

# Classifying events in transportation and road safety. Some algorithms and models

M. Aron<sup>a,1</sup> and R. Seidowsky<sup>a,2</sup>

<sup>a</sup> INRETS (National Research Institute for Transportation and Safety)  
2 av. du général Malleret-Joinville 94114 ARCUEIL CEDEX, France

**Abstract :** This paper describes a help to the diagnosis, which consists in classifying events (as road accidents) in pre-defined classes. Different sources of knowledge are combined, including the knowledge of the area, statistics, logical rules, questions to the operator. Uncertainty and inaccuracy are processed to some extent. The tool "CRIQUE" (Classifying from Rules, Informatics data, Questions), which has been used for several applications, is presented.

The interest of different algorithms coming from our team or from others are highlighted. A method for the creation of the classes is described. An algorithm for the choice of the next question is reported. An algorithm of reconstruction of missing data is used to propose default values for the user.

**Key Words :** Classifying, artificial Intelligence, operational research, road safety diagnosis.

---

<sup>1</sup>E-mail: [maurice.aron@inrets.fr](mailto:maurice.aron@inrets.fr), URL: [www.inrets.fr/ur/gretia/index.htm](http://www.inrets.fr/ur/gretia/index.htm)

<sup>2</sup>E-mail: [regine.seidowsky@inrets.fr](mailto:regine.seidowsky@inrets.fr), URL: [www.inrets.fr/ur/gretia/index.htm](http://www.inrets.fr/ur/gretia/index.htm)

# 1 Introduction

In the transportation or safety areas, the events (disturbances, incidents, accidents) never occur identically; however they often are quite similar to some of a few “typical” events or classes. Classifying an event in one or several classes is important in the road accident analysis, because these classes, if well-defined, summarize the essence of the accident process.

The actions required by these typical events or classes may have been elaborated and recorded offline. They constitute an historical basis for processing new events of the same class. The aim of this paper is to propose a method for assigning a class to an event, which is a help for the diagnosis.

It is first required to identify the classes, and secondly to classify the current event among those classes. Here this last point is addressed using a new computer tool, “CRIQUE” [11].

The existing diagnosis tools aim at analysing the whole set of accidents on a site, also the relevant geographic information is particularly detailed – Concerto [3] in France, Highway safety Analysis Software in the USA [7] use geographical information systems. Another French tool, SECAR [9] is an expert system for a safety analysis of the intersection, with a very detailed description on the infrastructure. The main feature of our method versus these systems is the taking into account of the uncertainty and the inaccuracy.

The types and features of available knowledge, which are processed by the proposed method, are presented in section 2. The tool CRIQUE is described in section 3, and some results in the diagnosis of road accidents are presented. Some complementary algorithms, issued from the work of different authors in Artificial Intelligence are reported in section 4. The conclusion discusses the interest of the model among other works.

## 2 The type and features of the available knowledge

Several available pieces of knowledge are generally available for a given event as a road accident - computer data files, historical statistics, expert knowledge, additional information available by the operator.

- The computer data files are the easier information to process. But these files never will contain the whole information on an event, they are other sources of information.
- Statistics may apply and complete the missing data or the unknown variables, but a two-fold uncertainty is attached to this type of information. An uncertainty comes from the limited power of the statistics, from its limited applicability. A second type of uncertainty comes from the fact that statistics provide a range of results, each with a probability.
- A lot of common knowledge, describing facts which are generally (but not always) true must be formalized; indeed the computer cannot guess them. Expert knowledge is also primordial.
- In the following, the presence of an available operator/user is assumed. This operator may bring, if asked, some additional information. For instance he can state that the speed (before the accident) was high, on the basis of the statements of witnesses, or on the aspect of the car. Two problems may come from his answers : the precision of this information may be less than asked. An uncertainty may exist (for instance, in police reports, some contradictory pieces of evidence appear).

Thus the classification of an event requires a model combining those types of knowledge. The original method presented here deals with those features, using operational research algorithms, advanced informatics techniques and a conversational interface via the questions which are asked to the operator.

### **3 Presentation of the tool “CRIQUE”**

The classifying method takes into account the different types of knowledge described above in the following way:

- the statistical or historical knowledge is formalized within a rule,
- the general knowledge, which describes the variables and specifies the links between the variables also is formalized within rules,
- the additional information coming from the user is collected through a conversational technique: some questions are selected then asked to the user.

A preliminary model describes the specific problem (here the set of road accidents) in variables and in relationships. This is the “semantic” model. Then the classifying method is developed on a particular event.

#### **3.1 The semantic model**

The semantic model consists in five parts: the variables dictionary, the interface between the data files and the variables, the logical rules, the statistical rules, the pre-defined classes.

