Mining Constrained Association Rules to Predict Heart Disease

Carlos Ordonez[1], Edward Omiecinski[1], Levien de Braal[1],
Cesar A. Santana[2], Norberto Ezquerra[1], Jose A. Taboada[3],
David Cooke[2], Elizabeth Krawczynska[2], Ernest V. Garcia[2]

[1]Georgia Institute of Technology [2]Emory University Hospital [3]Universidad de Santiago Compostela

***Abstract*—This work describes our experiences on discovering association rules in medical data to predict heart disease. We focus on two aspects in this work: mapping medical data to a transaction format suitable for mining association rules and identifying useful constraints. Based on these aspects we introduce an improved algorithm to discover constrained association rules. We present an experimental section explaining several interesting discovered rules.**

I. 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. On the other hand, medical data is generally analyzed with classifier trees, clustering, or regression, but rarely with association rules. A survey on these techniques is found in [10].

In this work we analyze the idea of discovering constrained association rules in medical records that include numeric, categorical, time and image data. This work is based on a long time joint research effort by Georgia Tech and Emory University to discover knowledge in medical data to predict coronary heart disease [7], [6], [5], [13], [14]. In [6] association rules are proposed and preliminary results are justified from the medical point of view. In [5] neural networks are used to predict reversibility images based on stress and myocardial thickening images. In [14] we explore the idea of constraining association rules in binary data and report preliminary findings from a data mining perspective.

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 discovered by algorithms that do not constrain associations are not useful because they may contain redundant information, may be irrelevant or describe trivial knowledge. The goal is then to find those rules that are medically significant or interesting, but which also have minimum support and confidence.

In our research project the discovered rules have two main purposes: validating rules used by an expert system to aid in diagnosing coronary heart disease (PERFEX [9], [7]) and discovering new rules that relate patient data to heart disease and thus can enrich the expert system knowledge base. At the moment all rules used by our expert system were discovered and validated by a group of domain experts, as described in detail in [9]. Since PERFEX is essentially a production rule system (i.e., composed of IF-THEN rules) used in conjunction with temporal and uncertainty reasoning models, the discovery of knowledge resulting from association rule mining would represent a potentially powerful and innovative way to validate and acquire knowledge to enhance the knowledge base. Importantly, the methods proposed herein are capable of inferring medical knowledge from a vast array of data that includes image and alphanumeric data that represent highly relevant, patient-specific clinical data (such as electrocardiographic information, patient history, symptoms and the results of clinical tests). Hence the methods described in this paper may provide a more efficient knowledge acquisition technique than classical approaches.

Throughout the paper we try to provide a general framework for understanding the approach underlying our research. We believe many of the problems we are facing (small data size, richness of content, high dimensionality, missing information, etc) are likely to appear in other domains. As such, this work tries to isolate those problems that we consider will be of greatest interest to the data mining community.

A. Contributions and paper outline

Our main contributions are the following. First, a justification is given for the use of association rules in the medical domain. We explain why mining medical data for association rules is an interesting and hard problem and we present the problem in an abstract manner so that this work can be applied to other domains. We introduce a simple mapping algorithm that transforms medical records into a binary format suitable to mine constrained association rules. We identify important constraints to make association rules useful for the medical domain and propose an algorithm to discover constrained association rules with very low support and relatively high confidence. Finally, we identify open problems that require further research.

This is an outline of the rest of the paper. Section II states the definition of association rules and describes our medical data. We use a small example to motivate the use of association rules in the medical field and explain the kind of rules that are sought. Section III addresses the problem of mapping medical data to binary attributes to be treated as items, emphasizes the main difficulties encountered using association rules, and introduces useful constraints for a customized A-priori type algorithm. Experimental results with medical data sets are described in Section IV. Finally, section V contains the conclusions of this article and directions for future research.

II. Definitions, data mapping and interesting rules

A. Association rules

The standard definition of association rules [1] is the following. Let $D = \{T_1, T_2, \ldots, T_n\}$ be a set of n transactions and let $\mathcal{I}$ be a set of items, $\mathcal{I} = \{i_1, i_2 \ldots i_m\}$. Each transaction is a set of items, i.e. $T_i \subseteq \mathcal{I}$. 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 $X \subset \mathcal{I}$, $support(X)$ is the fraction of transactions $T_i \in D$ such that $X \subseteq T_i$. The support of a rule is defined as $support(X \Rightarrow Y) = support(X \cup Y)$. The rule has a measure of strength called *confidence* defined as the ratio $confidence(X \Rightarrow Y) = support(X \cup Y)/support(X)$. The standard 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 [2].

