A Novel Approach for Incremental Uncertainty Rule Generation from Databases with Missing Values Handling: Application to Dynamic Medical Databases

Sokratis Konias[1], Ioanna Chouvarda[1], Ioannis Vlahavas[2], Nicos Maglaveras[1]

[1]Laboratory of Medical Informatics, Medical School, Aristotle University of Thessaloniki, P.O. Box 323, 54124 Thessaloniki, Greece

{sokratis, ioanna, nicmag}@med.auth.gr

[2]Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

vlahavas@csd.auth.gr

Corresponding Author: Nicos Maglaveras,

Tel: +302310999281

Fax: +302310999263

Abstract: Current approaches for mining association rules usually assume that the mining is performed in a static database, where the problem of missing attribute values does not practically exist. However, these assumptions are not preserved in some medical databases, like in a home care system. In this paper, a novel uncertainty rule algorithm is illustrated, namely URG-2 (Uncertainty Rule Generator), which addresses the problem of mining dynamic databases containing missing values. This algorithm requires only one pass from the initial dataset in order to generate the itemset, while new metrics corresponding to the notion of Support and Confidence are used. URG-2 was evaluated over two medical databases, introducing randomly multiple missing values for each record's attribute (rate 5-20% by 5% increments) in the initial dataset. Compared to the classical approach (records with missing values are ignored), the proposed algorithm was more robust in mining rules from datasets containing missing values. In all cases, the difference in preserving the initial rules ranged between 30% and 60% in favour of URG-2. Moreover, due to its incremental nature, URG-2 saved over 90% of the time required for thorough re-mining. Thus, the proposed algorithm can offer a preferable solution for mining in dynamic relational databases.

Keywords: Data Mining; Association Rules; Missing Values; Incremental Update

1. Introduction

Knowledge discovery is an important problem of increasing interest in a variety of fields where data are collected, stored and made widely available, like finance, marketing, medicine, engineering and other [1-5]. Mining association rules is one of the most popular tasks in KDD (Knowledge Discovery in Databases) procedure, introduced in [6]. The development of those rules was aiming at capturing significant dependencies among items in transactional datasets. The extraction of rules from market basket content is a classical example, where items are the objects bought and itemset the basket containing several items together. For instance, an association rule 'beer→chips (89%)' states that 89% of the times someone buys beer he or she also buys chips.

However, medical databases, and especially the ones related to telemedicine monitoring systems, have some unique features [7]:

- Medical databases are constantly updated. In such databases, for example in the database of a homecare system, the task of knowledge discovery is not merely a retrospective analysis task in a static completed database. On the contrary, while the database is being updated with patients' data, the medical rules and associations among patient's data have to be updated as well, in order to support the medical procedures of diagnosis and treatment on a personalized basis.
- The medical information collected in a database is often incomplete, e.g. some tests were not performed at a given visit, or imprecise, e.g. "the patient is weak or diaphoretic."
- Medical databases are often relational instead of transactional, and in that case, we encounter a problem of randomly missing values. Unlike the market-oriented

transactional databases, the medical databases, and specifically the electronic patient record databases, usually follow a relational structure, being organized around tables containing predefined lists of data that are requested for a patient, for which values may be present or randomly absent, and therefore the problem of missing values has to be addressed. A missing value may have been accidentally not entered, or purposely not obtained for technical, economic, or ethical reasons.

Therefore, a data mining algorithm addressing the problems arising in the field of dynamic medical databases would be of added value in the context of telemedicine systems.

1.1. Association rules

The problem of mining association rules can be formalized as follows. Let $I=\{i_1, i_2, \ldots, i_n\}$ be a set of items, while an itemset $X \subseteq I$ is a subset of items. Let D be a collection of transactions, where each transaction has a unique identifier and contains an itemset. The support of an itemset X in D, denoted as sup(X), is the ratio of number of transactions containing X to the total number of transaction. An association rule is an implication of the form $X \rightarrow Y$, where $X,Y \subseteq I$ and $X \cap Y = \varnothing$. The confidence of this rule is the ratio of $sup(X \cup Y)/sup(X)$. An itemset X is considered as frequent if $sup(X \cup Y)$ is no less than a user-specified minimum support threshold, while the corresponding rule "$X \rightarrow Y$" is strong if its confidence is also no less that a user-specified minimum confidence threshold.

Uncertainty rules have first been encountered in [6], where an in depth study of rule based systems, including the uncertainty management technique used in MYCIN was presented [8]. The structure of the rules “X→Y with CF” is very close to the association rules, as described above. For non-exact or relative sciences, like medicine, the main advantage of uncertainty rules compared to classical association rules is that they contain additionally a Certainty Factor (CF), which represents the efficiency of the corresponding rule.

Generally, association or uncertainty rules mining can be divided into two phases. During the first phase, the frequent itemsets are extracted, while during the second phase the strongest rules from all the frequent itemsets are mined. Since the second phase is considered less complex, almost all current studies for association discovery concentrate on efficient detection of the frequent itemsets [9-13]. However, most of them are based on the Apriori algorithm and require repeated passes over the initial dataset to determine the set of frequent itemsets, thus incurring high I/O overhead. Apriori algorithm first finds the sets with one item, based one these finds the sets with two items, and goes on in this manner. Therefore, the addition of a new record will affect the sets with 1 item, and analogously all the sets, requiring remining.

Most of the association rule algorithms were initially developed to capture significant dependencies among items in transactional databases (basket data analysis). In a transaction’s dataset, each transaction has a unique identifier and contains a set of items (i.e. variable length). In these cases, the missing values (i.e. values from the initial dataset that are not available) problem is not taken into account, and therefore most association rule algorithms do not provide for missing values.