##### **3.1.1 The variables dictionary**

The variables are qualitative; their modalities are exclusive and must be enough to describe the events characteristics and the classes.

##### **3.1.2 The interface between the data files and variables**

The computer data are “translated” into the language of variables/modalities.

##### **3.1.3 The logical rules between the variables**

A hierarchy (taxonomy) may exist between some variables. In certain contexts (defined by the value of a few variables), some modalities for an other variable are excluded; these relationships are modelled as “logical rules”. During the process of the current event, some rules are applied, which decreases the number of available modalities until the assignment of one modality by variable.

##### **3.1.4 The approximate or statistical rules, the default values**

Some hypotheses for rebuilding a missing data are formulated within “statistical” or approximate rules which provide the most frequent value of a variable in a given context. For certain couples of variables, a relationship between the variables generally exists but cannot be modelled in a logical rule because of counter-examples; in that case it will be modelled as a statistical rule. The context is modelled by the premises of the rule; the missing modality is the conclusion of the rule. This supposes that (approximate) quantitative values on the “power” of the rule and on the probabilities of the conclusions of the rules are given. The result of the application of the rule is a “default value” proposed to the validation of the user.

##### **3.1.5 The predefined classes, which are defined by a set of modalities of variables**

The classes used in the applications - see 3.3 - were defined by an expert. Another approach, investigated by an other searcher [8], is applicable when there is no available expert: the automatic definition of classes - see 4.1.

### 3.2 The classifying process

The classifying process includes several phases:

- The data file acquisition and analysis, which translates these data into the “variable” language: some modalities are excluded, some are assigned; a few classes are eliminated or selected.
- A series of iterations including the following steps:
  - the selection of the next question to the user, - the algorithm used consists in ranking each question with the number of classes that are related to this question; then the question with the maximum rank is asked. The selection of the relevant questions is an important topic - if too many irrelevant questions are asked, the operator will lose patience or lose his confidence in the model. A more sophisticated algorithm proposed by other searchers [4,5] on this topic is reported further, section 4.2,
  - the proposition of a default value for the corresponding variable, i.e. the most likely and robust after the application of the statistical rules. The algorithm used is close to a shortest path algorithm, in a {Modality, Probability} space where a node is a modality with a given probability, and where an arc links one node-premise to a node-conclusion of a given rule. The shortest path is obtained through the optimisation of two criteria: the minimisation of the uncertainty coming from the application of the rules (which is the “cost” of the link) and the maximisation of the probability of the resulting modality - see also [11]. The sequence of applications of successive rules requires an independence between them, this implicit hypothesis is not always verified. Nevertheless, a great part of the dependencies lies in the “variables-modalities” structure: when a modality is assigned to a variable, the other modalities of the variable are excluded. These dependencies are correctly processed by the method,
  - the response of the user, who either validates the default value, or selects a modality, or answers ‘Unknown’; if his knowledge is not so precise, he can select a few modalities, which induces the elimination of the others, see Figure 1,

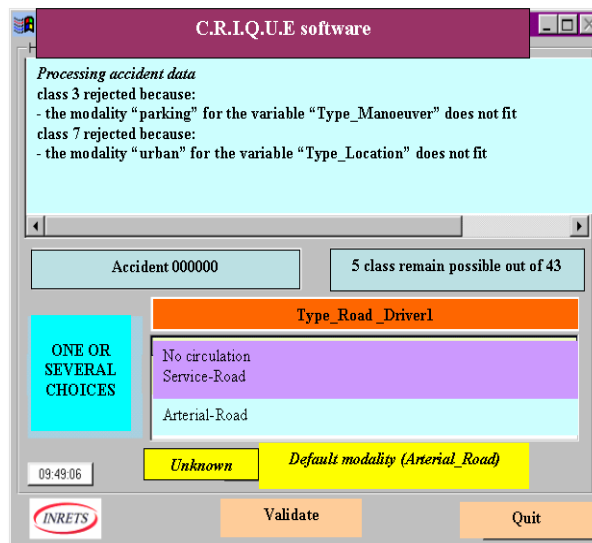


Figure 1: The user interface: Display of a question.

- the activation of the logical rules now applicable after the user response,
  - the elimination or selection of classes after the last step.
- This process is iterated until there is no more eligible class.

### 3.3 The applications of the tool

The tool has been validated on 35 road accidents [11]. The tool was also used to refine the analysis of 66 accidents before and after the road space management of a motorway weaving section [2]. The research is currently in progress on another weaving section.

## 4 Some complementary models

Here are reviewed two methods coming from the Artificial Intelligence area, addressing important topics that we did not worked on:

- the symbolic classification analysis, which aims to automatically create the classes and the criteria for assigning an event to a class. This approach has been applied in road safety diagnosis [8],
- an approach exploiting the taxonomies and causal relations [4,5] in the framework of the (Conversational) Case Based Retrieval [1]. This approach provides an algorithm for the selection of the “next” question.

