RULE MINING AND CLASSIFICATION OF ROAD TRAFFIC ACCIDENTS USING ADAPTIVE REGRESSION TREES

TIBEBE BESHAH TESEMA, AJITH ABRAHAM AND CRINA GROSAN

Department of Information Science, Addis Ababa University, Ethopia
Email: tibebe.beshah@gmail.com

[1]IITA Professorship Program, Department of Computer Science and Engineering,
Chung-Ang University, Korea,
Email: ajith.abraham@ieee.org

[2]Department of Computer Science, Faculty of Mathematics and Computer Science,
Babes-Bolyai University, Cluj-Napoca, Romania
Email: crina.grosan@ieee.org

Abstract: Road traffic accidents are among the top leading causes of deaths and injuries of various levels. Ethiopia is experiencing highest rate of such accidents resulting in fatalities and various levels of injuries. Addis Ababa, the capital city of Ethiopia, takes the lion's share of the risk having higher number of vehicles and traffic and the cost of these fatalities and injuries has a great impact on the socio-economic development of a society. This research is focused on developing adaptive regression trees to build a decision support system to handle road traffic accident analysis for Addis Ababa city traffic office. The study focused on injury severity levels resulting from an accident using real data obtained from the Addis Ababa traffic office. Empirical results show that the developed models could classify accidents within reasonable accuracy.

1. INTRODUCTION

Analyzing, interpreting and making maximum use of the data is difficult and resource demanding due to the exponential growth of many business, governmental and scientific databases. It is estimated that the amount of data stored in the world's database grows every twenty months at a rate of 100%. This fact shows that we are getting more and more exploded by data/information and yet ravenous for knowledge. Data mining therefore appears as a useful tool to address the need for sifting useful information such as hidden patterns from databases.
In today's world, where the accumulation of data is increasing in an alarming rate, understanding interesting patterns of data is an important issue to be considered to adjust strategies, to make maximum use of it, and find new opportunities. Organizations keeping data on their domain area takes every record as an opportunity in learning facts. But the simple gathering of data is not enough to get maximum knowledge out of it. Thus, for an effective learning, data from many sources must first be gathered and organized in a consistent and useful manner. Data warehousing allows the enterprise to recognize what it has noticed about its domain area. The data must also be analyzed, understood, and turned into actionable information. This is the point where the application of data mining is needed.

Although it is difficult to define precisely and delimit the range and limits of such scientific disciplines, many scholars try to indicate the basic tasks of data mining. In line with this, Hand [Hand et al, 2001] defines Data mining as the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Data mining can also be seen as a combination of tools, techniques and processes in knowledge discovery. In other words, it uses a variety of tools ranging from classical statistical methods to neural networks and other new techniques originating from machine learning and artificial intelligence in improving database promotion and process optimization. Six basic functions or activities of data mining are classified into directed and undirected data mining. Specifically classification, estimation and prediction are directed, where the available data is used to build a model that describes one particular variable of interest in terms of the rest of the available data. Affinity grouping or association rules, clustering, description and visualization on the other hand are undirected data mining where the goal is to establish some relationship among all variables.

Up to recent time, the only analysis made on data to get meaning out of it, is simple statistical manipulation that has no power to show all the necessary information content of a given data. But data mining technology, on the other hand has the greatest potential in identifying various interesting patterns for enabling organizations to control data

resources for strategic planning and decision-making in their domain area.

Traffic control system is one of the various areas, where critical data about the well-being of the society is recorded and kept. Various aspects of a traffic system like vehicle accidents, traffic volumes and concentration are recorded at different levels. In connection to this, injury severities resulted from road traffic accident are one of the areas of concern.

Out of all accidents registered in Ethiopia, Addis Ababa holds about 60% on average. This is partly because the city has great contact through its all gates with different regions every day. In addition to this, of the registered motor vehicles in Ethiopia, the city takes about 77% of it. All these facts reveal that Addis Ababa, having a great deal of concentration of vehicles and traffic, takes the lion's share in car accidents also. Statistical data from the office shows that Addis Ababa is experiencing around 700 accidents per month and the costs of such fatalities and injuries due to traffic accidents have a great impact on various aspects of the society.

In managing and controlling the city's traffic system, the Addis Ababa traffic office is structurally organized under three major departments' namely administrative support, accident investigation, security and control. The office has a total of 216 staff including the department head and traffic police officers.

In connection with this, Mekitew [Mekitew, 2000] tried to propose an automated traffic information system for Addis Ababa Region Traffic Office with the aim of helping the office in information handling. It has been seen that data especially in some regions where the traffic and number of vehicles are huge, does not get enough attention to use it as a base for decision-making. Identifying and knowing a given pattern of data in a given traffic office will help the decision makers in deciding on the specific future activities. Thus, through this research work an attempt has been made to apply data mining tools and techniques in analyzing and determining interesting patterns especially with respect to injury severity, on road accidents data at Addis Ababa Region Traffic Control System. In order to plan and implement effective strategies in reducing the severity of the injury and vehicle accident at large in Ethiopia, there is a need for actionable information which is obviously a result of a research work.

So, in the effort of alleviating the current problem of vehicle accidents, identifying factors leading to accidents through developing a capacity to design and implement an effective traffic information system that can provide timely and accurate traffic information is very crucial. Timely and reliable data collected about vehicle accidents can be used to identify major determinants and risk factors for vehicle accidents, severe injury and fatalities and to take preventive measures so that the effort of improving the quality of life will be enhanced. All the previous researches were conducted by using small proportion of the accumulated data. Besides, in those researches data analysis was conducted by using simple statistical methods.

Since the analysis made by using traditional methods focus on problems with much more manageable number of variables and cases than may be encountered in real world, they have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional databases .

The absence of significant attempt that has been made so far to this level in identifying the major determinants of car accidents and establishing the most important factors influencing the severity of an injury in Addis Ababa region justify the importance of this research. This research work will be groundwork for the effort of reducing vehicle accident in particular and improving the quality of life in general. Moreover it will also be an input for researches in the same area.

The costs of fatalities and injuries due to traffic accidents have a great impact on the society. In recent years, researchers have paid increasing attention to determining factors that significantly affect severity of driver injuries caused by traffic accidents 0. There are several approaches that researchers have employed to study this problem. These include neural network, nesting logic formulation, log-linear model, fuzzy ART maps and so on.

Applying data mining techniques to model traffic accident data records can help to understand the characteristics of drivers' behavior, roadway condition and weather condition that were causally connected with different injury severities. This can help decision makers to formulate better traffic safety control policies. Roh [Roh et al.,1998] illustrated how statistical methods based on directed graphs, constructed over data for the recent period, may be useful in modeling traffic fatalities by comparing models specified using directed graphs to a model, based on out-of-sample forecasts, originally developed by Peltzman [Peltzman, 1975]. The directed graphs model outperformed Peltzman's model in root mean squared forecast error.

