Discovering Associations in Clinical Data: Application to Search for Prognostic Factors in Hodgkin's Disease

N. Durand [1],[3], B. Crémilleux [1], and M. Henry-Amar[2]

[1] GREYC, CNRS - UMR 6072
Université de Caen
F-14032 Caen Cédex France
{ndurand,cremilleux}@info.unicaen.fr

[2] GRECAN, EA 1772
Centre François Baclesse
F-14076 Caen Cédex 5 France
m.henry.amar@baclesse.fr

[3] present address: France Télécom R&D,
42 rue des coutures F-14066 Caen Cédex 4 France,
nicola.durand@rd.francetelecom.fr

Abstract. The production of suitable clusters to help physicians explore data and take decisions is a hard task. This paper addresses this question and proposes a new method to define clusters of patients which takes advantage of the power of association rules method. We present different notions of association and we specify the notion of frequent almost closed itemset which is the most appropriate for applications in the medical area. Applied to Hodgkin's disease to help establish prognostic groups, the first results bring out some parameters for which classical statistic methods confirm that they are interesting.

1 Introduction

Important medical data are collected during the treatment strategies of patients suffering from serious diseases like, for example, cancer. The use of relevant and efficient methods to explore such large data sets is not easy and we notice that these data are often under-used. Statistics are often used to validate suspected models and we are facing today to a new challenge: how may new models be discovered? By extracting from large amounts of data non trivial "nuggets" of information, Knowledge Discovery in Databases (KDD) is a semiautomatic way which may help the user for this work. Association rules are one of the most popular methods in KDD and the aim of this paper is to show their contribution to define a new method to gather patients who are similar from a certain point of view. These "clusters" of patients, in our context, may help to establish prognostic groups in Hodgkin's disease.

2 Associations and clusters in medical area

2.1 Association rules and clusters

It is very common to try and structure a set of data in clusters in order to infer rules or other knowledge from them. In medicine, a cluster can group patients having similar and/or related features. The production of meaningful clusters in data mining is a hot topic. Similarity measures are often used. When data are complex and include categorical attributes, these measures are not easy to define [1]. Such situations are very usual in medical area. However, association rules method offers a background to produce clusters without requiring a similarity measure.

An association rule [2] is a statement of the form `‘‘95% of patients that have gender = male and mediastinum = enlarged also get platelets` $\in$ `[100, 600[`[1]`’’`, 27% of patients in the database match this rule. This last number is called *support* of the rule, 95% is the *confidence* (i.e. the percentage of data that contain the consequent among those containing the antecedent). `gender = male` is an item and both antecedent and consequent are sets of items (or *itemsets*).

We claim that an itemset brings an intuitive way to define a cluster: it gathers patients who are similar according to items that are making up. According to the number of features and their frequencies in the patients data, the definition of a cluster can be more or less specific (and a cluster can collect a few or a large amount of patients). A sound idea is to use the itemsets generated during the process of mining the rules.

2.2 Method: discovering clusters of patients

The search of clusters that we propose follows three main stages.

Frequent closed itemsets. The first stage provides all the frequent closed itemsets. The notion of frequency takes into account the “weight” of an itemset: if the frequency is not large enough, it means that the rule is not worth consideration. To generate clusters, for a given group of patients, we prefer to simply produce the sole and only itemset which is composed of the maximal number of items shared by the group: such an itemset is called *closed* itemset [3]. The idea is to catch the maximum amount of similarities among the data. We here propose to reuse this notion in a context of generation of clusters since a closed itemset catches the maximum amount of similarities among a set of data.

Almost closed itemsets. Nevertheless, in the medical area, it is well-known that data are particularly hard to explore: data are often noisy, missing values and redundancy often occur. From our point of view, one of the most difficult sides is the uncertainty intrinsically embbeded in the data. Physicians know that some examples escape from the rules and this fact makes difficult the use of data mining methods [4]. For association rules, we note that in practice, the “best” rules that we are able to extract, are not characterized by a confidence level of 100%. To be efficient in a domain where there are always exceptions we relax the constraint of closure which is on the core of the closed itemsets [5] to allow for few exceptions in a cluster. From a technical point of view, this means that we

[1] expressed in $10^{-9}/L$

mine the *almost* closed frequent itemsets. The number of exceptions is controlled by a parameter noted δ.

Constraints. The third stage of our method is the introduction of constraints. They are a way to take into account domain knowledge or expert know-how. A constraint can be about the domain or connected with the characteristics of the clusters (for instance, a cluster should contain a minimum number of items and patients). Only the clusters checking the constraints are kept. Moreover, it is also a pragmatic way to reduce the number of clusters.

