Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 12 (17): 1792-1800, 2012
ISSN 1812-5654 / DOI: 10.3923/jas.2012.1792.1800

Constrained Association Rules for Medical Data

[1]Bakheet Aldosari, [2]Ghada Almodaifer, [2]Alaaeldin Hafez and [2]Hassan Mathkour
[1]College of Public Health and Health Informatics,
King Saud Bin Abdulaziz University for Health Sciences, Saudi Arabia
[2]College of Computer and Information Sciences,
King Saud University, Saudi Arabia

Abstract: The aim of the study is to develop a system for discovering interesting association rules from medical data sets for the purposes of prediction. The system addresses three problems: Medical data contain a combination of categorical and numerical attributes, processing bottlenecks are caused by large search spaces and the discovery of useful association rules that connect textual information with medical image features. The medical data set used comprised 80 patients' records each containing 142 attributes (textual information, numeric data and mammogram image features). The system developed used constraint-based association rule mining with a frequent pattern growth algorithm, with rules being filtered using support, confidence and lift. Association rules were constrained to have a maximum number of attributes and covers were additionally used to produce a summarization by introducing a greedy algorithm. The number of rules prior to the rule cover was 4607; after applying the rule cover summarization, the number of rules was decreased to 4170. When a 90% minimum confidence value was specified, just five association rules resulted from the 4170 available. The constrained association rule mining technique was successfully applied to medical data that contain image features. Interesting and concise association rules were able to be discovered for prediction purposes, which should assist clinicians in their decision making.

Key words: Knowledge discovery, data mining, medical informatics, association rules

INTRODUCTION

During the last decade, with the increasing use of digital data collection techniques and the expanding amount of data, Knowledge Discovery and Data Mining (KDDM) has become a very important research field (Kurgan and Musilek, 2006). Data mining refers to the process of finding patterns and relationships among data using various techniques and knowledge discovery refers to the results of such a process becoming useful knowledge in a particular domain of application (Kurgan and Musilek, 2006; Delen, 2009). Data mining techniques are used in various diverse domains such as market analysis, business management, space exploration, financial data analysis and medical analysis (Antonie, 2002; Doddi *et al*., 2001). In the medical field, databases now contain vast accumulated quantities of information about patients and their medical conditions, tests and treatments. The relationships and patterns within these data, revealed by using data mining, are providing new medical knowledge and contributing to better clinician decision-making and patient care (Doddi *et al*., 2001; Lee *et al*., 2000; Sriraam, 2006; Ting *et al*., 2009). Generic applications of KDDM in the medical domain include diagnosis, prognosis, treatment planning, avoidance of unnecessary treatment, intervention programs and medical guideline compliance and clinical concepts of interest include such items as presence of a symptom, presence of a medical condition, existence of a test/exam and existence of a risk factor (Delen, 2009; Rao *et al*., 2006).

Data mining techniques can be divided into the two broad categories of unsupervised and supervised. Techniques in the former category include clustering and association rules and those in the latter include classification, artificial neural networks and support network machines (Ting *et al*., 2009). Knowledge gained via such techniques is being increasingly utilized to improve the efficiency and quality of health care, in which domain the number of data are huge. Diagnosis of medical diseases is improving and clinician time is being freed up, by computerizing aspects of tasks such as diagnostics and guideline compliance using computer-processable rules (Rao *et al*., 2006; Razavi *et al*., 2008).

Veen (2008) discussed the challenges facing medical data mining, focussing on issues, difficulties and discovered rules. Commercial data mining tools such as XLMiner and knowledge studio were used to run different data mining techniques (neural networks, naive bayes, association rules mining and classification trees) on a

Corresponding Author: Alaaeldin Hafez, College of Computer and Information Sciences, King Saud University, P.O. Box 1178, Riyadh 1143, Saudi Arabia Tel: +9 6 65052953 67

thrombosis medical database. Based on his results, naive bayes, association rules in XLMiner and decision tree in knowledge studio were the operations that were found to be most adaptable for using diverse data types without limitations on variables categorical or continuous.

Cios and Moore (2002) emphasized the unique characteristics of medical data mining, including ethical, security-related and legal aspects of medical data mining. They conclude that data from medical sources are voluminous, heterogeneous and different in structure or quality, compared with other types of data. They suggest that methods of medical data mining should address the heterogeneity of data sources, data structures and the pervasiveness of missing values, for both technical and social reasons.