Ossenbruggen [Ossenbruggen et al., 20010 used a logistic regression model to identify statistically significant factors that predict the probabilities of crashes and injury crashes aiming at using these

models to perform a risk assessment of a given region. These models were functions of factors that describe a site by its land use activity, roadside design, use of traffic control devices and traffic exposure. Their study illustrated that village sites are less hazardous than residential and shopping sites.

Abdalla [Abdalla et al., 1997] studied the relationship between casualty frequencies and the distance of the accidents from the zones of residence. As might have been anticipated, the casualty frequencies were higher nearer to the zones of residence, possibly due to higher exposure. The study revealed that the casualty rates amongst residents from areas classified as relatively deprived were significantly higher than those from relatively affluent areas.

Miaou [Miaou and Harry, 1993] studied the statistical properties of four regression models: two conventional linear regression models and two Poisson regression models in terms of their ability to model vehicle accidents and highway geometric design relationships. Roadway and truck accident data from the Highway Safety Information System (HSIS) have been employed to illustrate the use and the limitations of these models. It was demonstrated that the conventional linear regression models lack the distributional property to describe adequately random, discrete, nonnegative, and typically sporadic vehicle accident events on the road. The Poisson regression models, on the other hand, possess most of the desirable statistical properties in developing the relationships.

Abdelwahab [Abdelwahab and Abdel-Aty, 2001] studied the 1997 accident data for the Central Florida area. The analysis focused on vehicle accidents that occurred at signalized intersections. The injury severity was divided into three classes: no injury, possible injury and disabling injury. They compared the performance of Multi-layered Perceptron (MLP) and Fuzzy ARTMAP, and found that the MLP classification accuracy is higher than the Fuzzy ARTMAP. Levenberg-Marquardt algorithm was used for the MLP training and achieved 65.6 and 60.4 percent classification accuracy for the training and testing phases, respectively. The Fuzzy ARTMAP achieved a classification accuracy of 56.1 percent.

Yang used neural network approach to detect safer driving patterns that have less chances of causing death and injury when a car crash occurs [Yang et al, 1999]. They performed the Cramer's V Coefficient test to identify significant variables that cause injury to reduce the dimensions of the data. Then, they applied data transformation method with a frequency-based scheme to transform categorical codes into numerical values. They used the Critical Analysis Reporting Environment (CARE) system, which was developed at the University of Alabama, using a Backpropagation (BP) neural network. They used the 1997 Alabama interstate alcohol-related data, and further studied the weights on the trained network to obtain a set of controllable cause variables that are likely causing the injury during a crash. The target variable in their study had two classes: injury and non-injury, in which injury class included fatalities. They found that by controlling a single variable (such as the driving speed, or the light conditions) they potentially could reduce fatalities and injuries by up to 40%.

Sohn [Sohn and Lee, 2003] applied data fusion, ensemble and clustering to improve the accuracy of individual classifiers for two categories of severity (bodily injury and property damage) of road traffic accidents. The individual classifiers used were neural network and decision trees. They applied a clustering algorithm to the dataset to divide it into subsets, and then used each subset of data to train the classifiers. They found that classification based on clustering works better if the variation in observations is relatively large as in Korean road traffic accident data.

Mussone [Mussone et al., 1999] used neural networks to analyze vehicle accident that occurred at intersections in Milan, Italy. They chose feed-forward MLP using BP learning. The model had 10 input nodes for eight variables (day or night, traffic flows circulating in the intersection, number of virtual conflict points, number of real conflict points, type of intersection, accident type, road surface condition, and weather conditions). The output node was called an accident index and was calculated as the ratio between the number of accidents for a given intersection and the number of accidents at the most dangerous intersection. Results showed that the highest accident index for running over of pedestrian occurs at non-signalized intersections at nighttime.

Dia [Dia and Rose, 1997] used real-world data for developing a multi-layered MLP neural network freeway incident detection model. They compared the performance of the neural network model and the incident detection model in operation on Melbourne's freeways. Results showed that neural network model could provide faster and more reliable incident detection over the model that was in operation. They also found that failure to provide speed data at a station could significantly deteriorate model performance within that section of the freeway.

Shankar applied a nested logic formulation for estimating accident severity likelihood conditioned on the occurrence of an accident [Shankar et al.

1996]. They found that there is a greater probability of evident injury or disabling injury/fatality relative to no evident injury if at least one driver did not use a restraint system at the time of the accident.

Kim [Kim et al., 2004] developed a log-linear model to clarify the role of driver characteristics and behaviors in the causal sequence leading to more severe injuries. They found that alcohol or drug use and lack of seat belt use greatly increase the odds of more severe crashes and injuries.

Abdel-Aty used the Fatality Analysis Reporting System (FARS) crash databases covering the period of 1975-2000 to analyze the effect of the increasing number of Light Truck Vehicle (LTV) registrations on fatal angle collision trends in the US [Abdel-Aty and Abdelwahab, 2003]. They investigated the number of annual fatalities that resulted from angle collisions as well as collision configuration (car-car, car-LTV, LTV-car, and LTV-LTV). Time series modeling results showed that fatalities in angle collisions will increase in the next 10 years, and that they are affected by the expected overall increase of the percentage of LTVs in traffic.

Bedard applied a multivariate logistic regression to determine the independent contribution of driver, crash, and vehicle characteristics to drivers' fatality risk [Bedard et al., 2002]. They found that increasing seatbelt use, reducing speed, and reducing the number and severity of driver-side impacts might prevent fatalities. Evanco conducted a multivariate population-based statistical analysis to determine the relationship between fatalities and accident notification times [Evanco, 1999]. The analysis demonstrated that accident notification time is an important determinant of the number of fatalities for accidents on rural roadways.

Ossiander used Poisson regression to analyze the association between the fatal crash rate (fatal crashes per vehicle mile traveled) and the speed limit increase [Ossiander and Cummings]. They found that the speed limit increase was associated with a higher fatal crash rate and more deaths on freeways in Washington State.

Finally, researchers studied the relationship between drivers' age, gender, vehicle mass, impact speed or driving speed measure with fatalities and the results of their work can be found in [Buszeman et al, 1998; Kweon et al., 2003; Martin et al., 2000; Mayhew et al., 2003; Tavris et al., 2001].

The general objective of the research was to investigate the potential applicability of data mining technology in developing a model that can support road traffic accident severity analysis in the effort of preventing and controlling vehicle accident at the city of Addis Ababa.

