Discovering Interesting Association Rules in Medical Data

Carlos Ordonez
Georgia Institute of Technology
Atlanta, GA, USA

Cesar A. Santana
Emory University Hospital
Atlanta, GA, USA

Levien de Braal
Georgia Institute of Technology
Atlanta, GA, USA

ABSTRACT

We are presently exploring the idea of discovering association rules in medical data. There are several technical aspects which make this problem challenging. In our case medical data sets are small, but have high dimensionality. Information content is rich: there exist numerical, categorical, time and even image attributes. Data records are generally noisy. We explain how to map medical data to a transaction format suitable for mining rules. The combinatorial nature of association rules matches our needs, but current algorithms are unsuitable for our purpose. We thereby introduce an improved algorithm to discover association rules in medical data which incorporates several important constraints. Some interesting results obtained by our program are discussed and we explain how the program parameters were set. We believe many of the problems we come across are likely to appear in other domains.

1. INTRODUCTION

Data Mining is an active research area. One of the most popular approaches to do data mining is discovering association rules [1, 2]. Association rules are generally used with basket, census or financial data. Medical data is generally analyzed with classifier trees, clustering, or regression. For an excellent survey on these techniques consult [12].

In this work we explore the idea of discovering association rules in medical data, which we believe to be an untried approach. One of the most important features of association rules is that they are combinatorial in nature. This is particularly useful to discover patterns that appear in subsets of all the attributes. However, most patterns normally discovered by current algorithms are not useful since they may contain redundant information, may be irrelevant or describe trivial knowledge. The goal is then to find those rules which are medically interesting besides having minimum support and confidence. In our research project the discovered rules have two purposes: validate rules used by an expert system to aid in heart disease diagnosis (PERFEX [11]) and discover new rules that relate causes to heart disease and thus can enrich the expert system knowledge. At the moment all rules used by our expert system [11] were discovered and validated by a group of domain experts.

This paper is a continuation of previous joint research by Georgia Tech and Emory University to discover knowledge in medical data to predict heart disease [10, 6]. In [10] association rules are proposed and preliminary results are justified from the medical point of view. In [6] neural networks are used to predict reversibility images based on stress and myocardial thickening images.

Throughout the paper we try to provide a general framework to understand our approach. We believe many of the problems we are facing are likely to appear in other domains. As such this work tries to isolate those problems which we consider will be of most interest to the database community doing research on association rules.

1.1 Contributions and paper outline

The main contributions of our work include the following. Explain why mining medical data for association rules is an interesting problem. Justify the use of association rules for the medical domain. Phrase the problem in a general manner so that this work can be applied to other domains. Explain why many rules discovered by a straightforward approach are not useful as they contain redundant information, are trivial, are too complex or simply make no medical sense. Identify useful constraints to make association rules useful for the medical domain. Propose an algorithm to discover constrained association rules with very low support and fairly high confidence. Identify open problems that require further research.

Paper outline. Section 2 states the definition of association rules and addresses the problem of mapping medical data to binary attributes to be treated as items. We use a small example to motivate the use of association rules in the medical field and explain the kind of rules we are after. Section 3 outlines the classical algorithm to mine association rules, explains the main difficulties encountered using association rules, discusses useful constraints and presents an improved a-priori algorithm. Theoretical results and related work are briefly explained. Experimental results with medical data sets are described in Section 4. Section 5 contains the conclusions of this paper.

2. DEFINITIONS, DATA MAPPING AND INTERESTING RULES

2.1 Association rules

Here we give the classical definition of association rules. Let $\{t_1, t_2 \ldots t_n\}$ be a set of transactions, and let $\mathcal{I}$ be a set of items, $\mathcal{I} = \{i_1, i_2 \ldots i_m\}$. An *association rule* is an implication of the form $X \Rightarrow Y$, where $X, Y \subset \mathcal{I}$, and $X \cap Y = \emptyset$. X is called the antecedent and Y is called the consequent of the rule. In general, a set of items, such as the antecedent or the consequent of a rule, is called an *itemset*. Each itemset has an associated measure of statistical significance called *support*. For an itemset $X \subset \mathcal{I}$, *support*$(X) = s$, is the fraction of transactions in the database containing X. The rule has a measure of strength called *confidence* defined as the ratio *support*$(X \cup Y)$ / *support*(X).

