The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

A preprocessing method for improving data mining techniques. Application to a large medical diabetes database

A Duhamel[a], MC Nuttens[a], P Devos[a], M Picavet[a], R Beuscart[a]

[a]CERIM - Faculté de Médecine - 1, Place de Verdun – 59045 Lille Cedex – France

Abstract

The Knowledge Discovery in Databases (KDD) methodology seems to be attractive on the analyze of large clinical databases. In the KDD process, the preprocessing step (data cleaning and handling of missing values) is paramount since it conditions the quality of the results obtained by data mining procedures and represents about 80% of the whole project time. The aims of the present study were to analyze this step and provide tools to handle inconsistent data and missing values. We have broken down the process into 3 main stages : data cleaning – explanatory study of missing values – choice of the procedure used for handling missing values. The data cleaning stage was based on a system of logical rules to correct mistakes and on cluster analysis to discard the poorly filled files. The missing-data mechanism was analyzed by means of multivariate statistical procedures. Two methods to deal with missing values were compared : imputation by the most common value (mode) and imputation using decision trees. This study was performed on a large medical diabetes database (23601 patients) including numerous missing values. A system of logical rules allowed to correct mistakes on essential parameters (for example, the type of diabetes). Cluster analysis allowed to identify 10% of poorly filled files. After multivariate analysis, the missing-data mechanism could be considered as random. For variables with low number of missing values (<10%) and categories(<4), imputation using decision trees provided better results than imputation by mode.

Keywords:
KDD; Data mining; missing value; imputation; data cleaning; Health Care Systems; Diabetes

1. Introduction

Over the past decade, many organizations have begun to routinely capture huge volumes of data, describing their operations, products and customers. This is facilitated by the increasing memorization capacity of computers, now considered as unlimited. As a result, hospitals and health care institutions have taken advantage of the development of the new information and communication technology, and have built more and more advanced information systems. Where traditional medical researchers usually collect data with a specific design, which requires a large expenditure of time and resources, now, data are automatically collected and integrated into large information systems [1]. There is a growing demand from the healthcare community to dig into and transform the vast quantities of healthcare data into value added, 'decision-quality' knowledge. The situation of the healthcare enterprise can be summed up in 'data rich' but 'knowledge poor'.

Knowledge Discovery in Databases (KDD) might provide a solution. KDD is "the process of non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data" [2]. Data mining concerns the discovery stage in the KDD process. This approach is well suited to the analysis of a vast amount of data where traditional statistical analysis based on the "hypothesis and test" paradigm becomes a time consuming process. Data mining has demonstrated its efficiency in some application fields like marketing, bioinformatics or genetics, but it is still a research area for medical data bases and is primarily used for diagnosis or treatment selection rather than for knowledge discovery.

KDD involves different steps including data preparation (cleaning, filling in missing values, coding), data mining and validation of results. In the KDD process, data preparation is a crucial stage and represents about 80% of the whole project time [2]. In this stage, several peculiarities encountered in the medical field must be taken into account :

- medical data are heterogeneous, likely to be imprecise or subjective (self reported data), including erroneous data and missing values.
- each process on medical data such as recoding or imputation of missing values must be validated by experts.
- requested interpretability of results from data mining procedures.

In previous papers ([3],[4]), we have reported experiences of data mining in the medical field. In these studies, we concluded that data mining tools can be efficient in the context of complete and reliable data, but we also pointed out the necessity to develop tools for the data preparation step, especially in order to deal with inconsistent data and missing values. A literature review of existing studies revealed very little information about the data preparation step. Most of these studies concerned the application of data mining tools for diagnosis or treatment selection or the development of new data mining procedures [5]. In particular, data consistency was not analyzed and most studies concerned data without missing values or were confined to complete cases analysis. In other applications, missing values were treated using specific and internal methods for the data mining algorithm [6].

The present study concerns the data preparation step in the KDD process applied to a large medical diabetes data base.