The paper is organized as follows: Section 2 describes adaptive decision trees. In Section 3 an attempt has been made to review literatures and trends in road transport, traffic system and road safety with reference to Addis Ababa. Injury severity levels resulting from such road traffic accidents were also assessed. Section 4 reports the experiment results of the research. The experiment basically comprises training; building and validation of the models in addition to analysis and interpretation of the results. Some conclusions are provided towards the end.

2. ADAPTIVE REGRESSION TREES

Data mining by itself is a process of finding, understanding and interpreting interesting and useful knowledge. Usually, data mining tasks can be categorized into either prediction or description. Directed data mining, which is a top-down approach, is used when the ultimate result is expected. It is often also called predictive modeling. Undirected data mining, basically used to identify patterns in a given data and let the user to determine the interestingness of it. It is normally a bottom-up approach of describing the data. Different problem types in a data mining process to ultimately solve the business problems are often also called data mining tasks. The most widely used tasks are classification, association, clustering, dependency analysis, prediction, segmentation, description.

A decision tree is composed of hierarchically arranged nodes, growing from the top most nodes called root node to leaf node. So it can be thought as the tree growing upside down, splitting the data at each level to form new nodes. The resulting tree comprises of many nodes connected by branches. Nodes that are at the end of branches are called leaf nodes and play a special role when the tree is used for prediction. That means each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this specific attribute.

Tree-based models are useful for both classification and regression problems. In these problems, there is a set of classification or predictor variables (X_i) and a dependent variable (Y). The X_i variables may be a mixture of nominal and/or ordinal scales (or code intervals of equal-interval scale) and Y may be a quantitative or a qualitative (in other words, nominal or categorical variable) 0; Steinberg and Colla, 1995].

The Classification and Adaptive Regression Trees (CART) methodology is technically known as binary recursive partitioning. The process is binary because parent nodes are always split into exactly two child nodes, and recursive because the process can be repeated by treating each child node as a parent. The key elements of a CART analysis are a set of rules for splitting each node in a tree:

- deciding when a tree is complete
- assigning each terminal node to a class outcome (or predicted value for regression)

CART is the most advanced decision-tree technology for data analysis, pre-processing and predictive modeling. CART is a robust data-analysis tool that automatically searches for important patterns and relationships and quickly uncovers hidden structure even in highly complex data. CART's binary decision trees are more sparing with data and detect more structure before further splitting is impossible or stopped. Splitting is impossible if only one case remains in a particular node, or if all the cases in that node are exact copies of each other (on predictor variables). CART also allows splitting to be stopped for several other reasons, including that a node has too few cases 0.

Once a terminal node is found we must decide how to classify all cases falling within it. One simple criterion is the plurality rule: the group with the greatest representation determines the class assignment. CART goes a step further: because each node has the potential for being a terminal node, a class assignment is made for every node whether it is terminal or not. The rules of class assignment can be modified from simple plurality to account for the costs of making a mistake in classification and to adjust for over – or under – sampling from certain classes.

A common technique among the first generation of tree classifiers was to continue splitting nodes (growing the tree) until some goodness-of-split criterion failed to be met. When the quality of a particular split fell below a certain threshold, the tree was not grown further along that branch. When all branches from the root reached terminal nodes, the tree was considered complete. Once a maximal tree is generated, it examines smaller trees obtained by pruning away branches of the maximal tree. Once the maximal tree is grown and a set of sub-trees is derived from it, CART determines the best tree by testing for error rates or costs. With sufficient data, the simplest method is to divide the sample into learning and test sub-samples. The learning sample is used to grow an overly large tree. The test sample is then used to estimate the rate at which cases are misclassified (possibly adjusted by misclassification costs). The misclassification error rate is calculated for the largest tree and also for every sub-tree.

The best sub-tree is the one with the lowest or near-lowest cost, which may be a relatively small tree. Cross validation is used if data are insufficient for a separate test sample. In the search for patterns in databases it is essential to avoid the trap of over fitting or finding patterns that apply only to the training data. CART's embedded test disciplines ensure that the patterns found will hold up when applied to new data. Further, the testing and selection of the optimal tree are an integral part of the CART algorithm. CART handles missing values in the database by substituting surrogate splitters, which are back-up rules that closely mimic the action of primary splitting rules. The surrogate splitter contains information that is typically similar to what would be found in the primary splitter 0.

3. ROAD TRAFFIC ACCIDENT AND INJURY ANALYSIS

A road traffic accident is defined as any vehicle accident occurring on a public highway. It includes collisions between vehicles and animals, vehicles and pedestrians, or vehicles and fixed obstacles. Single vehicle accidents, which involve a single vehicle, that means without other road user, are also included (Safecarguide, 2004).

At all levels, whether at national or international level, road traffic accidents continue to be a growing problem. In connection with this, according to a World Health organization/World Bank Report, deaths from non-communicable diseases are expected to grow from 28.1 million a year in 1990 to 49.7 million by 2020, which is an increase in absolute number of 77%. Traffic accidents are the main cause of this rise. Road traffic injures are expected to take higher place in the rank order of disease burden in the near future.

The tragedy is more or less similar in Ethiopia, Addis Ababa. The rate of traffic accidents in Addis Ababa goes up together with the increase of motor vehicles and population size. The rise in automobile ownership together with the poor condition of the roads has resulted in the high level of traffic safety and congestion problems.

In Ethiopia, above 1,800 people died while above 7,000 were crippled or injured in 2003. Moreover the death rate is 136 per 10,000 vehicles and Ethiopia is loosing over 400 million birr yearly as a result of road traffic accidents. The share of Addis Ababa city in the total number of accidents was 60 percent in 1989 with annual average traffic accident growth 31.4 percent [Addis Ababa Transport Authority (AATA), 2004]. Nowadays Addis Ababa is

experiencing around 700 accidents per month resulting in various level of injury severity

The study conducted by Mekitew [Mekitew, 2000], revealed that the Institution of Highways and Transportations (IHT) attributes road traffic accidents to numerous factors such as weather, light conditions, faulty design and /or inadequate maintenance of the road infrastructure development, as well as some vehicles with mechanical defects and human error.

As to the cause of road traffic accidents in Ethiopia, the first four leading causes as identified by the Road Transport Authority (RTA) are not respecting speed limit, driver characteristics, not giving priority for pedestrian, and vehicle defects. Especially with respect to the vehicle defect although there is an annual program for technical investigation of vehicles, it is not enough when compared to the magnitude of the problem. And consequently conducting occasional technical investigation have got due attention now days.