The problem of mining association rules is to generate all rules that have support and confidence greater or equal than some prespecified minimum support and minimum confidence thresholds, respectively.

2.2 General description of our medical data

The medical data set we are mining describes the profiles of patients of a hospital being treated for heart disease. Each record corresponds to the most relevant information of one patient. This profile contains personal information such as age, race, smoker/non-smoker. Measurements on the patient such as weight, heart rate, blood pressure, are included. Preexistence or existence of certain diseases are stored. The diagnostics made by a medical doctor or technician are included as well. Time attributes mainly involve medical history dates. Then we have a complex set of measurements that estimate the degree of disease in certain regions of the heart, how healthy certain regions remain, and quality numbers that summarize the patient's heart effort under stress and relaxed conditions. Finally imaging (perfusion) information from several regions of the heart is stored as binary data. The image data is just a summarization of the heart divided into a few regions; these number of regions varies between 3 and 32. As we can see this type of data is very rich in information content.

2.3 Mapping attributes

Our medical data has to be transformed into a transaction format suitable to discover association rules. The medical data contains categorical, numerical, time and image attributes. To make the problem simpler we treat all the attributes as being either categorical or numerical. Let $A_1, A_2, \dots A_p$ be all the attributes, let $R = \{r_1, r_2, \dots r_n\}$ be a relation with n tuples whose values are taken from $A_1 \times A_2 \times \cdots \times A_p$, where A_i is either categorical or numerical. The data set size is n and its dimensionality is p. In the process described below we take one attribute at a time and map it to a series of consecutive integers that we will treat as items.

Here we explain missing information management. We reserve the first available integer to missing information for both categorical and numerical values for each A_i. It is important to create an item for missing information in each medical variable for two reasons: missing values are common and mapping would be incorrect without them. Besides there is an interest by doctors in analyzing missing information to track errors. This treatment of missing information is not complete. In some cases a missing value may mean that the person has no disease whereas in other cases it may be inapplicable or not available. For some of the records almost all fields have missing information and then it becomes a challenge to get reliable rules involving them. Also, not all attributes are equally likely to have missing information. In any case, since our data sets are so small and it is important to take into account every sample we do not discard records which have many missing values. We need to conduct further research to find rules which involve missing values.

Gender	Age	Smokes	LAD %	RCA %
F	53	Y	85	100
M	62	N	80	0
M	75	Y	70	80
M	73	Y	40	99
M	66	N	50	45

Table 1: Original medical data

M	F	$Age < 70$	$70 \leq Age$	S=Y	S=N	$LAD < 50$	$50 \leq LAD < 70$	$70 \leq LAD$	$RCA < 50$	$50 \leq$RCA
1	2	3	4	5	6	7	8	9	10	11

Table 2: Mapping table

The data table has categorical attributes that are easily mapped to items by associating an integer to each different categorical value. Each categorical value is a good candidate to appear in a rule. Obviously this may be a problem if the cardinality of the attribute domain is high; but that is uncommon with medical data. Binary attributes are a special case in which sometimes both the 1 and 0 occurrences may be interesting or only either of them is interesting. So we assume the medical doctor decides which categorical values are relevant.

The second important type of attributes is numerical. To simplify the problem time and image attributes are uniformly treated as numerical attributes. To use association rules on numerical data attributes must be partitioned into intervals, those intervals are indexed and the index is used to generate rules. An important work that deals with this problem is [20]. The authors prove equidepth partitioning minimizes information loss. In our case this technique was not used because medical doctors tend to have an idea how to partition ranges. That is, intervals have some semantics. Example: there are generally known cutoff points for weight to classify an adult patient as overweight or not. In a similar manner there are certain conventions to consider a person young, adult or elder with respect to age. Also, since these ranges are manipulated by domain experts they prefer to partition quantitative attributes into a few intervals to make result interpretation easier. However, an automatic partitioning could be helpful. So we do not discard that this task could be automated if it could be related back to domain knowledge.