2. The Datadiab project

In 1990, The European Diabcare project was established to develop instruments and mechanisms for assessing quality care in diabetes, and to reduce the divergence between real and the ideal quality of care [7]. For this purpose, a standardized questionnairė was developed, the Basic Information Sheet (BIS). BIS comprises about 150 items related to state of health, treatment of diabetic patients and existence of diabetes-related complications. The French DiabCare program federates the majority of French diabetologists. Data collection has been organized once a year since 1993. It is performed during a one-month period and allows for collecting data from all diabetic patients taken care of in hospital units taking part in the project. In 1998, all the French data collected from 1994 were integrated into an Oracle database called Diabcare database. This database contains 32551 records and about 300 items included numerous indicators of quality of care, evaluation of complications and risk factors. In a first step, data was analyzed by means of classical statistical methods. They have contributed to underlining major trends in the quality of diabetes care and confirmed the gravity of the current status of this population. The present configuration and the increase in the size of the Diabcare database induces the major problem of efficient use. A research project, supported by the French Ministry was initiated in 2001. The major purpose of this project, called Datadiab, was the development of data mining tools for medical data. This project involves several partners : statisticians, computer scientists, data management specialists and medical experts. The objectives of the project were :

1. to analyze the data preparation step and to provide tools for handling inconsistent data and missing values in order to minimize the calling in of experts.

2. to evaluate recent methods of supervised classification derived from boosting techniques on a large medical data base.

This paper concerns the first stage.

3. Methods

The data preparation process is summarized in figure 1.

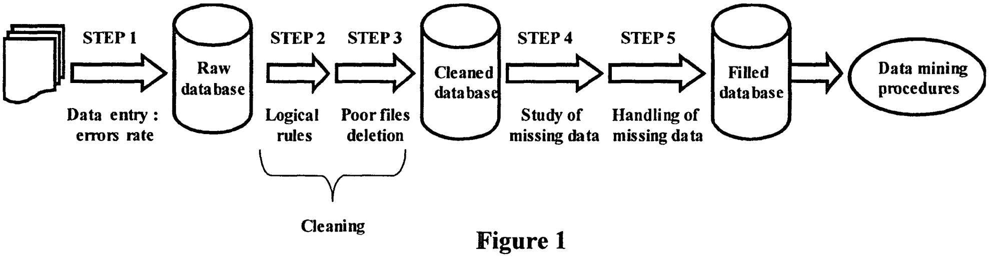

Figure 1

Step 1. The first source of errors in the large data bases results from the data recording process. This is generally done by a single person different from the health care provider. We estimated the rate of error committed by doing a single data entry process compared to the independent double data entry process, which we considered as the gold standard.

Step 2. Simple checks on the upper and lower limits of the parameters are effected at the time of the recording. However, incoherence can exist in an apparently well completed file and certain missing values can be deduced from other informed variables. We developed and evaluated a system of logical rules using several variables to detect and correct these problems.

Step 3. The poorly filled files must be discarded before using data mining procedures. These files constitute a "noise" which can distort the results obtained from analyses. The notion of missing value is more or less important depending on the variables. With the help of experts, we have organized the variables into 2 hierarchical groups :

- variables providing essential information regarding to project objective (group 1),
- the other variables (group 2)

The poorly filled files are records including numerous missing values on the group 1 variables. They were identified by means of cluster analysis using k-means algorithm.

Step 4. The handling of missing values is still a challenging task of the KDD methodology [8]. Two strategies are currently employed : (1) filling in the missing values (imputation) or (2) using methods which are specific and internal to the data mining procedure. For example : (1) one can replace a missing value by the most common value (mode) for categorical variables; (2) in C4.5 (decision tree algorithm) missing values are distributed to each subset with a weight proportional to the observed distribution in complete cases. The previous methods assume that the missing-data mechanism is random [9]. Imputation appears to be the most attractive method since it allows to obtain a complete database which can be analyzed using any data mining procedure.

Different reasons can be given to explain a missing value in a medical data base :

1. poor file completion,
2. information unreported by the clinician for reason of unavailability: ophtalmological examination results, for example,
3. examination considered unnecessary by the clinician (absence of guideline).

After step 3, missing values can be the result of poor file completion but also of causes 2 or 3, the mechanism of which is not random.

The objective in stage 4 was estimate the hypothesis of "missing at random" on the essential variables (group 1). In order to do this, we realised bivariate and multivariate analyses (stepwise logistic regression and decision trees).

Step 5. Two imputation methods for categorical variables were evaluated and compared in the Diabcare database : imputing by the mode and imputing using decision trees. The former is used in numerous software such as Sipina, Enterprise Miner (SAS) or Solas. The latter is analogous to imputation by multiple regression in the case of numerical variables. We proceeded in the following way : let Y be the variable that must be imputed. Cases with Y missing were discarded. A decision tree was built using Y as a dependent variable and all the others as independent variables. An extra category 'missing' was created for each of these. Analysis was performed by means of the Sipina software [10]. The Chaid method was employed and we chose a low significant level to restrict the size of the tree. Using terminal nodes (the leafs), rules were produced to impute the missing values. The imputed value corresponded to the most common value of the leaf.