4. EXPERIMENT SETUP AND RESULTS

The concept of traffic control system in Addis Ababa is dated back around 1900 with the introduction of motor vehicles. As the case is in any other areas, regulations and laws also govern road traffic. In connection with this, the first cited traffic control rule having 18 articles was in action in 1918. This rule was aimed at facilitating the traffic system, which involves movement of pedestrians, animals and motor vehicles. It was in 1935 that formal and legal driving license become in action and the present driving license issuance rule is formally implemented since 1960. With respect to comprehensiveness, it is the Transport Act numbered 361 /1961, which is more comprehensive in having most of the major regulations commonly used. The road traffic safety regulation number 5/1998 and number 4/2004 are also the recent rules and regulation in use at the City of Addis Ababa.

4.1 Accident Data Analysis at Addis Ababa Traffic Office

The study by Mekitew [Mekitew, 2000] lists the basic data items that should be included in an accident reporting. They are basic accident description, road type, environmental features, driver features, causality details and traffic characteristics related to time and location.

Similarly, according to the discussion with the traffic officers, upon receiving a notification about an accident, Addis Ababa traffic police department assigns an investigator to collect the necessary details about a given accident. Notifications are normally reported by the drivers or any party being involved or having interest on it because the low requires doing so. On site investigation and recording is done with the aim of finding detailed and accurate information as to its cause, determine whether or not there has been violation of the law and ultimately to prevent the re-occurrence of further accidents. But some times as reported by the officers, due to time gap between the accident and the arrival of traffic officers, some details like the severity level and cause of an accident may not be identified effectively.

This accident record is basically used for various purposes in the office and for other stakeholder. National and regional transport offices use the data in directing their focus of attention in decision and policymaking's with regard to road safety. Different health offices and non-governmental organizations working in this area use the data in determining and managing health problem in society.

Recent analysis proved that 81% of the accident all over the county is due to drivers fault and the other is due to vehicle, pedestrian and road faults. The main road safety problems are:

- drivers not respecting pedestrian priority ;
- over speeding;
- unsafe utilization of freight vehicles for passenger transportation;
- poor skill and undisciplined behavior of drivers;
- less engineering effort in road design to consider safety;
- poor vehicle conditions;
- pedestrian not taking proper precautions;
- week traffic law enforcement;
- lack of proper emergency medical services.

Road safety publicity, targeted traffic law enforcement, hazardous location identification, pedestrian awareness, upgrading drivers skill and behavior both technically and with respect to keeping rules should get due consideration.

Wondwossen [Wondwossen, 1999] had also tried to study correlation of car accidents in Addis Ababa. He used chi-square test and logistic regression analysis in identifying correlations of car accidents. The research indicates that in addition to other variables like light condition of the road, 'not giving priority to pedestrians' has the highest association with physical damage.

Taddele and Larson [Taddele and Larson, 1991] studied the occurrence and driver characteristics associated with in Addis Ababa. The purpose of the study was to determine the incidence density of hospital treated motor vehicle injuries and to identify

driver and vehicle characteristics placing them at increased risk of inflecting injuries. And he comes up with the result of finding the incidence density of motor vehicle injuries rate of 279 per 100,000 person-years exposure. In addition to this he point out that drivers involved in a motor vehicle injuries are more likely to be male, young and less experienced. With respect to vehicle characteristics associated with involvement in motor vehicle injuries, elevated odds ratios were found for newer and government owned vehicles and for taxis and buses.

4.2. Data Collection

Collecting, analyzing and understanding the content and structure of the data available is one of the most important tasks that need close attention. With respect to this specific research the sole source of data about an accident is the daily accident report form to be filled and reported by the traffic police officers. It consists of full details about a given accident. Through successive update, the office keeps this data in excel flat file format. This helps a great deal in understanding the data and viewing it from different angle in exploring its potential. This study used data from the Addis Ababa Traffic Office (AATO).

The initial data source for the study contained traffic accident records from September, 1995 up to March 1997 E.C a total number of 25,560 cases. The first attempt of the office to electronically record the accident data in September, 1995 was limited to some attributes like labels of *date and time, accident id , driver's name, driver's age, driver's gender, driver's license status, relation of the driver and vehicle, driver's experience, possession of the vehicle, vehicle type, accident area, accident road name, road segment separation, road direction, road surface type, roadway surface condition, vehicle maneuver, accident type, total vehicles involved, total number of victims, accident victims category, victims profession, victims health condition, pedestrian maneuver, vehicle plate number, cost estimate of the damage* and *injury severity*. Through time specifically starting from September 1995, attempt has been made to incorporate some more attributes like, *driver's educational level, vehicle defect, vehicle age, weather condition, light condition, cause for accident*. The accident record's database AATO is still under revision to incorporate as many attribute as possible so that to create a more comprehensive database.

Because of the unavailability of important attributes in the accident data of the previous years, the data set selected for this specific research covers the time from September 1995 to March 2005, which is a total of 5,207 records. This data was in an excel file format with 36 attributes to describe each record.

4.2.1 Accident Data Set: Data Understanding

A good understanding of the data at hand leads to a better success in achieving the data mining goal. The success criterion for this data mining research is the discovery of accident severity classification rules that would find out and differentiate accidents which are serious to those which are potentially not serious in different levels. Provided that reasonable accident severity classification rules are discovered, the office could device a means to reduce the number of fatal and serious injuries and be able to recognize the level of severity when an accident has occurred. In short, this can help decision makers to formulate better traffic safety control policies.

As indicated above the organization keeps detailed information about an accident through various attributes. Types of information that the office records about an accident includes driver and vehicle characteristics, road and weather conditions, date and time of the accident, type, injury severity and possible causes of such accidents.

The specific attributes by which a given accident can be described are *date and time, accident id , driver's name, vehicle type, driver's age, driver's gender, driver's educational level, driver's license status, relation of the driver and vehicle, driver's experience, possession of the vehicle, vehicle defect, vehicle age, accident area, accident road name, road segment separation, road direction, road surface type, roadway surface condition, light condition, weather condition, vehicle maneuver, accident type, total vehicles involved, total number of victims, accident victims category, victims profession, victims health condition, pedestrian maneuver, vehicle plate number, cost estimate of the damage and cause for accident*. In addition to the input variables mentioned above the output variable for this research that is injury severity is also another attribute of a given accident. The target attribute, *injury severity*, has four classes: *fatal injury, property damage (no injury), serious injury* and *slight injury*.

4.2.2. Defining the Data Mining Function

Each individual accident record in the data set is an input/output pair with each record having an associated output. The output variable, the injury severity, is categorical and as described above, has four classes. A supervised learning algorithm employed maps an input vector to the desired output class. Accurate results of such data analysis could provide crucial information for the road accident prevention policy.

4.2.3. Data Cleaning

A large number of errors are to be expected with any massive data set. That is to say, most of the time, the raw data to be used for data mining process are not clean, they have some errors, and irrelevant attributes which are not necessary for the goal of a data mining research at hand. Data may be missed due to equipment problems, deletion of related records, transcription error while keying the manual daily accident record in an electronic format.

