Uncertainty Rule Generation on a Home Care Database of Heart Failure Patients

S Konias, GD Giaglis, G Gogou, PD Bamidis, N Maglaveras

Laboratory of Medical Informatics, Medical School, Aristotle University of Thessaloniki, Greece

Abstract

In this paper we present the Uncertainty Rule Generator tool and the algorithm used. This data-mining tool generates uncertainty rules as a part of the Knowledge Discovery in Databases process and is tested upon a home-care database containing data from congestive heart failure patients over a period of approx. 10 months.

This algorithm can handle dynamic data without the need of recovering the itemsets from the beginning. This is highly appropriate for a home-care monitoring system, where new records are constantly added. Moreover it can deal with missing values, since it uses flexible metrics, similar to those of other Association Rule algorithms. Finally this algorithm computes a Certainty Factor for each extracted rule, which is representative of its efficiency. In a future step, this extracted rule can be used on newly entered data, in order to predict the missing values, while its Certainty Factor will allow the exact estimation of error in this prediction.

1. Introduction

Knowledge Discovery in Databases (KDD) is of increasing interest for medicine, because it can extract the knowledge hidden in big medical databases by discovering regularities and exceptions. A well-known KDD technique are the association rules [1,2], which relate the presence of items in transactions. One classic example is the rules extracted from the content of market baskets. In this example, items are the things bought in a market and transactions are the whole market baskets containing several items. The association rules describe which things are usually bought together with other things.

Common to the association rules are also the uncertainty rules. The form of those rules is easy to understand: "**IF** A **THEN** B **WITH** CF", where A, B are the itemsets and CF is the certainty factor. The role of the certainty factor is to represent the efficiency of each rule, that is how often it is applied. This feature makes similar rules more appropriate for relative sciences, like medicine.

Currently, most of the existing algorithms [1,2] for generating associations among the attributes describing the data in a database have a number of shortcomings: they mandate multiple passes over the initial database assuming a static database. However, many applications in medicine require generating associations in dynamic databases, where new records are constantly being added (e.g. in our case a home-care monitoring system). Furthermore, the aforementioned algorithms cannot deal with missing values.

In dynamic databases, it is hard to maintain the discovered rules, since the continuous updates may not only invalidate some already existing rules but also produce other rules relevant. To date, some solutions to this problem are presented in the literature [3-5], but the need for dealing effectively with missing values still exists.

In this paper we present the Uncertainty Rule Generator (URG) tool and the algorithm used, namely URG-2, which is an evolution of [6], for incrementally updating the uncertainty rules when new records are added, even if the database contains missing values.

2. Helpful definitions

Let I be a set of items and T a set of transactions with items in I, both assumed to be finite. Let "IF X THEN Y WITH CF" be an uncertainty rule, where $X,Y \subseteq I$, $X,Y \neq \emptyset$ and $X \cap Y = \emptyset$. This rule means "CF% of the transactions in T that contain X, contain also Y".

The main idea upon tackling the problem of missing values is to ignore records containing missing values for each corresponding itemset separately, in order to avoid missing important information.

Let B be a Database and Z a set of items.

Definition 1: We note B(Z) the subset of B containing Z, i.e $B(Z)=\{w \in B / Z \subseteq w\}$.

Definition 2: If a transaction $t \in T$ contains missing values for at least one item of Z, then t is disabled for Z in the database B. We note $B_{dis}(Z)$ the subset of B disabled for Z.

As a consequence, the metrics probability and conditional probability have to be calculated taking in account the transactions disabled for each itemset in a rule. Below, new definitions for the above metrics are given, fully compatible with [7].

Definition 3: The Probability P_X of an itemset X in a database B with missing values is:

$$P_X = P(X) = \frac{|B(X)|}{|B| - |B_{dis}(X)|}$$

Definition 4: The Conditional Probability $P_{X/Y}$ or Certainty Factor (CF) of the rule "IF X THEN Y" in a database B with missing values is:

$$P_{X/Y} = P(X/Y) = \frac{|B(X \cap Y)|}{|B(X)| - |B_{dis}(Y) \cap B(X)|}$$

3. Method

The used URG-2 algorithm from the URG tool generates relations between the data for a heart failure patient in an understandable form. The more important advantages of this algorithm for a home-care monitoring system are that it:

- Can handle dynamic data without the need of recovering the already existing itemsets from the beginning.
- Can deal with missing values, since it uses flexible metrics.
- Computes for each extracted rule a certainty factor, which represents its efficiency.

