Extracting Traffic Safety Knowledge from Historical Accident Data

Andreas Gregoriades*, Andreas Chrystodoulides**

*Cyprus University of Technology, Limassol, Cyprus, andreas.gregoriades@cut.ac.cy
**European University Cyprus, Nicosia, Cyprus

Abstract. This paper presents a method and a tool for analyzing historical traffic accident records using data mining techniques for the extraction of valuable knowledge for traffic safety management. The knowledge is distilled using spatio-temporal analysis of historical accidents records. Raw accident data, obtained from Police records, underwent pre-processing and subsequently integrated with secondary traffic-flow data from a mesoscopic simulation. Clustering analysis was performed with self-organizing maps (SOM) to identify accident black spots on the road network and visualize this on a map. Distilled knowledge is used to develop a prototype mobile application to warn drivers of accident risk in real time.

Keywords. Self-Organizing-Maps, Traffic Accidents Heat-maps, Accident Warnings

1. Introduction

Traffic Accidents are one of the most significant causes of fatalities and injuries worldwide. According to World Health Organization (WHO) Global Status Report on Road Safety 2015 (WHO, 2015) over 3400 people die daily and millions are involved in car accidents and left seriously injured. From a data mining perspective, there is a number of approaches to accident detection and prediction. These can be divided into: predictive and descriptive techniques (Berry et al, 1997). Descriptive techniques looks at historical data for insight, while, predictive use models and forecasting to understand the future. Example descriptive techniques include clustering and association rules. Some descriptive techniques use cluster analysis to divide heterogeneous data into several homogeneous classes or clusters and subsequently find patterns in these classes. Predictive models makes prediction about unknown data values by using the known values. For instance, Classification, Regression and Time series analysis are based on prediction (Aggarwal, 2015). This work presents a combination of descriptive and predictive analysis of historical traffic accidents, occurred between 2004 and 2014
in Nicosia-Cyprus, to distil patterns necessary for the development of a mobile accident warning application to alert drivers of imminent risks. The paper is organized as follows. Firstly a review of the literature is presented followed by an overview of the methodology, data pre-process and application of SOM. The paper concludes with a summary and future directions.

2. Related Work

Despite the fact that statistical models have been widely used to analyze road crashes, certain problems may arise when analyzing datasets with large number of dimensions (Chen, 2002). In such cases data mining is preferred as it allows extracting implicit, previously unknown, and potentially useful information from large amounts of data (Aggarwal et al, 2015). Clustering has been used extensively for market segmentation (Liu, 2012). It is a descriptive data mining technique, used in market segmentation (Smith, 1956) to provide a conceptual view of heterogeneous markets (Liu et al., 2012). Clustering approaches aim to classify data records into different groups. Numerous clustering methods exist and are divided into hierarchical and partitioning techniques i.e. DBSCAN, Expectation Maximization, K-means. The latter however was criticized on their accuracy of detecting clusters when these do not have spherical shape (Tan, 2005). Moreover, these techniques lack appropriate visualization metaphors, hence, make the analysis of their results less attractive. On the contrary SOM provides the analyst with an intuitive visualization that enables the interpretation of its results. Essentially, SOM is a special case of ANN, which can identify patterns and cluster data by identifying common features. SOM produces a low-dimensional representation of the input space of the training data, called a map and belong to the category of unsupervised competitive learning algorithms for which, no human intervention is required.

3. Methodology

The main steps in the data analysis methodology include the integration of accident data with traffic flow data from a traffic simulator as per our previous work (Gregoriades et al, 2013). The next step involves pre-processing the resulting dataset to eliminate outliers and reduce the dataset’s dimensionality as detailed next. Finally cluster analysis is performed using SOM for all combination of accident variables. The distilled knowledge was used to specify the accident prone locations on the road network. These are validated with safety experts. The recognition of the driver’s situation in combination with the environmental and traffic conditions is also required prior to the mobile application inferring the likelihood of accident.
3.1. Data Pre-processing
The original accident dataset contained 21179 records of accidents occurred in Nicosia, Cyprus between 2004 and 2014. An accident record contained 47 variables, each associated with multiple attributes. The variables were grouped in the following categories: environment, infrastructure, driver and vehicle. Pre-processing and data transformation was performed to convert the data in the desired format based on a set of dimensionality reduction rules. These rules have been specified by a traffic safety expert. The study focused on the town of Nicosia, hence the data was selected accordingly from the original dataset.

3.2. Map-matching accidents to Geolocation
Due to unavailability of geospatial coordinates of accidents’ locations (for accidents that occurred before 2007), the accident data had to undergo processing and map-matching onto a GIS system. The original dataset as was obtained from the Police, was plotted on a hardcopy map divided into squares and populated with accident locations. Hence, it had to be converted into an electronic form to enable its processing. Specifically, a variable X in the dataset, encoded the x/y coordinates of the accident on the hard-copy map. To avoid manual entry of each individual accident on the GIS system, accidents were grouped according to the box they belonged based on their XY coordinates. To achieve that, we used Google Earth to geotag the GPS coordinates of each box from the original hard copy version of the map. To do that the two end points of each square were used. Subsequently, the coordinates were exported from Google Earth in a KML format and imported in ArcMap from where it was again exported in an .xml format and subsequently converted in excel format. The resulting number of accidents modelled using this coordinate system was 13327. To enhance accident records with traffic flow information at the time of the accident, ArcMap software was used. Essentially, mapping accident location with the road-link on a simulation model and from there retrieving the traffic-flow for that link at the time of accident.