In order to get a relevant output, relevant input should get due consideration. In line with this, an in-depth exploration of the data and frequent consultation with the traffic officers has revealed that a good part of the variables were irrelevant to this specific research. Accordingly, based on the domain experts opinion and the researcher's own observation irrelevant attributes like *name and license level of the driver; possession of the vehicle; name and segment separation of the road; category, profession, health condition, maneuver and total number of victims; relation of the driver and the vehicle; total number of vehicles involved and their plate number; accident day, estimate of the damage cost*, which is a total of 16 were removed. In addition to the removal of irrelevant attributes for the research undertaking, attributes like driver's educational level, accident areas and road direction which has considerable amount of missing values – 27% 19% and 15% respectively – were also removed from the data set with the intent of finding meaningful classification of injury severity. Moreover, attributes like accident date and time, vehicle defect and vehicle maneuver during an accident are removed due to considerable inconsistencies which will affect the output a great deal. Missing values of nominal variables, from the remaining relevant data set, are filled based on the idea of observing neighboring record values. Specifically missing values under nominal attributes like 'sex' and 'road surface type' are filled with the values of the modal class 'M' and 'Asphalt' respectively. On the other hand, missing values for the numerical attributes are filled by using the average value of the corresponding attribute values. Accordingly, missing values of driver age, driver experience and vehicle age are filled by average values of 34, 12 and 11 respectively.

Specifically with respect to this research, the goal being the possible identification of fatal and serious injuries, records with missing values under one or more of the following independent variables namely *'accident type', 'accident cause', 'road condition', 'vehicle type', 'light condition' and 'weather condition'* were excluded in order to avoid compromising the result. This will account for the removal of 520 records which is 9.9% of the total records. Similarly, 29 records which is 0.56% of the total data set with missing value under the dependent variable *injury_severity* were removed. Consequently the size of the dataset was reduced to 4658 records.

4.2.4. Attribute/Variable Selection

Eliminating unwanted attributes, irrelevant for the research goal and introducing new attributes necessary for the target objective was considered in the construction of the final data set. A heuristic approach was followed in the selection of the relevant attributes for the classification task. Accordingly, the selected attributes with their description and data type are presented in Table 1.

Attribute Name	Type	Description
Accident_ID	N	A number to identify a given accident uniquely
Driv_Age	N	Age of the driver
Driv_Sex	T	Gender of the driver
Driv_Exp	N	Driving experience of the driver
Vehic_Age	N	Service year of the vehicle
Vehic_Type	T	The type of the vehicle
Road_Surf_Type	T	The surface type of the road
Road_Cond	T	Road surface condition at the time of accident
Light_Cond	T	Light condition at the time of Accident
Weather_Cond	T	The weather condition at the time of accident
Acci_Type	T	The type of the accident
Acci_Cause	T	Causes for the accident
Injury_Severity	T	The injury severity level due to an accident

Table 1. Selected attributes with their data type and description (N=numerical, T=Text)

4.2.5. Data Transformation and Aggregation

Constructive data preparation operations such as the production of derived attributes and new records or transformed values for existing attributes are the major tasks under this phase.

Accordingly, based on the office's classification of driver's experience, *Driv_Exp_Cat* is derived from the base attribute driver experience to categorize the input values as no experience, between 1 and 2, 3 and

5, 6and 10 and above 10 years. *Driv_Age_Cat* is also derived from driver's age attribute to classify the input values as less than 18, between 18 and 30, 31and 50 and above 50 years. Similarly *Vehic_Age_Cat* is also derived from the attribute vehicle age to classify the input values as less than 1, between 1 and 2, 3 and 5, 6 and 10 and above 10 years. This helped to reduce the cardinality of each attribute to a manageable size so as to make the result easily interpretable. Table 2 summarizes the transformed attributes. When the pre-processing was completed, the final dataset used for modeling had 4,658 records described by 16 attributes (13 base and 3 derived). And with respect to the dependent variable, injury severity, there were 341 (7.32%) records with fatal injury, 402 (8.6303134 %) records with serious injury, 709 (15.22 %) records with slight injury and 3206 (68.82%) records with property damage or no injury. As to the cause of the accidents, the statistics shows that, denying pedestrian's priority, not keeping appropriate distance while driving, driving on the left side and over speeding, ranks from first to fourth respectively. For instance from the existing data set of 4,658 accident records, the above four cause comprised of 21%, 19.13 %, 19.02 %, 16.70% percent respectively.

Attribute name	**Derived Attributes**	**Values**
Driver experience	Driv_Exp_Cat	A - F
Driver age	Driv_Age_Cat	A - D
Vehicle age	Vehic_Age_Cat	A - F

Table 2. Derived attributes with their base attributes and values.

4.3 Experiments

Since the dataset of this experiment has an unbalanced dependent class, injury severity, using a random sample will most likely result the minor class belonging in only one of the partitions. In order to overcome this problem the balanced partitioning option was used. To validate model, a dataset is usually divided into two partitions. One is used for training, the learn partition, and the other is held back for testing, the validation partition. Consequently the dataset was divided into the training set (75%) and validation set (25%). After partitioning the training dataset contained 3493 (75% from each class) records and the validation set contained 1165 (25% from each class) records.

4.3.1. Experiment One

The first experiment is to construct a decision tree model based on the training set which is 75% (i.e.3493) of the data. The variable *Accident-Severity*' was set as dependent variable and all others were set as independent variables.

Feature selection is done based on the contribution the input variables make to the construction of the decision tree. Feature importance is determined by the role of each input variable either as a main splitter or as a surrogate. Surrogate splitters are defined as back-up rules that closely mimic the action of primary splitting rules. Suppose that, in a given model, the algorithm splits data according to variable '*protocol_type*' and if a value for '*protocol_type*' is not available, the algorithm might substitute *'service'* as a good surrogate. Variable importance, for a particular variable is the sum across all nodes in the tree of the improvement scores that the predictor has when it acts as a primary or surrogate (but not competitor) splitter. For example, for node *i*, if the predictor appears as the primary splitter then its contribution towards importance could be given as $i_{importance}$. But if the variable appears as the n^{th} surrogate instead of the primary variable, then the importance becomes $i_{importance} = (p^n) * i_{improvement}$ in which p is the 'surrogate improvement weight' which is a user controlled parameter set between 0 and 1.

The first experiment shows that though each record in the dataset included 13 attributes with the inclusion of the dependent variable '*Accident-Severity*', the decision tree constructed had only used 10 attributes. Attributes namely *Vehic_Age_Cat* and *Driv_Exp_Cat* are considered as statistically insignificant. As depicted in Table 3, 86% of the records were classified correctly. And this accuracy does not necessarily show the adequacy of the model. So, review of variables used should be considered in the subsequent experiments.