In this paper, a novel algorithm, called URG-2 (Uncertainty Rule Generator), is proposed focusing on dynamic databases containing missing values. URG-2 needs one pass from the initial dataset in order to generate the itemset list and does not required fundamental itemset list reconstruction upon database update. Thus, it reduces significantly the I/O cost when applied to databases that are frequently updated, such as the patients' database in a medical monitoring system. In addition, URG-2 algorithm is applied to a database's table, where each record (i.e. fixed length and sorting) corresponds to a transaction. The difference with the transactional databases is that in a database's table each field corresponds to a specific predefined element (i.e. age, glucose and so on), whose value can be present or absent in a record; therefore the missing values problem can be defined in this case. Information of records containing some missing values is neither discarded nor estimated (e.g. by the most popular value or similar approaches which inevitably introduce noise), but internally treated by URG-2.

The remaining part of the paper is organized as follows. In Section 2 briefly related work is presented, while in Section 3 the URG-2 is demonstrated. In Section 4 the results and performances of experimental evaluations of our approach on two different medical databases coming from a home care system and the UCI Machine Learning Repository, are illustrated. In Section 5 a short discussion about the results is presented, and finally, Section 6 concludes with the future perspectives and improvements of the presented algorithm.

2 Related work

A few studies concerning association rules applied in the medical field can be found in the literature, probably due to the complexity of the field and due to the difference from the basket analysis approach classically adopted in the data mining domain. In order to deal with this problem, [14] proposed the transformation of the medical data from a relational to a transactional structure as a preparatory step, in order to apply the classical association rules algorithms.

As already mentioned, there are two basic features frequently met in medical databases: their dynamic nature and the randomly missing values. Regarding frequently updateable transactional databases, universal algorithms for efficient mining association rules have been proposed in the literature [13-18]. In [13], a new frequent pattern tree structure, namely FP-tree, is denoted, which is an extended prefix tree structure for storing compressed, crucial information about frequent patterns, thus achieving an improvement towards the problem of multiple passes from the initial dataset to create the itemsets, as most algorithms required. In [15] an alternative way for itemset generation is illustrated using an approximate tree structure similar to the previous study. The tree data structure is initially created by a single scan of the data and further updated whenever new records are added. Nevertheless, such tree structures are not able to deal with missing values, since all missing values would be mapped to a single new item. However, for a relational database, missing values may correspond to different attributes, and could not be considered as a single new item. An incremental technique for updating the sequential patterns is demonstrated in [16], which is quite different than updating association rules. In association discovery none of the new transactions appended is related to the old transactions, while in sequential pattern mining the newly appended database is merged together with the initial one into a data sequence whenever their ids are the same. However, it has to be noted that in all the above-mentioned studies, since they regard transactions as the items existing in the "basket", the missing values problem is not addressed.

On the other hand, universal approaches have been suggested, to address the problem created by missing values, either internally in the data mining technique or externally by replacing missing values prior to the data-mining step [17-20]. In [17], an approach towards increasing association rule's resistance against missing values is described, without attempting any estimation of the missing value, where the main idea is to split the transactional database under consideration into several valid databases (without missing values) for each rule separately. In [18], different approaches dealing internally with missing values are illustrated and compared, i.e. a) replacing the missing values of an attribute with most common value of that attribute found in the database, b) ignoring records that have at least one unknown attribute value, and other. Their approach focuses on finding a solution towards application of algorithms that are valid for complete datasets. In [19], a pre-processing method, called MVC (Missing Values Completion), for filling missing values using association rules, is proposed. Furthermore, numerous general well known approaches are illustrated in [20] towards estimating the missing values of a dataset and replacing them prior to the mining phase, i.e. a) estimate values using simple measures derived from means and standard deviations or regression, b) augment each feature with a special value or flag that can be used in the solution as a condition for prediction.

However, all previous algorithms are not capable to realize the aforementioned requirements of telemedicine databases, that is dealing with relational structures, dynamic databases, where records are frequently added, and addressing the problem created by missing values. The proposed URG-2 algorithm consists a novel approach appropriate for dynamic relational databases containing missing values, by single pass itemset generation and internal missing values treatment, hence reducing significantly the loss of information and the computational cost, offering a favourable data mining solution for medical databases such as the patients' database of a home care system.

3. The URG-2 algorithm

As all association rule algorithms, URG-2 consist of two main parts: The first part, namely IG (Itemset Generator), is a novel procedure that scans the data and finds the existing itemsets taking, into account the missing values information, while the second part, called RG (Rule Generator) consists a modification of the classic approach (where no missing values exist) in order to generate the uncertainty rules whose redefined support and confidence are higher than a user-specified threshold. The following subsections include a definition of the modified support and confidence criteria used in URG-2, taking into account the information of the records containing missing values, as well as a description of the IG and RG procedures.

3.1. Modified definitions for support and confidence

Uncertainty rules and association rules have been mentioned in literature since the '80s. The support and confidence measures are part of these techniques. The definition $X \cup Y$ refers to the set of common and non-common items of X and Y (i.e. their union). For instance, if X={a, b, d} and Y={a, b, c} then $X \cup Y$={a, b, c, d}. The role of support is to remove rules that do not apply often enough, if the frequency of the corresponding itemset is not greater than a user specified threshold, while confidence is needed to exclude rules that are not strong enough to be interesting (i.e. with lower than the specified threshold).

