# MERGING DATA ANALYSIS AND SYMBOLIC CALCULATION INTO A DIAGNOSTIC SYSTEM FOR NATURAL HAZARDS

Robert Bolognesi, Othmar Buser
Swiss Federal Institute for Snow and Avalanche Research (SFISAR)
7260 Weissfluhjoch / Davos, Switzerland
Voice : (41) 81 417 01 53
Fax : (41) 81 417 01 10
E-mail : bolo@slf.ch

**KEYWORDS** : diagnostic system, natural hazard, hybrid model.

## ABSTRACT

Much theoretical work has been done in the building of diagnostic systems to protect oneself from natural calamities. The two classical approaches to solve this typical non-algorithmic problem are case-based reasoning and knowledge-based reasoning. Each of these approaches has already given good results, but also shown its limitations. Following the idea that they may be synergetic, a research project with the goal of merging a data analysis system and an expert system into one single "hybrid system" was initialised at the Swiss Federal Institute for Snow and Avalanche Research (SFISAR) in 1991.

The first step of the project consisted in designing the theoretical model, by defining the different calculation processes and the resolution sequence. This begins with the determination by rules of the parameters of a k-nearest cases selection procedure. It proceeds with the extraction of the nearest solved cases from the database and concludes with an inference session giving the diagnosis from the data of the case to be analysed and the data of the nearest solved cases.

The second step of the project involved the implementation of an operational computer application usable in real time by practitioners. This tool is composed of a file manager including data control procedures on the one hand, and the diagnostic system on the other hand. It has been designed to run on a PC to be easily used in actual conditions and without high costs. An evaluative phase for the estimation of diverse natural hazards is now in process.

At the present time, the next step consists in adding optimized machine learning procedures to the model in the hope of giving it the capability of improving its performance by itself.

## INTRODUCTION

Forecasting an event (an avalanche, a wildfire, etc.) is always the first problem of natural hazard prevention.

Prediction is often based on previous experience, and these experiences are available to us in many different ways. The simplest form is a recording of measured values, observed facts and, of course, of the pertinent event. From them if we are lucky or a genius we may find a close relationship between these values and the event and call it a physical law (example: the hydrogen absorption lines found by Balmer, the Balmer series). In order to find such relations among variables, of which one is normally called the event, mathematical tools are available in statistical software packages. Assuming normal (Gaussian) distributions of the variables all goes well with classical parametric statistics and we may calculate correlations, find clusters and factors. From these results we may infer the chance of the occurrence of the event, given the values of the relevant variables. The advantage of such methods is the ease with which we can perform the calculations for new cases, which is a consequence of the mathematical formulation of the distributions and the very limited number of parameters of the statistic (mean value, standard deviation). It is even possible to forecast an event that has never occured so far.

If there is a good number of observations available we may use non-parametric statistics which is essentially a case-based reasoning : if all the circumstances causing the event are the same as already experienced at least once in the past, then the event must be the same as in the past. The logical issue of this idea leads to the method called "nearest neigbour" or "k-nearest neighbours". The advantage is : no assumption about distributions and their parameters concerning the variables. Both methods mentioned above need variables, measured values and observed facts, on which depends the event. However this is not the only way we store our experience. There are rules, knowledge. These ideas lead to expert systems and artifical intelligence models. Data of the past is "compiled", keeping the essential (for the expert). Details will be lost (such as date of occurrence) : even worse the essential may get lost if it is not recognized by the expert. Therefore the experience of more than one expert is needed.

As a simple conclusion we may say : the first two models need data of the past, the last mentioned models need rules and knowledge of the present (although having its roots in the past).

The usual question is what model is the best ? One answer is "try all of them" (Weiss et al, p.145-176). In fact, each resolution method has its own advantages and limits. According to the idea that they may be synergetic, we have decided to merge two different models into one single system. One uses the nearest neighbours method whereas the other one exploits classical artificial intelligence techniques. In fact, this merging simulates a human being reasoning using theoretical knowledge and practical experience at the same time. In every day life, such reasoning is very frequent and very efficient. One of the best illustration is given by a child learning to read : he "decodes" new or unusual words with the help of phonetic rules whereas he directly recognizes the words he is used to see. The diagnostic model we present here is designed to work like this child. Thus, two different aspects have been treated : data analysis and symbolic calculation.

