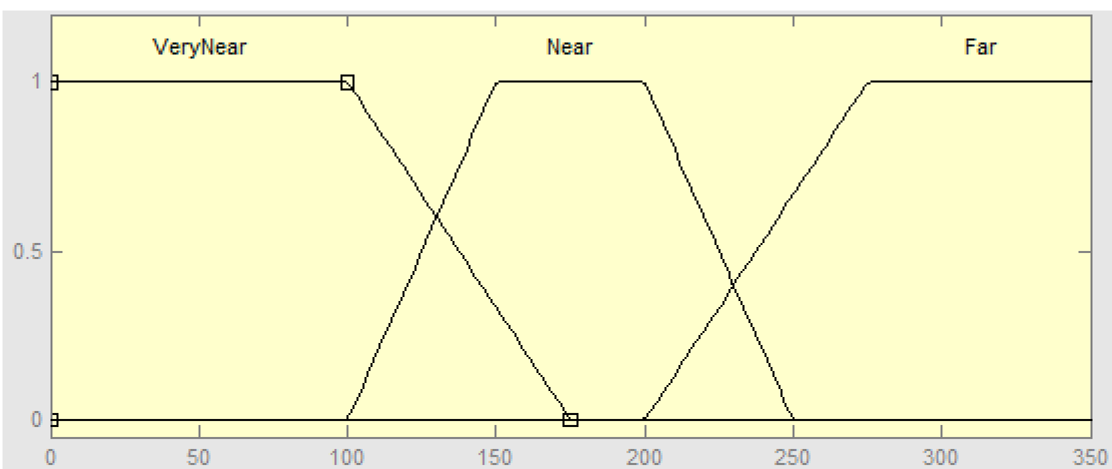


Figure 5.8 Membership Functions of the Distance from Intersection

The following equations express the membership functions of collision distance from intersection attribute:

$$\mu_{VeryNear} = \begin{cases} 1 & x < 100 \\ \frac{175-x}{75} & 100 \leq x < 175 \end{cases}$$

$$\mu_{Near} = \begin{cases} \dfrac{x-100}{50} & 100 \le x < 150 \\ 1 & 150 \le x \le 200 \\ \dfrac{250-x}{50} & 200 < x \le 250 \end{cases}$$

$$\mu_{Far} = \begin{cases} \dfrac{x-200}{50} & 200 \le x < 275 \\ 1 & x \ge 275 \end{cases}$$

- ***Distance from Population Centers***

This factor shows the distance of a crash location to the nearest population centre. Locations on roads closer to the population centres, such as cities, have higher collision potential. The distance from population centre attribute has continuous values; therefore, there was no need for conversion to digital values. In this study, the exact values of the distance from population centres attributes in the information table were applied in the construction of the FGDT. The designed membership functions are depicted in Figure 5.9.

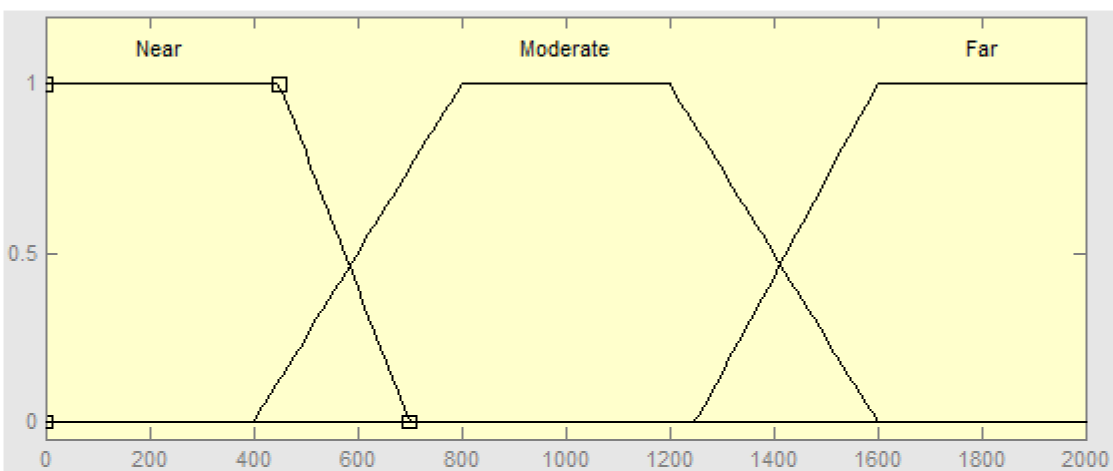


Figure 5.9 Membership Functions of the Collision Distance from Population Center

The following equations express the membership functions of collision distance from population centers attribute:

$$\mu_{Near} = \begin{cases} 1 & x < 450 \\ \dfrac{700-x}{350} & 400 \le x < 750 \end{cases}$$

$$\mu_{Near} = \begin{cases} \frac{x-400}{400} & 400 \leq x < 800 \\ 1 & 800 \leq x \leq 1200 \\ \frac{1600-x}{400} & 1200 < x \leq 1600 \end{cases}$$

$$\mu_{Far} = \begin{cases} \frac{x-1200}{400} & 1200 \leq x < 1600 \\ 1 & x \geq 1600 \end{cases}$$

- ***Severity***

This attribute expresses the final class of vehicle collision severity. The severity attribute is a decision attribute that has text values that need to be converted to digital values. Table 5.6 shows the values of severity in the information table and their corresponding digital values assigned based on the fuzzy membership

Table 5.6 Conversion of Text Value Format to Digital Representation of Severity Attribute

Text Value in the Dataset	**Assigned Digital Values**
PDO	1
Injury (Complaint of Pain)	2
Injury (Other Visible)	3
Injury (Severe)	4
Fatal	5

The membership function of Severity attribute represents in figure 5.10:

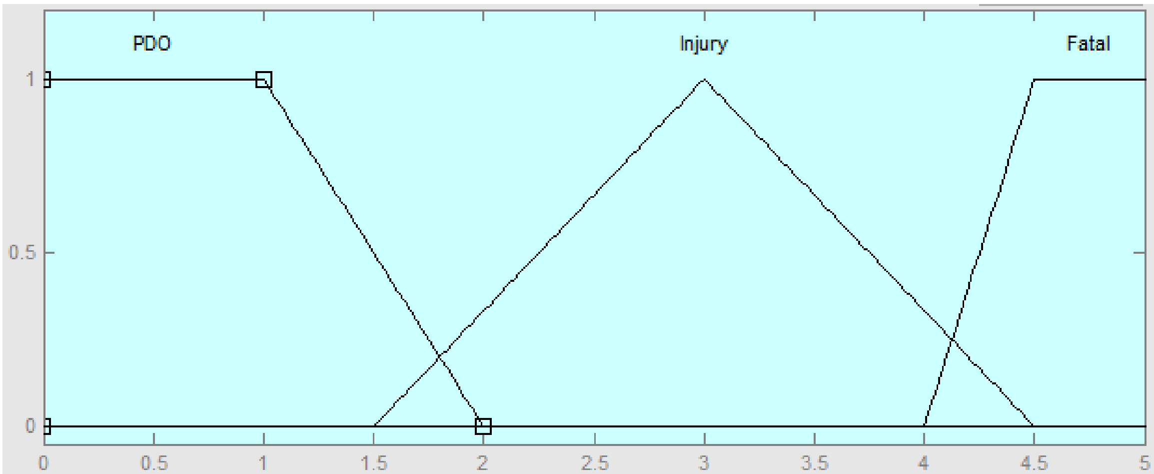


Figure 5.10 Membership Functions of the Collision Severity

5.5 Implementation of Fuzzy Granular Decision Tree for Vehicle Collision Events

In this section, the calculation of the proposed FGDT measures with a sample of information table, according to the mentioned methodology, is described. The information table of training dataset in this research was constructed from 48150 instances, and 20635 instances were used as testing data. Table 5.7 shows a sample of an information table with 16 rows of unique objects and nine columns of attributes.

To produce the FGDT, the proposed methodology was designed and implemented with MATLAB software. To select the proper nodes and construct the FGDT by computing the fuzzy generality, fuzzy coverage, fuzzy confidence and fuzzy granular conditional entropy based on a need for more information, more consistent and high-quality granules were needed, as well as less redundancy with the universe. Table 5.8 demonstrates the sample of the first level of FGDT measures, so that the universe was equal to $\{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8, O_9, O_{10}, O_{11}, O_{12}, O_{13}, O_{14}, O_{15}, O_{16}\}$, where is O_i are object numbers with $i = 1,\ldots,16$.

