Neurocomputing 72 (2009) 2902–2908

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A self-organizing feature maps and data mining based decision support system for liability authentications of traffic crashes

Pei Liu *

Department of Transportation & Traffic Engineering & Management, Feng–Chia University, Taiwan

A R T I C L E I N F O

Available online 9 April 2009

Keywords:
Self-organizing feature maps
Data mining
CART
Decision support system
Traffic crashes
Liability authentication

A B S T R A C T

This study develops a decision support tool for liability authentications of two-vehicle crashes based on generated self-organizing feature maps (SOM) and data mining (DM) models. Factors critical to liability attributions commonly identified theoretically and practically were first selected. Both SOM and DM models were then generated for frontal, side, and rear collisions of two-vehicle crashes. Appropriateness of all generated models was evaluated and confirmed. Finally, a decision support tool was developed using active server pages. Although with small data size, the decision support system was considered capable of giving reasonably good liability attributions and references on given cases.

1. Introduction

Liability attribution is closely related to insurance coverage and compensation matters; consequently it is an important issue to all parties involved in traffic crashes. Nonetheless, precise and fair liability authentication is difficult to achieve due to the complex nature of traffic crashes. Meanwhile, liability authentication can be highly influenced by personal belief and legal system structures. Hence, liability authentication of a crash case can be varied from man to man, and from country to country. In United States, a police officer can give ticket to the party that he/she judges violating traffic regulations right at the crash scene. However, police officers in Taiwan are not authorized to determine crash liabilities directly.

In Taiwan, before any agreement is reached, crash cases will normally be sent to the government founded local authentication committees (LAC) for suggestion on liability attribution. A system of five levels of liabilities, viz. full, major, even, minor, and none, is adopted by such committees. If the LAC suggested liability authentication for specific case is not acceptable to any party involved, that case can then be sent to the supreme authentication committee (SAC) for further review. The SAC suggestion is the final conclusion from the authentication committee system. Suggestions from either LAC or SAC play a role as references to judges, and are challengeable by prosecutors and counselors. However, they do have certain credit to general public through years of service.

Currently there are 16 LACs and 1 SAC in Taiwan. As mentioned, LAC and SAC are government founded organizations, yet their committee members, except chairperson, are invited civilian experts in crash related fields, such as law, mechanics, and transportation. These civilian members serve two years per term and may surrender their service anytime on their own will. Hence each committee may have new comers, thus new ways of thinking about authentication, once in a while. As a result, similar cases may have very different authentication results from the same LAC. Considering the consequence of financial and legal burden that involved ones may confront with, an appropriate authentication result is crucial. Hence, that there is evidently a need to construct a reference tool for such committees to give righteous liability authentication suggestions. Our motivation in this study is thus to establish analytical models for analysis of liability attributions of traffic crashes. Meanwhile, a decision support tool capable of providing previously authenticated crash cases similar to the questioned case for liability attribution reference is also attempted.

Statistical methods are the prevalent tools used in traffic safety related studies [2,10]. However, statistical methods do have their limits in dealing with some characters, such as great amount of zeros, of crash events. On the other hand, statistics, machine learning, and pattern recognition techniques are all needed for the current study. Hence, artificial neural networks (ANN) and data mining were selected in this study for model generations. ANN has been proven to be capable of modeling complicated multivariate phenomena. It has been applied to various areas of studies, but seldom in traffic safety studies. On the other hand, data mining is the practice of automatically searching large stores of data for patterns. Application of data mining on traffic safety studies have been attempted lately.

Self-organizing feature maps (SOM) is an unsupervised ANN which provides a topology preserving mapping from the high

* Tel.: +886 4 24524362; fax: +886 4 24520678.
E-mail address: peiliu@fcu.edu.tw

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2008.06.032

dimensional space onto a two-dimensional plane. Such a mapping preserves the relative distance between data points, i.e. points that are near each other in the input space are mapped to nearby map units in the SOM. SOM has the ability to characterize inputs it has never encountered before. A new input can be assimilated with the map unit it is mapped to. Consequently SOM is suitable for clustering analysis. Moreover, SOM can determine appropriate amount of clusters automatically.

Lagus [8] applied SOM algorithm to speed up text retrieval. He concluded that a document map created for interactive exploration of a text collection can be successfully utilized as a clustering in speeding up document retrieval. Ferran et al. [4] used SOM algorithm to cluster 1758 human protein sequences stored in the SwissProt database (release 19.0) into families. They found although network training is time consuming, classification of a new protein in the final ordered map is very fast. Moreover, Roussinov and Hsinchum [12] compared performance of SOM with that of Ward's clustering method. They concluded that their implementation of Ward's clustering is slightly more precise in detecting associations between documents, but that the performances of these techniques in terms of recall of those associations are not statistically different. For crash characterization, it is evidently high dimensional. Meanwhile, we have no idea how many clusters are suitable in crash authentication results. Hence, it was judged that SOM is an appropriate tool for this study.