B. Brief literature review

Literature on association rules has become extensive since their introduction in the seminal paper [1]. Our work shares some similarities with [4], [11], [17]. In [17] the authors propose a few algorithms that can incorporate 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 [15] 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 [11]. Their approach based on succinctness and 2-var constraints is different as it is more query oriented and does not deal with rule semantics, mapping, rule size or noisy data. Bayardo et. al. [3] show that support and confidence are fundamental interestingness metrics.

C. General description of our medical data

The medical data set we are mining describes the profiles of patients being treated for coronary heart disease. All medical information is put in one file having several records. Each record corresponds to the most relevant information of one patient. This profile contains personal information such as age, race, smoking habits and other relevant information. Measurements on the patient such as weight, heart rate, blood pressure and information regarding the preexistence or existence of certain diseases are also stored. The diagnostics made by a clinician 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 myocardium (heart muscle) is stored as boolean data.

Table I shows the 25 medical fields that will be used throughout this paper. For each attribute we give its usual abbreviation in the medical domain, its data type (DT), what type of medical information (MI) it contains and a complete description. Attributes are classified into three types according to the medical information they contain. 'P' attributes correspond to perfusion measurements on specific regions of the heart, 'R' attributes correspond to risk factors and 'D' attributes correspond to heart disease measurements. The goal is to relate perfusion measurements and risk factors to disease measurements. The image data represents the local degree of blood distribution (perfusion) in the heart muscle (myocardium). There are some fields that are commonly selected in data mining experiments. These fields include 4 fields that store the percentage of heart disease caused by a specific artery of the heart (LM, LAD, LCX, and RCA) and 9 fields that store a perfusion measurement which is a value in the range $[-1, 1]$. Closer to 1 indicates a more severe perfusion defect. Closer to -1 indicates absence of a perfusion defect. Each of the artery fields has a value between 0 and 100, and each heart region has a value between -1 and 1.

D. Alternative approaches

Here we explain why other data mining techniques are inadequate to solve our problem. Decision trees [10] produce rules to classify records from a data set minimizing classification error. This approach assumes there is a target variable indicating the class to which each record belongs. In our case it would have to be a categorical variable indicating if the patient is healthy or sick. However, patients cannot be classified in such a simple way because they have a degree of sickness. It could be argued that there could be several classes indicating the degree of sickness, but this would have to be done for each artery making many runs and significant analysis effort mandatory. Besides this does not cover the case that the patient has combinations of diseased arteries. There is even another worse drawback about decision trees: they automatically split numerical variables. The medical community has standard cutoffs used to understand numerical variables and these cutoffs are widely accepted (high blood pressure, high cholesterol, male overweight, etc). Therefore, the split points chosen by the decision tree may be of little use if they are different from the standard ones; experimental results interpretation becomes more difficult.

No	Name	DT	MI	Description
1	Age	N	R	Age of patient
2	LM	N	D	Left Main artery
3	LAD	N	D	Latero Anterior Desc.
4	LCX	N	D	Left CircumfleX
5	RCA	N	D	Right Coronary Art.
6	AL	N	P	Antero-Lateral
7	AS	N	P	Antero-Septal
8	SA	N	P	Septo-Anterior
9	SI	N	P	Septo-Inferior
10	IS	N	P	Infero-Septal
11	IL	N	P	Infero-Lateral
12	LI	N	P	Latero-Inferior
13	LA	N	P	Latero-Anterior
14	AP	N	P	Apical
15	Sex	C	R	Gender
16	HTA	C	R	Hyper-tension
17	Diab	C	R	Diabetes
18	HYPLD	C	R	Hyperlipidemia
19	FHCAD	C	R	Fam. hist. of disease
20	Smoke	C	R	Smoking habits
21	Claudi	C	R	Claudication
22	PAngio	C	R	Previous angina
23	PStroke	C	R	Prior stroke
24	PCarSur	C	R	Prior carotid surgery
25	Chol	N	R	Cholesterol level

TABLE I
ATTRIBUTES

Clustering [8], [10], [12] is another potential technique. In our case it was useful to have a global understanding of the data set. However, it was not adequate to produce rules relating a subset of all the variables. A constrained version of clustering focusing on projections of the data could be useful but that is an aspect that deserves further research.