4. Results

23601 records corresponding to the non-insulin-dependent type II diabetic patients were analysed.

Step 1. Comparison between the 2 modes of recording was based on 148 parameters (108 binary and 40 numerical) et 1079 subjects (25% of records were selected at random from the year 2000 campaign data). The total number of listed errors was 360, which corresponded to an actual rate of 0.23 %. Not surprisingly, rate of error was higher for the numerical parameters (0.47 %) than for the binary variables (0.14 %).

Step 2. 65 coherence rules involving 50 parameters were developed based on expert advice. For example, the rule "*IF (existence of a medicinal treatment to be taken orally AND age of diabetes onset > 30 years AND BMI > 25). THEN type of diabetes = 2*" verifies the coherence of treatment with type 2 diabetes. The before and after correction data were compared. For the parameters studied, the average rate of correction was of 3.7%. Certain important errors were discovered and corrected. For examples, 756 errors on the type of diabetes were corrected and 2327 for the parameter "foot examination realised".

Step 3. Cluster analysis identified 2367 (10%) of poorly filled files on the group 1 parameters. Additional analyses showed that these files contained a significantly higher number of missing values for all the variables. Furthermore, they did not correspond to any particular clinical profile. These files were then discarded.

Step 4. Logistic regression and decision trees gave the same results : The missing values for the important variables (group 1) were strongly correlated to the other uninformed variables. We did not find clinical profiles able to explain the missing values. This rather suggests a group of " bad screening " rather than different clinical profiles and consequently the hypothesis of missing at random was considered as reasonable for the data mining phase.

Step 5. Prediction using decision trees failed for 6 variables among the 18 variables analysed. Figure 2 gives trees built in order to impute the missing values for the variable "diabetes duration" (ddiab)(4.2% of missing). This variable contains 2 categories, in proportions of 49 and 51 %. If the imputation by the mode – was applied, then all the missing values would be classed as 2. Now the tree shows that the distribution of ddiab is very different from one sub-population to the other. For example, if a patient has no retinopathy and is young, then he has a 79 % probability of having ddiab = 1. It was thus necessary, for every parameter, to compare the 2 methods.

We proceeded as follows : Say Y, the variable to be imputed, m the number of categories, p the percentage of the mode and N the number of missing values for Y. If imputing is done using the mode, the average rate of errors is estimated by : $Tm = (1 - p)$

If a classification tree is used, the calculation is as follows : say k is the number of allocation rules corresponding to the number of terminal nodes on the tree for complete cases, p_1 , ... p_k the corresponding proportions and N_1 , ... N_k the respective strengths of k nodes on the missing values. The number and the average rate of error committed can be estimated by :

$Tr = \frac{\sum_{i=1}^{k} N_i(1-p_i)}{N}$. For the ddiab variable in figure 2, we obtain :

	N	m	k	Tm	Tr
Ddiab	875	2	8	48.83	37.26

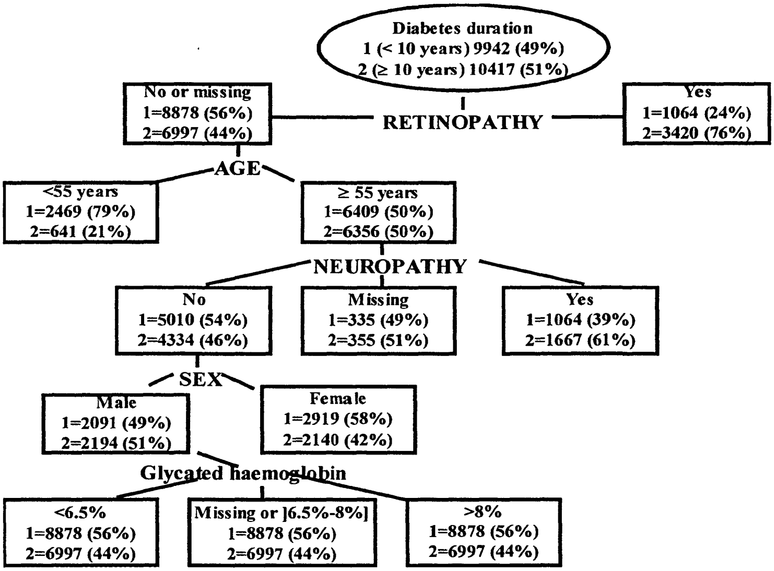

Figure 2 : imputation for diabetes duration (ddiab)

5. Discussion and conclusion

The data preparation stage is a crucial step since the results of data mining procedures depend on the quality of data analyzed. We showed that an important amount of incoherent data could be corrected by logical rules. Such rules could be integrated in the data acquisition software of information system to minimize mistakes.