In short, we assume the domain expert maps each attribute A_i to a series of items. For each categorical attribute A_i all relevant categorical values become an item including missing information. Each numerical attribute A_i is partitioned into a number of intervals and each of them is mapped to an item. This mapping algorithm discretizes the problem and does not increase data dimensionality.

2.4 Example

Here we provide a small example with 5 medical fields and 5 patients ($p = 5, n = 5$). Table 1 contains the original data, table 2 has all medical attributes mapped to items and and table 3 contains the data converted to items in transaction format. We also refer to the mapping table as the trans-

A'_1	A'_2	A'_3	A'_4	A'_5
2	3	5	9	11
1	3	6	9	10
1	4	5	9	11
1	4	5	7	11
1	3	6	8	10

Table 3: Mapped medical data to items

lation table. Note that identifying attributes such as name or SSN are not included because they are irrelevant to the mining process, and more importantly, there exist privacy policies which restrict access to us. *Age* is a numerical attribute and *Gender* and *Smokes* are categorical attributes. The columns LAD and RCA store the percentage of heart disease about specific arteries obtained from complex measurements. Their meaning is the following: LAD means Left Anterior Descending artery, RCA stands for Right Coronary Artery. The value 50% is a commonly accepted cutoff point to consider an artery diseased or not. However, not all arteries are equally important. In this case LAD is more important than RCA and that is why it is divided into 3 ranges to analyze it in greater detail.

By following the mapping procedure described above we map each tuple in the original data to a tuple that is a transaction containing items. The resulting table will have the same number of columns as the original one, but each column will contain only integers which correspond to the indexes of the mapping process. We must stress that a binary vector representation would be inefficient for two reasons: the dimensionality of our data which is already high and the potential high number of categorical values per categorical attribute and the number of intervals per numerical attribute would increase it further.

2.5 Discovering interesting association rules

The reader is asked to read this subsection carefully as it provides the motivation for the algorithm we will describe in the next section. Several medically important association rules are discussed as well as rules which are not interesting. In the previous example the last two columns in the original table contain information about the degree of heart disease. So, as the reader may guess we want to relate the age, gender, etc, with the possible absence or presence of disease in the heart. $70 \leq Age\ AND\ Smoke = Y\ AND\ Gender = M \Rightarrow 50 \leq RCA$ is an interesting rule with 40% support and 100% confidence because it provides a detailed profile of people having problems in the RCA artery. Although a rule like $70 \leq Age \Rightarrow Smokes = Y$ has 100% confidence, in general, it is not medically interesting. If the doctor wants to relate age to smoking habits then the rule can be helpful but if he is just trying to know the impact of these two factors together on heart disease then the rule is irrelevant. The rule $70 \leq LAD\ AND\ 50 \leq RCA \Rightarrow Smoke = Y$ is irrelevant because it does not make medical sense; we want to relate causes to disease but not viceversa. The association $70 \leq LAD\ AND\ 50 \leq RCA$ is irrelevant because medical doctors know this is a trivial case: when some heart region is diseased then the adjacent region has high likelihood of also being diseased; an example of such rule is $70 \leq LAD \Rightarrow 50 \leq RCA$. Therefore, any rule that includes these two items is of little interest. The rule $Male\ AND\ Age > 70\ AND\ Smoke = Y \Rightarrow 70 \leq LAD\ AND\ 50 \leq RCA$ is not interesting for two reasons. First, it involves two diseased heart regions and second, it is extremely selective as it involves all items and only one patient. Our data is high dimensional and then rules discovered by most current algorithms may involve many medical fields. Our experience has shown that rules with more than 5 medical variables are hard to interpret and slow down the mining task even though they are potentially relevant to the medical doctor.

3. DISCOVERING RULES IN MEDICAL DATA

Here we present our most important contributions. First we outline the widely known a-priori algorithm to find association rules, then we describe the problems we faced and finally an improved version of the algorithm.

