Assessing Association Rules and Decision Trees on Analysis of Diabetes Data from the DiabCare Program in France

Julie QUENTIN-TRAUTVETTER, Patrick DEVOS, Alain DUHAMEL,
Régis BEUSCART and the Qualidiab Group

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

Abstract. Recent advances in information technology have made it possible to solve increasingly complex problems, and also to collect and store huge amounts of information. These vast quantities of data further have to be transformed into relevant value-added and "decision-quality" knowledge. It is against this background that the KDD (Knowledge Discovery in Databases), a multidisciplinary field using computer learning, artificial intelligence, statistics, database technology, expert systems, and data visualization, appeared in the early 90's. In order to assess these technologies in the medical field, we have tested some of these techniques on a large database at our disposal, named DiabCare stemming from the WHO – DiabCare program for the application of the Saint-Vincent Declaration. It contains evaluation data on the health care of patients with diabetes, and in particular, its complications. So far, data analysis has been done using classical statistical methods, and we now intend to make use of such data-mining tools as Associations Rules and Decision and Classification Trees for further exploration of this database. The results presented here show that data mining techniques can be used successfully to extract knowledge from medical databases. The results obtained using Association Rules and especially Decision Trees are very promising.

Keywords (MeSH) KDD ; Data *Mining*; *Associations Rules*; *Decision Trees*; DiabCare; *Diabetes Database*; *Health* Care; *Decision Support*

1. Introduction

In hospitals and other healthcare settings, computers and electronic medical record systems have proliferated. When traditional medical researchers usually collect data with a specific design, which requires a large expenditure of time and resources, now, data are automatically collected. As a result, clinical databases have become more common. But whereas the data acquisition process and in particular, the acquisition of data from clinicians had previously been regarded as the difficult side, there has now been a gradual shift toward the need for effective tools to retrieve the relevant information [1]. The change in data collection has resulted in an overabundance of exhaustive data, so much so that methods of analysis have to change at the same time. There is presently a growing demand from the healthcare community to leverage upon and transform the vast quantities of healthcare data into value added, 'decision-quality' knowledge. The situation of the healthcare enterprise can be summed up as 'data rich' but 'knowledge poor' [2].

The KDD, also called Data Mining might provide a solution. KDD is linked with other research areas such as computer learning, artificial intelligence, expert system, database system, statistics, data vizualisation and data warehousing. KDD is the process of exploring the knowledge contained in databases by using data mining tools, techniques, algorithms and other methods for knowledge extraction [3]. The different steps of this KDD process are data selection, data cleansing and scrubbing, enrichment of data, data coding, then data mining and validation and visualization of results. There might be returns between different steps. Therefore, the data mining step represents 'only' 20% of the KDD process whereas data preparation (filling in missing values and correcting erroneous data ...) and then validation of results constitute the major part of the process [4].

Data mining is now very much in used and successfully applied in marketing, banks, insurance companies as well as in biology and genetics [5, 6, 7]. But it is still a research area for medical databases and primarily used for diagnosis and treatment selection, less for knowledge discovery [8, 9, 10]. Data mining aims to furnish a new generation of tools to assist humans in analyzing mountains of data intelligently. It has been recognized as opening up new areas for database research. These areas can be defined as efficiently discovering interesting rules from large collections of data.

The aims of our study are to apply data mining techniques to medical information concerning diabetes (DiabCare Database), and evaluate their feasibility with the coherence of the results compared to domain specific knowledge and recent classical studies.

2. Materials and methods

2.1. The DiabCare Database

In 1999, a large database was constructed with all the data collected, by the scientific comittee of DiabCare-France and the Department of Biostatistics and Medical Informatics of Lille [11, 12]. At the present time, DiabCare Database contains about 30.000 records. This collection is based on the 'Basic Information Sheet' which contains about one hundred items informing on the state of health and treatment of the diabetic patient and about the presence of severe complications which are termed the Saint-Vincent Complications : Blindness, Kidney insufficiency, Amputations, Myocardial infarction, Stroke. Data have been loaded, cleaned, controled and transformed (first steps of the KDD) and indicators have been elaborated according to the guidelines and recommendations for diabetes care [13, 14].