III. MINING CONSTRAINED ASSOCIATION RULES

In this section we introduce our most important contributions. First, we analyze the problem of mapping information from categorical and numerical attribute values to items. Second, we identify useful constraints on attributes and items to get interesting association rules.

A. Mapping attributes

The medical data records have to be transformed into a transaction format suitable to discover association rules. As noted above, there are categorical, numerical, time and image attributes. To make the problem simpler all attributes are uniformly treated as categorical or numerical. In numerical attributes there is a natural order among values as opposed to categorical attributes where there does not exist such order.

Let $A_1, A_2, \ldots A_p$ be all the attributes, let $R = \{r_1, r_2, \ldots r_n\}$ be a relation with n tuples whose values are taken from $dom(A_1) \times dom(A_2) \times \ldots \times dom(A_p)$, where $dom(A_i)$ is either a categorical or numerical domain. The data set size is n and its dimensionality is p. Let $D = \{T_1, T_2, \ldots, T_n\}$ be a set of n transactions containing subsets of m items, resulting from the mapping process. Items are identified by consecutive integers starting in one, i.e. $1, 2, \ldots, m$.

The following mapping algorithm discretizes medical data records transforming numerical and categorical values into binary data. The mapping process is divided in two phases. In the first phase a mapping table M is constructed based on user's requirements. In the second phase attribute values in each tuple r_j are mapped to items based on M. Each tuple r_j becomes a transaction T_j, suitable for association rule mining. For a categorical attribute A_i each categorical value is mapped to one item. If negation is desired for categorical attribute A_i then each negated value is mapped to an item. The domain expert specifies k cutoff points for each numerical attribute A_i producing $k+1$ intervals. Then each interval is mapped to one item. If negation is desired then $k+1$ additional items are created corresponding to each negated interval. In general negation significantly increases the potential number of associations. Therefore, it must be used on a per attribute basis after careful consideration. Once the M mapping table has been constructed the second phase is straightforward.

B. Constraining association rules

This is a summary of the main difficulties faced when trying to discover interesting association rules in medical data. For each problem we propose a solution that is generally in the form of a constraint. Problems are described in an abstract manner.

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 discovered associations is a practical bottleneck for algorithm performance. In our case even $k > 5$ 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.

Items restricted to appear only in the antecedent, only in the consequent or in either place. Remember that by rule definition an item appears only once in a rule and therefore it appears either in the antecedent or in the consequent of the rule. 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 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 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.

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 say anything 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 then 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.

Low support. It has been shown that support is the performance bottleneck for association rules [11]. It is desirable to run the algorithm once with a very low support avoiding repeated runs with decreasing supports. We are interested in rules involving at least two transactions; this is a very low minimum support level.

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 [16] 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.

Important constraints: Based on the difficulties outlined above we introduce the following improvements. Extend items with two fields indicating constraints. Let $\mathcal{I} = \{i_1, i_2, \ldots i_m\}$ be the set of items to be mined obtained by the mapping process from the attributes $A_1, \ldots, A_p$.

Let $\mathcal{C} = \{c_1, c_2, \ldots c_p\}$ be a set 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 out of 3 values: 1 if item A_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 antecedent/consequent $ac : \mathcal{R} \to \mathcal{C}$ as $ac(A_j) = c_j$ to make reference to one such constraint.

Let $\mathcal{G} = \{g_1, g_2, \ldots g_p\}$ be a set of group constraints for each attribute A_j; g_i is a positive integer if A_j is constrained to belong to some group or 0 if A_j is not group constrained at all. We define the function $group : \mathcal{R} \to \mathcal{G}$ as $group(A_j) = g_j$. Since each attribute belongs to one group then the group numbers induce a partition on the attributes. This will induce a partition on the attributes. Attributes belonging to some group and attributes not constrained to belong to any group. Note that if the group is > 0 then there must be two or more attributes with the same group value, otherwise, the attribute would appear as not constrained.