The first section of this paper presents the theoretical approach to the problem and the second one describes one practical application. The conclusion explores the different fields of utilization and the development perspectives.

## 1. THEORETICAL MODEL

### Data Analysis

In the nearest neighbours model we need some prescriptions of how to find them. There is the problem of the relevant variables. In most practical situations we have to adapt the method to available data. Next problem is the strength of influence of the variables on the event. For instance, it is well known that the amount of fresh snow has a great effect on the occurrence of avalanches. Other variables have minor influence. Therefore a weighting matrix is introduced. There are statistical means for calculating the weights, one of which is very elegant since it leads at the same time to the definition of the statistical Mahalanobis distance (McClung, 1994). However this excludes the introducing of expert knowledge. Experts put different weights on the variables, depending on their value. Mathematically speaking, they use nonlinear relations between variables and event. In order to account for that, different weighting matrices will be used. In our present model, we can consider different kinds of situations and attribute to each of them a specific weighting matrix. But we are still faced with the problem of calculating the distance needed to select the nearest neighbours of the case to be analysed. There are several definitions : the Mahalanobis distance mentioned above, the Euclidian distance commonly known and others. Using the Euclidian distance, the components are assumed to be orthogonal, that is the variables describing cases should be independent. Then the distance between two vectors in the space of cases is :

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} [\alpha_k . \beta_k . (x_{i_k} - x_{j_k})]^2}$$

with :

$x_i$, $x_j$ : vectors of space
$\alpha$ : weighting factor (real)
$\beta$ : normalization factor (real)
$n$: dimension of space

We could introduce a weighting factor $\alpha' = \alpha . \beta$. However $\alpha$ and $\beta$ have a different meaning : $\alpha$ is considered the true weight, whereas $\beta$ is chosen such that the range of the numerical values is about the same for each variable, that is each one is given the same weight for a start. Thus the expert's weight ($\alpha$) directly shows the importance he attributes to a variable.

### Symbolic Calculation

The inference phase simulates deductive reasoning. We have chosen to represent the experts' knowledge with the help of production rules because it seems easier to formulate their experience according to this formalism. But natural hazards are dependent on such complex phenomena that many of the rules used by experts are empirical and uncertain. Moreover, they have sometimes to work with doubtful values coming from field measurements. So we had to take into account these uncertainties when modeling reasoning.

The first question is : how to represent uncertain knowledge ? For our problem we can admit that the facts we manipulate are observable and that the notion of occurrence frequency has a meaning. With this assumption the mathematical theory of probabilities brings answers : we have chosen to qualify each fact used by the model by a probability coefficient. Thus, for the system each fact is an expression the formalism of which is :

fact = record
v : real (value)
p : real Î [0,1] (probability coefficient)

This allows to deal with uncertain measurements or deductions. But there is still another kind of uncertainty. Experts often make statements such as : "if the value of... is *around*... then the value of... *may be*...". This indicates that they make inferences from uncertain facts by using uncertain rules ! To represent this process we have also quantified the likelihood of each rule by a probability coefficient. This represents the supposed frequency of validity of the rule.
In order to simulate the reasoning more easily, we have adhered to one condition when writing knowledge basis : when several facts are needed to trigger a rule, they must be independent. According to this prescription, the process for producing uncertain deductions is :

$$F1_{(P1)} \xRightarrow{R_{(P)}} F2_{(P1.P)}$$

$$F1_{(P1)} \wedge F2_{(P2)} \xRightarrow{R_{(P)}} F3_{(P1.P2.P)}$$

$$F1_{(P1)} \vee F2_{(P2)} \xRightarrow{R_{(P)}} F3_{(P.(P1+P2-P1.P2)}$$

with :

Fi : fact
Pi : probability of fact i
R : rule
P : likelihood of R

Generally we try to decompose knowledge as much as possible in order to write simple rules. For example, if we have :