Performance of data mining methods have been evaluated according to three different criteria: rapidity, comprehensibility and coherence of results. Two data mining techniques have been tested: Association Rules [15, 16] and Decision Trees [17, 18, 19].

2.2. Methods

The algorithms for Association Rules mining in large databases have been developed by Rakesh Agrawal [15, 16]. Association Rules mining finds all rules in the database that satisfy some minimum support and minimum confidence constraints. An association rule is like:

(A and B) => C where A and B are conditions and C is a result. It is characterized at least by two parameters: the support, S = P(A and B and C) and the confidence, C = P(C/A and B). (P denotes probability). The main characteristics of this method are: the discovery target is not pre-determined, there is no hierarchy between items and this method is suited to very large database and numerous attributes. The CBA (Classification Based on Association) software has been used to perform analysis.

Several models of Decision Trees are known, CART, C4.5, CHAID [17, 18, 19]. These models share the process of selecting the most explicative factors for one and only one pre-determined target in a step by step fashion. They differ by the criterion of selection. The

CHAID model uses an adjusted Chi-Square measure (p-value) as the split (or criterion). We opted to test this model using the Sipina software for the analysis.

3. Results

3.1. Association Rules

First analysis was performed on the 29165 cases and a selection of 51 binary variables related to diabetes complications. The support (S) and confidence (C) were fixed at 10% and 60% respectively. With all the data, 128150 rules were generated. With such a huge quantity of rules, it is very difficult if not impossible to detect interesting rules. A second analysis was therefore carried out using a limited number of variables (Saint-Vincent complications, diabetes type, sex, diabetes duration, glycaemia, HbA1C, angina, BMI, HTA, micro-albuminuria, cholesterol and creatinin clearance) and the same values for support and confidence. About 900 rules were generated this time. Some show interesting associations between diabetes characteristics and angina :

Rule 1: If type 2 diabetes and female sex then angina presence (S = 23.7% and C = 68.2%)
Rule 2: If BMI > 30kg/m2 and female sex then angina presence (S = 16,9% and C = 68.5%)
Rule 3: If cholesterol >5,2mmol/l and type 2 diabetes then angina presence (S = 31,6% and C = 71.9%)

The rule 1 shows that women with type 2 diabetes present angina in 23.7% with a confidence of 68.2%. In rule 2, women which BMI is over 30 kg/m2 present angina in 16.9% with a confidence of 68.5%. And rule 3 shows that type 2 diabetics with high cholesterol present angina in 31.6% with a confidence of 71.9%.

3.2. Decision Trees

For experimenting Decision Trees, we used the CHAID model. The results were obtained with the data concerning only type 2 diabetes which represents 13592 subjects. All missing values were purged. Our target was to verify the presence or absence of the Saint-Vincent complication. We selected risk factors (diabetes antecedent, HTA, angina, cataract, tobacco and alcohol) and the numerical parameters (glycaemia, glycated haemoglobin, age, diabetes duration and BMI) were transformed into binary variables. The splits were determined according to the recommendations for diabetes care and calculated on general population [16, 17, 18]. A first Decision Tree was generated using all data, the 13 variables and a Chi-Square p-value=0.001. It showed that the most explicative variables for the presence of Saint-Vincent complications are cataract, followed by angina, sex, and diabetes duration. That way, on a branch the Saint-Vincent complications risk increased from 18.7% to 29.3% in the presence of a cataract , then 50.6% when added to angina and 60.9% if it's a man (with cataract and angina) versus 42.2% if it's a woman.

The Sipina Software gives deducted rules:

If angina=yes and cataract=no and sex=man then complication with a risk of 53%
If angina=yes and cataract=yes and sex=man then complication with a risk of 61%

Another tree (Fig1), was then generated, where angina and cataract attributes were ignored and the ChiSquare p-value was 0.001. In the results, the most explicative factors are diabetes duration, then sex and age. It's interesting to note that the complication risk is always higher for men compared to women. One branch is noteworthy: duration > 10 years, male sex and age >= 65 years : the complication risk rises from 18.7% (root nod) to 37% !