These measures have to be adapted when the dataset contains missing values. The main idea of addressing the difficulties coming from missing values is to ignore records containing missing values for each rule separately, which is based on the redefined metrics of support and confidence. Let B be a database and "X→Y with CF" an uncertainty rule, where X, Y are sets of items. In our approach, an attribute value is considered as an item.

Definition 1: We note B(X∪Y) the subset of B containing X and Y.

Definition 2: If an itemset X contains missing values for at least one item of X, then it is disabled for X in the database B. $B_{dis}(X)$ denotes the subset of B disabled for X.

Therefore in the proposed work the metrics support and confidence are not used according to their initial definitions [9] (mentioned in the introduction). New definitions have been proposed by [17], in order to include the missing values information (taking into account the two abovementioned definitions). These definitions are adopted in URG-2, however adapted to the database records instead of transactions. Specifically, in the definitions 3 and 4 the denominator includes additionally the term $B_{dis}(X)$ and $B_{dis}(Y) \cap B(X)$ correspondingly, which accounts for the records containing missing values.

Definition 3: The support S_X of an itemset X in a database B with missing values is denoted as:

$$S_X = \frac{|B(X)|}{|B| - |B_{dis}(X)|} \tag{1}$$

Definition 4: The confidence $C_{X/Y}$ or Certainty Factor of the rule "X→Y with CF" in a database B with missing values is denoted as:

$$C_{X/Y} = \frac{|B(X \cup Y)|}{|B(X)| - |B_{dis}(Y) \cap B(X)|} \tag{2}$$

An itemset X with support higher than a minimum threshold is frequent, while a rule "X→Y with CF" is efficient if its confidence is also higher than a specified minimum confidence. In this study, the certainty factor CF can be interpreted as the confidence metric.

Following, a very simple example is presented to illustrate the efficiency of the above novel metrics of support and confidence. In figure 1, the same database, without and with missing values, is shown. In database 2 we introduced one missing value to show how it would affect the metrics.

[insert figure 1]

For both databases (figure 1), the classical support and confidence, as well as their redefinitions (definitions 3 and 4) are calculated. The same thresholds for frequency and efficiency are considered (minimum support of 40% and minimum confidence of 75%). Among the itemsets generated using the classical metrics for database 1, itemsets {c,g} and {a,c} are also included, with supports 3/6 and 4/6 correspondingly. When database 2 is considered, itemset {c,g} has a support of 2/6 (less than 40%, thus not frequent) and itemset {a,c} has a support of 3/6 by use of the classical metrics. According to the redefinitions, support is 2/(6-1) and 3/(6-1) respectively, which are both higher than the minimum threshold of 40%. Furthermore, the corresponding rule "a→c", which is mined in database 1 with a confidence of 4/5, in database 2 applying the classical metrics, rule's confidence is reduced to 3/5 (less than 75%), whereas using the redefinition of confidence results in 3/(5-1), which is equal to the minimum threshold of 75%. Therefore, these definitions help preserve more patterns when missing values are also included in the database.

3.2. Itemset generation procedure

The scope of IG (Itemset Generation) procedure is to pre-process the database and create the itemsets with the information needed for the uncertainty rule's generation, needing only one pass from the initial database. Once the itemsets are created, there is no further need to access the original database again. IG procedure is capable for incrementally handling data containing multiple missing values by producing two itemset lists (IL_1 and IL_2). Besides the itemset list IL_1 containing elements without missing values the extra itemset list IL_2 is required in order to include the additional missing values' information.

During this phase, each record of the initial database is compared in pairs with each element of the itemset list IL_1. Itemsets not found in IL_1 are added in the list, while the already existing itemsets are updated. In case the current record contains missing values, a corresponding procedure is also executed for the second itemset list IL_2, which includes the missing values information. The main scheme of the IG algorithm is described in the flow chart of figure 2.

[insert figure 2]

Finally, figure 3 provides a simple example, where a new record containing missing value is currently updating both itemset lists, which are assumed to be previously generated.

[insert figure 3]

An important advantage of the IG procedure is that it creates a smaller itemset list than other algorithms based on the Apriori algorithm, by rejecting itemsets that are subsets of other identified itemsets occurring with the same frequency. For instance, let the itemset {a, c, e} be found in 5 records; then the itemset {a,c} will not be extracted, unless it is found in more than 5 records. Furthermore, in this example if {a} is found in 8 records and {a,e} in 6 records, then "{a,e}→{c}" has confidence 5/6, while the confidence of the redundant rule "{a}→{c}" is 5/8 (less than the previous). Thus, mining of redundant rules is avoided, since these rules would additionally have a confidence less or equal than the confidence of the more specific rules.

3.3. Rule generation procedure

The aim of the RG (Rule Generation) procedure, which follows the IG procedure, is to mine from the already generated itemsets the uncertainty rules whose support and confidence (definition 3 and 4) are higher than the predefined thresholds. Each itemset coming from the list IL_1 (without missing values information) is utilized to mine all the corresponding rules, while the needed information of missing values for definitions 3 and 4 is taken from the list IL_2. In figure 4 the main structure of the RG procedure is described as a flow chart. The symbols used in figure 4 have been defined in Section 3.1

[insert figure 4]

4. Experimental results and performance comparisons

In this section, a comparison between the URG-2 algorithm and the classical approach, where the itemset have to be regenerated in case new records are added into the database and records containing missing values are prior removed, is presented. All the experiments were performed on a 450 MHz Pentium III PC with 192 MB memory. Two medical databases have been used for the evaluation: a) a database which is a subset of the UCI Machine Learning Repository (Heart -Disease Database with 303 records and 17 attributes) and b) the CHS database, coming from the Laboratory of Medical Informatics in the Aristotle University of Thessaloniki, Greece (Citizen Health System Database with 141 records and 10 attributes). Both databases contain medical data including records with missing values and have a dynamic nature. Following the typical approach in the association rule literature, the variables have to be categorical. Therefore, the continuous values were first transformed to categorical before any further processing.