Abed and Zaoui (2011) investigated the problem of mining interesting association rules from large databases and suggested that background knowledge provides useful domain information that may help to subjectively evaluate and interpret the extracted rules. A framework was developed for mining interesting rules from the medical domain by incorporating background knowledge from the PubMed bibliographic database and UMLS. Background knowledge helped to distinguish obvious rules from spurious or probable ones and to aid in rule interpretation. The framework was tested by using electronic medical data records from the medical quality improvement consortium.

Siadaty and Knaus (2006) identified the problem of the bottleneck that can occur when pruning the usually large number of uninteresting rules and patterns. They used a dual-mining method, whereby the strengths of patterns (measured by association scores of tuples) mined from the database of interest (in the cited case the University of Virginia's clinical data repository) were compared with those of a relevant knowledge base (PubMed), using the same mining algorithm in each case. Differences between the pattern strengths highlighted the more novel or interesting patterns.

Williams (2003) developed an efficient association-rule mining algorithm called Partitioned Frequent Pattern growth (PFP growth). The algorithm is suitable for large data sets with long patterns. When parallel processing techniques are applied to the FP growth algorithm (Gupta and Garg, 2001), it reduces the processing bottleneck that arises when extremely large data sets are mined sequentially. Using test data extracted from medical images (mammograms), a typical example of such large data sets, Williams showed that the PFP growth algorithm improved the mining efficiency (speed) of the FP growth algorithm by 23-45%. The results comprised a set of association rules that provide a framework for an image classifier. The classification of new images with the resulting image classifier had a detection accuracy of approximately 80%.

Huang and Lee (2003) investigated the use of artificial neural networks and association rule mining (*a priori* rule algorithm) for mining medical images (digital mammograms) for anomaly detection and classification. The results showed that the two approaches performed well, with the classification accuracy reaching more than 70% for both techniques. Generating association rules is much faster than training a neural network. Therefore, in addition to its simplicity, using association rule mining is more efficient, compact and consistent for mining medical images than is using neural networks.

Antonie (2002) proposed a new classification method that uses association rule mining to discover interesting patterns in data that can be further used in building a classifier. First, input data were modelled as transactions. The association rules were generated from the transactional database using an *a priori*-like method and the set of generated rules or pruned rules was then used to classify new tuples. The classification method used was based on evaluating multiple sets of rules that match the new tuple to be classified, with the prediction being set by using the confidence of the rules as well as a dominance factor.

Ordonez *et al.* (2006) used constraint association rules in the medical domain to reduce the number of discovered patterns and provided a greedy algorithm to compute rule covers. Such rules are summarized to obtain a concise set of rules. The authors focused on using association rules for predicting combinations of several target attributes. The significance of association rules was evaluated by using support, confidence and lift. The proposed constraints included maximum association size, an attribute grouping constraint and an antecedent/ consequent rule filtering constraint. Ordonez *et al.* (2006) showed that the problem of mining association rules in a high dimensional data set with numeric and categorical attributes is challenging, due to the large number of patterns rather than to dataset size.

The number of examples of data mining applications to medical decision-making is increasing rapidly. Cerrito and Cerrito (Kharrousheh *et al.*, 2011), using SAS Enterprise miner and SAS Text miner, explored electronic records in a hospital emergency department with respect to the initial and changing diagnoses made during the course of assessment and to the triage values of emergent, urgent and non-urgent. Sriraam (2006) used association rule mining of a Malaysian hospital dataset to provide information to help clinicians decide on the level

of kidney dialysis treatment needed for particular patients. Razavi *et al.* (2008) used decision tree induction to discover patterns of non-compliance with a post-mastectomy radiotherapy guideline, information which should assist guideline authors and help improve the quality of oncological care. Using classification, clustering and association algorithms, Theodoraki *et al.* (2010) studied patterns in data regarding trauma cases in 30 Greek hospitals to determine factors and combinations of factors determining the outcome (the probability of death) of injured patients. Delen (2009) used a variety of data mining techniques, including artificial neural networks, to identify factors affecting survivability from prostate cancer using data from the US National Cancer institute. Mena *et al.* (2009) used a classification algorithm to discover new patterns of abnormal blood pressure variability as a cardiovascular risk factor, potentially providing decision support for cardiovascular disease prognosis. Imamura *et al.* (2007) developed a technique using association rule analysis for determining the three most useful clinical findings (diagnostic triads) for chronic diseases based on data mining of 477 patients' records.