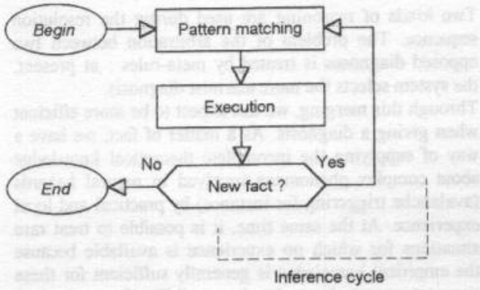
$$A \vee (B \wedge C) \Rightarrow D$$

then we write :

$$B \wedge C \Rightarrow E$$
$$A \vee E \Rightarrow D$$

Using fact E is not only simplification for calculation. It also facilitates releases of knowledge bases. The latter one may consist of thousands of facts and rules, and the inference engine has to exploit all these fragments of knowledge (got from experts with the help of Protocol Analysis method) to build the diagnosis. We have chosen to make it work by forward chaining until the knowledge basis is stable.

The principal process is :




The session stops when the inference engine cannot produce new facts.
The main characteristics of our inference engine are : forward chaining (only), order 0+, and irrevocable strategy (no backtracking).

**Merging Data Analysis And Symbolic Calculation**

In the course of a resolution sequence, the model uses rules (that is to say the knowledge basis) to drive the data analysis procedure and then the results of the latter to complete the knowledge basis. This is the reason why we speak about "merging". The following table gives an overview of the process by showing the input and output of each main step of the resolution sequence.

| Step | Input / Process / Output |
|---|---|
| 1 | Input<br>**-Data of the problem**<br>-Meta-rules<br>Process<br>Symbolic calculation<br>Output<br>Data analysis parameters<br>(weighting coefficients) |
| 2 | Input<br>-Data of the problem<br>-Output Step 1<br>-Database (solved cases)<br>Process<br>Data analysis (k-NN method)<br>Output<br>Solutions for analogous problems |
| 3 | Input<br>-Data of the problem<br>-Output Step 2<br>-Expert diagnostic rules<br>Process<br>Symbolic calculation<br>Output<br>**Solution of the problem** |

Two kinds of reasoning are used during the resolution sequence. The problem of the arbitration between two opposed diagnoses is treated by meta-rules : at present, the system selects the more alarmist diagnosis.
Through this merging, we can expect to be more efficient when giving a diagnosis. As a matter of fact, we have a way of supplying the incomplete theoretical knowledge about complex phenomena involved in natural hazards (avalanche triggering for instance) by practical and local experience. At the same time, it is possible to treat rare situations for which no experience is available because the empirical knowledge is generally sufficient for these cases (at any rate for avalanche or wildfire forecasting).

## 2. PRACTICAL APPLICATION

The model has been implemented so as to be widely tested in actual situations. The software is called NXLOG because it uses principles and some functions from two systems previously designed for avalanche forecasting : NXD (Buser, 1989) and AVALOG (Bolognesi, 1993). It is easily configurable for different similar diagnostic problems. So we may use it to help in forecasting various natural hasards. Of course this is possible only if strong expert knowledge and numerous actual cases described by relevant variables are available.
The system is designed to run on a personal computer. It proposes to the user two functions : data management and events forecasting.

### Data Management

The system permits the storing of cases and other data required. Let us remember that a "case" should be a record of data reporting or describing the event (an avalanche, a fire...) as well as its circumstances : a "case" associates the event with its assumed causes.
Example :

< precipitation, temperature, inclination..., size of avalanche,... >

circumstances *(causes)* — event *(consequences)*

In practice, it is more efficient to store the fields of such a record in different files according to the frequency which they are observed. So the system does not work with one file of cases, but with several files :
-one gathers cause variables (e.g. : precipitation, temperature...)
-one gathers cause constants (e.g. : inclination...)
-one gathers consequence variables (e.g. : size of avalanche...)

As each record of these files has at least one common identifier field (date and location of the observation in our application) there is no problem in reconstituting the cases.
When using the data of the past to build a diagnosis with the k-nearest neighbours method, one assumes that these data are quite reliable. But we know that many mistakes may be made when entering data in a computer... So we have implemented many data checks which inform the user if the entry is outside its possible range or outside its usual range. These different ranges are parameters : this makes the system more adaptable.