3.1 Algorithm to mine association rules

The problem of mining association rules is to generate all rules that have support and confidence greater or equal than some prespecified minimum support and minimum confidence thresholds, respectively. The classical algorithm [2] has two phases:

- **Phase 1**: All itemsets that have support above the user specified minimum support are generated. These itemset are called the *frequent* itemsets. All others are said to be *infrequent*. We refer to itemsets having k elements as k-itemsets. First all 1-itemsets are generated as candidates and then those that turn out to be frequent after computing their support are used to generate and check for support 2-itemsets, then 3-itetmsets are generated and tested and so on. So the size of itemsets is incremented by one at each iteration and each iteration requires one scan over the transactions. This phase stops when there is no frequent itemset.
- **Phase 2**: For each frequent itemset, all the rules that have minimum confidence are generated as follows: for a frequent itemset X and any $Y \subset X$, if $support(X)/support(X - Y) \geq minimum\_confidence$, then the rule $X - Y \Rightarrow Y$ is a valid rule.

3.2 Problems with association rules

Here we summarize the main difficulties we have isolated so far trying to discover interesting association rules in the medical domain. For each problem we propose a solution that is generally a constraint. We describe the problems in an abstract manner. Hence, whenever we mention the word ”item” we are referring to a mapped medical field as described in the previous section. Whenever we say ”transaction” we mean the mapped record containing all medical information about one patient. And when we say irrelevant, trivial, redundant, complex and the like we imply that the given association or rule was not interesting for the domain expert (medical doctor).

Items that can appear only in the antecedent, only in the consequent or in either. Note that given the interesting rule $X \Rightarrow Y$ no matter where an item appears the association $X \cup Y$ must be a frequent itemset, but where the item appears prunes out many uninteresting rules. In other words, support is still needed to prune uninteresting associations but confidence is not enough 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.

Association size. Associations and rules that involve many items are hard to interpret and can potentially generate a very high number of rules. And further, they slow down the interactive process by the user. 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 our domain that produces a huge amount of rules. The biggest size of found associations is a practical bottleneck for algorithm performance. If for a given support the k-itemset X is frequent then all $Y \neq \emptyset$ s.t. $Y \subset X$ are frequent and then there are $O(2^k)$ frequent itemsets included. It is easy to see that no matter how efficient the algorithm the approach above will be slow for a large k. In our case even $k > 8$ produces too many rules rendering the results useless. Another reason to limit size is that if there are two rules $X_1 \Rightarrow Y$ and $X_2 \Rightarrow Y$ s.t. $X_1 \subset X_2$ the *first* rule is more interesting because it is simpler and it is more likely to have higher support. Or if $Y_1 \subset Y_2$ and $X \Rightarrow Y_1$ and $X \Rightarrow Y_2$ then the 2nd rule is likely to have higher confidence but lower support.

Associations having uninteresting combinations of items. This is the case where certain combinations are known to be trivial or have such a high support that do not really tell something new about the data set. Consider items i_j and $i_{j'}$. If the association $X_1 = \{i_j, i_{j'}\}$ is not interesting then any other association X_2 s.t. $X_1 \subset X_2$ will not be interesting. Therefore, many of the items (if not all) can be grouped by the domain expert to discard uninteresting associations. If no grouping is done that means that item i_j is always relevant no matter which other items $i_{j'}$ appear together with it. We assume small groups can be identified either automatically by running a straight association rules algorithm or by previous knowledge.

Maximal frequent itemset. There are schemes [3] that efficiently search for the longest frequent itemset without looking at all subsets. In our case limiting the size of associations automatically discards the possibility of finding all longest itemtsets. But all subsets are precisely what we are searching. Thus this strategy, even though being efficient, is not helpful in this domain.

Low support. It is well known that support is the performance bottleneck for association rules. It should be desirable to run the algorithm with a very low support so that repeated runs with decreasing supports are avoided. In the best case, it is desirable that the algorithm could run without support using the other constraints to prune the search space but avoiding finding rules that involve only one transaction.

Noisy data. This problem is never tackled by association rule programs since they are run over transactional data that is assumed to be complete and correct. However, in our case many transactions have missing information. The algorithm should avoid at all cost returning prediction rules involving missing information and the justification is evident: those rules are not reliable. However, there is an interest in tracking errors by mining tuples having missing values; we are currently investigating this problem.