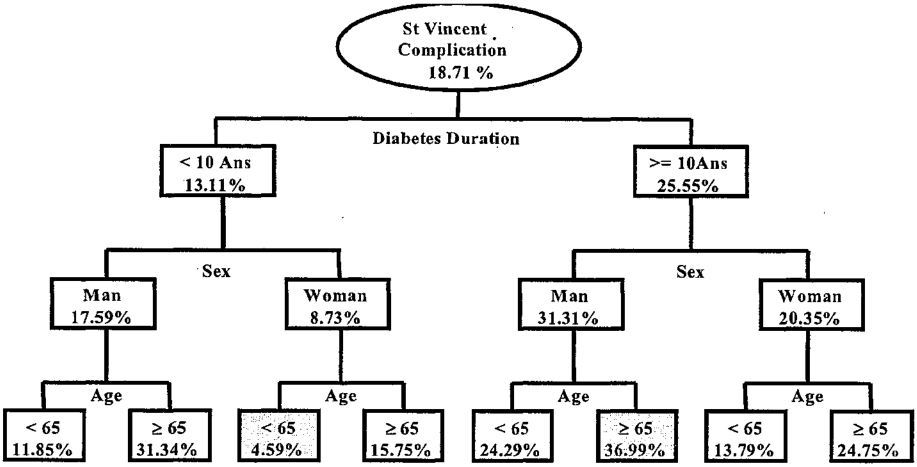

Figure 1: Decision tree with CHAID model and Sipina Software.

4. Discussion and conclusion

In this work, we applied two of the main data mining methods. The obtained results demonstrate the interest of these techniques. They are quickly done, easily comprehensible and coherent compared to the domain specific knowledge. But depending on the objectives of the analysis, each method presents advantages and disadvantages.

A major advantage of Association Rules is that there is no pre-determined target and no hypothesis. An extensive database is a necessary condition. This exploratory method is able to determine interesting associations which can constitute new research hypotheses for the construction of a regression model. This strength, however, comes with a major drawback. It often produces a huge number of associations particularly if attributes are highly correlated, which is the case for diabetes. The huge number of associations makes it very difficult, if not impossible, for a human user to analyze in order to identify those interesting or useful ones, even with strong support and confidence (when there are many parameters, there is a huge quantity of rules, whatever support and confidence may be). Those two mesures are too short to limit the number of associations. So, research to improve the Association Rules pruning is important in particular for medical data [20]. Another problem with the Association Rules mining is the absence of logical reasoning. This may not matter for supermarket data, but in medicine the advance of reasoning is necessary. Association Rules mining is a good exploratory method for very large databases but complementary analysis techniques are necessary.

Decision Trees and in particular CHAID model main interest is the logical reasoning, easy to follow, step by step, which is fundamental with medical data. Instantaneously, explicite rules are deducted. With CHAID model, there's only one parameter to fix, the Chi-Square p-value. According to this p-value, number of ramifications is limited with a better legibility and improving relevance of results. Results are quickly and simply obtained, easily interpretable, and can be used as such or studied with other methods. This method is well adapted for medical studies. Its qualities are often emphasized [19, 21]. In this work, we used binary variables. In future analysis, it would be interesting not turning data into binary variables and let the algorithm fix the different split values for different subgroups.

Although data collection has improved, the problem of missing values is still worrying. Important rates of missing values in the database corrupt the validity of generated rules. In this first approach, incomplete records have been canceled. But we envisage to use missings values in future analysis in order to test stability of results according to the missing values proportion.

In conclusion, data mining tools provide simple and effective methods of extracting knowledge from general medical information but we have to bear in mind that data preparation, in particular treatment of missing values, right choice of techniques used and validation of results are fundamental steps to improve the KDD.