The comparison between the two approaches was based on the following criteria:

a) to what extent the algorithms have the capability to preserve the initial rules when mining a database containing missing values,
b) how much time overhead is required for updating the rules when database updating takes place, and
c) how much the total execution time is increased due to introduction of multiple missing values.

In Section 4.1 the second dataset is described, since it is not an already published database, while in Sections 4.2 and 4.3 experimental results of the URG-2 algorithm and a performance study are presented, correspondingly.

4.1. The CHS database

CHS database consists of records of diabetic and congestive heart failure (CHF) patients who participated in clinical trials within the CHS project between September 2001 and

January 2003 [21]. CHS is a home care system built around an automated contact centre functioning as a server. Patients' communication with the contact centre is supported by a variety of interfaces, like public telephone, Internet or a mobile device. In this project, patients record the values of their vital parameters, such as pulse or blood pressure, (continuous variables) with the help of electronic microdevices and transmit them to the contact centre along with some yes/no answers to simple questions regarding mostly the occurrence of certain symptoms (dichotomous variable). The main idea behind the clinical trials was that the frequent recording of vital signs and symptoms could help the medical personnel monitor the medical condition of the patients, regulate them more efficiently and eventually help avoid hospital readmissions.

Data entry procedure takes place by the patient via an unsupervised telemedicine application. Consequently, missing values are due mainly to improper use of the various interfaces or technical problems (e.g. telephone communication interruption) and are considered random. The chosen dataset consists of 141 congestive heart failures' records with 10 attributes, both numerical and categorical, which are shown in table 1. The records correspond to the contacts of one particular patient chosen because a) the number of records was adequate b) initial missing values were few.

[insert table 1]

4.2. Experimental results

In these experiments, we used both databases described above intending to show the advantage of URG-2 when internally dealing with missing values. Initially, the URG-2 algorithm was applied to both datasets without missing values to obtain a rule set, called the original rule set. Following, we randomly introduced into the initial databases multiple missing values for each attribute (rate 5-20% by 5% increments) and in each case we rerun URG-2 twice, firstly after removing all records containing at least one missing value (classic approach) and secondly directly to the database containing records with missing values (new approach). A minimum support of 10% and a minimum confidence of 70% have been defined as thresholds for the CHS database, while a minimum support of 5% and minimum confidence of 70% have been applied to the database coming from UCI Machine Learning Repository. The thresholds were heuristically chosen, while the difference in the support threshold between the two databases is due to the bigger inhomogeneity of the data in the UCI compared to the CHS database.

Results of comparison between the classical and the new approach were compared regarding the percentage of retrieved rules, i.e. rules that were present in the original rule set, and the percentage of new rules, i.e. rules which did not appear in the original rule set (table 2). In all cases tested, the percentage of losing rules using URG-2 algorithm was reduced compared to the usual approach more than 30%, where records containing missing values were removed before applying URG-2 algorithm. Moreover, new rules discovered by URG-2 were connected with the missed rules, in other words new relations among items included in the missed rules were mined. For example, if the rule "$\{a, d\} \rightarrow \{e\}$" is lost, then a candidate rule could be "$\{a\} \rightarrow \{e\}$". Thus, the meaningfulness of the new rules mined by URG-2 is explained.

[insert table 2]

4.3. Performance study

In order to indicate the performance improvements of URG-2 algorithm in the procedure of mining uncertainty rules, we first compared the efficiency of incremental updating with that of remining from the beginning (classical approach). The time elapsed for the completion of the mining was used as performance metric. For that purpose, initially the performance of URG-2 applied to the entire databases was measured, and used as the basis for comparison. Afterwards, new records were added (rate 5-20% by 5% increments) and the additional time needed from URG-2 to update the mined rules was measured. The same thresholds of support and confidence, as in the previous Section have been chosen.

The results of this comparison between the time needed for remining the rules from the beginning and the additional time required for updating the already mined rules are demonstrated in table 3, for both databases used in this study. The difference between the ranges of time execution for the CHS and Heart Diseases databases could be explained by the bigger number of records (being almost double) and the bigger number of attributes in the latter. In any case, applying URG-2 when new records are added, instead of remining all the rules again, saves important execution time. In a realistic scenario, the insertion rate for the CHS home care would be less than 5% and the execution time saved would be more than 90%.

[insert table 3]

Furthermore, the total execution time of URG-2 and the classical approach have been calculated and compared, when these algorithms were applied to the databases after removing the records containing missing values. This procedure was repeated while randomly multiple missing values were introduced in a rate of 5-20% by 5% increments in both initial databases. Graphs of the execution times are presented in figure 5(a) and 5(b) for the two databases respectively. It is expectable that the classical approach needs less time, since it deals with fewer records, after removing incomplete records. The extra time required by URG-2 to deal with databases where multiple missing values exist, is small compared to the usual approach total execution time, and therefore not prohibitive at all for the mining phase.