However, despite the aforementioned contributions to knowledge discovery, the use of data mining techniques in the medical domain still has many challenges, given that medical data sets are usually very large, complex, heterogeneous, hierarchical and variable in quality (Delen, 2009; Ting *et al.*, 2009; Imamura *et al.* 2007; Veen, 2008). The purpose of the research presented in this study is to add further insight into how medical database information can be mined for knowledge discovery. The goal of this knowledge discovery effort is to discover significant and concise medical rules for the purposes of prediction to assist medical decision-makers. Medical data are often analysed with classifier trees, clustering, or regression (Lee *et al.*, 2000; Othman *et al.*, 2009). However, in that research, target classes are not considered and also there is not a global understanding of the data set. For those reasons, association rule mining is used here because of its combinatorial nature. Association rule analysis involves discovering association rules that frequently occur together in a given data set (Doddi *et al.*, 2001; Imamura *et al.*, 2007).

Therefore, the objective of this research is to design an efficient technique to discover significant and concise medical association rules to be used in obtaining medical knowledge. The proposed constraints-based medical data mining system could be used to solve some problems in the medical domain, where early detection of diseases and proper care management can make a difference. Three different medical problems are considered, which are: (1) Medical data sets contain patients' records that are a combination of categorical and numerical attributes; (2) Large search spaces in intermediate processing steps create bottlenecks and (3) The prediction of association rules that are useful and that connect the textual information of patients with their medical image features. These three problems are exemplified using patient record data that include mammogram images. The knowledge discovery process is more complex when image features are involved (Leu *et al.*, 2009).

MEDICAL DATA MINING PROBLEMS AND APPROACH TAKEN

A constraints-based medical data mining system is proposed which can be used in solving various problems in the medical domain.

Problem 1: A medical data set comprises a set of patients' records, where each record is a combination of categorical and numerical attributes of that patient (textual information, numerical data and extracted image features). This problem appears in medical data set representations that are not suitable for data mining. This problem could be solved by preparing such data before the data mining process starts. The data preparation process in the proposed system includes the following activities:

Data cleaning: Only complete and unified data are mined. By using a bin boundary technique, any outlier data are detected. A unified integer is used to represent any missing or outlier values. The resulting data set is reduced in size by discarding attributes and records that contain more than 50% outliers or missing data.

Data transformation: Since, association rule mining algorithms work only with binary variables, a mapping from the data set information to binary data is devised. Each binary variable is referred to as an 'item'. The medical data are transformed into a transaction format suitable for discovering association rules. To simplify the problem, all the attributes are treated as being either categorical or numerical, with image features being treated as numerical attributes. Data are mapped as follows: (i) Categorical attributes are mapped to items by associating an integer to each different categorical value and (ii) Numerical attributes are partitioned into intervals; these intervals are indexed and the index is used as an item to generate rules.

Problem 2: A large search space is needed when applying association rule mining algorithms to medical data sets.

The association rule mining algorithm has two main steps:

- Search for items (attribute values) with frequency exceeding a user-defined threshold (the minimum support). Such sets are referred to as frequent item sets
- Discover all large item sets (patterns) with frequency exceeding a second user-defined threshold (the minimum confidence)

Medical data sets are large sets with high dimensionality. Most existing association rule algorithms require an intermediate stage, while searching for frequent item sets, for processing. For instance, *a priori*-like algorithms generate candidate item sets. The search space for large item sets grows exponentially when the number of items is increased.

These intermediate data sets often become huge, resulting in a bottleneck in processing. In addition, the required multiple scans of the entire database also contribute to the bottleneck as blocks of data are transferred from disk to memory and back. The frequent pattern growth (FP growth) algorithm provides a compact representation using an FP tree and drastically reduces the input-output bottlenecks by scanning the entire database only twice. In the first scan, it counts the large item sets of length 1. In the second scan, it builds the FP tree 0. The FP tree approach scales much better than the *a priori* because as the support threshold goes down, both the number and length of frequent item sets increase dramatically. The FP Growth algorithm has proven to be significantly more efficient than *a priori*-like algorithms in mining large data sets that have a large number of relevant attribute values. For these reasons, the FP Growth algorithm is adopted in the system presented herein.