3.3. Self-Organizing maps
The general idea of a SOM is to take an input matrix NxM of N variables and M occurrences of each variable, and parse it into the SOM topology (usually a two dimensional grid or map). Using a neighborhood function, neurons organize themselves forming clusters on the output SOM topology. In SOM algorithm, the output neurons compete between themselves against the properties of an input vector that describe the variables of the problem. Only one neuron is activated at any given time during a SOM process cycle, the neuron that is most similar to the input vector. Each occurrence of the variables-set (input vectors) is assigned to a cluster. Input vec-
tors that are similar are grouped into clusters on the output SOM topology. To achieve this competition, there are feedback paths between the neurons which in return force neurons to organize themselves. The aim in SOM learning is to cause different parts of the network to respond similarly to certain input patterns.

4. Results

SOM analysis was performed with all combinations of variables in the cleaned and preprocessed dataset, to identify the combination of variables that yield significant clusters. Hence, a permutation algorithm was devised in Matlab to perform SOM analysis with all possible combinations between the dependent variable Accident-Type and the remaining 46 accident related variables, to find which sets of variables yielded significant clusters. Significant clusters where identified based on their Hits and Significance plots results (Figure 1). An example of the clusters that yielded from the analysis is depicted in Figure 2, for “accident-type” and “gender” variables. The “SOM Neighbor Weight Distances” depicts the distances between the neighboring neurons. Grey dots represent the neurons (clusters), while red lines connecting neighboring neurons and the colors that surrounds the red lines represent how similar a neuron is to its neighbor. Dark colors represent large distances between neurons which indicate dissimilarity and lighter colors represent closer distances which indicate similarity. Continuous lines with dark colors (borders) indicate that the network has segmented inputs into groups of clusters where each group has different features. It can be said that the yellow parts represent different clusters and the dark lines represent the division of the clusters. The “SOM Hits plot” indicates how many instances (vectors) of the input data are associated with each neuron (cluster center), as well as the neuron location. Specifically for the Accident Type and Gender variables, SOM created 3 distinct clusters. The ‘SOM Hits plot’ show the various number of hits per cluster for each SOM. Matlab provided an output array list for the cluster ID for each accident. These lists were exported in two datasets and then utilized in excel. These were used as the baseline for the creation of the heat map (Figure 3) by filtering each cluster according to day and time and subsequently importing these into Fusion Tables for the creation of the Maps.
The generated fusion maps were used to visualize the accident prone areas on a map. This were used as a preliminary validation of the results by safety experts, prior to being imported into the mobile applications rules-repository. Specifically, experts examined if the depicted accident prone points, were known black spots or if they could be classified as accident prone based on traffic safety knowledge and variable characteristics at visualized map-points. Figure 2, illustrates an example heat-map of the accident-prone areas in Nicosia, using three variables: Day, Time and Gender. The analyst can select any combination of variables and type of accident from the UI, to visualize the relevant heat-map. The safety expert can zoom-in and inspect in more detail the black spots that are revealed. Only the significant clusters that emerged from the SOM analysis are stored in the repository of heat-maps.

5. Specification of an application for improved driver safety

The main usage of the application is to warn drivers of imminent accident risk on the road network of Nicosia. The application utilized the build-in capabilities of mobile phones to recognize geolocation and travelling behavior (speed, acceleration, deceleration patterns), and accordingly in combination with information regarding: the driver (age, gender, years of driving experience), the time and day of the week, analyses the accident risk and present it to the user in the form of warnings. For the generation of the accident black spots the pre-processed data was analyzed using SOM and the
output for each input set was imported into the tool to create a series of heat maps coordinates for all combination of variables, for the Nicosia network. Geolocation data was used as input, along with accident related information. SOMs that emerged from the analysis of different combination of variables (from the accident dataset), yielded heat-maps, the coordinates of which are utilized by the mobile application. Hence, the tool dynamically selects the results of the most appropriate SOM, depending on driver’s features, entered by the user, and information inferred by the tool, regarding the environmental and traffic conditions. Properties that refer to the environment are dynamically inferred from the situation, such as: day, time and weather conditions. Properties that refer to the traffic conditions can be obtained from the simulation model of the road network of Nicosia.

6. Conclusions

This study demonstrated the use of SOM for the analysis on traffic accident data, to identify black spot for the city of Nicosia/Cyprus. The output of this analysis was used to develop a prototype mobile application to dynamically warn drivers of potential accident-prone locations. Future work aims to fully realize and test the mobile application and integrate its functionality with web-services such as weather service and driving behavior data, to enhance the recognition of the situation. The main limitation of this work is the conversion of not geospatial data into accident geolocations. This limits the accuracy of the methodology. However, a new dataset from the Police, currently being analyzed, has complete accident geolocations which will improve the accuracy of the results significantly.

References


Aggarwal, Charu C. (2015), Data Mining, Springer


Tan et al. (2005) Introduction to data mining. Boston, MA, USA: Addison-Wesley