The algorithm consists of two parts. The first part scans the data and finds the existing itemsets. The second part generates the uncertainty rules whose probability and conditional probability are higher than a user-specified threshold.

3.1. Itemset generation algorithm

Our idea is to preprocess the database and create the itemsets with the information needed for uncertainty rule generation. Once the itemsets are created there is no further need to access the original database again.

The main aim of the first part, the Itemset Generation algorithm (IG) is to be able to deal with dynamic databases. As mentioned above, in a home-based care monitoring system, where new records are added very often, the need for this kind of algorithm is high. Updating the itemsets when the database is changed is fast and simple without the need of recovering the itemsets from the beginning. In addition, the facts that it makes only one pass over the initial data set and that it deals with missing values are some other important advantages for the IG algorithm.

We present a novel idea for generating the itemset. The IG algorithm stores the needed information into two list structures with itemsets, one for the itemsets with missing values and one for the itemsets without missing values. In Figure 1 we show the main idea for the IG algorithm.

```
Itemset_Generation_algorithm
for all records R in Database B do
 if R is the first record
  then add R as the first node $n_1$
  else scan each node in the list to
    find common values with R
    if a node with the common values already exists
       then increase by one $n_v$.count
       else add a new node $n_m$
    if R contains missing values then
       if a node already exists that has the same
         common values and the same missing values
         then increase by one $m_k$.count
         else add a new node $m_p$
    if  there is no node exactly matching R in the list
      then if R contains no missing values
              then add at the end of the list a new
                node $n_m$ with all values of R
            else add at the end of the list a new
                node $m_p$ with all values of R
endfor
```

Figure 1. IG Algorithm

The itemset generation algorithm creates smaller itemsets than other algorithms, for example the ones based on the apriori algorithm. Thus, if the itemset {a, b, c} is found in 5 records, then the itemset {a, b} will not be mined unless it is found in more than 5 records. This way, no redundant rules are mined.

3.2. Rule generation algorithm

The main aim of the second part is to mine from the already generated itemsets those uncertainty rules whose probability and conditional probability are higher than thresholds specified by the user. The used metrics are those in Definitions 3 and 4. The needed information is taken from the list data structure. In Figure 2 the main structure of the Rule Generation algorithm (RG) is described. The definitions of the symbols that are used in Figure 2 are in section 2.

```
Rule_Generation_algorithm
1 for each node $n_v$ in the Itemset List with
     no missing values do
2  find $|B| - |B_{dis}(X)|$
3  if the Probability $P_X$ of $n_v$ is high enough
4   then number of items i←1
5   for all possible combinations of the attributes,
      that have i items at right, find those with
      high enough Conditional Probability $P_{X\backslash Y}$
6   endfor
7   if at least one item was found
8    then i++ and goto 5
9 endfor
```

Figure 2. RG Algorithm

4. The home care system

The database that was used for this study was collected at the Lab of Medical Informatics in Thessaloniki, Greece. It consists of records of patients who participated in the Citizen Health System (CHS) project between September 2001 and January 2003 [8]. CHS is a home care system constructed around an automated Contact Center functioning as a server. Patients can communicate with it via a variety of interfaces, like public telephone, Internet or a mobile device. In this project patients record with the help of electronic microdevices and transmit to the Contact Center the values of their vital parameters (continuous variables), and yes/no answers to simple questions regarding mostly the occurrence of certain symptoms (boolean variables). In such a database, missing values are due mainly to technical problems or improper use of the various interfaces and are considered random.

During the period of this study, 11 Congestive Heart Failure (CHF) patients participated for 8-13 months and were sending once a week their ECG and three times a week values of the parameters described in Table 1. The primary purpose was to monitor the condition of the patients and help them avoid hospital readmissions.