Problem 3: The goal here is to discover association rules that are medically significant and interesting and which will predict the association between the textual information of patients and their medical image features. Association rules discovered by most algorithms that do not constrain associations are not useful because they may contain redundant or irrelevant information, or describe trivial knowledge. Here, a constraints association rule mining system is proposed to satisfy the goal. The resulting association rules must have the following characteristics: (1) Textual features appear only on the antecedent side of association rules, while image features could appear on either side, (2) For simple interpretation, association rules have only a few features, i.e., the number of items is limited and (3) There are interesting combinations between association rule items.

If the straightforward approach for association rule mining is used, these previous characteristics may not be able to be satisfied. The association rules algorithm could be modified as follows: (1) Evaluate the significance of association rules using support, confidence and lift, (2) Constrain the length of association rules to reduce the number of discovered frequent patterns by using the maximum association size constraint and the antecedent/consequent rule filtering constraint and (3) Summarize discovered association rules to obtain a concise set of medical rules by using rule covers with a greedy algorithm.

DESCRIPTION OF THE MEDICAL ASSOCIATION RULE MINING SYSTEM USED

System specification: The system works with a specific format of metadata to simplify the resulting medical rules interpretation. For each attribute, the following fields are identified: Attribute ID#, Attribute description (attribute name), Attribute type (categorical or numeric), Attribute place in the rules (antecedent, consequent, or both of them), Attribute ranges number, which means (a) The number of discrete values of categorical attributes and (b) The number of bins (intervals) for numerical attributes and the description of each range (the range number, the minimum and maximum value of range (for numerical attributes only) and the meaning of the range).

The user must enter the following: The minimum support; the maximum association rule items and the minimum confidence.

System implementation: The system implementation consists of the following: (1) An initial attribute selection module that selects the irrelevant attributes manually, (2) A preparation module that prepares the medical raw data set by cleaning the selected attributes and transforms them to binary format suitable for the mining process, (3) A mining module that uses an FP growth algorithm and a constraints association rule generation process and (4) A summarizing module that calculates the rule cover.

Initial attribute selection for mining: The target data set structure stores all of a patient's information in a single record. Each record is a combination of both textual information (personal and medical) attributes and image feature attributes. The data set may contain hundreds of attributes, many of which may be redundant or irrelevant to the mining task. Only relevant attributes for the mining task are selected, so. that the mining task discovers those association medical rules that could assist medical

decision-makers. For examples, an attribute such as the patient's name is likely to be irrelevant to such decision-making, unlike attributes such as age or race, which bear some influence on the health and medical conditions of populations. Relevant attributes from the medical data set are selected manually. Such attributes include: personal information, such as age, gender, race and smoker/non-smoker and medical information such as blood pressure, heart rate and blood analysis.

Preparing the medical data set for mining

A cleaning and reduction of the data set: The medical data set is subjected to cleaning and reduction techniques applied for each attribute value, as follows:

- If there is a missing value, a global constant value (9999) is used to fill in the missing value. All missing attribute values are replaced by the same (9999) value, then this item is identified in the mining program as a missing value
- If there are noisy data (outliers):

- **1:** For categorical attributes: if the attribute value is not one of the attribute discrete values specified in the metadata, it is an outlier
- **2:** For numerical attributes: A binning technique (bin boundary method) is used to detect the outliers (outside of bin boundary) for the numerical attributes.

- The bin boundary method is as follows:

- **1:** The input metadata of the medical data set are read and a number of bins are constructed according to the number of ranges specified in the input metadata, with each bin representing one range
- **2:** If the medical attributes value is less than the minimum boundary of the first bin or larger than the maximum boundary of the last bin, then it is out of range
- **3:** Outlier values are replaced by the number (9999).