References

[1] Nigrin DJ, Kohane I., Data mining by clinicians. Proc AMIA Symp, 1998: p. 957-61.

[2] Abidi SSR, Applying Data Mining in Healthcare: An Info-Structure for Delivering "Data-Driven" Strategic Services. Stud Health Technol Inform, 1999. 68: p. 453-6.

[3] Lee IN, Liao SC, Embrechts M, Data mining techniques applied to medical information. Med. Inform.(2000) vol.25, N°2,81-102.

[4] Adrians P and Zantinge. Data Mining. Edinburgh : Addison Wesley, 1996.

[5] Piatetsky-Shapiro, G., The Data-Mining industry coming of age. IEEE Intell System, 1999. Nov-Dec: p. 32-34.

[6] Salzberg SL, Gene Discovery in DNA Sequences. IEEE Intell System, 1999: p. 44-48.

[7] Kasif S, Datascope : Mining Biological Sequences. IEEE Intell System, 1999. Nov-Dec: p. 38-43.

[8] Shun Ngan P, Leung W.M., Lam W, Sak Leung, Cheng JCY, Medical data mining using evolutionnary computation. Artif Intell Med, 1999. 16(1): p. 73-96.

[9] Berridge EJ, Roudsari A., Taylor S, Carey S,, Computer-aided learning for the education of patients and family practice professionals in the personal care of diabetes. Computer Methods and Programs in Biomedicine, 2000. 62: p. 191-204.

[10] Montani S et Al., Diabetic patients management exploiting case-based reasoning techniques. Computer Methods and Programs in Biomedicine, 2000. 62: p. 205-218.

[11] Kleinebreil, L. and V. Durlach, [Five years of DiabCare--France: assessment and outlook]. Diabetes Metab, 1998. 24 Suppl 3: p. 8-12.

[12] Beuscart R, VanHoecke MP, Devos P, Allouche R, Kleinebreil L and the QUALIDIAB group, QUALIDIAB : implementation of the DIABCARE project in the French speaking environment : regional , national and international issues, in Proceedings of Medical Informatics Europe '99 , August 22 - 26, 1999 : 799-800, Ljubljana, Slovenia.

[13] Hypoglycemia in the Diabetes Control and Complications Trial. The Diabetes Control and Complications Trial Research Group. Diabetes, 1997. 46(2): p. 271-86.

[14] Cost effectiveness analysis of improved blood pressure control in hypertensive patients with type 2 diabetes: UKPDS 40. UK Prospective Diabetes Study Group [see comments]. Bmj, 1998. 317(7160): p. 720-6.

[15] Agrawal R, Imielinski T., Swami A, Mining associations rules betweens sets of items in large databases. SIGMOD- 1993, 1993: p. 207-216.

[16] Agrawal R, Srikant R., Fast discovery of association rules. VLDB-1994, 1994.

[17] Breiman L, Friedman.J., Olshen RA, Stone CJ,, Classification and regression trees. Wadsworth Int Group, Belmont, CA, 1984.

[18] Quinlan JR, Inductionof decision trees, Machine Learning, n°1, pp.81-106, 1986.

[19] Okell J, Neural networks versus CHAID. White paper from smartFOCUS, June 1999. Site web : http://www.crm-forum.com/crm_forum_white_papers/.

[20] Liu B, Hsu W., Ma Y,, Pruning and Summarizing the Discovered Associations. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1999. KDD-99, August 15-18(San Diego, CA, USA).

[21] Ganti V, Gehrke J., Ramakrishnan R,, Mining very large Databases. Computer, 1999. August: p. 38-45

This work has been supported by the DiabCare Group and the Ministery of Research in France. The authors wish to thank all those who gave much of their time for participation, particularly Dr Omolade Alao.

Adress for correspondance Julie Quentin, CERIM, Faculty of Medicine, pole seeks, 1, place of Verdun, 59045 Lille Cedex, France Tel. +33 3 20 62 69 69 E-mail : jquentin@univ-lille2.fr