Table 5.7 The Road Collision Decision Table

Objects	*Weather*	*Surface*	*Lighting*	*Time*	*Radius (m)*	*Slope (%)*	*Distance from Intersection(m)*	*Distance from Population Centers(m)*	*Severity*
O_1	*Clear*	*Dry*	*DayLight*	*11*	*800.00*	*3.00*	*360.00*	*1700.00*	*PDO*
O_2	*Clear*	*Dry*	*Dusky/Dark*	*10*	*850.00*	*7.00*	*180.00*	*2800.00*	*PDO*
O_3	*Clear*	*Dry*	*Dusky/Dark*	*20*	*900.00*	*8.00*	*280.00*	*1000.00*	*Injury*
O_4	*Clear*	*Not Dry*	*DayLight*	*9*	*300.00*	*4.00*	*200.00*	*1900.00*	*Injury*
O_5	*Clear*	*Not Dry*	*Dusky/Dark*	*13*	*550.00*	*11.00*	*90.00*	*900.00*	*Injury*
O_6	*Clear*	*Not Dry*	*Dusky/Dark*	*21*	*880.00*	*3.50*	*170.00*	*1800.00*	*Injury*
O_7	*Raining*	*Not Dry*	*DayLight*	*12*	*890.00*	*4.00*	*420.00*	*1100.00*	*PDO*
O_8	*Raining*	*Not Dry*	*Dusky/Dark*	*14*	*350.00*	*10.00*	*220.00*	*750.00*	*Injury*
O_9	*Raining*	*Not Dry*	*Dusky/Dark*	*15*	*350.00*	*18.00*	*75.00*	*420.00*	*Fatal*
O_{10}	*Fog*	*Dry*	*DayLight*	*11*	*770.00*	*6.00*	*320.00*	*1450.00*	*PDO*
O_{11}	*Fog*	*Dry*	*Dusky/Dark*	*10*	*500.00*	*9.00*	*240.00*	*350.00*	*Injury*
O_{12}	*Fog*	*Dry*	*Dusky/Dark*	*21*	*780.00*	*13.00*	*130.00*	*980.00*	*Injury*
O_{13}	*Fog*	*Not Dry*	*DayLight*	*10*	*1500.00*	*4.00*	*170.00*	*1200.00*	*PDO*
O_{14}	*Fog*	*Not Dry*	*Dusky/Dark*	*20*	*450.00*	*3.00*	*110.00*	*650.00*	*Fatal*
O_{15}	*Fog*	*Not Dry*	*Dusky/Dark*	*22*	*600.00*	*7.00*	*110.00*	*650.00*	*Fatal*
O_{16}	*Fog*	*Not Dry*	*Dusky/Dark*	*21*	*850.00*	*14.00*	*110.00*	*650.00*	*Fatal*

Seven granules of formulas (road lighting = daylight, distance to intersection = far, weather = clear, distance to intersection = very near, collision time = morning, road radius = small, road slope = low) had minimum entropy values. They were chosen first as candidate granules to be nodes of a decision tree based on the value of entropy and generality. They were sorted in order of their value of entropy. The formula of road lighting = daylight with the granule of $\{O_1, O_4, O_7, O_{10}, O_{13}\}$ was chosen as the first node of the FGDT, given its lowest entropy value. The other six granules could not cover the universe; therefore; they were not a covering solution to reduce redundancy.

The algorithm searched and analyzed other granules, in order to find a set of granules that cover the whole universe. The algorithm considered the non-redundant covering and removed those six addition candidates, since they could not form a non-redundant covering. As a consequence, these granules could not be selected, even if other measures were in favour of this granule. Granules were to cover the universe and were chosen accordingly.

Table 5.8 The Measures of Fuzzy Granule

Formula	*Granular*	*Generality*	***Confidence***			***Coverage***			*Granular Fuzzy Conditional Entropy*
			PDO	***Injury***	***Fatal***	***PDO***	***Injury***	***Fatal***	
Weather = Clear	$\{O_1,O_2,O_3,O_4,O_5,O_6\}$	*0.38*	*0.33*	*0.67*	*0.00*	*1.00*	*1.00*	*0.00*	*0.92*
Weather = Unclear	$\{O_7,O_8,O_9\}$	*0.19*	*0.33*	*0.33*	*0.33*	*1.00*	*1.00*	*1.00*	*1.58*
Weather = Fall	$\{O_{10},O_{11},O_{12},O_{13},O_{14},O_{15},O_{16}\}$	*0.44*	*0.29*	*0.29*	*0.43*	*1.00*	*1.00*	*1.00*	*1.56*
Road Surface = Dry	$\{O_1,O_2,O_3,O_{10},O_{11},O_{12}\}$	*0.38*	*0.50*	*0.50*	*0.00*	*1.00*	*1.00*	*0.00*	*1.00*
Road Surface = Wet	$\{O_4,O_5,O_6,O_7,O_8,O_9,O_{13},O_{14},O_{15},O_{16}\}$	*0.63*	*0.29*	*0.14*	*0.57*	*0.67*	*1.00*	*1.00*	*1.38*
Road Lighting = Day Light	$\{O_1,O_4,O_7,O_{10},O_{13}\}$	*0.31*	*0.80*	*0.20*	*0.00*	*1.00*	*1.00*	*0.00*	*0.72*
Road Lighting = Dusk-Down	$\{O_2,O_3,O_5,O_6,O_8,O_9,O_{11},O_{12},O_{14},O_{15},O_{16}\}$	*0.69*	*0.17*	*0.17*	*0.67*	*0.25*	*1.00*	*1.00*	*1.25*
Collision Time = Morning	$\{O_1,O_2,O_4,O_5,O_7,O_8,O_{10},O_{11},O_{13}\}$	*0.56*	*0.63*	*0.38*	*0.00*	*1.00*	*1.00*	*0.00*	*0.96*
Collision Time = Night	$\{O_3,O_6,O_9,O_{12},O_{14},O_{15},O_{16}\}$	*0.44*	*0.00*	*0.44*	*0.56*	*0.00*	*0.57*	*1.00*	*0.99*
Road Radius = Small	$\{O_4,O_8,O_9,O_{11},O_{14}\}$	*0.31*	*0.00*	*0.60*	*0.40*	*0.00*	*1.00*	*1.00*	*0.97*
Road Radius = Moderate	$\{O_5,O_{15}\}$	*0.13*	*0.00*	*0.50*	*0.50*	*0.00*	*0.34*	*0.44*	*1.00*
Road Radius = High	$\{O_1,O_2,O_3,O_6,O_7,O_{10},O_{12},O_{13},O_{16}\}$	*0.56*	*0.53*	*0.33*	*0.14*	*1.00*	*1.00*	*1.00*	*1.41*

Road Slope = Low	$\{O_1,O_4,O_6,O_7,O_{13}\}$	0.31	0.62	0.38	0.00	1.00	1.00	0.00	0.96
Road Slope = Moderate	$\{O_2,O_3,O_8,O_{10},O_{11}\}$	0.31	0.57	0.43	0.00	1.00	1.00	0.00	0.99
Road Slope = High	$\{O_5,O_9,O_{12},O_{14},O_{15},O_{16}\}$	0.38	0.00	0.50	0.50	0.00	0.80	1.00	1.00
Dist to Intersection = Very Near	$\{O_5,O_9,O_{12},O_{14},O_{15}\}$	0.31	0.00	0.37	0.63	0.00	0.96	0.76	0.95
Dist to Intersection = Near	$\{O_2,O_4,O_6,O_8,O_{11},O_{13},O_{16}\}$	0.44	0.40	0.56	0.04	1.00	0.82	0.33	1.18
Dist to Intersection = Far	$\{O_1,O_3,O_7,O_{10}\}$	0.25	0.75	0.25	0.00	1.00	0.56	0.00	0.81
Dist to Population Centers = Near	$\{O_9,O_{11}\}$	0.13	0.00	0.50	0.50	0.00	1.00	0.63	1.00
Dist to Population Centers = Moderate	$\{O_3,O_5,O_7,O_8,O_{12},O_{13},O_{14},O_{15},O_{16}\}$	0.56	0.26	0.50	0.24	0.84	1.00	0.97	1.50
Dist to Population Centers = High	$\{O_1,O_2,O_4,O_6,O_{10}\}$	0.31	0.56	0.44	0.00	1.00	1.00	0.00	0.99

It can be verified that the union of two chosen formula granules $\{O_1,O_4,O_7,O_{10},O_{13}\}$ and $\{O_2,O_3,O_5,O_6,O_8,O_9,O_{11},O_{12},O_{14},O_{15},O_{16}\}$ satisfied universal coverage with no redundancy. The objects in road lighting = daylight and road lighting = dusk-down could not belong to the same decision classes, because they were active nodes (as mentioned in methodology section); therefore, further granulation to this granule needed be conducted in order to find smaller definable granules.