Finally, let us note that it is hard to evaluate quantitatively the quality of the discovered clusters because, unlike classification, there is no predefined "correct value" that can be used to compute measures such as precision or recall.

3 Experiments

The used data set has been collected by the Lymphoma Group of the EORTC. It describes more than 4000 patients with early stage Hodgkin's disease (HD) treated with various protocols. Since protocols have changed with time, experiments where done on a selected subset (protocol H7, 816 patients, 1988-1993) [6]. Patients are divided in "favourable" (360 cases) and "unfavourable" (456 cases) prognostic groups described through 66 items. The data set contains 2825 missing values (five attributes concentrating 88% of the missing values).

In a first attempt, we searched for associations on all data (with a minimum support of 4% (i.e. 33 patients) and a maximum number of two exceptions by itemset). A large number of almost closed itemsets (850 417) were obtained and the selected clusters forming the "pseudo-partition" are difficult to interpret: in general, they contain a mixture of the prognostic groups or tend to be identified by the attributes corresponding to the definition of the prognostic groups.

Items belonging to the definition of the prognostic groups were then removed in order to better understand the role of the others (since experts were especially interested in biological data). 42 items remained. In order to distinguish subgroups of patients within each prognostic group, the data set was split according to the "favourable" and "unfavourable" values. We mined associations with the same parameters as before. From the "favourable" group, 4224 almost closed itemsets are extracted and 997 from the "unfavourable" group. Clusters were ranked on their treatment-failure associated rate. Are associated with a high risk of failure, in the "favourable" group, a low lymphocyte count, and in the "unfavourable" group, a low lymphocyte count and disease localized in the right supraclavicular region.

More investigation was made on the "unfavourable" group, after patients were separated according to the chemotherapy administrated: MOPP/ABV versus EBVP. For EBVP patients, a high level of white blood count (WBC) is associated with a high risk of failure while a low lymphocyte count is associated with a high risk of failure in MOPP/ABV patients. In order to confirm these observations, a classical survival analysis was performed that assesses the relevance of the WBC in EBVP patients only.

4 Conclusion

The production of meaningful clusters in data mining is a hot topic. We have presented a new approach to generate clusters of patients. This method is based on the search of closed itemsets and we have introduced the relaxation of the constraint of closure to be suitable in uncertain domains such as clinical data.

Applied to HD, using a clean database including patients with sufficient follow up, our method was able to discriminate few itemsets that may be prognostic. It was particular pertinent in the subgroup of patients experiencing the worse treatment failure free survival (i.e. EBVP patients of the "unfavourable" group). In the future, we will consider patients with early stage HD and unfavourable prognostic features together with patients with advanced stage disease and favourable prognostic features whose clinical outcome are similar thanks to the modern therapeutic strategies applied in order to highlight new itemsets potentially prognostic that could be used in clinical practice providing they are reproducible.

Considering the method to produce clusters, work has to be done to settle the best definition of an almost closed itemset in domains like medicine. Another way is to specify a proper strategy to obtain a clustering of the data.

Acknowledgements. The authors thank Artur Bykowski and Jean-Franois Boulicaut for stimulating discussions and the use of `min-ex` software.

References

[1] G. Das and H. Mannila. Context-based similarity measures for categorical databases. In *the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 00*, number 1910 in Lecture notes in artificial intelligence, pages 201–210, Lyon, F, 2000. Springer-Verlag.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. *Fast discovery of association rules*, chapter 12. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[3] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, Elsevier, 1999.

[4] B. Crémilleux and C. Robert. A theoretical framework for decision trees in uncertain domains: Application to medical data sets. In E. Keravnou, C. Garbay, R. Baud, and J. Wyatt, editors, *6th Conference on Artificial Intelligence In Medicine Europe (AIME 97)*, volume 1211 of *Lecture notes in artificial intelligence*, pages 145–156, Grenoble (France), 1997. Springer-Verlag.

[5] J. F. Boulicaut and A. Bykowski. Frequent closures as a concise representation for binary data mining. In *proceedings of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 00*, volume 1805 of *Lecture notes in artificial intelligence*, pages 62–73, Kyoto (Japan), 2000. Springer-Verlag.

[6] E. M. Noordijk, P. Carde, A. M. Mandard, W. A M. Mellink, M. Monconduit, H. Eghbali, U. Tirelli, J. Thomas, R. Somers, N. Dupouy, and M. Henry-Amar. Preliminary results of the EORTC-GPMC controlled clinical trial H7 in early stage Hodgkin's disease. *Ann. Oncol.*, 5 (Suppl. 2):S107–S112, 1994.