Let $X = \{i_1, i_2, \ldots, i_k\}$ be a k-itemset. X is said to be antecedent-interesting if $\forall i_j \in X \ ac(i_j) \neq 2$. X is said to be consequent-interesting if $\forall i_j \in X \ ac(attribute(i_j)) \neq 1$. X is said to be group-interesting if $\forall i_j \forall i_{j'} \in X \quad i_j \neq i'_j \Rightarrow group(attribute(i_j)) \neq group(attribute(i_{j'}))$. We will use $group(i)$ and $ac(i)$ for item i to simplify notation.

Lemma 1 Itemset interestingness has the downward closure property in both $ac(i)$ and $group(i)$ constraints.
Proof: this is straightforward to prove since these properties are defined on sets. □

Lemma 2 The $ac(i)$ constraint cannot be used to prune away associations because of the rule generation phase.
Proof: Assume we have a rule $X \Rightarrow Y$. X and Y must respect the ac constraint for each of their items, but $X \cup Y$ will not. $ac(i)$ is an antimonotic constraint, but it cannot be used to discard $X \cup Y$ because the support for the rule is computed on $X \cup Y$. □

Lemma 3 Let X be a frequent k-dimensional itemset. Assume $\kappa < k$ then there are $2^k - \binom{k}{\kappa} 2^\kappa$ pruned associations.
Proof: We just need to substract the number of itemsets of size κ, which is the right term, from the powerset on k items. □

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.
Proof: Consider the powersets of X and Y. Every union of X and one or more elements of Y is invalid. Every union of Y and one or more elements of X is invalid. Counting all these cases gives the stated bound. □

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 states that the number of pruned associations is big when the maximal frequent itemset is large. In our case this produces significant speedup to make computation more interactive. Lemma 4 gives an idea about the number of discarded rules.

C. Algorithm to mine constrained association rules

We propose the following algorithm based on the well-known A-priori algorithm [2]. All the basic notation and definitions are taken from section 2. Let κ 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. We require a minimum support allowing us to mine associations referring to only two transactions. This number will be fixed. Pruning will be based mostly on constraints. Minimum confidence will vary from run to run.

1) Mapping algorithm
   - Construct mapping table M
   - For each tuple $r_1, r_2, \ldots, r_n$ do the following. Map attribute values of $A_1, A_2, \ldots, A_p$ to items $1, 2, \ldots, m$ based on M producing transactions $T_1, T_2, \ldots, T_n$ (section III-A).
2) Constrained association rule algorithm
   - 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 κ 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 $group(attribute(i_j)) \neq group(attribute(i_{j'}))$ and $group(attribute(i_j)) * group(attribute(i_{j'})) > 0$ 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. 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 $(ac(attribute(i)) \neq 2 \ \forall i \in X)$ and

$(ac(attribute(i)) \neq 1 \; \forall i \in Y)$ and $(minconfidence \leq support(X \cup Y)/support(X))$ then $X \Rightarrow Y$ is a valid rule.

IV. Experimental evaluation

In this section we present important association rules discovered by our algorithm. Our experiments were run on a Sun computer. Our algorithm implementation was done in the C language.

A. Medical data set used

All our experiments were based on a real data set obtained from a hospital. The data set consisted of 655 patients having 113 attributes. We selected the 25 most important medical attributes for mining listed in table I. So $p = 25$ and $n = 655$. These attributes include perfusion measurements for 9 regions of the heart and heart vessel disease for 4 vessels and attributes relating high risk factors for heart disease. 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. Each of the LM, LAD, LCX, and RCA numerical attributes refer to vessel measurements.

B. Setting program parameters

To automatically map attributes to items we did the following. The LAD, RCA, LCX and LM numbers represent the percentage of vessel narrowing and they are split into ranges as follows. LAD, LCX and RCA were partitioned by cutoff points 50% and 70%. The 70% value indicates significant coronary disease. The 50% value indicates borderline disease. Less than 50% means the patient is considered healthy. The most common cutoff value used by the cardiology community is 50%. LM was partitioned by cutoff points at 30% and 50%. Both the LAD and the LCX arteries branch from the LM artery and then a defect in it is more likely to cause a larger diseased heart region. That is why its cutoff values are set lower. The 9 heart regions (AL, IL, IS, AS, SI, SA, LI, LA, AP) were partitioned into 2 ranges at a cutoff point of 0.2. CHOL was partitioned with cutoff points 200 (warning) and 250 (high). These values correspond to known medical settings. Since the clinicians were interested in getting rules involving healthy and sick patients these 4 attributes were chosen for negation. Missing values were assigned one item but were ignored for rule generation.