The road collision FGDT that was constructed with this algorithm is demonstrated in Figure 5.11. Rules extracted from the sample constructed FGDT are illustrated in Figure 5.12.

Considering fuzzy granular entropy, fuzzy generality and non-redundant covering, the other nodes for fuzzy decision tree would be chosen until all nodes were non-active. By running the algorithm on 12 major highways with 25,000 collision events as objects of a collision information table, the numbers of rules were inferred at three levels of the decision tree, which were labeled by a granular set; and, each branch was labeled by an attribute value of the parent. Table 5.9 demonstrates the number of rules that were extracted from the FGDT algorithm.

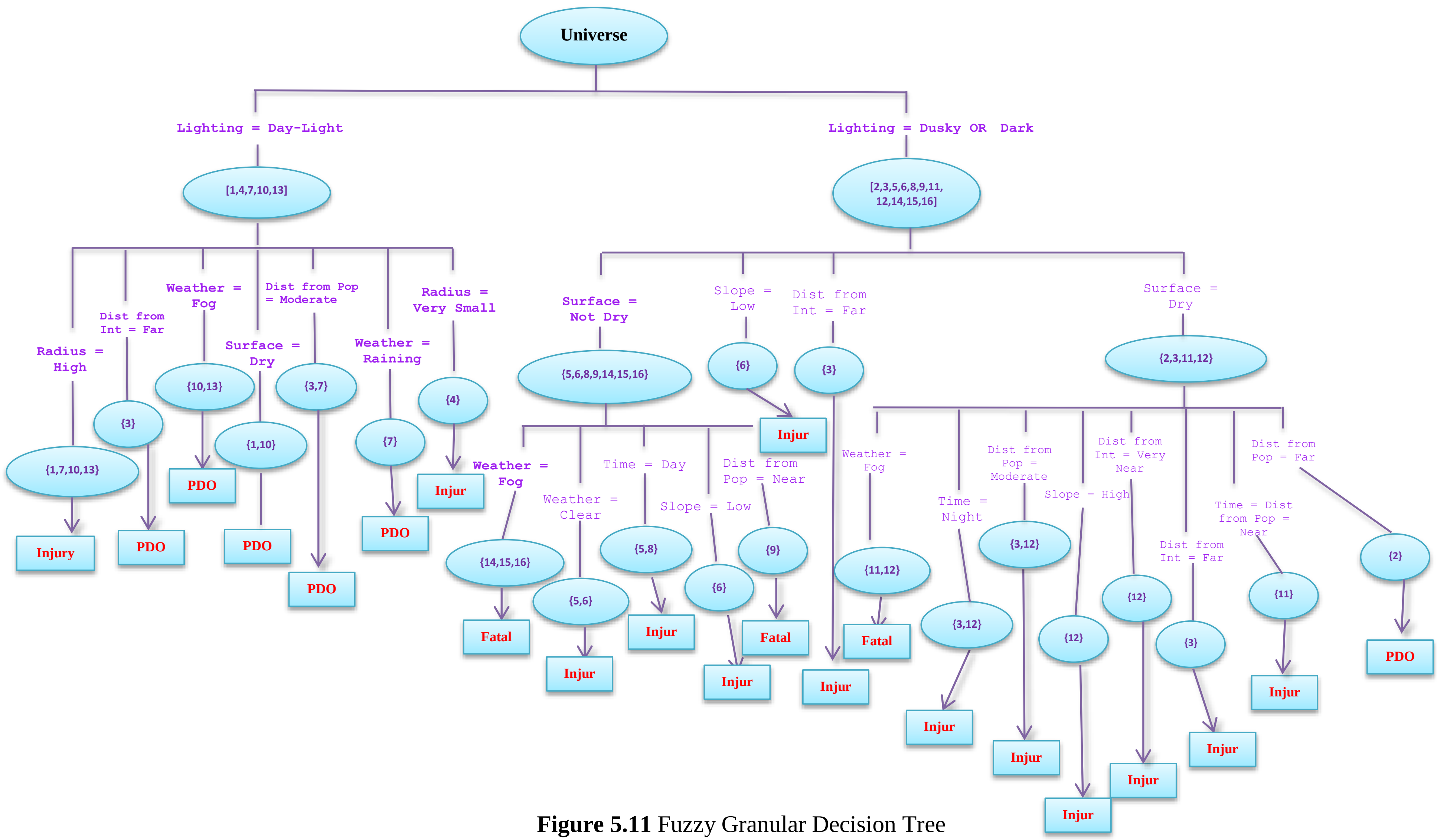


Figure 5.11 Fuzzy Granular Decision Tree

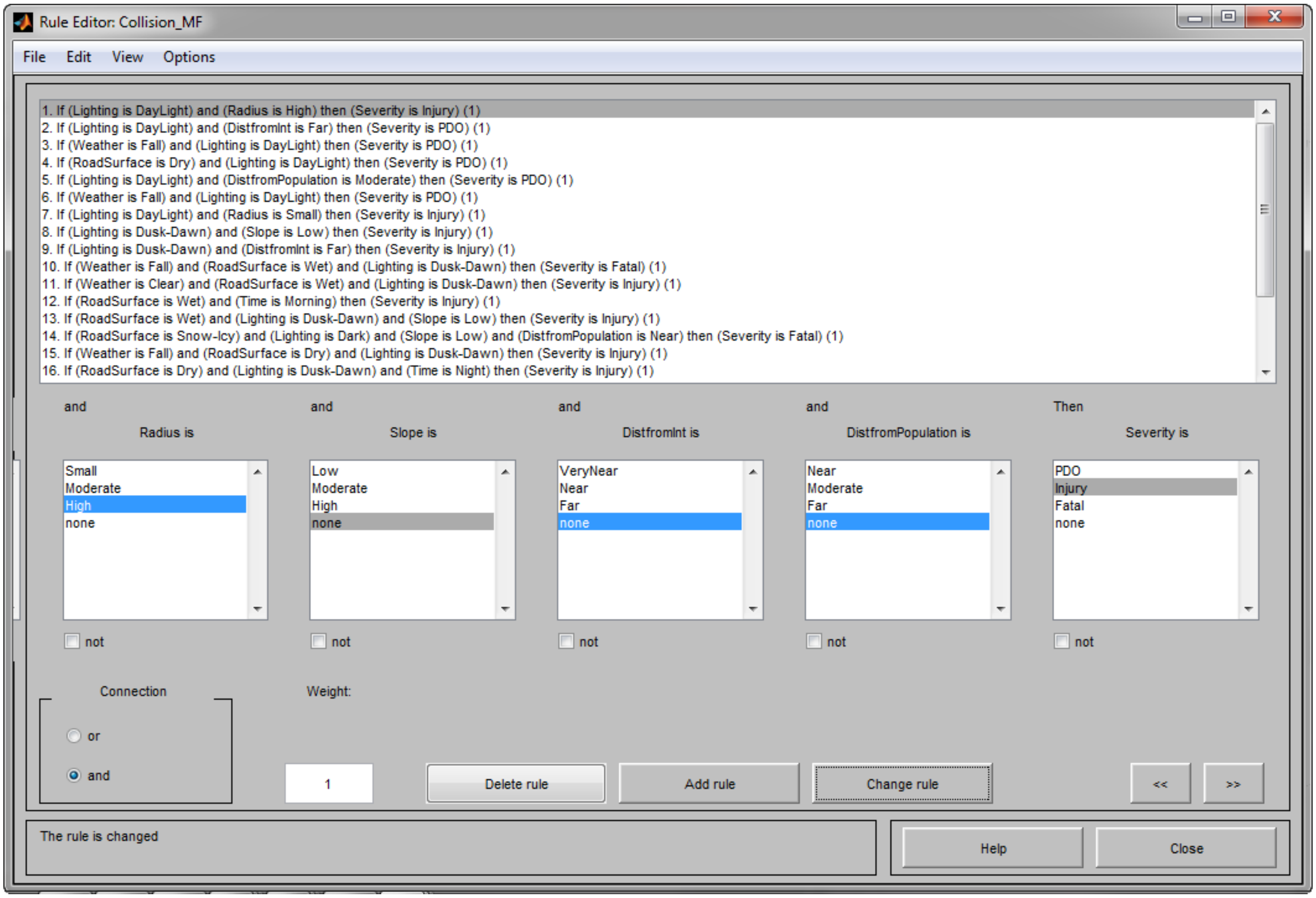


Figure 5.12 The extracted rules from fuzzy granular decision tree from sample information table

5.6 Implementation of the Reasoning with Fuzzy Granular Decision Tree

As mentioned in the methodology section, the fuzzy inference process comprises four parts fuzzification, rule evaluation, aggregation of the rule outputs, and defuzzification which were applied on testing data using MATLAB programing. The first step was the fuzzification interface, which transformed crisp data into fuzzy sets

Table 5.9 The number of extracted rules from FGDT method

Road Name	The Number of Events	Number of rules
All Roads	25000	87
US Highway 50	466	55
US Highway 101	2809	52
US Highway 395	187	23
Interstate Highway 5	16227	77
Interstate Highway 8	270	21
Interstate Highway 10	1292	48
Interstate Highway 15	850	61
Interstate Highway 40	85	19
Interstate Highway 80	450	56
Interstate Highway 580	100	22
State and County Highway 14	480	61
State and County Highway 99	1784	56

Rule evaluation was then applied on the all extracted rules belonging to the 12 major highways, and the strengths of the rules were computed based on the extracted rules and inputs. They were applied to antecedents of the fuzzy rules. In our study, the minimum (AND) fuzzy value was applied as the strength of rules.