- Data inconsistency errors are not considered in the cleaning step because the data set is collected from one source (hospital database)
- The results of the cleaning process may identify some unavailable attributes with many missing values, or a high rate of outliers. Those attributes should be removed from the data set

The pseudo code of the proposed cleaning and reduction process is as follows:

Step 1: Scan the medical database vertically by checking attribute values (cleaning):

- The attribute value is missing
- For categorical attributes, the value is an outlier when it is not equal to (the attribute distinct values)
- For numerical attributes, the value is an outlier or out of range if it is less than the minimum boundary of the first bin or larger than the maximum boundary of the last bin (minvalue>value or maxvalue<value)

Step 2: Replace all missing or outlier values by 9999
Step 3: Increment the invalid counter by one
Step 4: An attribute is deleted if its invalid counter comprises more than 50% of its number of values (reduction)
Step 5: Scan the database horizontally (record by record) and attribute values as follows (cleaning): If the attribute value is 9999, then increment the invalid counter by one
Step 6: Delete the patient record if its invalid counter constitutes more than 25% of the number of attributes (reduction)

Transformation of the data set: Before applying the data mining methodology to extract useful rules from data, the data must be transformed into a form acceptable for the mining method. The "flattened table" format is most common and is required for most methods. In the flattened table format, each row represents an instance for training and/or testing a model (often containing relevant data for an individual patient); each column represents the values for a variable across the instances. Due to the characteristics of medical practice and to the data structures required in medical databases, some form of data transformation is invariably necessary to convert data from their original format to a flattened table.

Since, the association rule mining algorithm works only with binary variables, a mapping from the database information to binary data is devised. Each binary variable is referred to as an 'item'. The medical data contains categorical, numerical and image feature attributes. All attributes are treated as either categorical or numerical, with image feature attributes being treated as numerical attributes. In short, each transaction is a set of items and each item corresponds to the presence or absence of one categorical value or one numeric interval.

Mining the medical data set (constraints association rule mining): In this study, constraint-based association rule mining has been selected and the frequent pattern growth

(FP growth) algorithm has been chosen. A new significance measure for association rules called lift (also termed 'interest') is used. Lift is defined as:

$$\text{Lift}(X \Rightarrow Y) = P(Y|X)/P(Y) = \text{confidence}(X \Rightarrow Y)/\text{support}(Y) \quad (1)$$

Lift measures how much the presence of Y depends on the presence or absence of X, or vice versa. Lift values greater than 1 indicate that the presence of Y depends on the presence of X. Support, confidence and lift were used as the main filtering parameters in the system used. Support was used to discard low probability patterns. Confidence was used to look for reliable prediction rules. Lift was used to compare similar rules with the same consequent and to select rules with higher predictive value. Confidence, combined with lift, was used to evaluate the significance of each rule. Several problems appear when trying to discover association rules in a high dimensional data set. For each problem, a solution is proposed that is generally in the form of a constraint.

Association size: Associations and rules that involve many items are hard to interpret and can potentially generate a very high number of rules. Therefore, there should be a default threshold for association size. Most approaches are exhaustive in the sense that they find all rules above the user-specified thresholds, but in the particular domain of interest, such an approach produces a huge number of rules. The largest number of discovered associations is a practical bottleneck for algorithm performance. In the case presented, even k>5 produces too many rules, rendering the results useless. Another reason to limit the size is that if there are two rules $X_1 \rightarrow Y$ and $X_2 \rightarrow Y$ so that $X_1 \subset X_2$, the first rule is more interesting because it is simpler and it is more likely to have higher support. If $Y_1 \subset Y_2$ and $X \rightarrow Y_1$ and $X \rightarrow Y_2$, then the second rule is likely to have higher confidence but lower support.

Items restricted to appear only in the antecedent, only in the consequent, or in either place: By rule definition, an item appears only once in a rule and therefore appears either in the antecedent or in the consequent of the rule. Given the interesting rule $X \rightarrow Y$, no matter where an item appears, the association $X \cup Y$ must be a frequent item set because this association is precisely the rule support, but where the item appears prunes out many uninteresting rules that have useless combinations of items. In other words, support is still needed to prune uninteresting associations. Confidence is not enough on its own to prune out uninteresting rules because there may be many rules having high confidence containing forbidden items in the antecedent or in the consequent. Therefore, items need to be constrained to appear in a specific part of the rule.

The mining algorithm is applied in two phases:

Phase 1: Find all frequent item sets (using the FP growth algorithm). By definition, each of these item sets will occur at least as frequently as a user-specified minimum support count.

