Proceedings of the 5th WSEAS Int. Conf. on SIMULATION, MODELING AND OPTIMIZATION, Corfu, Greece, August 17-19, 2005 (pp214-218)

Discovering Characteristics of Aberrant Driving Behavior

LOUKAS TSIRONIS, Lecturer,
Department of Production and Management Engineering,
Democritus University of Thrace,
Xanthi 67100
Greece,
http://www.duth.gr/

VASSILIS MOUSTAKIS, Associate Professor
Department of Production Engineering & Management
Technical University of Crete
73100 Chania
GREECE
http://www.tuc.gr

HARRY MAVROPOULOS,
Department of Production Engineering & Management
Technical University of Crete
73100 Chania
GREECE
http://www.tuc.gr

EMMANUEL MARAVELAKIS, Lecturer
Department of Natural Resources Engineering
Technological Educational Institute of Crete
73133 Chania
GREECE
http://www.chania.teicrete.gr

NICHOLAS BILALIS, Associate Professor
Department of Production Engineering & Management
Technical University of Crete
73100 Chania
GREECE
http://www.tuc.gr

Abstract: - : Recent studies have shown that unsafe driver acts can be classified into two distinct categories (i.e. errors and violations) entailing different measures for reducing road traffic accidents [1],[2]. A survey of over 1400 drivers in Greece is reported in which a variety of aberrant driving behaviors were identified. Factor analysis was performed to the data collected and seven groups of violations were found. Further statistical analysis showed correlations between those groups and accident liability. Data mining software SEE5 was then applied to reveal the tendencies of the Greek drivers and the descriptions of "dangerous" drivers. The algorithm traced the violations that are responsible for the risky driving acts and brought out useful, but yet hidden, information.

Key-Words: - ***driver behavior, violations, errors, data minng, SEE5***

1 The Method

A properly formed, two-section questionnaire was distributed to the main cities of Greece, containing general items like drivers age, gender, marital status, etc. at its first section, while the second section consisted of 112 items based on the Driver Behavior Questionaire [3] and the extensions to it, introduced in a similar swedish study [4]. Participants were asked to indicate on a six-point scale (never=1, very seldom=2, rather seldom=3, sometimes=4, often=5, very often=6) how often they committed the behaviour described in each item. More than 1450 questionnaires were completed and collected for further analysis. The analysis performed contained 2 stages. At first a factor analysis was performed to identify the main groups of violations and then a machine learning approach using the SEE5 tool was applied in order to discover the interesting patterns and trends and bring out the hidden information contained in our database.

2 Factor Analysis of the Questionnaire Items

The questionnaire, including 112 items, was submitted to a principal components analysis using oblimin rotation to allow for correlations among factors [5]. The scree plot suggested a seven factor solution. The seven factors found by the analysis are the follows:

1. Mistakes
2. Highway Code Violations
3. Low Alertness
4. Aggressive Violations
5. Inexperience
6. Lack of Consideration
7. Parking Violations

3 Predictors of accident involvement

Hierarchical multiple regression analysis was used to predict accident rates using as independent variables: age, gender, mileage and the seven classes of behavior. The variables that independently and significantly predicted accident involvement were found to be: mileage, gender, age and highway code violations (HCV). At this point let‘s see the violations that the HCV consist of :

1. Exceed speed limit during low traffic (Sign5)
2. Disregard speed limit to follow traffic (Sign6)
3. Forget the speed limit (Sign7)
4. Deliberately exceed speed limit when overtaking (Take15)
5. Crossing solid line when changing lane (Take8)
6. Drive at the speed other drivers do (Guy3)
7. Accelerate at a green / yellow phase (Lit1)
8. Cross on lights that have just turned red (Lit2)
9. Disregard red lights at night (Lit7)

The names in brackets are the code names which were used for each violation.

4 Data Mining Concepts

Data mining is the discovery of interesting, yet hidden, knowledge in very large databases [6]. Corporate databases often contain unknown trends, patterns and relationships among objects (e.g. clients and products) that are of strategic importance to the organization. This knowledge cannot be discovered easily with conventional query tools or statistical packages, because they either lack support for handling very large data sets or expect the user to have some idea of the form of the hidden relationships from the beginning of the search process. Data mining tools in general, apply algorithms to large amounts of data in such a way that the data reveal hidden patterns and relationships and uncover correlations that were previously invisible to workers and the business [7]. Data mining tools help the enterprise understand customer behavior, predict events and expose the linkages between events and trends. It is important to realize that data mining is not so much a new technique as a new way to deal with information. A data mining environment can be realized on many different levels using several different techniques. The basic steps of a data mining project are shown in the following diagram:

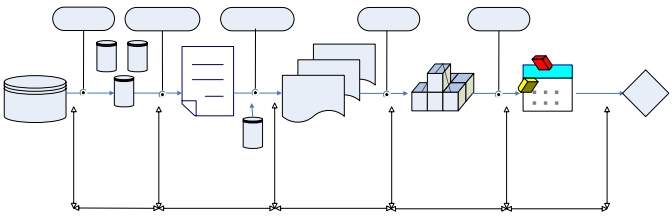

Figure 1: The data mining process

5 The Data Mining Tool – SEE5 / C5.0

Data mining is all about extracting patterns from an organization's stored or warehoused data. These patterns can be used to gain insight into aspects of

the organization's operations, and to predict outcomes for future situations as an aid to decision-making. Patterns often concern the categories to which situations belong. For example, is a loan applicant creditworthy or not? Will a certain segment of the population ignore an incoming mail or respond to it? Will a process give high, medium, or low yield on a batch of raw material? See5 (Windows) and its Unix counterpart C5.0 are sophisticated data mining tools for discovering patterns that delineate categories, assembling them into classifiers, and using them to make predictions. See5/C5.0 has been designed to analyze substantial databases containing thousands to hundreds of thousands of records and tens to hundreds of numeric or nominal fields. To maximize interpretability, See5/C5.0 classifiers are expressed as decision trees or sets of if-then rules, forms that are generally easier to understand than neural networks. The algorithm that the program is using is the same as the previous edition, C 4.5 [8], which is one of the most popular classifiers. It was produced by J.R. Quinlan as an extension of the ID3 tree classifier [9]. Due to the widely acknowledged efficiency of ID3 and C4.5, the results generated by these algorithms have been used in comparative tests in numerous papers and have become characteristic benchmarks for efficiency in the field of machine learning. One of the strongest aspects of the C4.5. algorithm is the information gain – an information based consistency measure used by the method to evaluate partitioning of the examples into disjoint subjects. The measure is defined as follows. Let U denote a set of examples, n the number of different classes of examples in U and p(U,j) the proportion of those examples in U that belong to the j-th class. The information content of the set U is expressed as:

$$Info(U) = -\sum_{j=1}^{n} p(U, j)\log(p(U, j)) \quad (1)$$

6 The Data Mining Tool – SEE5 / C5.0

From the records collected, the attributes that were previously found that have strong correlation with the accident involvement were selected, properly formatted and imported to the SEE5 software in order to extract a ruleset that properly describes our database and is able to predict accident involvement efficiently. The attributes used, their price range and the target attribute which is no other than the accident involvement are shown in table 1 :

ATTRIBUTE	CODE NAME	VALUES
Gender	**Gender**	1, 2
Age	**Age**	<25, 26_35, 36_45, 46_55, >56
Mileage	**Mileage**	0_5, 5_10, 10_20, 20_30, 30_50, >50
Exceed speed limit during low traffic	**Sign5**	1, 2, 3, 4, 5, 6
Disregard speed limit to follow traffic	**Sign6**	1, 2, 3, 4, 5, 6
Forget the speed limit	**Sign7**	1, 2, 3, 4, 5, 6
Deliberately exceed speed limit when overtaking	**Take15**	1, 2, 3, 4, 5, 6
Crossing solid line when changing lane	**Take8**	1, 2, 3, 4, 5, 6
Drive at the speed other drivers do	**Guy3**	1, 2, 3, 4, 5, 6
Accelerate at a green / yellow phase	**Lit1**	1, 2, 3, 4, 5, 6
Cross on lights that have just turned red	**Lit2**	1, 2, 3, 4, 5, 6
Disregard red lights at night	**Lit7**	1, 2, 3, 4, 5, 6
Target Attribute : Accident Involvment	**B6**	**0 = NO 1 = YES**

Table 1 : Accident Involvement Prediction Attributes

7 Results

After importing the data to SEE5 and running the algorithm, 16 rules were extracted that predictit the class of each record :

```
Rule 1: (119, lift 2.4)
age = 46_55
Take8 <= 2
-> class 0 [0.992]
Rule 2: (183/1, lift 2.4)
Take8 <= 1
-> class 0 [0.989]
Rule 3: (47, lift 2.4)
mileage = >50
-> class 0 [0.980]
Rule 4: (126/6, lift 2.3)
mileage = 0_5
-> class 0 [0.945]
Rule 5: (210/28, lift 2.1)
Take15 <= 1
Lit2 <= 2
-> class 0 [0.863]
Rule 6: (62/10, lift 2.0)
mileage = 5_10
Sign7 <= 1
-> class 0 [0.828]
Rule 7: (336/62, lift 2.0)
Take15 <= 1
-> class 0 [0.814]
Rule 8: (282/66, lift 1.8)
Sign5 <= 1
-> class 0 [0.764]
Rule 9: (379, lift 1.7)
```