High support. Even though the algorithm may prune out many rules by the above criteria, since we are working with high dimensional data there may still be lots of rules involving a few items having a high support. This problem is duly identified in [20] for quantitative association rules; and it basically appears because of the high number of combinations of partitioned intervals. So this idea is helpful: the algorithm should have a maximum support threshold $maxsupport$.

3.3 Theoretical results

Proofs are omitted for brevity. Extend items with two attributes as constraints. Let $\mathcal{I} = \{i_1, i_2, \ldots i_m\}$ be the set of items to be mined. Let $\mathcal{C} = \{c_1, c_2, \ldots c_m\}$ be a set antecendent and consequent constraints for each item. Each constraint c_j can have one out of 3 values: 1 if item i_j can only appear in the antecedent of a rule, 2 if it can only appear in the consequent and 0 if it can appear in either. We define the function antecendent/consequent $ac : \mathcal{I} \rightarrow \mathcal{C}$ as $ac(i_j) = c_j$ to make reference to one such constraint. Let $\mathcal{G} = \{g_1, g_2, \ldots g_m\}$ be a set of group constraints for each item; g_i is a positive integer if the item is constrained to belong to some group or 0 if the item is not group contrained at all. We define the function $group : \mathcal{I} \rightarrow \mathcal{G}$ as $group(i_j) = g_j$. Note that we make the simplifying assumption that items belong only to one group. We have faced cases in which one item can belong to several groups but that complicates the algorithm and obscures user understanding; this aspect requires further research. Let $X = \{i_1, i_2, \ldots, i_k\}$ be a k-itemset. X is said to be antecendent-interesting if $\forall i_j \in X\ ac(i_j) \neq 2$. X is said to be consequent-interesting if $\forall i_j \in X\ ac(i_j) \neq 1$. X is said to be group-interesting if $\forall i_j \forall i_{j'} \in X\ \ i_j \neq i'_j \Rightarrow group(i_j) \neq group(i_{j'})$.

Lemma 1 Itemset interestingness is antimonotonic in both $ac(i)$ and $group(i)$ constraints. □

Lemma 2 The $ac(i)$ constraints cannot be used to prune away associations because of the rule generation phase. □

3.4 Improved association rule algorithm

With all the above requirements we propose the following algorithm. All the basic notation and definitions are taken from section 2. Let Δ be the maximum number of items appearing in one rule. Let $X_1, X_2 \ldots X_M$ be all frequent itemsets obtained in phase 1.

- Phase 1:

  Generate all 1-itemsets as candidates and make one pass over $t_1, t_2, \ldots, t_n$ to compute their supports.

  for $k = 2$ to Δ do

  Extend frequent $(k-1)$-itemsets by one item belonging to any frequent $(k-1)$-itemset. Let $X = \{i_1, i_2, \ldots, i_k\}$ be a k-itemset. If $group(i_j) \neq group(i_{j'})$ and $group(i_j) * group(i_{j'}) > 0$ for $j \neq j' \wedge 1 \leq j, j' \leq k$ then X is a candidate. Check support for all candidate k-itemsets making one pass over the transactions. Those itemsets X s.t. $minsupport \leq support(X) \leq maxsupport$ will be the input for the next iteration. If there is no frequent itemset stop (sooner) this phase.

- Phase 2:

  for $j = 1$ to M do for $k = 1$ to M do

  Let $X = X_j, Y = X_k$,

  if $X \cap Y = \emptyset$ and $minsupport \leq support(X \cup Y) \leq maxsupport$ and $(ac(i) \neq 2\ \forall i \in X)$ and $(ac(i) \neq 1\ \forall i \in Y)$ and $(support(X \cup Y)/support(X) \geq minconfidence)$ then $X \Rightarrow Y$ is valid.