### 4.1 The analysis of symbolic data

Aurélien De Reyniès [8] develops the analysis of *symbolic* data to take into account the inaccuracy and the uncertainty of the accidents descriptions. As for the traditional data analysis, classes have first to be created, and every (new) event must be assigned to a class.

The classification of “i” events, which are defined by a set of variables “j” requires to define a distance between two events. First the  $\delta_j$  components of this distance are defined for every “j” variable, then these components are aggregated with a function of aggregation.

Then, an iterative algorithm will create the classes, optimizing a function. At last a discriminant analysis points out the assignment criteria of an event to a class; these criteria will be used to classify a new event.

#### 4.1.1 Distance

For the quantitative variables, the inaccuracy is characterised by an interval around the given value. Thus the chosen distance makes it possible to compare two intervals.

For the qualitative variables, the taking into account of the inaccuracy is visible through the definition of “more general” variables, with less detailed modalities which result in taxonomies (that means relations between the modalities of the detailed variable and those of the general variable). When the information on two different events corresponds to different detailed variables and thus cannot be directly compared, the comparison (which enables to calculate the distance between both events) is possible on a more general variable, if it was previously defined – for instance the presence of the “fatigue” is not measured, but it can be supposed during the journeys by night, or for the very aged persons; in that way the comparison between a “very aged person” and a “driver by night” will concern the “fatigue” variable.

The basic data sometimes is uncertain, made of a “histogram”. For a given event, there is no determined modality, but probabilities for different modalities of the variable. The chosen distance should make it possible to compare two histograms, especially to take into account some contradictory aspects in the police reports.

Note : in CRIQUE, the operator can select several modalities if he doesn’t know the precise modality, but no credibility percentage (even if it is available) is assigned to the various selected modalities; the distance between an event and a class takes this inaccuracy into account in a very simplified way (a unit is added for every ambiguous variable).

### 4.1.2 The function of aggregation

Let's call "p" the number of the variables, "j" their index  $j=1 \dots p$ .

Four main functions of aggregation are used in the analysis of symbolic data.

The aggregated distance between event "i" defined by the variables  $\{x_j^i\}$  and event or class "l" defined by the variables  $\{x_j^l\}$ , taking into account  $\alpha_j$  weightings, will be:

$$\text{A weighted average of order 1 as the index of de Minkowsky: } \sum_{j=1}^p \alpha_j \cdot \delta_j(x_j^i, x_j^l) \quad (1)$$

$$\text{A weighted average of order 2 } \left[ \sum_{j=1}^p \alpha_j \left[ \delta_j(x_j^i, x_j^l) \right]^2 \right]^{\frac{1}{2}} \quad (2)$$

$$\text{Or the maximum } \text{Max}_{j=1..p} \left\{ \alpha_j \cdot \delta_j(x_j^i, x_j^l) \right\} \quad (3)$$

$$\text{Or the minimum } \text{Min}_{j=1..p} \left\{ \alpha_j \cdot \delta_j(x_j^i, x_j^l) \right\} \quad (4)$$

### 4.1.3 The algorithm of classification

The algorithm of *symbolic* classification used in [8] is similar to the traditional algorithms of classification – "kmeans", "dynamic kernel" – which minimize a fonction, the sum of the intra-classes variances [6].

### 4.1.4 The classification from the results of the symbolic discriminant analysis

After having identified the classes – which leads to assign every event to a class -, the obtained classes then have to be interpreted in order to be able to classify a new event. This is obtained through the symbolic discriminant analysis, which identifies, for every class a number of criteria on the variables: an event belongs to the class if the criteria are satisfied; these criteria are used to classify a (new) event. However, some work is done on these criteria. First the criteria which correspond to too small a number of events are abandoned. Then, among the set of the parts of these criteria, a sub-set of criteria is selected, which is the one producing the smallest number of wrongly classified events. This method is qualified as "automatic labelling", because the classes and the criteria have been automatically obtained.

The analysis of the symbolic data enables correctly to process the uncertainty and inaccuracy ; however the main problem remains, which is to choose the distance, which implies a previous knowledge of the significant variables and of their weighting. We have not use this approach because an expert was available for defining the classes: the defined classes through such an algorithm are not yet as relevant as the defined classes by an expert.

## 4.2 The selection of the relevant questions considering the taxonomies

Gupta [4] completes the information asking the user a series of questions. To improve the relevance of the questions and to reduce their number, he introduces the Taxonomic Conversational Case Retrieval approach. This approach optimises the choice of the next question.

These works are not directly linked with the field of classification but with the field of the Case Base Retrieval [1]. However, the choice of the best “case” being analogous to the choice of the best “class”, the algorithm which the author proposes for the choice of the next question very well could be used for the classification.