The third step was the aggregation of the extracted rules, which was the process of unification of the outputs of all extracted rules. In this step, the implication operator was applied, and the consequent was reshaped using a function associated with the antecedent. The input for the implication process was a single number given by the antecedent, and the output was a fuzzy set. In this study, multiple rules can fire simultaneously for the same collision event as a testing object; therefore, the implication was implemented for each rule extracted by the FGDT.

Figure 5.13 presents the rule viewer display, which shows a roadmap of the whole fuzzy reasoning process. It is based on the fuzzy inference. There are 198 plots nested in the roadmap of Figure 5.13. The three plots across the top of the figure represent the antecedent and consequent of the first rule. Each rule is a row of plots, and each column is a variable. The rule numbers are displayed on the left of each row. In the designed system, we can click on a rule number to view the rule in the status line.

In the last step that is called defuzzification, the fuzzy rule which in turn translates the results back to the real world values. The testing data with the 10700 events are used in the deffuzification step and then the final collision severity class are determined based on the final value of defuzzification step. The below show the code implementation of the last step in MATLAB :

```
Checkdata = xlsread('CheckData.xlsx');
ConditionalAttCheckdata = Checkdata(:,2:5);
FuzzyParameters = readfis('Collision_MF.fis');

for j=1:size(ConditionalAttCheckdata,1)

    Object = Checkdata(j,:);
    FinalClasses(j,1) = evalfis(Object,FuzzyParameters);
End
```

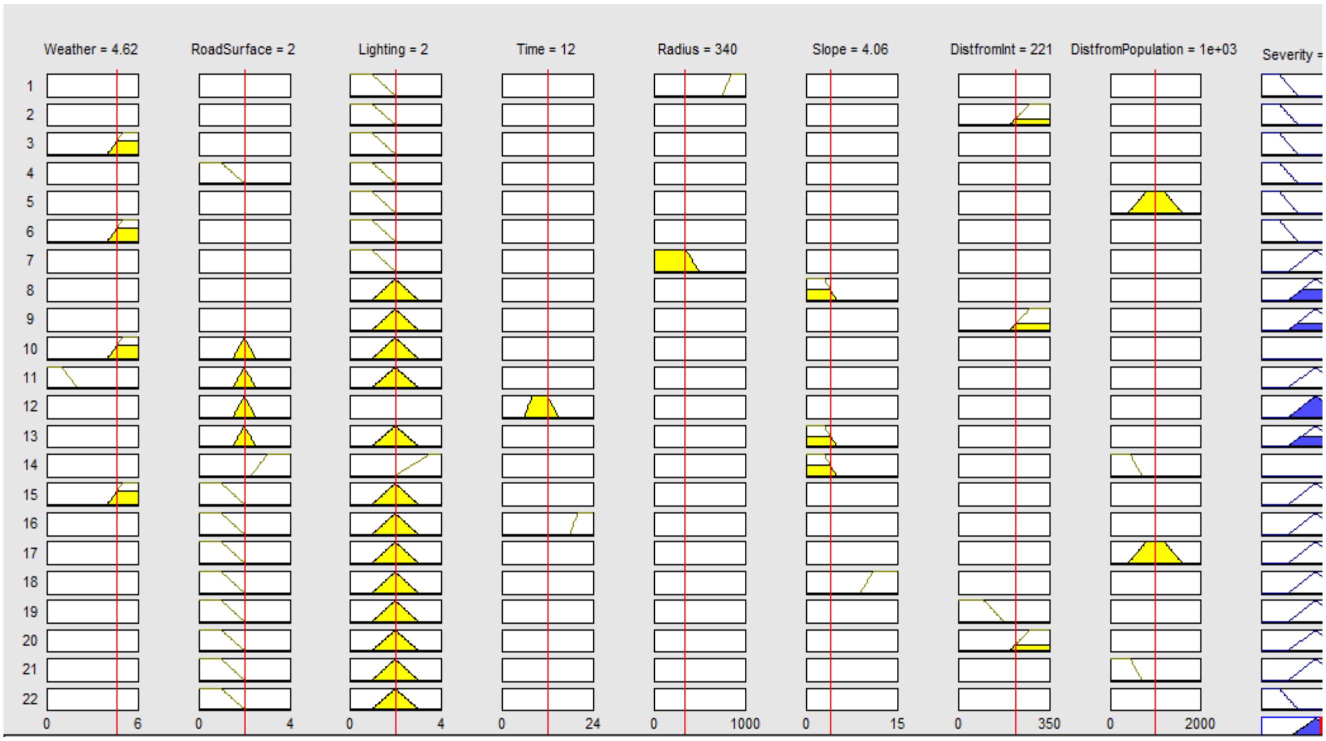


Figure 5.13 Rule Viewer Displays of a Roadmap for the Whole Fuzzy Reasoning Process

5.7 Accuracy Assessment of Fuzzy Granular Decision Tree

To demonstrate the suitability of the FGDT over that of the existing problem-solving approaches, such as ID3, C4.5, CART, SLIQ and Random Tree, all these decision tree algorithms were implemented in MATLAB using the road vehicle collision event dataset. Moreover, the accuracy of the classification methods was compared with respect to all and selected conditional attributes. To assess the accuracy of the result classes, the confusion matrix was employed in the classification accuracy assessment (Zhao, Yao et al. 2007). The diagonal elements in this matrix indicated the numbers of collision events for which the classification results agreed with the testing data, and the off-diagonal elements in each row presented the numbers of collision events that had been misclassified.

In order to properly generate the confusion matrix, testing data were needed. The overall accuracy was derived from the confusion matrix for all classification methods and was the sum of the major diagonal elements (i.e. correctly classified sample units) divided by the total number of sample units in the confusion matrix. This value was the most commonly reported accuracy assessment statistic, which is figure out in the below equation 7.

$$OveralAccuracy = O.A. = \frac{\sum_{i=1}^{c} a_{ii}}{N} \tag{7}$$

where a_{ii} is the number of collision events in class *i* in row *i*, which is classified by classifiers, and class *j* in column *j*, which are labeled in reality.

To assess the effect of the fuzzy rough set feature selection (FRFS) method in the process of constructing the decision tree, the confusion matrix of instances considering all and selected features was employed in the assessment of classification accuracy and time consumption of the decision tree method. The results obtain from various classification algorithms are given in Table 5.10. As shown in in this table, FRFS outperformed other algorithms with higher accuracy and shorter running times.

Table 5.10 Comparison of Overall Accuracy based on the Fuzzy Rough Set Feature Selection

Decision Tree Methods	All Features (Accuracy)	Total Time (s)	Selected Features (Accuracy)	Total Time (s)
ID3	65.2	36.21	65.2	24.14
C4.5	68.6	120.47	68.5	71.70
Rnd Tree	61.3	116.93	61.9	67.59
SLIQ	60.2	99.98	59.9	66.63
CART	71.3	1160.27	71.1	651.81
FGDT	**84.7**	**29.38**	**84.7**	**19.56**

From Table 5.10, it can be concluded that the FRFS algorithm decreased the time consumption of the decision tree construction. Although the time consumption decreased, the accuracy did not change considerably in all methods. Therefore, the FRFS method can have a positive effect on the performance of the mentioned decision tree methods. The following assessments are based on features selected by the FRFS method.

Table 5.11and Figure 5.14 represent the overall accuracy of different decision tree algorithms which is estimated for the major highways of California. It can be observed that the FGDT gives the most accurate results in the all instances and the instances belong to each road separately rather than the other decision tree algorithms in classifying the instances based on the manner of collision. All extracted rules are used in the fuzzy reasoning process to classified the instances belong to testing dataset to determine the severity classes. It can be concluded that the extracted rules from FGDT are more accurate than the other decision tree algorithms.