Lemma 3 Let X be a frequent k-dimensional itemset. Assume $\Delta < k$ then there are $2^k - \binom{k}{\Delta}2^\Delta$ pruned associations. □

Lemma 4 Let $X \Rightarrow Y$ be a valid rule where all items are $ac(i)$ constrained. Then there are $O(2^{|X|+|Y|})$ discarded rules. □

Lemma 1 is used to prune out associations based on the $group(i)$ constraint. Lemma 2 states that the algorithm cannot take advantage of $ac(i)$ constraints in Phase 1. Lemma 3 gives the number of pruned associations when the maximal frequent itemset is big. In our case this produces significant speedup to make computation more interactive. Lemma 4 gives an idea about the number of discarded rules; it is not a tight bound.

3.5 Related work

Literature on association rules has become extensive since their introduction in the seminal paper [1]. Given space constraints it is impossible to compare our approach against everybody else's. Most of the proposed approaches are used with basket data. Medical data sets are more complex and thus present many new challenges. This paper incorporates some ideas from our previous work to mine rules on segmented images [17]. Most papers published in the database literature concentrate on optimizing the first phase [8, 9, 13, 14, 15, 19, 18] but a few look at the problem of also improving rule generation (2nd phase) [7, 8, 16]. For instance, [15] proposes an algorithm to summarize associations when they are too many. [9] attacks the problem of inserting transactions on an already mined set and proposes an algorithm that incrementally maintains associations. [7] proposes a scheme to identify true correlations and [8] proposes a new metric called conviction to identify strong implications; we find this approach interesting.

Our work shares some similarities with [5, 16, 21]. In [21] the authors propose a few algorithms that can incoporate constraints to include or exclude certain items in the association generation phase; they focus only in two types of constraints: items constrained by a certain hierarchy [19] or associations which include certain items. This approach is limited for our purposes since we do not use hierarchies and excluding/including items is not enough to mine medically meaningful rules. The work which addresses the constraining problem in the most general way is [16]. Their approach based on succintness and 2-var constraints is different being more query oriented and not dealing with rule semantics, mapping, rule size, noisy data. Some authors support the idea of finding more complex rules discarding their simplifications for frequent itemsets [5].

Perhaps most of the effort on improving rules has been in developing new interestingness metrics; Bayardo et. al. [4] give a good overview on this theme and show that support and confidence are still fundamental metrics. So, instead of developing yet another metric we decided to constrain association rules to our needs, but maintaining their simplicity. Most of the improvements we propose were easy to incorporate but essential for our problem, and then do not change the basic framework described in [2].

4. EXPERIMENTAL EVALUATION

This work is preliminary. Therefore, we do not present extensive experiments to assess quality of results and performance. Because of space constraints we present experiment results in tabular form rather than graphs. Our experiments were run on a Sun multiprocessor computer having 4 Sparc Processors, each running at 125MHz. This computer has 128 Mb of main memory and a disk array with several gigabytes of available storage space. Our algorithm implementation was done in the C language.

4.1 Medical significance of association rules

The goal of the following experiment was to relate perfusion measurements to vessel disease (a.k.a. stenosis) to validate actual diagnosis rules used by an expert system [11]. In this case the purpose was not to find new rules but to confirm the validity of medical knowledge.

The data set consisted of $n = 655$ patients having 113 attributes. First of all the 12 most important medical attibutes were selected for mining ($p = 12$). These attributes included perfusion measurements for 9 regions of the heart and heart vessel disease for 3 vessels. The perfusion measurements quantify the deviation each heart region has from the corresponding region of a normal heart. The normal values for the 9 regions are taken as the means from which deviations are computed. The three vessels are identified by the acronyms LAD, RCA and LCX. LAD means Left Anterior Descending artery. RCA stands for Right Coronary Artery. LCX is the Left CircumfleX artery. The corresponding numerical attributes measure the percentage of disease in the vessel. Each of these numerical attributes referring to vessels were partitioned into $< 50\%$ and $\geq 50\%$. As we mentioned before a finer partitioning is used sometimes for more important arteries, such as LAD, but in this case it was not required.