Kuhnert et al. [7] used classification and regression tree (CART), multivariate adaptive regression splines (MARS), and logistic regression to study severity of motor vehicle injuries. They found that logistic regression gave less accurate results than CART and MARS. Meanwhile, better predictions were obtained when they combined these methods together. Chen [3] applied CART and negative binomial model (NBM) to analyze highway crashes and importance of corresponding factors. He concluded that secondary crashes are prone to occur when AADT greater than 4677 is observed; whereas crash rate will be lower when AADT is less than 2096. He also observed that CART has better accuracy than NBM. Wang [13] also reported that CART is a better tool than multinomial logit model. These studies imply that data mining is also suitable for data clustering and classification.

Decision tree is one of the most popular and powerful tools in data mining. Algorithms such as classification and regression tree, C4.5 and C5.0 are popular algorithms for tree generation. Basically, CART and C4.5/C5.0 induct trees in a top-down recursive divide-and-conquer manner. They both use post-pruning method to obtain final trees too. However, they are different in several ways. In C4.5, attribute with the maximum gain ratio is selected as the splitting attribute. Yet CART adopts the one with smallest Gini index. C4.5 allows more than two branches at each node, while CART allows only two branches. C4.5 prunes branches based on predicted error rate of each node, where CART does it based on entire error rate. Moreover, C4.5 inducts and prunes trees with the same set of data, whereas CART divides data for tree induction and pruning. For the current study, a binary decision tree is sufficient. Meanwhile, a training/validation process is preferred for model validation. Consequently CART is selected for tree induction in this study.

As to the proposed decision support tool, due to the fact that crash cases and their authentication results would not be fully identical to each other, the purpose of the tool is to provide an interface for inquiry and reference only. That is, this passive type decision support tool is intended to retrieve and reuse data only. Hence this application is somewhat similar to the case based reasoning (CBR) algorithm. However, it does not complete the 4-step process, viz. "retrieve", "reuse", "revise, and "retain" of CBR.

2. Analytical framework

2.1. Factor determination and data processing

To construct usable analytical models, the required input factors have to be critical, clearly defined, and easy to collect. In this study, common factors from both statistical approaches and expert questionnaires were selected as model input factors.

Due to the fact that influence factors to traffic crashes adopted by various researchers differ vastly, factors ever appeared in all reviewed literatures were first summarized [1,9,11]. Up to 164 different factors were listed in the original database; whereas most of them are similarly defined. Hence, only 29 frequently and commonly used factors, including 22 nominal and seven continuous ones, stayed on the list after subjective judgment and discussion. Meanwhile, those continuous factors were converted to nominal ones for further analysis. Chi square test and *F* test were performed respectively for each factor to identify its significance. For Chi square test, significance level 0.05 was used for all factors, whereas degree of freedom varied with respect to each factor. As to the *F* test, Wilks' lambda, *F* value, and *P* value were calculated for each factor, meanwhile same significance level was also selected for testing significance. Factors identified significant by both Chi square test and *F* test results were considered as statistically determined critical factors. Table 1 lists test results from Chi square test and *F* test. It can be seen that only 14 factors are statistically significant.

To avoid impractical factor selection from pure statistics, including professional opinion on critical factor determination was attempted. Questionnaires were sent to LAC and SAC members for determination of critical factors practically considered in authentication committee meetings. Fuzzy Delphi method was

Table 1
Selection of critical factors through statistically and expert determined factors.

Factors studied	κ^2 test	*F* test	Determined factors		Critical factors
			Stat.	Exp.	
Right-of-way	✸	✦	☆	✡	◎
Weather	–	–	–	–	–
Driver injury	–	–	–	–	–
Driver fatality	✸	✦	☆	–	–
Perception	✸	✦	☆	✡	◎
Passed center of intersection	✸	✦	☆	–	–
Brightness of sky	–	–	–	–	–
Speeding	✸	✦	☆	✡	◎
Lane changing	✸	✦	☆	✡	◎
Artery or minor road	✸	–	–	✡	–
Signal status	✸	✦	☆	✡	◎
Signal types	✸	–	–	✡	–
Phase of signals	–	–	–	✡	–
Driving direction	–	–	–	–	–
♯ of lanes	✸	–	–	–	–
Maneuver	✸	✦	☆	✡	◎
Irregularity	✸	✦	☆	✡	◎
Vehicle type	✸	✦	☆	–	–
Drunk driving	✸	–	–	✡	–
Mutual position	✸	✦	☆	✡	◎
Prevention measures adopted	✸	–	–	–	–
Roadway pattern	✸	✦	☆	–	–
Driver age	✸	–	–	–	–
Speed limit	–	–	–	✡	–
Perception distance	✸	✦	☆	✡	◎
Driver claimed impact speed	✸	✦	☆	–	–
♯ of injured passengers	–	–	–	–	–
♯ of dead passengers	–	–	–	–	–

Note: ✸(κ^2 test significant), ✦(*F* test significant), ☆(stat. significant), ✡(experts selected), ◎(critical factors).