Table 5.11 Vehicle Event Classification Accuracy for the Main Highways of California

Highway Names	**Overall Accuracy**					
	ID3	C4.5	Rnd Tree	SLIQ	CART	FGDT
All Roads	65.2	68.5	61.9	59.9	71.1	84.76
US Highway 50	67.6	70.9	64.2	62.2	73.9	87.9
US Highway 101	58.9	61.8	55.9	54.2	64.2	82.8
US Highway 395	69.8	73.2	66.3	64.2	76.1	90.7
Interstate Highway 5	68.3	71.7	64.9	62.8	74.4	88.8
Interstate Highway 8	68.5	71.9	65.1	63.1	74.7	89.1
Interstate Highway 10	63.2	66.7	60.1	58.1	68.9	82.2
Interstate Highway 15	65.3	68.6	62.2	60.2	71.8	84.9
Interstate Highway 40	73.2	76.8	69.5	67.3	79.9	95.2
Interstate Highway 80	60.2	63.2	57.2	55.9	65.6	78.3
Interstate Highway 580	63.3	66.7	60.1	58.2	68.9	82.3
State and County Highway 14	65.3	68.6	62.1	60.1	71.2	84.9
State and County Highway 99	66.9	70.2	63.5	61.5	72.9	86.0

Other measures that could affect the interpretability and performance of the decision tree are the time consumed in constructing the decision tree and the number of rules extracted from the decision tree. Table 5.12 and Figure 5.15 present the time consumption of different decision tree algorithms used for collision database of the major highways of California. As Table.5.12 and Figure 5.15 show, the time consumption of the ID3 algorithm was less to build a model among the six classifiers for small datasets. However, the CART algorithm provided better accuracy than the other methods, except for the FGDT, but the time consumption of this method was high.

A measure of the usefulness of the decision tree can be based on the discriminatory power of the leaves, with leaf nodes being more desirable if they have low ambiguity with regards to the class to which a case is to be assigned. In our study, the discriminatory power of a decision tree was considered as the number of leaves in which more than 50% of the objects belonged to one class of severity. Table 5.13 and Figure 5.16 present the discriminatory power percent of each decision tree algorithm for the major highways of California.

Figure 5.12. Different decision tree methods time running based on California major

Highway Names	Time(s)					
	ID3	C4.5	Rnd Tree	SLIQ	CART	FGDT
All Roads	24.14	71.70	67.59	66.63	651.81	19.56
US Highway 50	0.45	1.34	1.26	1.24	4.56	0.82
US Highway 101	2.71	8.05	7.89	7.48	210.15	2.85
US Highway 395	0.18	0.53	0.50	0.49	0.91	0.19
Interstate Highway 5	15.67	46.54	43.88	43.25	456.89	14.85
Interstate Highway 8	0.26	0.78	0.73	0.72	0.85	0.32
Interstate Highway 10	1.25	3.71	3.50	3.45	105.30	0.98
Interstate Highway 15	0.83	2.46	2.32	2.29	87.35	0.89
Interstate Highway 40	0.08	0.24	0.22	0.22	0.32	0.12
Interstate Highway 80	0.43	1.28	1.20	1.19	3.50	0.53
Interstate Highway 580	0.09	0.28	0.52	0.25	0.85	0.18
State and County Highway 14	0.46	1.37	1.29	1.27	4.98	0.58

State and County High	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.03

Table 5.13 Total Percentage of Discriminatory Power of Decision Tree Methods

Highway Names	Discriminatory of Decision Tree (%)					
	ID3	C4.5	Rnd Tree	SLIQ	CART	FGDT
All Roads	43	58	60	65	73	82
US Highway 50	45	55	59	65	70	91
US Highway 101	39	54	51	52	70	82
US Highway 395	46	52	52	58	65	85
Interstate Highway 5	48	61	58	65	69	96
Interstate Highway 8	38	66	56	67	75	95
Interstate Highway 10	44	55	56	68	75	86
Interstate Highway 15	48	54	57	68	81	85
Interstate Highway 40	42	52	48	61	82	83
Interstate Highway 80	40	51	49	59	72	92
Interstate Highway 580	35	61	51	62	73	81
State and County Highway 14	45	50	48	55	69	80
State and County Highway 99	12	48	51	61	69	86