The constraints for the association rule mining program were set as follows. The 9 perfusion regions were constrained to appear in the antecedent of the rules, i.e. $ac(i) = 1$. The 3 heart arteries were constrained to appear in the consequent of the rule, that is, $ac(i) = 2$. Since we wanted to study the three vessels separately they were further group constrained. That is, LAD, RCA and LCX are attributes belonging to the same group (say group 1) and therefore should never appear together (in pairs or the three of them). Rule size Δ was set to 2. We are after simple rules having a single item in both the antecendent and the consequent. Since our data set is small and we want to find any possible association involving 2 or more patients $minsupport$ was set to 0.2%. We want to find out confidence for rules so we set $minconfidence$ to 30%. Rules having a confidence lower than this value are considered irrelevant. Rules whose confidence is greater than 80% are considered reliable. Note that the values for the thresholds are very low. Without the constraints above it would be impossible to set the thresholds to such low values (the number of potential associations would be 2^{12}).

We ran the association rule program with the parameters previously described. The program discovered 89 associations and 27 rules in less than 5 seconds. All rules were interesting but not equally important. Here we explain in detail the three most relevant rules in order of significance. Each of these rules relate a different vessel. The first important rule is $SeptoAnterior \Rightarrow LAD \geq 50\%$ with $s = 18\%$ and $c = 80\%$. This rule has a normal support and high confidence, and confirmed medical opinion. When there is a heart defect in the Septo Anterior region then it is very likely that the LAD artery is diseased. The second rule was $InferoSeptal \Rightarrow RCA \geq 50\%$ which had lower support equal to 12% and lower confidence, equal to 65%. This rule also confirmed the expected relationship between the Infero Septal region of the heart and the RCA vessel. The third rule was $InferoLateral \Rightarrow LCX \geq 50\%$. This rule was surprising because it had a relatively higher support (20 %) but its confidence was only 53%. This confidence was much lower than what medical doctors expected and opened a new

minsupp	minconf	# assocs	# rules	time in secs
0.40	0.20	8	0	3
0.20	0.30	336	12	10
0.10	0.40	2939	60	95
0.05	0.50	20852	301	2018

Table 4: Experimental results with medical data

n	time in secs
655	11
6550	88
65500	643

Table 5: Times for large files

set of questions.

In short, our program discovered 27 rules out of which 3 were considered very important. Two of the rules confirmed prediction rules stored in an expert system. Their support and confidence measures were more or less close to that was expected. The third rule had higher support and lower confidence than expected. This rule surprised medical doctors and challenged the Confidence Factor (CF) used in a specific rule of the expert system. Thus data mining helped validating old knowledge.

4.2 Performance evaluation

Our medical data set only has $n = 655$ records but $p = 113$ dimensions. In practice, we work with projections of the data focusing in subsets of medical attributes as described in the previous subsection. So the following experiments are performed under pessimistic conditions. According to domain expert opinion we tentatively set values for $\mathcal{C}$ and $\mathcal{G}$ that are too long to describe here. We ran experiments with $maxsupport = 100\%$, threshold for association size $\Delta = 4$. We varied minimum confidence and minimum support to measure execution time. The program cannot be run without constraining because the number of association even for 60% *minsupport* was >500,000. Increasing Δ for low supports grows time/associations exponentially and it is not reported. Look at tables 4 and 5. Note the ladders of values for *minsupport* and $minconfidence$. These numbers show what we are after: rules with very low support having high confidence. Note the increasing number of associations. This points out to several optimizations that can be done in Phase 2.

Performance results with larger data sets. Setting for parameters was $minsupport = 0.20$, $minconfidence = 0.30$, $\Delta = 5$. Results are summarized in table 5. Getting access to more medical data sets is difficult for a number of reasons. Since all we have at this moment is a patient file containing only 655 records we decided to replicate it several times to get larger files. This keeps the problem complexity constant. Medical data sets, as described, are very different from typical transaction files used to benchmark association rules programs. Generating synthetic data to benchmark our approach is something worthy of future research. In any case fairly large medical data sets exist but they are not available to us. Also, it would be interesting to apply these ideas in other domains where large complex data sets are available.