A constraint is added in this phase: The constraint is the user-specified maximum association size κ. Associations are generated up to size κ, eliminating the search for associations of size $\kappa+1$, $\kappa+2$ and so on. This constraint is simple, but essential in order to obtain simple rules and reduce output size. Note that each discovered rule will also have, at most, κ items.

Phase 2: Generate strong association rules from the frequent item sets. By definition, these rules must satisfy minimum support, minimum confidence and minimum lift.

The antecedent/consequent rule filtering constraint is added in this phase:

Let $C = \{c_1, c_2, \ldots c_p\}$ be a set of antecedent and consequent constraints for each attribute A_j. Note that constraints are specified on attributes and not on items. Each constraint c_j can have one of three values: 1 if item A_j can appear only in the antecedent of a rule; 2 if item A_j can appear only in the consequent; or 0 if item A_j can appear in either place.

So, a 0 value constraint is specified for the image feature attributes and a 1 value constraint for all other attributes.

Summarization of the resulting medical association rules: If the number of discovered rules is large, even having incorporated filtering constraints, interpretation becomes difficult. It is desirable to obtain a few representative rules so that many rules can be derived from them. Therefore, it is proposed that a cover of association rules with the same consequent be generated in order to form a rule summary that can be interpreted in less time. Given two rules with the same consequent:

R_1: $X_1 \Rightarrow Y$ and R_2: $X_2 \Rightarrow Y$, such that $X_1 \subseteq X_2$, it is said that R_1 covers R_2. Then, R_2 can be included and R_1 omitted. When there are several rules covered by R_2, this will produce a concise summary. This approach is applied to each set of rules with the same consequent but different antecedent item sets.

EXPERIMENTAL RESULTS

Target medical data set description: A medical data set consisting of 80 patients' records each with 142 attributes was chosen. Each patient profile contains personal information and some image-related information (date of study, film type) and the mammogram image feature variables (64 mean, variance, skewness and kurtosis values for right and left breasts, respectively) from the output of the image processing program.

For each attribute, the following were assigned: Its ID#, its description (the attribute name), its type (categorical or numerical), its place in the rules (antecedent, consequent or both), the number of ranges (the number of discrete values for the given attribute) and the description of each range (the range number, the minimum and maximum value of range and the meaning of the range). Each of the four image feature variables (mean, variance, skewness and kurtosis) has the same metadata fields.

Initial attribute selection: The goal is to discover the associations between the 128 mammogram image features and between the image features and other medical attributes. The medical data set consist of some attributes with no relevant value to the goal of the data mining. For instance, the file name does not provide any medical information for mining. Selection of relevant attributes was applied manually. In the experiments, 142 attributes were initially chosen for the mining and after the initial selection, 129 attributes were used. For space considerations, Table 1 reports only a small sample of the data.

Data set transformation results: The resulting transactional medical data set was converted into a format where each attribute becomes several binary variables as required by the data mining algorithm. This conversion resulted in 1157 binary variables for the attributes in the database. A binary vector representation would be inefficient for two reasons: (1) The dimensionality of the data, which is already high and the potential high number of categorical values per categorical attribute and (2) The number of intervals per numerical attribute would further increase the dimensionality of the data.

The resulting frequent item sets: If all 129 attributes had been used for the original data mining algorithm, there would have been a very large set of association rules found (approximately 2129). Therefore, the FP growth algorithm with the maximum association rule size constraint was used to reduce the number of possible associations and to reduce the number of discovered association rules. In the experiments, a 50% minimum support and a maximum of four association rule items were used. The algorithm found a large number, 26419, of frequent item sets.

The resulting medical association rules: The antecedent and consequent constraints in phase 2 were used and the greedy algorithm with minimum confidence 0.9 and a minimum lift of 1.0, in order to include only rules with high predictive value and to produce a manageable number of rules. The resulting number of rules prior to the rule cover was 4607; after applying the rule cover summarization, the number of rules was decreased to 4170, meaning that 437 rules were deleted. It is a good achievement

Table 1: A sample of initial attribute selection.