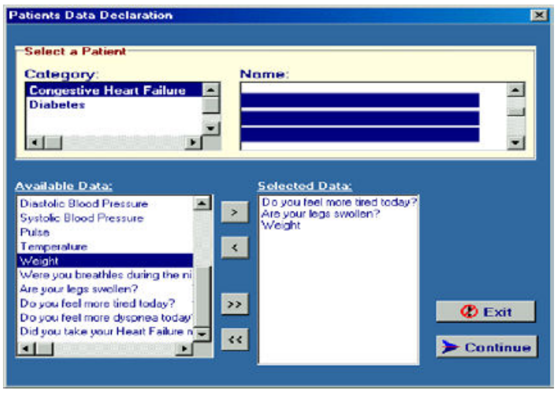
Table 1. Values transmitted to the Contact Center.

Vital parameters (continuous)
Systolic blood pressure
Diastolic blood pressure
Pulse
Weight
Temperature
Questions asked (boolean)
Did you feel breathless during the night?
Are your feet swollen?
Do you feel more tired today?
Do you have dyspnoea today?
Did you take your heart failure medication?

5. The URG tool

Since medical data is highly confidential, transferring it from the Contact Center was avoided, so as not to compromise its safety. For this purpose, a client/server architecture was used in the implementation of the URG tool, were the use of the client is password-protected, as well. More specifically, the queries are defined on the client but are executed on the server. Only the results are returned from the server to the client so as to be viewed by the user.

An intuitive Graphical User Interface (Figure 3) was designed to help the physician-user choose one of the available patients and also select among the corresponding parameters the ones that will be used in the rules to be mined.



Patients Data Declaration

Select a Patient

Category:

Congestive Heart Failure
Diabetes

Name:

[redacted]

Available Data:

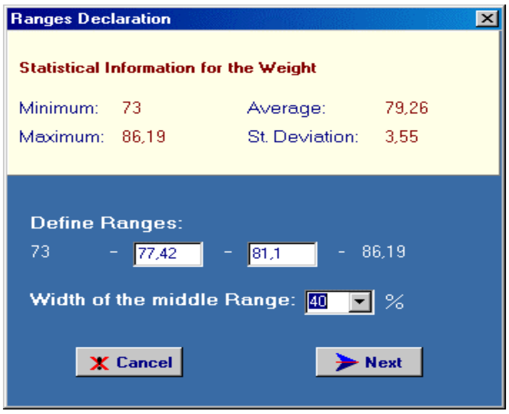
Diastolic Blood Pressure
Systolic Blood Pressure
Pulse
Temperature
Weight
Were you breathles during the ni
Are your legs swollen?
Do you feel more tired today?
Do you feel more dyspnea today
Did you take your Heart Failure n

Selected Data:

Do you feel more tired today?
Are your legs swollen?
Weight

Figure 3: Selection of patients and parameters

Since the URG algorithm works with categorical data, the answers to the questions were used as such, but the numeric data were transformed into 3 categories (low, medium, high) in order to correspond with the usual categorization physicians use in their everyday practice. The cut-off scores applied are customized for each patient and each parameter and can be selected by the physician. The latter can either choose specific values for each categorization or can rely on statistical information (like average, st. deviation, and percentage of values in a normal distribution) provided by the tool in order to define the ranges of the categories in a uniform way (Figure 4).



Ranges Declaration

Statistical Information for the Weight

Minimum: 73 Average: 79,26
Maximum: 86,19 St. Deviation: 3,55

Define Ranges:
73 - 77,42 - 81,1 - 86,19

Width of the middle Range: 40 %

Figure 4. Definition of categories

The user can also set the probability and conditional probability thresholds that will decide which of the rules mined will be finally viewed. The results screen presents the corresponding rules in an easily comprehensible way along with a summary of the data, the categories and the metrics used.

6. Results

The standard metrics of 10% minimum probability and 75% minimum conditional probability were applied to generate rules for the values transmitted by the 11 CHF patients of the study. For comparison purposes, the data of each numeric variable of each patient was categorized so that the middle 40% of the values, according to the normal distribution, would fall in the medium category, while the low and high categories would cover 30% of the values respectively.