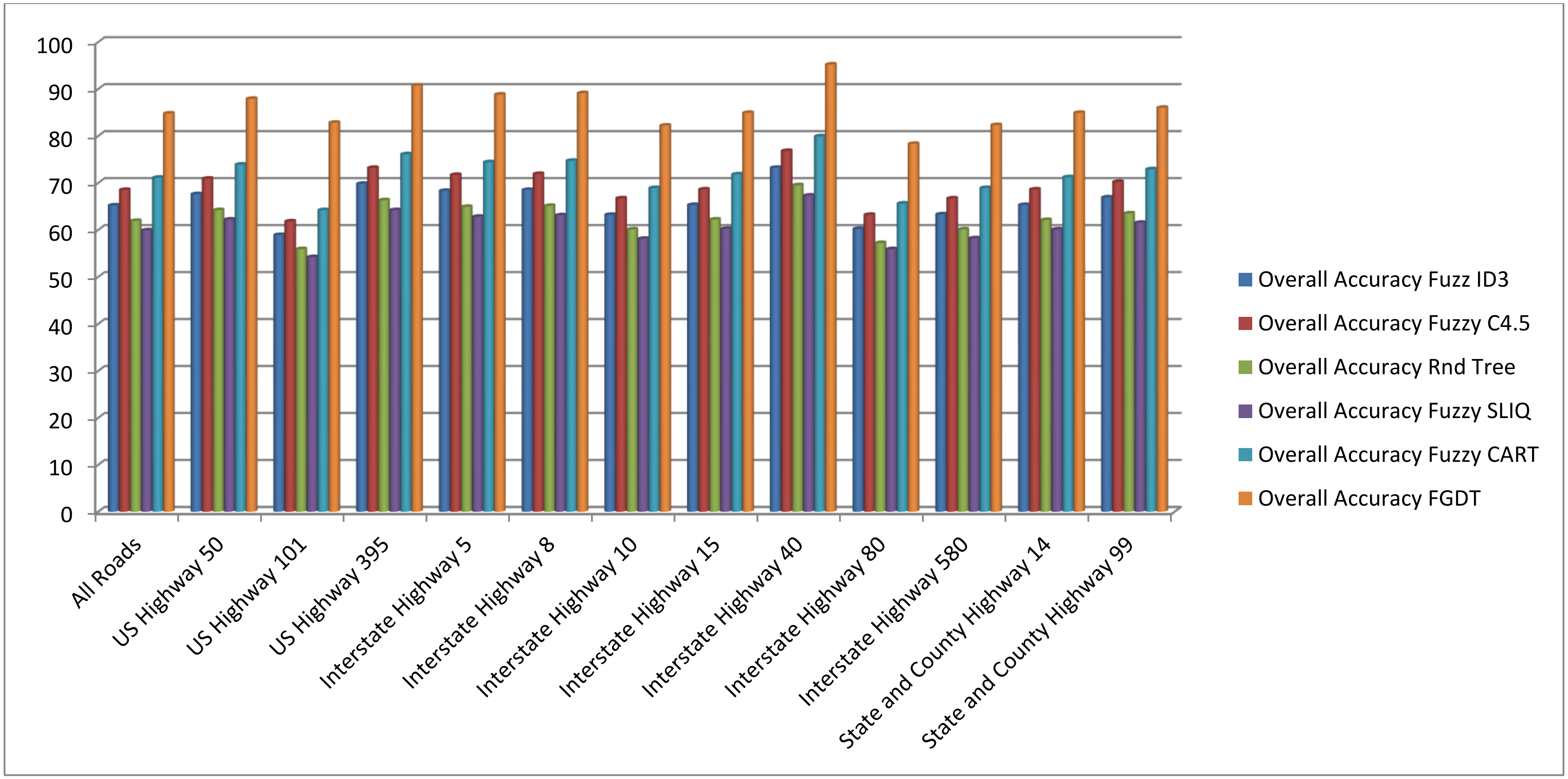


Figure 5.14 Accuracy of 6 Decision Tree Methods

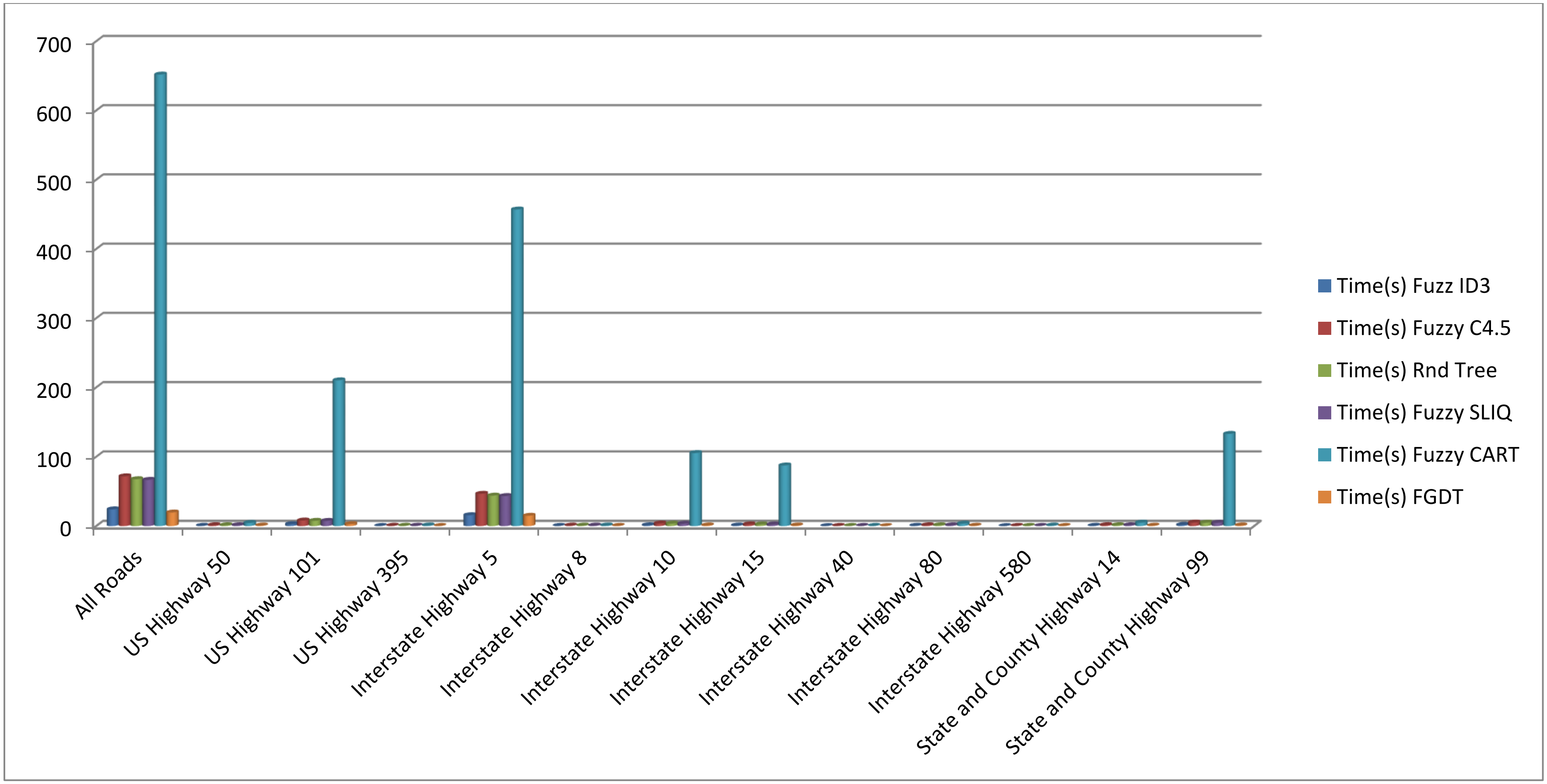


Figure 5.15 Computational Time of 6 Decision Tree Methods

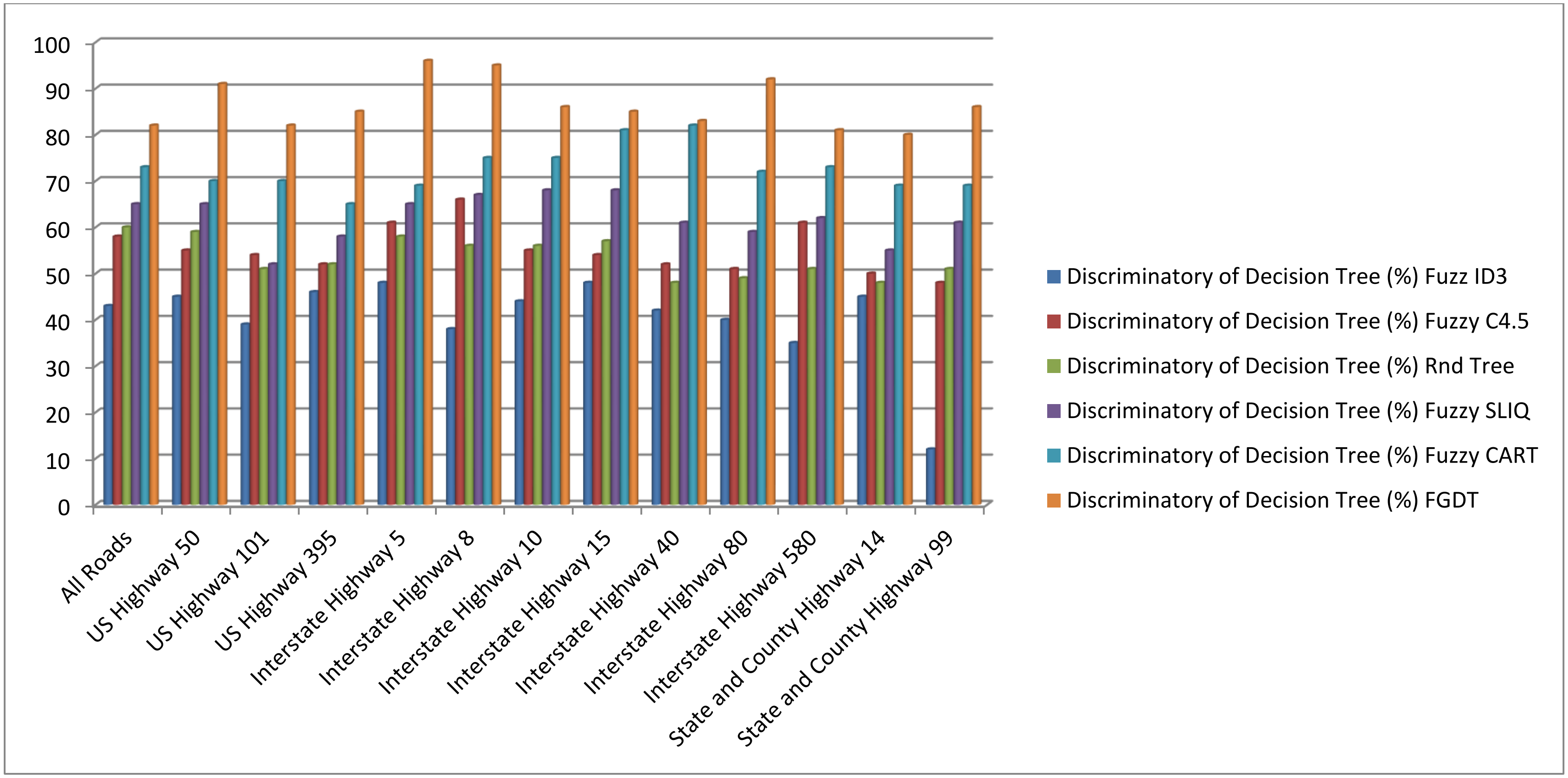


Figure 5.16 Discriminatory Power of six decision tree methods

By observing the experimental analysis, FGDT algorithm yields better accuracy compared to the other five decision trees for both small and large data sets. The time complexities to build a decision tree model using FGDT are better than the other methods for large data sets. Also, more than 80% of nodes in the FGTD are useful in the tree. It causes constructing smallest tree with useful rules and more accurate result. Hence, FGDT serves as a powerful model for solving the problem of balancing accuracy, size and time complexity of a decision tree.

Chapter Six: Conclusion and Future Works

6.1 Conclusion

This thesis proposes a new approach to extract rules for predicting the severity of vehicular road collisions. The results indicate that the fuzzy granular decision tree (FGDT) finds the most suitable granules defined by an attribute-value pair that is selected considering the continuous and discrete values in the database. Moreover, using fuzzy data input and fuzzy entropy impacts the performance of the learning by efficiently involving the discrete and continuous values in the database.

The fuzzy rough set selection is applied to a vehicle collision dataset to select the conditional features with minimum correlation and improve the performance of a constructed decision tree. As vehicle collision events databases include inconsistent objects (uncertainty and vagueness), the fuzzy rough set feature selection algorithm is applied to deal with uncertainty and vagueness.