The set of the variables (thus the corresponding questions) is divided into several independent trees, every tree corresponding to the taxonomy of a factor, the factor being a sub-set of variables linked with each other and *independent* from the variables outside from the tree. In every tree the question at the root is more general and logically should be asked before the other questions from the same tree. The reply then makes it possible to select a branch of the tree and thus to eliminate the other branches and the corresponding questions.

To obtain this result, the author defines a function of similarity between every question and every case. Every case is described by replies to some characteristic questions. The similarity between the case and each of its characteristic questions is “1”. The similarity between the case and any other question depends on the presence (in the tree) of a link (ancestor or descendant) between the question and one of the characteristic questions of the case. The similarity of an “ancestor” question of the characteristic question also is equal to 1. The similarity of a “descendant” of the characteristic question is less than 1; it is inversely proportional to the number of arcs between both questions. Then these similarities are aggregated (on all of the cases and taxonomies) into a “score”; the question of maximum score is selected. This algorithm leads to give priority to both general and linked with numerous cases questions.

In a further paper, Gupta [5] goes back over the hypothesis of the independence between the taxonomies. For two given taxonomies, he authorises a causal relation between the root nodes. The basic hypothesis being weaker, the addressed problem is similar to the actual problems. The root node score from the dependant taxonomy is propagated to the previous taxonomy, which increases the score of this previous taxonomy.

The present programmed algorithm in CRIQUE doesn’t give priority to the general questions because these don’t play any direct role in defining the classes. So it would be interesting to test the algorithm which is presented in the above section.

## 5 Conclusion

This paper proposes a model which processes different types of available knowledge, even if inaccurate or uncertain, in order to diagnose an event - to classify it into one or several classes. In our model the classes are pre-defined - the class definition was not investigated, but an overview on a work addressing the automatic class definition, using *symbolic* data analysis [8], is given.

The structure of the preliminary model of the phenomenon describing the variables and their relationships are presented. The class assignment is performed with the model CRIQUE, which allows to combine computer data files, logical rules, “statistical” rules, and the answers brought by the operator. The logical rules, specifying obvious facts or expert knowledge, avoid asking boring questions for the computer operator. The proposition of “default” values, based on the statistical rules, may help the operator/user. The choice of the “next” question is very simple in CRIQUE, an overview on a more sophisticated algorithm is reported: this algorithm decreases the dialog length and improves its quality, by focusing first on more general questions [4,5].

Several applications of this method have been performed to classify the accidents within the French accident data files (computer data) and police reports. Other potential applications

might appear, in the problems where the information is shared between general knowledge (which is the base of logical or approximate rules), computer data, and a user.

## References

- [1] Aha, D.W., T. Maney & L. A. Breslow (1998), Supporting Dialogue Inferencing in Conversational Case-Based Reasoning, EWCBR-98.
- [2] Cohen, S., M. Aron & R. Seidowsky (2003) Safety ITS Tools For Roadspace Management. An example on the Paris Motorway Network. 10 th World Congress on ITS, Ertico Ed, Madrid.
- [3] Concerto (1999), Outil de connaissance de l'accidentologie (Tool of knowledge for accidentology). In: Software series CERTU ed.
- [4] Gupta, K.M. (2001), Taxonomic Conversational Case-Based Reasoning (2001).Proc. of the fourth ICCBR, D.W Aha & I. Watson (eds.). Springer Verlag, Germany, pp219-233.
- [5] Gupta, K.M., D.W Aha & N. Sandhu (2002), Exploiting Taxonomic and Causal Relations in Conversational Case-Retrieval, 2002 European Conference on Case-Based Reasoning.
- [6] MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5<sup>th</sup> Berkeley Symposium on Mathematical Statistics and Probability. *Vol 1*, Statistics, University of California Press.
- [7] Mirsakov, A. (2002), Highway safety Analysis Software. In: Proceedings of "e-safety", ITS solutions for Safety and Security in Intelligent Transport, ERTICO ed. Lyon.
- [8] De Reyniès, A. (2002), Classification et discrimination en analyse de données symboliques, thèse, [Université Paris IX-Dauphine.
- [9] SECAR, (1997). Etudes de sécurité concernant les carrefours plans en rase campagne. French DoT, Direction des Routes.
- [10]Seidowsky, R., M. Aron & G. Scemama (1999), Assigning Urban Accidents to predefined Scenarios, 2<sup>nd</sup> European Road Research Conferences, Proceedings, Session N° 20 "Urban Safety", Bruxelles.
- [11]Seidowsky, R. & M. Aron (2003), Using operational research and advanced informatics for classifying: application for road accidents. Proceedings of IFAC CTS 2003, session "traffic safety", Tokyo.