Using these metrics a mean of 3668.4 rules/patient were mined (ranging from 2606 to 5528). The number of rules depended on the variance of the data transmitted by each patient, especially their answers to the questions, while it was not correlated with the number of contacts, due to the categorization method.

The most relevant, i.e. logical or interesting, of those rules were selected by one of the authors, a medical doctor. Some of the induced rules represent everyday knowledge in medicine or can be explained by common sense. For example, in patient number 1 (#1) low systolic blood pressure (SBP) occurred together with low diastolic blood pressure (DBP) or with low pulse, as well (CF 75% and 81.8% respectively); while patient #4 used to report together breathlessness during the night with dyspnoea and tiredness during the day (CF of their combinations from 80% to 100%).

Other rules were more interesting, in the way that they predict subjective symptoms based on objective signs or vice versa for a specific individual, e.g. patient #7 reported feeling tired, when either his SBP, DBP or pulse were high (CF 87.8%, 86.7% and 86%) or a combination of the three at once. Patient #11 in the same way, complained about daytime dyspnoea whenever his DBP was up, his pulse was rapid or his feet where swollen (CF 78.6%, 75% and 90.9%). On the other hand, for patient #8 the elevation of her SBP could be “predicted” by her tiredness and the swelling of her feet (CF 88.9% for both), which of course tended to occur simultaneously (CF 80.6%). These rules may indicate to the patients a way to avoid or understand the appearance of disturbing symptoms or “foresee” relatively elevated values of their vital parameters.

Of course, given the limitations and specifics of the available data, and the biases of the expert choosing the importance of the rules mined, the induced rules should not be considered as applicable to all CHF patients. This paper should only be viewed as a methodological study suggesting how to use this new algorithm on similar databases.

7. Conclusions

Contrary to previous algorithms mentioned before, the ability of URG to generate associations in dynamic databases combined with its effective manipulation of missing data makes it ideal for medical applications, like home care centers, as it is already demonstrated. At the present time, work is done on several points to improve the use of rules in URG. One point is to try to determinate the optimum thresholds and categorization criteria. As the URG tool will become more automated, the physician will have the opportunity to concentrate more on the interpretation and evaluation of the rules. Another point will be the use of this algorithm to fill the missing values on newly entered data in dynamic databases, enabling URG to become a tool for the data cleaning step of the KDD process.

Acknowledgements

This work was supported in part by the EC projects IST-1999-13352-CHS, IST-2001-33369-PANACEIA-ITV and by the program PROMESIP funded by the Ministry of Education (EPEAEK I).

References

[1] Agrawal R, Imielinski T, Swami A. Mining Association Rules between Sets of Items in Large Databases. Proc. of the ACM Intl. Conf. on Management of Data, May 1993.

[2] Agrawal R, Swami A. Fast Algorithms for Mining Association Rules. Proc. of the 20th Intl.Conf. on Very Large Data Bases, September 1994.

[3] Veloso A, Possas B, Meira W, Cavalho M. Knowledge Management in Association Rule Mining. IEEE International Conference on Data Mining, 2001.

[4] Cheung D, Han J, Ng V. Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique. Proc. of the 12th Intl. Conf. on Data Engineering, 1996.

[5] Cheung D, Lee S. A General Incremental Technique for Maintaining Discovered Association Rules. Proc. of the 5th Intl. Conf. on Databases Systems for advanced Applications, 1997.

[6] Konias S, Bamidis P, Maglaveras N, Chouvarda I. Treatment of Missing Values Using Uncertainty in Medical Data. Proc. 7th International Conf. on the Medical Aspects of Telemedicine, Regensburg, September 2002.

[7] Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI. Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, Chapter 12, AAAI/MIT Press, 1996.

[8] Maglaveras N, Koutkias V, Chouvarda I, Goulis DG, Avramides A, Adamidis D, Louridas G, Balas EA. 'Home Care Delivery through the Mobile Telecommunications Platform: The Citizen Health System (CHS) Perspective', International Journal of Medical Informatics 2002; 68: 99-111.

Address for correspondence.

Sokratis Konias
Aristotle University of Thessaloniki
Lab of Medical Informatics, POB 323
Thessaloniki, 54124, Greece
E-mail address: sokratis@med.auth.gr