Results	
Total records	1165
Correctly predicted	1009
Percentage	86.61%

Table 3. Prediction statistics of the first experiment

4.3.2. Experiment Two

Considering the relative importance of each variable for the classification of the accident severity, assessing the variables that are used for splitting the tree was the next important task done. Accordingly, based on experts opinion the second variable, i.e. *accident_id*, used to split the tree at the node next to the root was not considered as being the most important variable to consider the level of injury severity, hence the next most statistically significant variable was taken. In this experiment, the number of variables used was reduced to nine by excluding *accident_id* which is not significant in this specific research. Thus the tree was splited first by the variable accident type, having 5 distinct values and out of which '*vehicle_ped*' and '*turn over*' contributed much for fatality and serious injuries.

		Predicted			
		Fatal Injury	**Partial damage**	**Serious Injury**	**Slight Injury**
Actual	Fatal Injury	37	31	1	22
	Partial damage	7	744	2	35
	Serious Injury	9	11	8	67
	Slight Injury	12	37	1	141

Table 4. Confusion matrix of injury severity

Predicted					
		Fatal Injury	**Partial damage**	**Serious Injury**	**Slight Injury**
Actual	Fatal Injury	32	31	12	16
	Partial damage	7	776	5	0
	Serious Injury	2	7	39	47
	Slight Injury	3	7	9	172

Table 7. Confusion matrix of injury severity

The next most important variable used for splitting was '*accident cause*' and it was revealed from the decision tree nodes value, '*denying pedestrians priority*' and '*over speeding*' were the most significant cause for such fatality and serious injuries.

Assessment and validation of the result was the next task after building the tree. An evaluation of the results was made on both the training and testing set. The result of the validation of the decision tree built is presented in Table 4 in the form of confusion matrix.

Statistics	
Total records	1165
Correctly predicted	930
Percentage	79.83%

Table 5. Validation results from the second decision tree.

Out of the total records for testing (i.e. 1165), 37, 8, 141 and 744 records were classified correctly in the class of fatal injury, property damage, serious injury and slight injury respectively. On the other hand, 54 records were incorrectly classified as serious injury (1), slight injury (22) or property damage (31) while actually they were supposed to be in the fatal injury class and 44 records were classified incorrectly as fatal injury, serious injury, and slight injury while actually they were in the property damage class. This result reveals that from the total records (i.e. 1165), 930 were classified correctly while the remaining 235 records were classified incorrectly (Table 5). Hence this indicates that records whose class is property damage were classified with a minimum error as compared with the records in the class fatal and serious injury.

Although the performance of this training scheme was not bad in terms of accuracy careful consideration of the variables need more attention to get better model that can provide sound rule with better accuracy. Hence, further tests were conducted by building several decision trees through varying the number and combination of the variables. Some of the results obtained from the experiments conducted are presented in Table 6.

Number of variables used	**Accuracy**
10 (ten)	87.47%
9 (nine)	79.83 %
8 (eight)	84.56%
7(seven)	87.47%
6 (Six)	85.37%

Table 6. Results of the tests conducted

As depicted in Table 6, the 'best' decision tree was obtained by using 7 variables out of 13 variables that were considered for model building.

It is also found that two types of accidents namely *vehicle_ped* and *Turn_over* are contributing a lot to fatalities and serious injuries. Next to the accident type, *Acci_cause* was the next important attribute in splitting the tree. Again, taking into consideration the two important accident types, denying pedestrian priority, driving with alcohol and over speeding are determinant cause of turn over and vehicle-ped accident types.

To illustrate this out of 391 *vehicle_ped* accidents caused by denying pedestrians priority, 208(53%) results in either fatal injury or serious injury. And out of 53 accidents caused by driving with alcohol, 45 (85%) results in either fatality or serious injury. The detail rules generated from this tree are attached as an annex. The confusion matrix for the decision tree constructed in this manner is depicted in Table 7 and the performance is illustrated in Table 8.

Statistics

Total records	1165
Correctly predicted	1019
Percentage	87.47%

Table 8. Validation results from the 'best' decision tree.

Although several attempts were made, it was not possible to obtain a good tree with a sound rule and an accuracy of more than 87%.

4.3.3. Generating Rules from Decision Tree

After successive experiments in building the best decision tree model, the next step was to generate, rules by tracing through the branches up to leafs. A rule is a correlation found between the main variable (dependent) and the others (independent. Some of the rules extracted from the decision tree are selected and presented below.

Generic rules

RULE #1
(Whole Tree)
Injury_Severity = F-Injury 0.0781115879828
Injury_Severity = P-damage 0.676394849785
Injury_Severity = Serious_Injury 0.0815450643777
Injury_Severity = Sli-Injury 0.163948497854

RULE #2
if
Acci_Type = Turn Over
Acci_Cause = Denying pedestrians priority, Driving with alchool, Inappropriate Preceding and Turning or Over Speeding
then
Injury_Severity = F-Injury 0.535714285714
Injury_Severity = P-damage 0.321428571429
Injury_Severity = Serious_Injury 0.0357142857143
Injury_Severity = Sli-Injury 0.107142857143

RULE #3
if
Acci_Type = vehicle_Peds
Acci_Cause = Denying pedestrians priority or Not keeping appropraite distance
then
Injury_Severity = F-Injury 0.155405405405
Injury_Severity = P-damage 0.00675675675676
Injury_Severity = Serious_Injury 0.371621621622
Injury_Severity = Sli-Injury 0.466216216216

RULE #4
if
Acci_Type = vehicle_Peds
Acci_Cause = Driving with alchool
then
Injury_Severity = F-Injury 0.333333333333
Injury_Severity = P-damage 0.0833333333333
Injury_Severity = Serious_Injury 0.375
Injury_Severity = Sli-Injury 0.208333333333

RULE #5
if
Acci_Type = vehicle_Peds
Acci_Cause = Over Speeding
then
Injury_Severity = F-Injury 0.137254901961
Injury_Severity = P-damage 0.21568627451
Injury_Severity = Serious_Injury 0.137254901961
Injury_Severity = Sli-Injury 0.509803921569

RULE #6
if
Acci_Type = vehicle_Peds
Acci_Cause = Denying pedestrians priority or Not keeping appropriate distance
Road_Cond = Dry
then
Injury_Severity = F-Injury 0.112781954887
Injury_Severity = P-damage 0.00751879699248
Injury_Severity = Serious_Injury 0.383458646617
Injury_Severity = Sli-Injury 0.496240601504

RULE #7
if
Acci_Type = vehicle_Peds
Acci_Cause = Denying pedestrians priority or Not keeping appropriate distance
Road_Cond = slippery
then
Injury_Severity = F-Injury 0.533333333333
Injury_Severity = P-damage 0
Injury_Severity = Serious_Injury 0.266666666667
Injury_Severity = Sli-Injury 0.2

From the general rule, it is easy to see that, there is a 7.8 % chance that *Injury_Severity* will be a fatal injury, a 67.63% chance that Injury_Severity will be partial damage, a 8.15% chance that Injury_Severity will be serious injury and a 16.39% chance that Injury_Severity will be slight injury and this reveals that about 33% of the accidents, results in injury of different level half of which is either fatal or serious injury.