In general we are interested in rules that involve at least two patients. Obviously rules that refer to only one patient are not reliable and some of those may have 100% confidence. Then the minimum support was always fixed at $2/n \approx 0.2\%$. Note that this is in fact the lowest support discarding rules for one transaction. The minimum confidence was set at 70%. The maximum support was set at 30%.

In the past we attempted using association rules without constraints [14], but results were useless. The number of rules went over 1 million, and most of them involved the same medical variables. So, a post-processing approach did not work. That is, mining association rules with minimum support and confidence and then filtering out unwanted rules was not practical. This made constraints a required ingredient both from a performance point of view and from a practical standpoint. Note that we require a very low support allowing us to mine associations referring to only two transactions.

Now we explain what constraints were set for association rule finding. This set of constraints is by no means definitive or optimal, but it represents what our experience has shown to be most useful. Please refer to table I to understand the attribute meanings. The constraints for the association rule mining program were set as follows. The 4 main coronary arteries LM, LAD, LCX, and RCA were constrained to appear in the consequent of the rule, that is, ac(i) = 2. All the other attributes were constrained to appear in the antecedent, i.e. ac(i) = 1. In other, words R (Risk factors) and P (Perfusion Measurements) should appear in the antecedent, whereas D (disease) medical fields should appear in the consequent of a rule.

The 9 regions of the heart (AL, IS, SA, AP, AS, SI, LI, IL, LA) were constrained to be in the same group. Sex, HTA, HYPLPD, FHCAD, Smoke and Chol were constrained to be in the same group. Claudi, PANGIO, PSTROKE, PCARSUR were constrained to be in the same group. Age, Sex, LAD, LCX, RCA were not group constrained. Remember that combinations of items in the same group are not considered interesting.

C. Medical significance of association rules

The goal of the experiment was to relate perfusion measurements and risk factors to vessel disease (also known as stenosis) to validate and improve actual diagnosis rules used by an expert system [9]. Some rules were expected, confirming valid medical knowledge, and some rules were surprising, having the potential to enrich the expert system knowledge base. This is an analysis of our most interesting results.

There are two main measurements to quantify the quality of medical findings: sensitivity and specificity. Sensitivity refers to the probability of correctly identifying patients with disease. Specificity is the probability of correctly identifying healthy individuals. These measures rely on a gold standard, that is, a measurement that tells with very high accuracy if the person is sick or not. Getting such ideal measurement may involve doing invasive medical procedures on the patient. In the context of this paper the gold standard was catheterization. In a few cases a clinician reading was taken, but in general it was not available.

The data mining algorithm produced a total of 2987 rules, almost all having a medical significance of some sort. All of them could be used in answering medical questions. Most, however, were addressing issues that were not being examined at this time. Reducing the number of rules found to the point where the results can be easily interpreted by a clinician was done in two steps.

The first step reduced the total number of rules to 850. This was achieved by removing rules that are, in whole or in part, counter-intuitive to medical knowledge. These rules can be useful in confirming, disproving, and further quantifying what is already considered established fact. Although interesting in their own right, these investigations fall outside the scope of our present research. Doing this requires filtering out combinations for specific values of the variables combined in the rules. Although our extensions to the association rules algorithm allow filtering out combinations of fields, this does not extend to combinations of specific values.

The second step reduced the number left for further examination down to 73. To achieve this, rules were further subdivided into categories. Examples of these categories are 'relating Age to CAD', 'relating smoking habits to CAD', and 'predicting CAD from image data'. In each category, only the rules with the highest support and/or confidence were selected to represent the results in that category.

The 73 rules were analyzed by a domain expert (clinician). In the following paragraphs we will discuss the most interesting rules. The discovered rules were classified into 2 groups: First, those that express that if there no risk factor then there is no heart disease. Second, those that express that if there exists a risk factor then there is heart disease. It is important to observe that the rules below involve several attributes in the antecedent or in the consequent, negation and attributes being in different ranges.