[insert figure 5a][insert figure 5b]

5. Discussion

In medicine we are interested in creating understandable to human descriptions of medical concepts, or models. Machine learning, conceptual clustering, genetic algorithms, and fuzzy sets are the principal methods used for achieving this goal, since they can create a model in terms of intuitively transparent "if…then…" rules. Most current data mining (DM) applications in medical information systems (MIS) use only clean databases [22], while missing data can be a problem. However, dealing with missing data is crucial when applying DM in MIS. There are several ways to address the problem of missing data, e.g. statistical estimates, neural networks estimates, and simulation models. Most software packages e.g SAS (Statistical Analysis System), MetaNeural, and SPSS (Statistical Product & Service Solution) cannot accommodate missing data. IBM Intelligent miner® either ignores missing values or replaces them by a zero value, in order to make the dataset consistent. Furthermore, association rule algorithms, even when they are incremental and appropriate for dynamic databases are in general applied under the notions of transactional databases.

The application of a data mining procedure robust in cases of missing values and performing fast in dynamic relational databases can be of high interest in a telemedicine database, such as the CHS database depicted in the present work. In such a home care database URG-2 could assist physicians towards monitoring patient status. Specially, the medical areas of interest, as reported by clinicians who evaluated the use of URG-2 and reviewed the generated rules for CHS database, can be:

- Identification of conflicting patterns, which may indicate cases when patients do not give "honest" answers. For instance, a rule stating "*when someone obtained his/her medications, then his/her pressure was increased*" indicates that either there is a crisis situation for the specific patient or the patient's statements are not true.
- Detection of personalized thresholds for patients' vital signs, for example when diastolic blood pressure can be considered as high for an individual, according to his/her general medical status. These thresholds could be further used in alerting mechanisms.
- Possible relation among specific vital signs and symptoms or lifestyle on a personalized basis i.e. a patient reported feeling tired, when either his systolic blood pressure, diastolic blood pressure or pulse were high (CF 87.8%, 86.7% and 86%),

while another patient complained about daytime dyspnoea whenever his diastolic blood pressure was up, his pulse was rapid or his feet where swollen (CF 78.6%, 75% and 90.9%). On the other hand, for a third patient the elevation of her systolic blood pressure could be "predicted" by her tiredness and the swelling of her feet (CF 88.9% for both).

Therefore, the incorporation of URG-2 in the backend intelligence of a home care system could offer additional functionality towards mining useful rules concerning the behaviour or the medical status of patients on a personalized basis, which might be difficult to extract merely by visual observation.

6. Conclusion

Uncertainty rules are very useful towards exploring large databases in many fields, such as in medicine. Two important characteristics of such databases generated by real-world applications are the dynamic nature (constantly new records are added) and the frequency of missing values. Unlike aforementioned approaches found in literature, the novel URG-2 algorithm proposed in this paper, combines interesting features reducing significantly the lost information due to missing values and the execution time, when new records are added. Additionally, unlike most association rule algorithms, which are developed for transaction datasets, the URG-2 is appropriate for records of sorted and fixed length data (like a table in a database). The experimental results presented, show that this novel approach is much more robust for random missing values than the previous ones based on the classic approach (of prior ignoring records with missing values), while important execution time is saved in dynamic databases applications due to the incremental approach of URG-2.

The incorporation of URG-2 in a home care application could support the medical personnel towards recognizing personalized medical patterns and thus adapting the interventions and treatment on a patient basis, however, saving some of the time cost of visual inspection. A mechanism for the automated evaluation of the rules extracted and the pruning of the redundant/meaningless ones would then be of added value.

Considering the perspectives for future development, an interesting idea is the introduction of adaptive thresholds in the mining procedure, an issue of increased interest in the relative literature. In association rules mining, minimum support and minimum confidence thresholds are assumed to be available for mining the most frequent and efficient rules correspondingly. However, setting such thresholds is typically hard. In case the thresholds are set too high, almost nothing will be discovered; on the other hand, if they are set too low, too many rules will be generated. In order to avoid examining many mined rules, which are mostly redundant, a relatively high minimum support is typically selected. This way, efficient rules with lower support than the specified threshold are lost, as well. The prior definition of these thresholds is not trivial and has to take into account the characteristics of the dataset. A possible solution could be the use of the weighted average of the confidence according to their support. Another future perspective could be the development of a pre-processing method, combining the mined uncertainty rules from URG-2 using the uncertainty theory, in order to fill missing values in new added records, enabling URG-2 to become a method for the data-cleaning step of the KDD process.

Acknowledgments

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

References

1. Lavrac, N., 1999, Machine learning for data mining in medicine, *AIMDM'99, LNAI 1620*, pp. 47-62.

2. Chun, S.H., Kim, S.H., 2004, Data mining for financial prediction and trading: Application to single and multiple markets. *Journal of Expert Systems with Applications*, 26, 131–139.

3. Konias, S., Giaglis, G.D., Gogou, G., Bamidis, P.D., Maglaveras, N., 2003, Uncertainty rule generation on a home care database of heart failure patients. *Proceedings of Computers in Cardiology*, *IEEE Comp. Soc. Press*, September 2003, 30, pp. 765-768.

4. Konias, S., Bamidis, P.D., Maglaveras, N., 2004, An uncertainty rule generation algorithm for dynamic medical data. *Proceedings of 11th World Congress on Medical Informatics (medinfo)*, San Francisco USA, September 2004, In Press

5. Rygielski, C., Wang, J.C., Yen, D.C., 2002, Data mining techniques for customer relationship management. *Journal of Technology in Society*, 24(4), 483-502.

6. Buchanan, B.G., Duda, R.O., 1983, Principles of rule-based expert systems. *Journal of Advances in computers, ed. M. C. Yovits*, 22, 164-216.