In addition to this, accident types involving pedestrians and single vehicle turn over mostly results in either fatal or serious injuries. Denying pedestrian priority and over speeding were also the top most determinant factors in injury severity.

It is also apparent from rule 4 that if accident cause is driving with alcohol and accident type is vehicle_peds there is high probability of an accident resulting in fatalities or injuries. Over speeding is also another important factor especially if associated with vehicle_peds type of accidents.

In general, the rules presented above indicate the possible conditions in which an accident will result in either of the injury severity classes. Moreover, the rules generated have indicated that attributes such as *'accident cause', 'accident type', 'driver age',' road surface type', 'road condition', 'vehicle type'*, and *'light condition'* are found to be important variables for classification of accident severity. They are playing a significant role in all experiments being placed at the higher level of the tree which indicates their statistical significance than other variables like *'sex', 'weather condition'* and *'accident_id'*.

The analysis, which was closely undertaken with domain experts, revealed that the variables that were used were good attributes to classify injury severity into the predefined classes (i.e. fatal, serious, slight, and property damage). Decision of selecting the best decision tree was based on the soundness of the rules generated as well as the number of misclassified records of different levels of injury severity.

Decision makers at the traffic office can easily detect and explain the level of injury severity of a given accident with decision trees because all the reasons for the different levels severity can be traced back on the decision tree. For example, attributes like accident type and accident cause are most important in determining the severity of an accident. That means these decision trees can provide the possible combinations of determinant factors in classifying injuries at a different level of injury severity. But further experiments need to be conducted to assess if the use of a combination of the techniques could yield better results. However, since no analysis technique can replace experience and knowledge of the domain expert consideration of such expertise should get due attention.

As it is apparent from the above discussion, application of data mining technology in road traffic accident analysis or in modeling traffic data records can help to understand driver's behavior, roadway conditions, and weather conditions that were causally connected with the different injury severity. This can in turn help decision makers to formulate better traffic safety control policies.

5. CONCLUSIONS

The objective of this research undertaking was to explore the possible application of data mining technology at Addis Ababa Traffic Office, for developing a classification model. Such a classification model could support the traffic officers at Addis Ababa Traffic Office in making decisions in traffic control activities. Specifically it helps decision makers to understand driver's behavior, road and weather conditions and other related issues causing accidents resulting in fatalities or serious injuries so as to formulate better traffic safety control policies.

Addis Ababa traffic control division has a function of reducing accidents as one of its major activities. In doing so periodic registration of accidents and a small scale analysis has been done so far. But the severity of the case, that is the magnitude of accidents, needs some more advanced techniques in order to reach at measurable and actionable recommendations to help decision makers in strategic planning and decision-making. Moreover, the identification of major determinants of road traffic accident and severity nature and fatality of the injury will help save lives and support the effort of National Police Commission in reducing crime and improving the quality of life at large.

In order to support traffic control system of Addis Ababa City, several models were built by employing decision tree approaches and extracting rules. The best performing decision tree classifier was chosen taking into account the soundness of the rules it

generated and also the number of false negatives it reduced, then its predictive accuracy was evaluated and analyzed. The classification accuracy of the decision tree was tested, and it showed an accuracy of 87.47%.

REFERENCES

Abdalla, I.M., Robert, R., Derek, B. and McGuicagan, 1997, D.R.D. "An investigation into the relationships between area social characteristics and road accident casualties". *Accidents Analysis and Preventions,* 5, pp. 583–593.

Abdel-Aty, M., and Abdelwahab, H. 2003, "Analysis and Prediction of Traffic Fatalities Resulting From Angle Collisions Including the Effect of Vehicles' Configuration and Compatibility". *Accident Analysis and Prevention.*

Abdelwahab, H. T. and Abdel-Aty, M. A. 2001, "Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections". *Transportation Research Record 1746*, Paper No. 01-2234.

Bedard, M., Guyatt, G. H., Stones, M. J., and Hireds, J. P. 2002, "The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities". *Accident analysis and Prevention*, 34, pp. 717-727.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. J. 1984, "Classification and Regression Trees", Chapman and Hall, New York.

Buzeman, D. G., Viano, D. C., and Lovsund, P. 1998, "Car Occupant Safety in Frontal Crashes: A Parameter Study of Vehicle Mass, Impact Speed, and Inherent Vehicle Protection". *Accident Analysis and Prevention*, 30 (6), pp. 713-722.

City Government of Addis Ababa, Transport Authority. Facts about Addis Ababa City Transport. http://www.telecom.net.et/~aata/

Chong M., Abraham A., Paprzycki M. 2004, "Traffic Accident Data Mining Using Machine Learning Paradigms", In *Proceedings of Fourth International Conference on Intelligent Systems Design and Applications (ISDA'04)*, pp. 415-420, Hungary.

Chong M., Abraham A., Paprzycki M. 2004, "Traffic Accident Analysis Using Decision Trees and Neural Networks", In *Proceedings of International Conference on Applied Computing*, (*IADIS*), Nuno Guimarães and Pedro Isaías (Eds.), pp. 39-42, Portugal.

Dia, H., and Rose, G. 1997, "Development and Evaluation of Neural Network Freeway Incident Detection Models Using Field Data". *Transportation Research C*, 5(5), pp. 313-331.

Evanco, W. M.1999, "The Potential Impact of Rural Mayday Systems on Vehicular Crash Fatalities". *Accident Analysis and Prevention*, 31, pp. 455-462.

Hand, D., Mannila, H., and Smyth, P. 2001, 'Principles of Data Mining". The MIT Press.

Kim, K., Nitz, L., Richardson, J., and Li, L. 1995, "Personal and Behavioral Predictors of Automobile Crash and Injury Severity". *Accident Analysis and Prevention*, 27(4), pp. 469-481.

Kweon, Y. J., and Kockelman, D. M. 2003, "Overall Injury Risk to Different Drivers: Combining Exposure, Frequency, and Severity Models". *Accident Analysis and Prevention*, 35, pp. 441-450.