#	Attribute	Accepted	Reason for rejection
12	LEFT_MLO LINES 4616 PIXELS_PER_LINE 2952 BITS_PER_PIXEL 12 RESOLUTION 50 NON_OVERLAY	No	Relevant to image processing but not to the mining goal
13	RIGHT_CC LINES 4640 PIXELS_PER_LINE 2984 BITS_PER_PIXEL 12 RESOLUTION 50 NON_OVERLAY	No	Relevant to image processing but not to the mining goal
14	RIGHT_MLO LINES 4608 PIXELS_PER_LINE 2984 BITS_PER_PIXEL 12 RESOLUTION 50 NON_OVERLAY	No	Relevant to image processing but not to the mining goal
15	MeanRightbreast1	Yes	-
16	VarianceRightbreast1	Yes	-
17	SkewnessRightbreast1	Yes	-
18	KurtosRightbreast1	Yes	-
19	MeanRightbreast2	Yes	-
20	VarianceRightbreast2	Yes	-
21	SkewnessRightbreast2	Yes	-
22	KurtosRightbreast2	Yes	-
23	MeanRightbreast3	Yes	-
23	VarianceRightbreat3	Yes	-
24	MeanRightbreast1	Yes	-

Table 2: Association rules resulting from the mining process

Rule No.	Association rule
1	{KurtosRightbreast9=Averagely Peaked}, {KurtosLeftbreast8=Averagely Peaked}=>{KurtosLeftbreast4=Averagely Peaked}
2	{KurtosRightbreast9=Averagely Peaked}, {KurtosLeftbreast1=Averagely Peaked}, {KurtosLeftbreast8=Averagely Peaked}=> {KurtosLeftbreast4=Averagely Peaked}
3	{KurtosRightbreast9=Averagely Peaked}, {KurtosLeftbreast8=Averagely Peaked}, {SkewnessLeftbreast13=Average}=> {KurtosLeftbreast4=Averagely Peaked}
4	{KurtosRightbreast9=Averagely Peaked}, {KurtosLeftbreast5=Averagely Peaked}, {KurtosLeftbreast8=Averagely Peaked} => {KurtosLeftbreast4=Averagely Peaked}
5	{SkewnessRightbreast3=Average}, {KurtosRightbreast9=Averagely Peaked}, {KurtosLeftbreast8=Averagely Peaked} => {KurtosLeftbreast4=Averagely Peaked}

that rule covers produce a reduction in the number of rules in addition to the constraints. All the discovered frequent item sets were used to generate association rules. When a 90% minimum confidence value was specified, only five association rules resulted from the 4170 available. All resulting rules had 100% confidence and greater than 1.25 lift. The association rules thus produced are presented in Table 2.

CONCLUSION

This study used association rules for predicting combinations of several target attributes. The main reason behind using association rules for predictive purposes is that, in contrast to other data mining techniques, they are adequate to discover combinatorial patterns that exist in subsets of the data set attributes. In the proposed technique, association rules were filtered using support, confidence and lift. Lift helped to select rules with high predictive power and was used in conjunction with confidence to evaluate the significance of discovered rules.

In the study, several computational problems related to association rule mining were presented to motivate the introduction of search and filtering constraints. Such constraints have two main purposes: Finding only predictive rules and reducing the number of discovered rules to a manageable size. All attributes were constrained to appear either in the antecedent or in the consequent to discover predictive rules. Association rules were constrained to have a maximum number of attributes in order to produce fewer and simpler rules. But, because the desire was to discover very concise rules, the constraints on their own were not sufficient. Therefore, covers were used to produce a summary, by introducing a greedy algorithm that computed a cover for each set of rules having the same consequent. Covers included rules whose antecedent was a superset of the antecedent of simpler rules. This resulted in a smaller set of representative rules.

A simple approach was provided for solving the problem of mapping information from categorical and numerical attribute values to binary format. In the experiments performed with medical data sets, each patient's record comprised a combination of textual information and extracted image features. Constraints were shown to significantly reduce the number of discovered rules. In addition, interesting association rules were predicted with high confidence and high lift. Rule covers were used to summarize rules with a higher number of attributes and higher lift. The proposed system could be applied in other domains where prediction based on subsets of attributes is required for combinations of several target attributes.

Overall, it was found that the constrained association rule mining technique is of use when applied to mining medical data that contain image features. With this technique, interesting and concise association rules are able to be discovered for prediction purposes, which should help medical decision-makers to accomplish their work more quickly and with a greater degree of accuracy than may otherwise be the case.