This research solves some challenges in traditional decision tree methods, which have problems with a large number of branches, causing duplication and repetition of sub-trees within the tree when dealing with the large number of instances. Both repetition and duplication result in redundancy in the decision tree. In this case, the tree needs to be pruned, while maintaining the accuracy of the tree.

Another advantage of the proposed decision tree is its selection of the proper splitting entropy to dealing with over-fitting. The proposed decision tree solves this problem by creating a decision tree with the selection of the appropriated granular in each node of the tree and the construction of a decision tree with high performance. An attribute in this method is chosen based not solely on information about the node causes redundant attributes at different levels. This solution also prevents redundancies in the decision tree.

The vehicle collision events' database contains two different kinds of attributes: discrete and continuous. The proposed entropy in the FGDT algorithm considers both discrete and continuous attributes, such as spatial measures, which are some of the important factors in vehicle incidents. With the use of the granular computing concept in the FGDT method, the nodes of the decision tree include the granules with high discrimination. This leads to the creation of an efficient decision tree with the most accuracy and a logical running time for large datasets.

6.2 Future Works

This research applied 12 conditional attributes in the decision table to extract the vehicle collision rules for predicting the vehicle collision severity. In future research, more conditional attributes may be used in a huge database; and, the different feature selection algorithm may be applied to have a comparison between different feature selection algorithms and the fuzzy rough set feature selection (FRFS) algorithm.

Moreover, the concept of lower and upper approximations may be applied to create an FGDT, because rough set theory provides a new mathematical method to deal with uncertainty and vagueness in a dataset in the construction of the decision tree.

In this study, the membership functions of vehicle collision event input data were defined based on experimental knowledge. We used the Mamdani model as a decision engine for decision-making. In the future, the Sugeno model can be tested as a fuzzy reasoning engine, as it is a more compact and computationally efficient representation than the Mamdani model. The Sugeno model lends itself to the use of adaptive techniques for constructing fuzzy models. This adaptive model can be used to customize the membership functions, so that the fuzzy system best models the data.

The proposed decision tree has been applied to make rules for prediction in vehicle collision severity classification, often obtaining good results. Different decision-making methods for rule extraction can be used to solve the vehicle collision severity problem; however, one decision tree cannot often obtain a well-supported decision, and a single method presents weaknesses for rule extraction or classification. In future research, we can use the information fusion technique based on different decision tree methods at the rules level, which can integrate the respective strengths from different decision tree methods to improve the prediction accuracy rate. Different decision tree results can be fused to gain the prediction results for reliable decision.

References

Abdalla, I. M., et al. (1997). "An investigation into the relationships between area social characteristics and road accident casualties." Accident Analysis & Prevention **29**(5): 583-593.

Ameratunga, S., et al. (2006). "Road-traffic injuries: confronting disparities to address a global-health problem." The Lancet **367**(9521): 1533-1540.

Anderson, T. K. (2009). "Kernel density estimation and K-means clustering to profile road accident hotspots." Accident Analysis & Prevention **41**(3): 359-364.

Bargiela, A. and W. Pedrycz (2003). Granular computing: an introduction, Springer.

Beynon, M. (2001). "Reducts within the variable precision rough sets model: a further investigation." European journal of operational research **134**(3): 592-605.

Beynon, M. J. (2004). "Stability of continuous value discretisation: an application within rough set theory." International Journal of Approximate Reasoning **35**(1): 29-53.

Blatt, J. and S. M. Furman (1998). "Residence location of drivers involved in fatal crashes." Accident Analysis & Prevention **30**(6): 705-711.

Chang, R. L. and T. Pavlidis (1977). "Fuzzy decision tree algorithms." Systems, Man and Cybernetics, IEEE Transactions on **7**(1): 28-35.

Chen, Y.-l., et al. (2009). "A survey of fuzzy decision tree classifier." Fuzzy Information and Engineering **1**(2): 149-159.

Cummings, P., et al. (1995). "Methodological challenges in injury epidemiology and injury prevention research." Annual review of public health **16**(1): 381-400.

Downarowicz, T., et al. (2015). "Shearer's inequality and Infimum Rule for Shannon entropy and topological entropy." arXiv preprint arXiv:1502.07459.

Dubois, D. and H. Prade (1992). Putting rough sets and fuzzy sets together. Intelligent Decision Support, Springer**:** 203-232.

Dumbaugh, E. and R. Rae (2009). "Safe urban form: revisiting the relationship between community design and traffic safety." Journal of the American Planning Association **75**(3): 309-329.

Edwards, P., et al. (2008). "Deprivation and road safety in London-a report to the London Road Safety Unit."

Elvik, R., et al. (2009). The handbook of road safety measures, Emerald Group Publishing.

Erdik, T. (2014). "Mamdani approach for 2% wave run-up on breakwaters." Journal of Intelligent and Fuzzy Systems **26**(1): 1-9.

Ewing, R., et al. (2003). "Urban sprawl as a risk factor in motor vehicle occupant and pedestrian fatalities." American Journal of Public Health **93**(9): 1541-1545.

Gazis, D. and C. Liu (2003). "Kalman filtering estimation of traffic counts for two network links in tandem." Transportation Research Part B: Methodological **37**(8): 737-745.

Geurts, K., et al. (2003). "Profiling of high-frequency accident locations by use of association rules." Transportation Research Record: Journal of the Transportation Research Board **1840**(1): 123-130.

Gooder, P. and M. Charny (1993). "The difficulties of investigating motor vehicle traffic accident mortality in a district." Public health **107**(3): 177-183.

Guo-jiang, S. (2010). An intelligent hybrid forecasting model for short-term traffic flow. Intelligent Control and Automation (WCICA), 2010 8th World Congress on, IEEE.

Gupta, Y., et al. (2015). "A new fuzzy logic based ranking function for efficient Information Retrieval system." Expert Systems with Applications **42**(3): 1223-1234.

Han, S. W. and J.-Y. Kim (2007). Rough Set-based Decision Tree using the Core Attributes Concept. Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on, IEEE.

Han, S. W. and J. Y. Kim (2008). "A new decision tree algorithm based on rough set theory." International Journal of Innovative Computing, Information and Control **4**(10): 2749-2757.

Ho, T. B., et al. (2003). Documents clustering using tolerance rough set model and its application to information retrieval. Intelligent exploration of the web, Springer**:** 181-196.

Holz-Rau, C. and J. Scheiner (2013). "Geographical Patterns in Road Safety: Literature Review and a Case Study from Germany." EJTIR **13**(2): 99-122.

Hori, K., et al. (1999). Fuzzy C4. 5 for generating fuzzy decision trees and its improvement. Proc. 15 th fuzzy system Symposium.

Ilczuk, G. and A. Wakulicz-Deja (2007). Visualization of rough set decision rules for medical diagnosis systems. Rough sets, fuzzy sets, data mining and granular computing, Springer**:** 371-378.

Janecek, A., et al. (2008). "On the Relationship Between Feature Selection and Classification Accuracy." Journal of Machine Learning Research-Proceedings Track **4**: 90-105.

Janikow, C. Z. (2004). FID4. 1: an Overview. Fuzzy Information, 2004. Processing NAFIPS'04. IEEE Annual Meeting of the, IEEE.

Jassbi, J., et al. (2006). A comparison of mandani and sugeno inference systems for a space fault detection application. Automation Congress, 2006. WAC'06. World, IEEE.

Jensen, R. and Q. Shen (2002). Fuzzy-rough sets for descriptive dimensionality reduction. Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on, IEEE.

Jensen, R. and Q. Shen (2009). "New approaches to fuzzy-rough feature selection." Fuzzy Systems, IEEE Transactions on **17**(4): 824-838.

Jing-ti, H. and G. Yu-jia (2009). Study on Handling Range Inputs Methods on C4. 5 Algorithm. Computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on, IEEE.

Jing, S.-Y. (2015). "Keyword Reduction for Text Categorization using Neighborhood Rough Sets." International Journal of Computer Science Issues (IJCSI) **12**(1): 22.

Joly, M.-F., et al. (1991). "Geographical and socio-ecological variations of traffic accidents among children." Social science & medicine **33**(7): 765-769.

Joshi, M. V., et al. (1998). ScalParC: A new scalable and efficient parallel classification algorithm for mining large datasets. Parallel processing symposium, 1998. IPPS/SPDP 1998. proceedings of the first merged international... and symposium on parallel and distributed processing 1998, IEEE.

Kass, G. V. (1980). "An exploratory technique for investigating large quantities of categorical data." Applied statistics: 119-127.

Kiavarz , H. and X. Wang (2014). Vehicle Accident Severity Rules Mining Using Fuzzy Granular Decision Tree for Inconsistent Data Based Fuzzy Rough Set Feature Selection. 2014 Joint Rough Set Symposium, Granada & Madrid, Spain, Springer.