5. CONCLUSIONS

Our research effort goes into making association rules more useful for medical data rather than proposing a novel scheme for mining them. One of the main appeals of association rules is their simplicity. Most of the improvements we propose are simple but useful. Association rules have a combinatorial nature; in that spirit we isolated those combinations that are interesting for our domain. We briefly address the problem of mapping complex medical data to items. We constrain associations to exclude certain combinations of items. We constrain rules to have certain items in the antecedent and certain items in the consequent. We limit rule size to get higher confidence and higher support rules. Our modified algorithm is then faster and finds fewer rules; but those rules tend to be concise and relevant.

Aspects which deserve further research. Automate mapping of attributes relating machine-generated partitions back to domain kowledge. Examine problems with noisy data more closely. Identify other useful constraints besides grouping and antecedent/ consequent. Extend grouping contraints to include several groups in a user-friendly manner. Run without support as a pruning strategy but reporting support and confidence always. Optimize the rule generation phase. We believe association rules can be used in more domains besides transactional data (basket or financial data). We presented a case in the medical domain in which association rules are useful, but we expect many of our research issues to appear in other complex real life domains.

Acknowledgments

This work was supported by grant LM 06726 from the National Library of Medicine.

6. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Conference*, pages 207–216, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB Conference*, pages 487–499, 1994.

[3] R. Bayardo. Efficiently mining long patterns from databases. In *ACM SIGMOD Conference*, 1998.

[4] R. Bayardo and R. Agrawal. Mining the most interesting rules. In *ACM KDD Conference*, pages 145–154, 1999.

[5] R. Bayardo, R. Agrawal, and D. Gounopolos. Constraint-based rule mining in large, dense databases. In *Proc. IEEE ICDE Conference*, 1999.

[6] L. Braal, N. Ezquerra, E. Schwartz, and Ernest V. Garcia. Analyzing and predicting images through a neural network approach. In *Proc. of Visualization in Biomedical Computing*, pages 253–258, 1996.

[7] S. Brin, R. Motwani, L. Page, and T. Winograd. Beyond market basket analysis: Generalizing association rules to correlations. In *ACM SIGMOD Conference*, pages 265–276, 1997.

[8] S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proc. ACM SIGMOD Conference*, pages 255–264, 1997.

[9] D. Cheung, J. Han, Vincent Ng, and C. Wong. Maintenance of discovered association rules in large databases: An incremental technique. In *IEEE ICDE Conference*, 1996.

[10] D. Cooke, C. Ordonez, E.V. Garcia, E. Omiecinski, E. Krawczynska, R. Folks, C. Santana, L. de Braal, and N. Ezquerra. Data mining of large myocardial perfusion SPECT (MPS) databases to improve diagnostic decision making. *Journal of Nuclear Medicine*, 40(5), 1999.

[11] N. Ezquerra and R. Mullick. Perfex: An expert system for interpreting myocardial perfusion. *Expert Systems with Applications*, 6:455–468, 1993.

[12] U. Fayyad and G. Piateski-Shapiro. *From Data Mining to Knowledge Discovery.* MIT Press, 1995.

[13] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *VLDB Conference*, pages 420–431, 1995.

[14] M. Houtsma and A. Swami. Set-oriented mining of association rules. Technical Report RJ 9567, IBM, October 1993.

[15] B. Liu, Wynne Hsu, and Yiming Ma. Pruning and summarizing the discovered associations. In *ACM KDD Conference*, pages 125–134, 1999.

[16] R. Ng, Laks Lakshmanan, and J. Han. Exploratory mining and pruning optimizations of constrained association rules. In *Proc. ACM SIGMOD Conference*, pages 13–24, 1998.

[17] C. Ordonez and E. Omiecinski. Discovering association rules based on image content. In *IEEE Advances in Digital Libraries Conference (ADL'99)*, pages 38–49, 1999.

[18] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules. In *VLDB Conference*, pages 432–444, September 1995.

[19] R. Srikant and R. Agrawal. Mining generalized association rules. In *VLDB Conference*, pages 407–419, 1995.

[20] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proc. ACM SIGMOD Conference*, pages 1–12, 1996.

[21] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In *Proc. ACM KDD Conference*, pages 67–73, 1997.