7. Cios, K.J. Moore, G.W., 2002, Uniqueness of Medical Data Mining. *Journal of Artificial Intelligence in Medicine*, *26 (1-2), 1-24.*

8. Sortliffe, E. H., 1976, Computer-based medical consultations: MYCIN. *New York: Elsevier.*

9. Agrawal, R., Imielinski, T., Swami, A., 1993, Mining associations between sets of items in massive databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington, DC, May 1993, pp. 207-216.

10. She, L., Shen, H., Cheng, L., 1999, New algorithms for efficient mining of association rules. *Journal of Information Sciences*, 118, 251-268.

11. Mannila, H., Toivonen, H., Verkamo, A.I., 1994, Efficient algorithms for discovering association rules. *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases*, Seattle, Washington, July 1994, pp. 181-192.

12. Zaki, M.J., Parthasarathy, S., Ogihare, M., Li, W., 1997, New parallel algorithms for fast discovery of association rules. *Journal of Data Mining and Knowledge Discovery: Special Issue on Scalable High- Performance Computing for KDD,* 1(4), 343-373.

13. Han, J., Pei, J., Yin, Y., 2000, Mining frequent patterns without candidate generation. *Proceedings of ACM-SIGMOD*, Dallas Texas, May 2000, pp. 1-12.

14. Ordonez C., Santana, C.A., Braal L., 2000, Discovering interesting association rules in medical data. *Proccedings of ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery,* Dallas Texas, USA, May 2000, 78-85.

15. Amir, A., Feldman, R., Kashi, R., 1997, A new and versatile method for association generation. *Journal of Information Systems*, 2, 333-347.

16. Lin, M.Y., Lee, S.Y., 2004, Incremental update on sequential patterns in large databases by implicit and efficient counting. *Journal of Information Systems*, 29, 385-404.

17. Ragel, A., Crémilleux, B., 1998, Treatment of missing values for association rules. *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Berlin, April 1998; pp. 258-270.

18. Grzymala-Busse, J.W., Hu, M., 2000, A comparison of several approaches to missing values in data mining, *Proceedings of International Conference on Rough Sets and Current Trends in Computing*, LNAI 2005, October 2000, pp. 378-385.

19. Ragel, A., Crémilleux, B., 1999, MVC- a preprocessing method to deal with missing values. *Journal of Knowledge-Based Systems*, 12, 285-291.

20. Pyle, D., 1999, Data preparation for data mining, (San Francisco; Morgan Kaufmann).

21. Maglaveras, N., Koutkias, V., Chouvarda, I., Goulis, D.G., Avramides, A., Adamidis, D., Louridas, G., Balas, E.A., 2002, Home care delivery through the mobile telecommunications platform: The Citizen Health System (CHS) Perspective. *International Journal of Medical Informatics*, 68, 99-111.

22. Lee, I.N., Liao, S.C., Embechts, M., 2000, Data mining techniques applied to medical information. Journal of Med. Inform. Internet Med., 25 (2), 81- 102

Figure Legends

Figure 1. a) A database without missing values, b) The same database with a missing value (denoted as '?').

Figure 2. Flow chart describing IG procedure. MV stands for "Missing value".

Figure 3. An example of IG procedure application, showing how the itemsets are incrementally updated.

Figure 4. Flow chart describing RG procedure.

Figure 5. a) Total execution times over the CHS database, after introducing multiple missing values for each attribute. b) Total execution times over the Heart Disease database (UCI), after introducing multiple missing values for each attribute.

X_1	X_2	X_3
a	c	g
a	c	f
a	c	g
b	d	f
a	c	g
a	e	f
Database 1		

(a)

X_1	X_2	X_3
a	c	g
a	c	f
a	?	g
b	d	f
a	c	g
a	e	f
Database 2		

(b)

Figure 1.