Rules predicting no heart disease: In this case new medical fields not previously included in the expert system are mined for association rules. All these rules have the potential to improve the expert system. In this example we can see there is less incidence of coronary disease in the patients who do not smoke, and in those who have lower cholesterol. The second rule has very high support, compared to the other rules. It states that non smokers have a lower chance of having a diseased RCA artery; note that there is the chance that some of these patients are in the 50-70% range, i.e. being borderline cases. The fourth rule is particularly interesting as it involves two arteries in the consequent. Basically if a person is young (regardless of sex) and does not smoke the risk for heart disease is low. There are more complex rules relating two heart arteries. The last two rules say that an adult female patient with no diabetes is very likely to be healthy, that is, having no heart disease.

1. $[Sex = F] \Rightarrow ([0.0 <= LCX < 50.0])$ $s = 0.229, c = 0.728$
2. $[Smoke = n] \Rightarrow [not(70.0 <= RCA < 100.1)]$ $s = 0.290, c = 0.714$
3. $[0.0 <= CHOL < 200.0] \Rightarrow [not(70.0 <= LAD < 100.1)])$ $s = 0.078, c = 0.708$
4. $[0.0 <= Age < 40.0][Smoke = n] \Rightarrow [0.0 <= LCX < 50.0][0.0 <= RCA < 50.0]$ $s = 0.008, c = 0.714$
5. $[0.0 <= Age < 40.0][Diab = n] \Rightarrow [0.0 <= LAD < 50.0]$ $s = 0.027, c = 0.818$
6. $[0.0 <= Age < 40.0][Diab = n] \Rightarrow [0.0 <= LAD < 50.0]$ $s = 0.027, c = 0.818$
7. $[40.0 <= Age < 60.0]and[Sex = F][Diab = n] \Rightarrow [0.0 <= LCX < 50.0]$ $s = 0.084, c = 0.917$
8. $[40.0 <= Age < 60.0]and[Sex = F][Diab = n] \Rightarrow [0.0 <= RCA < 50.0]$ $s = 0.073, c = 0.800$

Rules predicting heart disease: These rules relate risk factors to heart disease. Heart disease can be detected by tomography or coronary catheterization. Tomography corresponds to myocardial perfusion studies. Catheterization involves inserting a tube into the coronary artery and injecting a substance to measure which regions are not well irrigated. These rules characterize the patient with coronary disease. There are three basic elements for analysis: perfusion defect, coronary stenosis and risk factors.

Most of the rules below refer to older patients with localized perfusion defects in specific heart regions. Rule 1 says that if the patient has a perfusion defect and had a previous carotid surgery then he has a high probability of having heart disease. The number of patients for this rule is low, but when conditions hold the disease probability will be high. These rules relate more information such as age, smoking habits, cholesterol levels. Rules 4,5,6 are outstanding as they confirm medical knowledge for very high risk of heart disease with high accuracy. Basically if a person is old, has high cholesterol levels and has a perfusion defect then it is almost sure that person has a serious heart condition. All these aspects have an impact on the risk for heart disease. Rules 5 and 6 state that high cholesterol levels and age are determinant factors to have a diseased RCA artery; these rules have 100% confidence. Rule 11 has relatively high support and very high confidence; it relates a specific defect in a heart region (SA) with a chance of having a diseased LAD artery. We conclude observing that according to medical knowledge the LAD artery has a higher chance of being diseased than the other arteries [6]. As can be seen the rules that involve the LAD artery confirm this fact since they have higher support and almost 100% confidence.