mileage = 10_20 Take8 > 2 Take15 > 1 -> class 1 [0.997]
Rule 10: (167, lift 1.7) mileage = 10_20 age = 36_45 Take8 > 1 Take15 > 1 -> class 1 [0.994]
Rule 11: (104, lift 1.7) mileage = 10_20 age = 26_35 Take8 > 1 Take15 > 1 -> class 1 [0.991]
Rule 12: (49, lift 1.7) mileage = 10_20 age = <25 Take8 > 1 Take15 > 1 -> class 1 [0.980]
Rule 13: (57/2, lift 1.6) mileage = 30_50 Take8 > 1 Take15 > 1 -> class 1 [0.949]
Rule 14: (114/6, lift 1.6) mileage = 20_30 Sign7 > 1 Take8 > 3 -> class 1 [0.940]
Rule 15: (439/36, lift 1.6) mileage = 10_20 Sign7 > 1 Take8 > 1 -> class 1 [0.916]
Rule 16: (181/15, lift 1.6) mileage = 5_10 Sign7 > 1 Take8 > 1 -> class 1 [0.913]
Rule 17: (135/11, lift 1.6) mileage = 20_30 Sign5 > 1 Take8 > 1 Take15 > 1 -> class 1 [0.912]

Table 2 : Rules of Accident Involment Prediction

Each rule consists of:

- A rule number -- this is quite arbitrary and serves only to identify the rule.
- Statistics (n, lift x) or (n/m, lift x) that summarizes the performance of the rule, where n is the number of training cases covered by the rule and m, if it appears, shows how many of them do not belong to the class predicted by the rule. The lift x is the estimated accuracy of the rule divided by the prior probability of the predicted class.
- One or more conditions that must all be satisfied if the rule is to be applicable.
- A class predicted by the rule.
- A value between 0 and 1 that indicates the confidence with which this prediction is made

This ruleset classifies correctly 1371 of the 1453 records, achieving accuracy of 94.4 %. Specifically, the general performance of the algorithm is shown in table 3:

Class (0)	**Class (1)**	**← Classified as**
585	20	**Class (0)**
62	786	**Class (1)**

Table 3 : Algorithm Performance

8 Conclusions

In this study, data mining is proposed as an operational decision tool for the prediction of accident involvement in Greece. This method, especially conceived for multi-attribute classification problems, suits the problem well. The prediction model has the form of decision rules. The derived decision rules reveal the most relevant attributes that should be considered by the analyser in order to evaluate the risk of accident of a driver. It is important to mention that the rules were derived from a particular data set and as such they represent a generalized description of the experience of it. Following this, these rules cannot be applied uncritically to other databases. If such a need arises, however, a new data set may be created and the same method can be used to analyze it and generate the appropriate rules. Concerning the classification of drivers, the data mining approach produced very satisfactory results. This result is very important because this approach becomes, for the future, a strong alternative tool for the analysis of similar problems.

Finally, compared to other existing methods, this approach offers the following advantages:

- It discovers important facts hidden in data and expresses them in the natural language of decision rules.
- It accepts both quantitative and qualitative attributes

- It can contribute to the minimization of the time and cost of the decision making process as it is an information processing system in real time.
- It offers transparency of classification decisions, allowing for their argumentation.
- It takes into account background knowledge of the decision maker.

References:

[1]Evans L. (1991): *Traffic Safety and the Driver*, Van Nostrand Reihnold, New York

[2]Rumar K. (1985): The role of perceptual and cognitive failures in observed behavior. In Human Behavior and Traffic Safety, Plenum Press, New York

[3]Reason J.T., Manstead A., Stradling S., Baxter J. and Campbell K. (1990): Errors and violations on the road: a real distinction? , Ergonomics, 33, 1315 – 1332

[4]Aberg L. and Rimmo P.A. (1998): Dimensions of aberrant driver behavior, Ergonomics, 41, 39 – 56

[5]Kontogiannis, T., Kossiavelou, Z. and Marmaras, N. (2002). Self-reports of aberrant behaviour on the roads: errors and violations in a sample of Greek drivers. Accident Analysis and Prevention, 34, 381-399.

[6]Adriaans P. and Zantinge D. (1996): Data Mining, Addison-Wessley

[7]Michalski R., Bratko I. & Kubat M. (1999): Machine Learning and Data Mining – Methods and Applications. John Wiley and Sons, NY,USA.

[8]Quinlan J. R. (1993): C4.5: Programs for Machine Learning, Morgan Kaufmann

[9]Quinlan J. R. (1986): Induction of decision trees , Machine Learning, vol. 1