Martin, P. G., Crandall, J. R., and Pilkey, W. D., 2002, "Injury Trends of Passenger Car Drivers In the USA". *Accident Analysis and Prevention*, 32, pp. 541-557.

Mayhew, D. R., Ferguson, S. A., Desmond, K. J., and Simpson, G. M. 2003, "Trends In Fatal Crashes Involving Female Drivers", 1975-1998. *Accident Analysis and Prevention*, 35, pp. 407-415.

Mekitew Mola. 2000, "Traffic Information System: The case of Addis Ababa Traffic Police Department". Master's Thesis. Addis Ababa University. Addis Ababa.

Miaou, S.P. and Harry, L. 1993, "Modeling vehicle accidents and highway geometric design relationships". *Accidents Analysis and Prevention*, 25 (6), pp. 689–709.

Mussone, L., Ferrari, A., and Oneta, M. 1999, "An analysis of urban collisions using an artificial intelligence model". *Accident Analysis and Prevention*, 31, pp. 705-718.

Ossenbruggen, P.J., Pendharkar, J. and Ivan, J. 2001, "Roadway safety in rural and small urbanized areas". *Accidents Analysis and Prevention,* 33 (4), pp. 485–498.

Ossiander, E. M., and Cummings, P. 2002, "Freeway speed limits and Traffic Fatalities in Washington State". *Accident Analysis and Prevention*, 34, pp. 13-18.

Peltzman, S. 1975, "The effects of automobile safety regulation". *Journal of Political Economy* 83, pp. 677–725.

Roh J.W., Bessler D.A. and Gilbert R.F. 1998, **"**Traffic fatalities, Peltzman's model, and directed graphs**".** *Accident Analysis and Prevention*, 31 (1-2), pp. 55-61.

SafeCarGuide, 2004, *International Injury & Fatality Statistics*: http://www.safecarguide.com/exp/statistics/statistics.htm

Shankar, V., Mannering, F., and Barfield, W. 1996,"Statistical Analysis of Accident Severity on Rural Freeways". *Accident Analysis and Prevention*, 28(3), pp.391-401.

Steinberg, D. and Colla, P. L. 1995, "CART: Tree-Structured Nonparametric Data Analysis", San Diego, CA: Salford Systems.

Tavris, D. R., Kuhn, E. M, and Layde, P. M. 2001, "Age and Gender Patterns in Motor Vehicle Crash injuries: Importance of Type of Crash and Occupant Role". *Accident Analysis and Prevention*, 33, pp. 167-172.

Taddele, D. and Larson, C.P . 1991, "The occurrence and driver characteristics associated with motor vehicle injuries in Addis Ababa, Ethiopia". *Journal of Tropical Medicine and Hygiene*, 94, 395-400.

Yang, W.T., Chen, H. C., and Brown, D. B. 1999, "Detecting Safer Driving Patterns By A Neural Network Approach". In *Proceedings of Smart Engineering System Design Neural Network, Evolutionary Programming, Complex Systems and Data Mining*, pp 839-844.

Wondwossen Mulugeta. 1999, "C*orrelates of car traffic accident: the case of Addis Ababa in 1990*". Addis Ababa University, Addis Ababa.

Sohn, S. Y., and Lee, S. H. 2003, "Data Fusion, Ensemble and Clustering to Improve the Classification Accuracy for the Severity of Road Traffic Accidents in Korea". *Safety Science*, 4(1), pp. 1-14.

Authors Biographies

Tibebe Beshah Tesema received Master of Science degree from the Department of Information Science, Addis Ababa University, Ethopia. Tesema's main research ares are data mining, neural networks and machine learning.

Ajith Abraham currently works as a Distinguished Professor under the South Korean Government's Institute of Information Technology Assessment (IITA) Professorship programme at Chung-Ang University, Korea. He is also a visiting researcher of Rovira i Virgili University, Spain and an Adjunct Professor of Jinan University, China and Dalian Maritime University, China. His primary research interests are in computational intelligence with a focus on using evolutionary computation techniques for designing intelligent paradigms. Application areas include several real world knowledge-mining applications like information security, bioinformatics, Web intelligence, energy management, financial modelling, weather analysis, fault monitoring, multi criteria decision-making etc. He has authored/co-authored over 200 research publications in peer reviewed reputed journals, book chapters and conference proceedings of which three have won 'best paper' awards.

He is the Editor of The International Journal of Hybrid Intelligent Systems (IJHIS), IOS Press, Netherlands; Journal of Information Assurance and Security (JIAS), USA; International Journal of Computational Intelligence Research (IJCIR), Neurocomputing Journal, Elsevier Science, The Netherlands; International Journal of Systems Science (IJSS), Taylor & Francis, UK; Journal of Universal Computer Science (J.UCS), Austria; Journal of Information and Knowledge Management, World Scientific, Singapore; Journal of Digital and Information Management (JDIM), Digital Information Research Foundation, India and International Journal of Neural Parallel and Scientific Computations (NPSC), USA. Since 2001, he is actively involved in the Hybrid Intelligent Systems (HIS) and the Intelligent Systems Design and Applications (ISDA) series of annual International conferences. He was also the General Co-Chair of The Fourth IEEE International Workshop on Soft Computing as Transdisciplinary Science and Technology (WSTST05), Japan and the Program Co-Chair of the Inaugural IEEE Conference on Next Generation Web Services Practices, Seoul, Korea. He received PhD degree from Monash University, Australia. More information at: http://ajith.softcomputing.net

Crina Grosan currently works as an Assistant Professor in the Computer Science Department of Babes-Bolyai University, Cluj-Napoca, Romania. Her main research area is in Evolutionary Computation, with a focus on Evolutionary Multiobjective Optimization and applications and Genetic Programming. Crina Grosan authored/co-authored over 50 papers in peer reviewed international journals, proceedings of the international conferences and book chapters. She is co-author of two books in the field of computer science. She proposed few Evolutionary techniques for single and multiobjective optimization, a genetic programming technique for solving symbolic regression problems and so on. Dr. Grosan is the co-editor for a book on Swarm Intelligence for Data Mining, which will be published by Springer Verlag, Germany. She is member of the IEEE (CS), IEEE (NN) and ISGEG. She received her PhD degree from Babes-Bolyai University, Romania.

4. utilised for regression analysis and classification, SVM is a supervised learning model that additionally has linked learning algorithmsâ€”which identify patterns and examine data (Olivo et al., 2011); 5. MCAR can be described as a supervised learning model that assimilates between classification and association rule (Thabtah et al., 2005).Â Rule mining and classification of road traffic accidents using adaptive regression trees. International Journal of Simulation, 6(10-11), pp.80-94. 7. Geurts, K., Wets, G., Brijs, T. and Vanhoof, K., 2003.