1. $[0.2 <= AP < 1.1][PCARSUR = y] \Rightarrow [not(0.0 <= LAD < 50.0)][not(0.0 <= RCA < 50.0)])$ $s = 0.012, c = 0.800$
2. $[60.0 <= Age < 100.0][0.2 <= AP < 1.1][Smoke = y] \Rightarrow [not(0.0 <= LAD < 50.0)])$ $s = 0.107, c = 0.833$
3. $[0.2 <= LA < 1.1][Sex = M]and[250.0 <= CHOL < 500.1] \Rightarrow [not(0.0 <= LCX < 50.0)])s = 0.014, c = 0.750$
4. $[60.0 <= Age < 100.0][0.2 <= IL < 1.1][250.0 <= CHOL < 500.1] \Rightarrow [not(0.0 <= RCA < 50.0)])$ $s = 0.017, c = 0.917$
5. $[60.0 <= Age < 100.0][0.2 <= IS < 1.0][250.0 <= CHOL < 500.1] \Rightarrow [not(0.0 <= RCA < 50.0)])$ $s = 0.015, c = 1.000$
6. $[60.0 <= Age < 100.0][0.2 <= IS < 1.0][250.0 <= CHOL < 500.1] \Rightarrow [not(0.0 <= RCA < 50.0)])$ $s = 0.015, c = 1.000$
7. $[60.0 <= Age < 100.0][0.2 <= SA < 1.0][FHCAD = y] \Rightarrow [not(0.0 <= LAD < 50.0)])s = 0.015, c = 1.000$
8. $[0.2 <= SA < 1.0]and[PANGIO = y] \Rightarrow [not(0.0 <= LAD < 50.0)])s = 0.023, c = 0.938$
9. $[60.0 <= Age < 100.0][0.2 <= AP < 1.1][Sex = F] \Rightarrow [not(0.0 <= LAD < 50.0)])s = 0.049, c = 0.941$
10. $[60.0 <= Age < 100.0][0.2 <= SA < 1.0][Claudi = y] \Rightarrow [not(0.0 <= LAD < 50.0)])s = 0.029, c = 0.950$
11. $[60.0 <= Age < 100.0][0.2 <= SA < 1.0][HYPLPD = y] \Rightarrow [not(0.0 <= LAD < 50.0)])s = 0.070, c = 0.939$

V. Conclusions

This article presented our experiences mining association rules from medical data to predict heart disease. We explained the motivation and validity of using association rules on medical data. Association rules are useful for our purpose given their combinatorial nature. We described all information contained in medical records. We introduced a simple mapping algorithm to transform medical records to a transaction format. We then presented an improved algorithm to mine constrained association rules. Medical data records contain numerical, categorical, time and image attributes. The mapping algorithm uniformly treats attributes as numerical or categorical. Numerical attributes are split into intervals. Negation is used on a per attribute basis to avoid an explosion in the number of associations. A mapping table is constructed and based on this table attribute values are mapped to items. The algorithm to mine association rules uses several important constraints to reduce the number of rules and speed up the mining process. It uses a constraint to exclude combinations of attributes eliminating trivial or useless associations. Certain attributes are constrained to appear only in the antecedent, only in the consequent or in both to get medically meaningful rules. Rules are constrained to include a maximum number of items to make them simpler and more general. Maximum support is a constraint used to eliminate trivial rules. These constraints allowed us to mine medical records at a minimum support involving only two transactions. The experimental section discussed several important association rules predicting absence or presence of heart disease.

This is a summary of issues for future research. We would like to examine problems with missing information more closely. We want to identify other useful constraints besides grouping and antecedent/consequent. We want to compare the discovered association rules with classification rules obtained by a decision tree algorithm. We plan to process the data with a clustering algorithm [12] to explain why certain rules have low confidence and to find high confidence rules in subsets of medical records. Finally, we want to assess the impact on performance of each constraint.

Acknowledgments

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

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

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

[5] 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.

[6] 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.

[7] D. Cooke, Cesar Santana, Tahia Morris, Levien de Braal, C. Ordonez, E. Omiecinski, N. Ezquerra, and Ernest V. Garcia. Validating expert system rule confidences using data mining of myocardial perfusion SPECT databases. In *Computers in Cardiology Conference*, pages 116–119, 2000.

[8] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. J. Wiley and Sons, New York, 1973.

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

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

[11] 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.

[12] C. Ordonez and P. Cereghini. SQLEM: Fast clustering in SQL using the EM algorithm. In *Proc. ACM SIGMOD Conference*, pages 559–570, 2000.

[13] 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.

[14] C. Ordonez, C.A. Santana, and L. Braal. Discovering interesting association rules in medical data. In *Proc. ACM SIGMOD Data Mining and Knowledge Discovery Workshop*, pages 78–85, 2000.

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

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

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