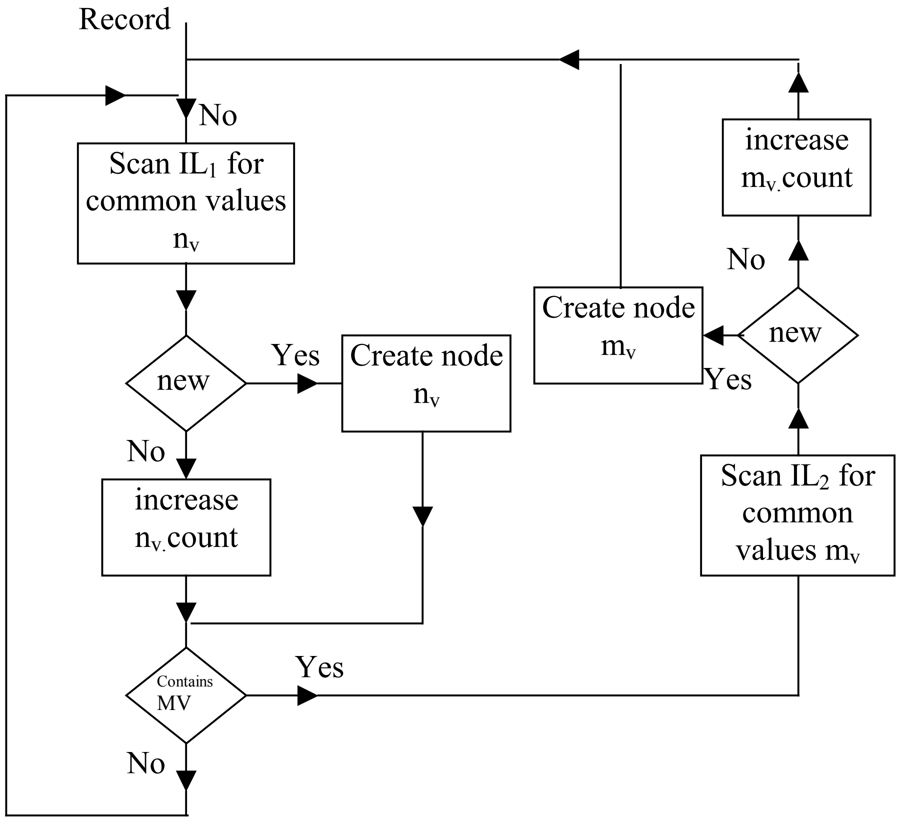

Record
No
Scan IL_1 for common values n_v
new
Yes
Create node n_v
No
increase n_v.count
Contains MV
Yes
No
Scan IL_2 for common values m_v
new
Yes
Create node m_v
No
increase m_v.count

Figure 2.

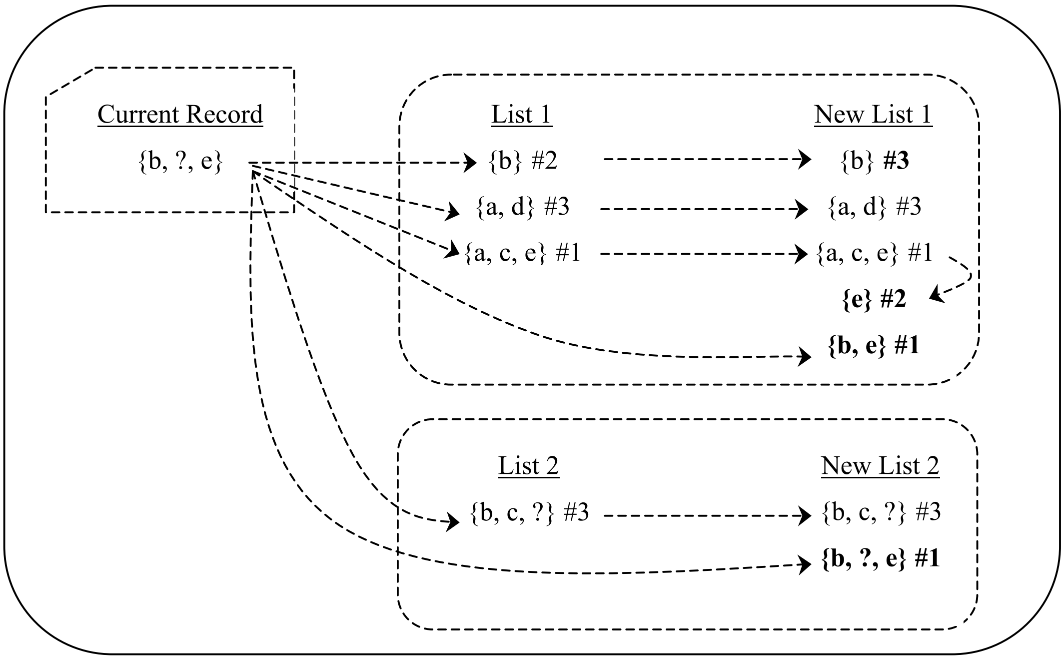
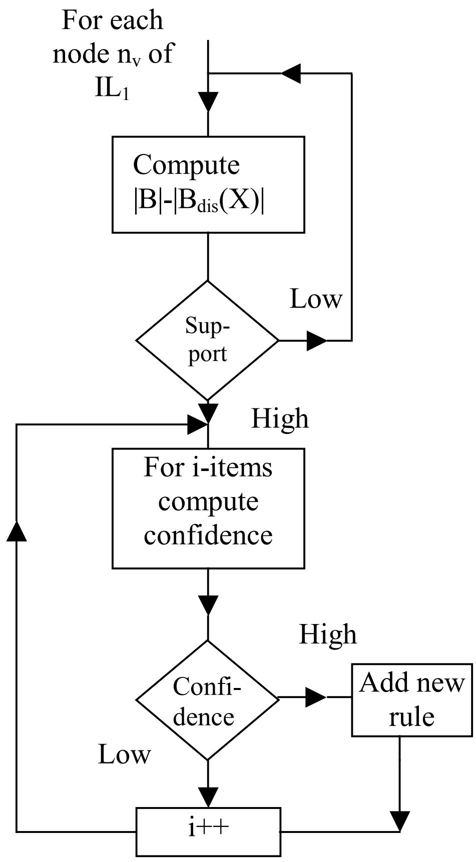

Current Record
{b, ?, e}
List 1
New List 1
{b} #2
{b} #3
{a, d} #3
{a, d} #3
{a, c, e} #1
{a, c, e} #1
{e} #2
{b, e} #1
List 2
New List 2
{b, c, ?} #3
{b, c, ?} #3
{b, ?, e} #1

Figure 3.



For each node n_v of IL_1
Compute $|B|-|B_{dis}(X)|$
Sup-port
Low
High
For i-items compute confidence
Confi-dence
High
Add new rule
Low
i++

Figure 4.