### Events Forecasting

NXLOG gives the user the probability of occurrence of the event to be forecast. As natural hazards are generally spatial events depending on geographical parameters, it is important to have different diagnoses according to the different locations involved. The system is able to apply the model to all the different locations requiring control. For instance, in quantifying avalanche hazards, the system determines the probability of an avalanche for each described slope of the considered mountain area.
As additional informations, the system also displays some intermediate results like the 3 nearest cases (circumstances and events).

## CONCLUSION

Even if our researches have theoretical aspects, our ultimate goal is to develop diagnostic systems usable in practice. Therefore, we have established many contacts between the Institute and the practitioners in order to design the products needed and to test them in actual situations. At present more than 15 European safety services work successfully with our systems for avalanche hazard diagnosis.
Although the model has been developed for avalanche forecasting, it may have become evident to the reader that it should work for any event which depends on measurable and observable facts. Thus applying the model for the prediction of forest fires was so encouraging that a project has been initiated in cooperation with a group working on fire prevention in the southern part of Switzerland.
At the present time we are working towards integrating machine learning procedures in the diagnostic process. This may be a way of increasing the power of the system. This new research project is just now beginning at SFISAR with the cooperation of the Artificial Intelligence Laboratory of the Swiss Federal Institute of Technology at Lausanne. We hope for first results before 1997.

## REFERENCES

Bolognesi R., 1993,
*Artificial intelligence and local avalanche forecasting : the system AVALOG.* Proceedings, International Emergency Management and Engineering Conference (Arlington, Virginia, 1993), p.113-116, Society for Computer Simulation, San Diego, California.

Bolognesi R., Buser O., 1995,
*Local avalanche forecasting in Switzerland : strategy and tools. A new approach.* Proceedings, International Snow Science Workshop (Snowbird, Utah, 1994). To be published.

Buisson L., 1993,
*Snow and computer - a survey of applications for snow hazards protection in France.* Proceedings, International Emergency Management and Engineering Conference (Arlington, Virginia, 1993), p.175-180, Society for Computer Simulation, San Diego California.

Buser O., 1989,
*Two years experience of operational avalanche forecasting using the nearest neighbours method,* Annals of Glaciology, vol.13, p.31-34, International Glaciological Society., Cambridge.

Good W., Ammann W., 1994,
*Modelling local avalanche forecast, a review.* Proceedings, International Symposium on Snow and Related Manifestations (Manali, India, 1994). To be published.

Good W., Obled C., 1980,
*Recent developpements of avalanche forecasting by discriminant analysis techniques : a methodological review and some applications to the Parseen area (Davos, Switzerland).* Journal of Glaciology, Volume 25, n°92, p.315-346, Cambridge.

McClung D.M., Tweedy J., 1994,
*Numerical avalanche prediction : Kootenay Pass, British Columbia.* Journal of Glaciology, vol 40, p.350-358, Cambridge.

Schweizer M., Föhn P.M.B., Schweizer J., 1994,
*Integrating neural network and rule based systems to build an avalanche forecasting system.* Proceedings, International Conference "Artificial Intelligence, Expert Systems and Neuronall Networks " (Zürich, Switzerland, 1994), International Association of Science and Technology for Development.

Weiss S.M., Kulikowski.C.A., 1991,
*Computer systems that learn.* Morgan Kaufmann Publishers, San Mateo, California.

## BIOGRAPHIES

### Robert Bolognesi

Graduate from the University of Sciences of Grenoble, France (doctor's degree, June 1991 and diploma of "Research Management", June 1994). Has worked as avalanche forecaster in the safety service of a large ski resort. Now in the staff of the SFISAR, leading researches about Machine Learning in order to try to improve decision support systems for avalanche control.

### Othmar Buser

Graduate from the University of Zürich (expèrimental physics) and doctor es sciences from the University of Neuchâtel (Switzerland). Joining the SFISAR in 1965, first engaged in hail research, then in physics of snow research. Now working on computer assisted avalanche forecast, trying to find practical ways for application.