In many studies, data mining procedures are applied to complete data, after deletion of missing values. This procedure is inadequate in large clinical data bases because the loss in sample size can be considerable, particularly when the number of variables with missing values is large. We propose a 3 stages process : elimination of poorly filled files, – explanatory study of missing values – choice of the procedure used for handling missing values.

In our study, 10% of files were considered lost and were discarded. After cleaning, the database still includes numerous missing values. Before applying a missing data handling procedure, it is necessary to explain the missing-data mechanism [9]. Explaining missing data is equally important for assessing quality of care and for the identification of different medical working practices. In our study, the existence of missing values regarding essential parameters for diabetes follow-up could be explained by missing values for other parameters. That suggests a bad screening of risk factors and complications for all patients and not for specific populations. Theses results are similar to other French and international studies [11]. If, on the contrary, distribution of missing values depended on different subgroups, data mining procedures would need to be done separately within each subgroup.

The imputation using decision trees performed better than imputation by mode for 6 of the 18 variables analyzed. For other variables, whether no model could be found or, if such a model existed, it provided results similar or poorer than imputation by mode. When the missing values rate was high (>10%) or when the number of categories was high (≥4), imputation by decision trees failed.

Filling in the missing values is considered as a general and flexible method for handling missing values. However, the use of this method raises several difficulties. First, it can induce substantial statistical biases (underestimation of variance, distortion of correlations between variables). Multiple imputation [12] could correct theses biases but it use in the data mining process must be studied. A second difficulty arises when a multivariate procedure is used for imputation. A predictive model must be identified for each variable to be imputed and missing data for independent variables must be handled. The same problem arises when using the method proposed in CART (surrogate split). In our study, an extra category 'missing' was created since this missing characteristic could be considered as a risk factor. However, trying to fill in missing values using other missing values is a serious problem. To resolve this problem and to improve the reliability of imputed values, an iterative procedure similar to the EM algorithm could be used.

This study concerning the handling of missing values in the KDD process is a preliminary study. In near future, we will perform simulations on a complete database to obtain a gold standard and thus be able to compare different methods (C4.5, imputations, extra category 'missing').

References

[1] Lavrac N, Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine*, 1999, 16: 3-231

[2] Adriaans P and Zantinge. *Data Mining*. Edinburgh : Addison Wesley, 1996.

[3] Duhamel A, Picavet M, Devos P, Beuscart R, From data collection to knowledge data discovery : a medical application of datamining, *Proceedings of Medinfo 2001*, N°10(Pt 2): 1329-33.

[4] Quentin J, Devos P, Duhamel A, Beuscart R and the Qualidiab group, Assessing association rules and decision trees on analysis of diabetes data from the DiabCare program in France, *Proceedings of MIE 2002*, Budapest.

[5] Konenko I, Machine learning for medical diagnosis : history, state of the art and perspective, *Artificial Intelligence in Medicine*, 2001, 23: 89-109

[6] C4.5 : Quinlan JR, Induction of decision trees, *Machine learning*, 1986, 1: 81-106

[7] Piwernetz K, Home PD, Snorgaard O, Antsiferov M, Staehr-Hohansen K, Krans M, Monitoring the targets of the St Vincent declaration and the implementation of quality management in diabetes care : the DiabCare initiative, *Diabetic Med*, 1993, 10 : 371-377.

[8] Fayyad U M, Piatetsky-Shapiro, Smyth P, From data mining to knowledge discovery : an overview. *Advances in Knowledge Discovery and Data Mining*, MIT Press 1996 : 1-36

[9] Little R. J. A, Rubin D. B, *Statistical analysis with missing data*, Wiley, 1987.

[10] Zhighed DA and Rakotomalala R. *Graphes d'induction : Apprentissage et Data Mining*. Paris : Hermès, 2000 : 55-92

[11] Wittchen H U, Hypertension and diabetes risk screening and awareness study, preliminary results : *http://www.lifescan.com/care/news/dn070502-1.html*

[12] Rubin D. B, Multiple imputation after 18+ years, *Journal of American Statistical Association* 1996, 91 (434) : 473-489.