ACKNOWLEDGMENT

The Authors wish to thank the Research Center-College of Computer and Information Sciences-King Saud University for their support and fund of this work (Grant # RC10 427-428).

REFERENCES

Abed, H. and L. Zaoui, 2011. Partitioning an image database by *K*-means algorithm. J. Applied Sci., 11: 16-25.

Antonie, L., 2002. Categorizing digital documents by associating content features. Masters Thesis, Computer Science Department, University of Alberta, Canada.

Cios, K.J. and G.W. Moore, 2002. Uniqueness of medical data mining. Artif. Intell. Med., 26: 1-24.

Delen, D., 2009. Analysis of cancer data: A data mining approach. Expert Syst., 26: 100-112.

Doddi, S., A. Marathe, S.S. Ravi and D.C. Torney, 2001. Discovery of association rules in medical data. Inform. Health Social Care, 26: 25-33.

Gupta, B. and D. Garg, 2001. FP-tree based algorithms analysis: FP growth, COFI-tree and CT-PRO. Int. J. Comput. Sci. Eng., Vol. 3.

Huang, P.W. and C.H. Lee, 2003. An efficient method of organizing bit-string signatures for searching symbolic images. Inform. Technol. J., 2: 159-172.

Imamura, T., S. Matsumoto, Y. Kanagawa, B. Tajima, S. Matsuya, M. Furue and H. Oyama, 2007. A technique for identifying three diagnostic findings using association analysis. Med. Biol. Eng. Comput., 45: 51-59.

Kharrousheh, A., S. Abdullah and M.Z.A. Nazri, 2011. A modified tabu search approach for the clustering problem. J. Applied Sci., 11: 3447-3453.

Kurgan, L.A. and P. Musilek, 2006. A survey of knowledge discovery and data mining process models. Knowledge Eng. Rev., 21: 1-24.

Lee, I.N., S.C. Lioa and M. Embrechts, 2000. Data mining techniques applied to medical information. Med. Inform. Internet Med., 25: 81-102.

Leu, J.H., C.Y. Lo and C.H. Liu, 2009. Development and test of fixed average K-means base decision trees grouping method by improving decision tree clustering method. J. Applied Sci., 9: 528-534.

Mena, L., J.A. Gonzalez and G. Maestre, 2009. Extracting new patterns for cardiovascular disease prognosis. Expert Syst., 26: 364-377.

Ordonez, C., N. Ezquerra and C.A. Santana, 2006. Constraining and summarizing association rules in medical data. Knowledge Inf. Syst., 9: 1-2.

Othman, M.L., I. Aris, S.M. Abdullah, M.L. Ali and M.R. Othman, 2009. Discovering decision algorithm from a distance relay event report. J. Applied Sci., 9: 3433-3453.

Rao, R.B., R. Rosales, S. Niculescu, S. Krishnan, L. Bogoni, X.S. Zhou and B. Krishnapuram, 2006. Mining medical records for computer aided diagnosis. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2006, Philadelphia, PA., USA., pp: 1-3.

Razavi, A.R., H. Gill, H. Ahlfeldt and N. Shahsavar, 2008. Non-compliance with a postmastectomy radiotherapy guideline: Decision tree and cause analysis. BMC Med. Inform. Decis. Making, Vol. 8.

Siadaty, M.S. and W.A. Knaus, 2006. Locating previously unknown patterns in data-mining results: A dual data- and knowledge-mining method. BMC Med. Inform. Decis. Making, Vol. 6.

Sriraam, N., V. Natasha and H. Kaur, 2006. Data mining approaches for kidney dialysis treatment. J. Mech. Med. Biol., 6: 109-121.

Theodoraki, E.M., S. Katsaragakis, C. Koukouvinos and C. Parpoula, 2010. Innovative data mining approaches for outcome prediction of trauma patients. J. Biomed. Sci. Eng., 3: 791-798.

Ting, S.L., C.C. Shum, S.K. Kwok, A.H.C. Tsang and W.B. Lee, 2009. Data mining in biomedicine: Current applications and further directions for research. J. Software Eng. Appl., 2: 150-159.

Veen, R.S., 2008. Medical data mining issues and experiments. Masters Thesis, Computer Science Department, The American University, USA.

Williams, A., 2003. Mining association rules in medical image data sets. Masters Thesis, Computer Science Department, University of Manitoba, Canada.