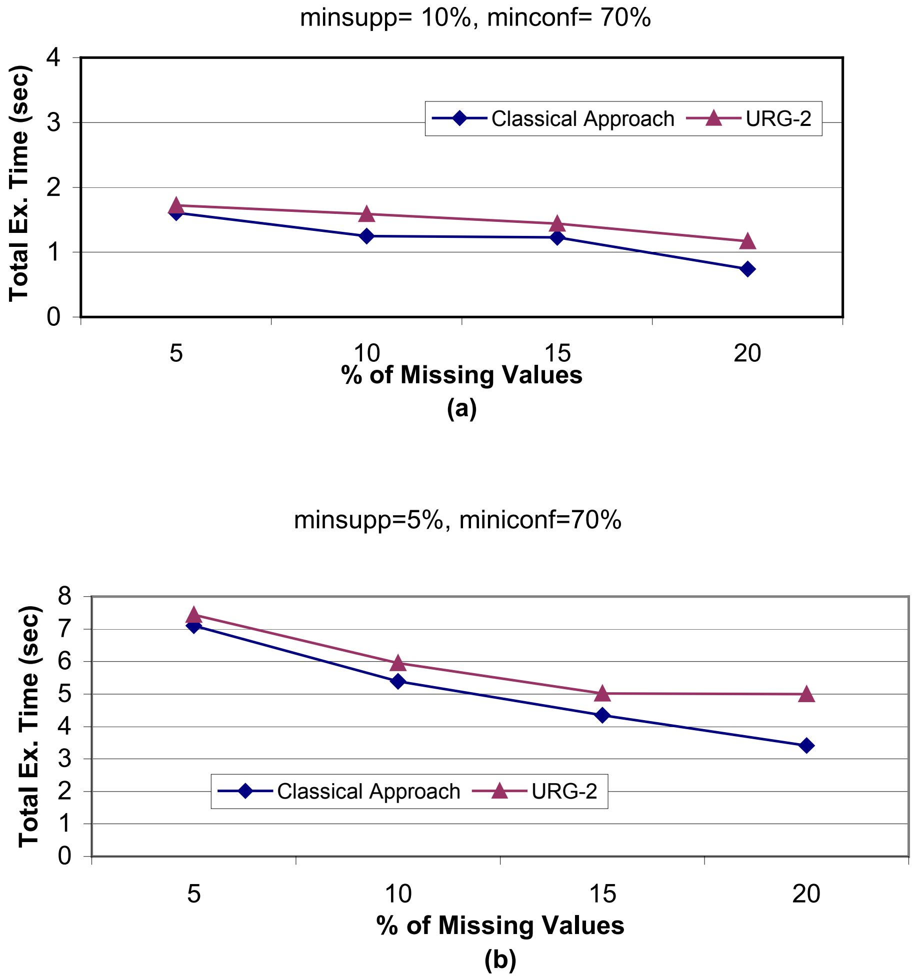

minsupp= 10%, minconf= 70%
Total Ex. Time (sec)
4
3
2
1
0
Classical Approach
URG-2
5
10
15
20
% of Missing Values
(a)
minsupp=5%, miniconf=70%
Total Ex. Time (sec)
8
7
6
5
4
3
2
1
0
Classical Approach
URG-2
5
10
15
20
% of Missing Values
(b)

Figure 5.

Table 1. Attributes included in the CHS database for the CHF patients.

Vital parameters (continuous)
Diastolic Blood Pressure
Systolic Blood Pressure
Pulse
Body Temperature
Weight
Questions asked (dichotomous)
Did you feel breathless during the night?
Are your legs swollen?
Do you feel more tired today?
Do you feel more dyspnoea today?
Did you take your Heart Failure medication?

Table 2. Comparison between the usual and the new approach concerning the percentage of mined rules from databases containing missing values with respect to the rules initially mined from the databases (before introducing missing values).

Percentage of multiple Missing Values[1] (%)	CHS Database[2]			
	Percentage of retrieved Rules		Percentage of new Rules	
	URG-2	Classical Approach	URG-2	Classical Approach
5	92.9	61.9	9.5	2.3
10	85.7	45.2	11.9	2.3
15	78.6	26.2	30.9	0.0
20	70.1	16.7	50.0	2.3
Percentage of multiple Missing Values[1] (%)	**UCI Machine Learning Repository Heart Disease Database[3]**			
	Percentage of retrieved Rules		Percentage of new Rules	
	URG-2	Classical Approach	URG-2	Classical Approach
5	81.2	49.4	35.8	9.2
10	77.8	29.0	54.9	3.7
15	74.7	15.4	67.3	1.8
20	64.2	6.8	240.7	1.2

[1] For each record there is possibility for more than one missing attribute value

[2] 10% minimum support and 70% minimum confidence

[3] 5% minimum support and 70% minimum confidence

Table 3. Comparison of the performance, in terms of execution time, between the approach of itemsets' incremental updating (implemented in URG-2 procedure) and that of re-mining.

Time in sec	Without MV[1]	5% MV	10% MV	15% MV	20% MV
CHS Database[2]					
Initial record set	2.06	1.72	1.59	1.44	1.17
5% additional	0.17	0.14	0.12	0.09	0.07
10% additional	0.32	0.28	0.23	0.17	0.13
15% additional	0.52	0.46	0.36	0.26	0.20
20% additional	0.67	0.59	0.47	0.33	0.26
Heart Disease Database[3]					
Initial record set	9.67	7.44	5.95	5.02	5.00
5% additional	0.93	0.67	0.52	0.40	0.33
10% additional	1.81	1.38	1.09	0.82	0.65
15% additional	2.96	2.03	1.63	1.22	0.93
20% additional	3.58	2.73	2.19	1.60	1.21

[1] Missing Values

[2] 10% minimum support and 70% minimum confidence

[3] 5% minimum support and 70% minimum confidence