Kulkarni, A. D. (2001). Computer vision and fuzzy-neural systems, Prentice Hall PTR.

Lavanya, D. and K. U. Rani (2011). "Performance evaluation of decision tree classifiers on medical datasets." International Journal of Computer Applications (0975–8887 (, Volume 26–No. 4, 1-4.

Li, T., et al. (2007). "A rough sets based characteristic relation approach for dynamic attribute generalization in data mining." Knowledge-Based Systems **20**(5): 485-494.

Lin, T. (1997). Granular computing, announcement of the bisc special interest group on granular computing.

Liu, Y., et al. (2013). "A self-adapting fuzzy inference system for the evaluation of agricultural land." Environmental Modelling & Software **40**: 226-234.

Mamdani, E. and S. Assilian (1999). "An experiment in linguistic synthesis with a fuzzy logic controller." International journal of human-computer studies **51**(2): 135-147.

Mehta, M., et al. (1996). SLIQ: A fast scalable classifier for data mining. Advances in Database Technology—EDBT'96, Springer**:** 18-32.

Mendel, J. (2001). "Uncertain rule-based fuzzy inference systems: Introduction and new directions." Prentice-Hall. November **27**(2012): 23.

Minz, S. and R. Jain (2003). Rough set based decision tree model for classification. Data Warehousing and Knowledge Discovery, Springer**:** 172-181.

Minz, S. and R. Jain (2005). "Refining decision tree classifiers using rough set tools." International Journal of Hybrid Intelligent Systems **2**(2): 133-148.

Mousakhani, M., et al. (2013). Administrative corruption: Providing a fuzzy inference system of good governance to combat corruption. Fuzzy Systems (IFSC), 2013 13th Iranian Conference on, IEEE.

Okutani, I. and Y. J. Stephanedes (1984). "Dynamic prediction of traffic volume through Kalman filtering theory." Transportation Research Part B: Methodological **18**(1): 1-11.

Olaru, C. and L. Wehenkel (2003). "A complete fuzzy decision tree technique." Fuzzy sets and systems **138**(2): 221-254.

Organization, W. H. (2013). WHO global status report on road safety 2013: supporting a decade of action, World Health Organization.

Patil, N., et al. (2012). Comparison of C5. 0 & CART classification algorithms using pruning technique. International Journal of Engineering Research and Technology, ESRSA Publications.

Pawlak, Z., et al. (1995). "Rough sets." Communications of the ACM **38**(11): 88-95.

Peden, M., et al. (2004). World report on road traffic injury prevention, World Health Organization Geneva.

Peng, Y. and P. A. Flach (2001). Soft discretization to enhance the continuous decision tree induction. Proceedings of ECML/PKDD-2001 Workshop IDDM-2001, Freiburg, Germany.

Pradhan, B. (2013). "A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS." Computers & Geosciences **51**: 350-365.

Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. Machine learning, Springer**:** 463-482.

Quinlan, J. R. (1986). "Induction of decision trees." Machine learning **1**(1): 81-106.

Quinlan, J. R. (1993). C4. 5: programs for machine learning, Morgan kaufmann 2**(1):** 101-1011.

Roisenberg, M., et al. (2009). "A hybrid fuzzy-probabilistic system for risk analysis in petroleum exploration prospects." Expert Systems with Applications **36**(3): 6282-6294.

Rokach, L. (2008). Data mining with decision trees: theory and applications, World scientific.

Rui-Min, C. and W. Miao (2010). A more efficient classification scheme for ID3. Computer Engineering and Technology (ICCET), 2010 2nd International Conference on, IEEE.

Sahatier, U. P. (1992). "PUTTING ROUGH SETS AND FUZZY SETS TOGETHER." Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory **11**: 201.

Scheiner, J. and C. Holz-Rau (2007). "Travel mode choice: affected by objective or subjective determinants?" Transportation **34**(4): 487-511.

Scutari, M., et al. (2014). "Multiple Quantitative Trait Analysis Using Bayesian Networks." arXiv preprint arXiv:1402.2905.

Setiono, R. and H. Liu (1998). Fragmentation problem and automated feature construction. Tools with Artificial Intelligence, 1998. Proceedings. Tenth IEEE International Conference on, IEEE.

Shanthi, S. and R. G. Ramani (2011). "Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms." International Journal of Computer Applications **35**.

Shen, Q. and R. Jensen (2004). "Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring." Pattern Recognition **37**(7): 1351-1363.

Sikder, I. U. and T. Munakata (2009). "Application of rough set and decision tree for characterization of premonitory factors of low seismic activity." Expert Systems with Applications **36**(1): 102-110.

Stefanowski, J. and A. Tsoukias (2001). "Incomplete information tables and rough classification." Computational Intelligence **17**(3): 545-566.

Straccia, U. (2011). "Reasoning within fuzzy description logics." preprint :1106.0667.

Sugeno, M. (1985). Industrial applications of fuzzy control, Elsevier Science Inc.

Tahriri, F., et al. (2014). "The application of fuzzy Delphi and fuzzy inference system in supplier ranking and selection." Journal of Industrial Engineering International **10**(3): 1-16.

Thakur, D., et al. (2010). Re optimization of ID3 and C4. 5 decision tree. Computer and Communication Technology (ICCCT), 2010 International Conference on, IEEE.

Thangavel, K. and A. Pethalakshmi (2009). "Dimensionality reduction based on rough set theory: A review." Applied Soft Computing **9**(1): 1-12.

Thomson, J. and A. Tolmie (2001). "Road accident involvement of children from ethnic minorities." Road Safety Research Report **19**.

Tokumaru, M. and N. MURANAKA (2010). Kansei impression analysis using fuzzy c4. 5 decision tree. Kansei engineering and emotion research international conference. KEER-2010, Paris.

Vandenbulcke, G., et al. (2009). "Mapping bicycle use and the risk of accidents for commuters who cycle to work in Belgium." Transport Policy **16**(2): 77-87.

Wang, C. and F. Ou (2008). An algorithm for decision tree construction based on rough set theory. Computer Science and Information Technology, 2008. ICCSIT'08. International Conference on, IEEE.

Wang, Y. and H. Liu (2008). A practical intelligent navigation system based on travel speed prediction. 2008 11th International IEEE Conference on Intelligent Transportation Systems.

Yao, J. and Y. Yao (2002). "A Granular Computing Approach to Machine Learning." FSKD **2**: 732-736.

Yao, Y. (2004). Concept Formation and Learning: A Cognitive Informatics Perspectiv. IEEE ICCI.

Yao, Y. and J. Yao (2002). Granular computing as a basis for consistent classification problems. Proceedings of PAKDD, Citeseer.

Yao, Y. Y. (2001). On Modeling data mining with granular computing. Computer Software and Applications Conference, COMPSAC 2001. 25th Annual International, IEEE.

Yao, Z., et al. (2005). R-C4. 5 decision tree model and its applications to health care dataset. Services Systems and Services Management, 2005. Proceedings of ICSSSM'05, International Conference on, IEEE.

Yellasiri, R., et al. (2007). "DECISION TREE INDUCTION USING ROUGH SET THEORY--COMPARATIVE STUDY." Journal of Theoretical & Applied Information Technology **3**(4).

Zadeh (1965). "Fuzzy sets." Journal of Information and Control **8**(3): 338-353.

Zadeh, L. A. (1997). "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic." Fuzzy sets and systems **90**(2): 111-127.

Zhao, Y. and Y. Yao (2007). Interactive classification using a granule network. Cognitive Informatics.(ICCI 2005). Fourth IEEE Conference on, IEEE.

Zhao, Y., et al. (2007). ICS: an interactive classification system. Advances in Artificial Intelligence, Springer**:** 134-145.

Zhou, X., et al. (2008). A new credit scoring method based on rough sets and decision tree. Advances in Knowledge Discovery and Data Mining, Springer**:** 1081-1089.

Zhu, A., et al. (2014). "An expert knowledge-based approach to landslide susceptibility mapping using GIS and fuzzy logic." Geomorphology **214**: 128-138.

Ziarko, W. (2015). Dependency Analysis and Attribute Reduction in the Probabilistic Approach to Rough Sets. Feature Selection for Data and Pattern Recognition, Springer**:** 93-111.

Appendix

The following publication has been produced from the thesis research:

Peer Review Conference Proceeding:

Kiavarz Moghaddam. H., Wang, X.: Vehicle Collision Severity Rules Mining Using Fuzzy Granular Decision Tree. 2014 Joint Rough Set Symposium conference, Granada & Madrid, Spain (2014). PP.280, Springer.