Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction

Carlos Ordonez

***Abstract*—Association rules represent a promising technique to improve heart disease prediction. Unfortunately, when association rules are applied on a medical data set, they produce an extremely large number of rules. Most of such rules are medically irrelevant and the time required to find them can be impractical. A more important issue is that, in general, association rules are mined on the entire data set without validation on an independent sample. To solve these limitations, we introduce an algorithm that uses search constraints to reduce the number of rules, searches for association rules on a training set, and finally validates them on an independent test set. The medical significance of discovered rules is evaluated with support, confidence, and lift. Association rules are applied on a real data set containing medical records of patients with heart disease. In medical terms, association rules relate heart perfusion measurements and risk factors to the degree of disease in four specific arteries. Search constraints and test set validation significantly reduce the number of association rules and produce a set of rules with high predictive accuracy. We exhibit important rules with high confidence, high lift, or both, that remain valid on the test set on several runs. These rules represent valuable medical knowledge.**

***Index Terms*—Association rules, heart disease, search constraint, train and test.**

I. Introduction

ASSOCIATION rules represent a data mining technique that has great potential in the medical domain to improve disease prediction. The use of association rules on medical data records with heart perfusion measurements is explored for the first time in [1]. Discovering association rules on segmented images is studied in [2]. Neural networks are used to predict heart response based on exercise stress and heart muscle thickening images [3]. A basic set of constraints is introduced in [4] and [5], and preliminary results stress the importance of search constraints. Given the remarkable similarity between confidence and confidence factors, association rule confidence is used to validate confidence factors of production rules in an expert system [6]. Validated discovered rules have been added to an expert system knowledge base [7].

In this research work, association rules are mined on a medical data set to improve heart disease diagnosis. Each rule represents a simple predictive pattern that describes a subset of the data set projected on a subset of attributes. From a medical perspective, association rules relate combinations of binary target attributes (absence/existence of artery disease) and subsets of independent attributes (risk factors and heart muscle health measurements). Association rules have important advantages over traditional supervised machine learning or statistical algorithms (e.g., decision trees [8], [9], logistic regression [10], and support vector machines [10]): they have a straightforward interpretation based on the probability of occurrence of a pattern and the conditional probability between two patterns (medical measurements and risk factors relationship to specific artery narrowing); they can link combinations of predicted attributes (predict coexistence of disease in two or more arteries); they can handle several predicted attributes simultaneously without the need of having separate data subsets or separate runs (predicting disease of all arteries with a single data set); they can find patterns that exist in small subsets of attributes (which is particularly challenging for data sets that are small but also high dimensional); and finally, each rule can refer to overlapping subsets of the data set with respect to other rules (avoiding data set fragmentation when possible). In contrast, for our medical data set, supervised or predictive algorithms would require a large set of models (e.g., many decision trees or regression models), extra data preparation and data selection steps to handle infrequent cases or to create adequate target attributes, and complex feature selection or stepwise procedures to find small subsets of attributes related to a specific degree of disease. Nevertheless, we discovered two main issues when using association rules on a medical data set. First, the number of discovered rules becomes extremely large at low support thresholds, and most rules are medically irrelevant. Second, previous approaches [11], [12] focus on mining association rules on the entire data set, without validation on an independent data set. Another reason supporting a validation procedure is the fact that collecting new medical records with similar characteristics is difficult, due to privacy regulations and management complexity of medical data [13]. With those issues in mind, we define search constraints [14], [5] to reduce the number of rules and introduce a rule learning algorithm that searches for constrained association rules on a training sample, and validates them on an independent (disjoint) test sample. This approach helps finding rules with high predictive accuracy and reduces the number of unreliable or medically irrelevant rules. Our work advances previous research on incorporating constraints for association rules, including [14], [15], [16], and [17].

A. Contributions and Outline

We present an algorithm that incorporates search constraints to find medically relevant association rules, and validates them with the well-known *train and test* approach [10], [18] to get rules with high predictive accuracy. Search constraints include maximum association size, item filtering dependent on predictive goal (absence or existence of disease), attribute grouping (discard irrelevant combinations), and antecedent/consequent

Manuscript received June 9, 2005; revised September 29, 2005.

The author is with Teradata (NCR), San Diego, CA 92127 USA (e-mail: carlos.ordonez@teradata-ncr.com).

Digital Object Identifier 10.1109/TITB.2005.864475

rule filtering (find predictive rules). Support, confidence, and lift are the metrics used to evaluate the medical significance and reliability of association rules. Experiments study the importance of each constraint individually. We show there is a high proportion of rules that cannot be generalized after validation on the test set. Our proposed algorithm produces a set of rules that remain valid in several training/test cycles.

The paper is organized as follows. Section II introduces definitions. Section III explains a transformation process from medical record to transaction format, introduces search constraints to find predictive rules, and proposes a *train and test* algorithm to find and validate rules. Section IV presents experiments with a real data set containing medical records of patients with heart disease. Section V explains related work, and Section VI concludes the paper.

II. Definitions

We use the standard definition of association rules [19], [2], [5]. Let D be a set of n transactions such that (s.t.) $D = \{T_1, T_2, \ldots, T_n\}$, where $T_i \subseteq \mathcal{I}$ and $\mathcal{I}$ is a set of items, $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$. A subset of $\mathcal{I}$ containing k items is called a k-itemset. Let X and Y be two itemsets s.t. $X \subset \mathcal{I}, Y \subset \mathcal{I}$, and $X \cap Y = \emptyset$. An association rule is an implication denoted by $X \Rightarrow Y$, where X is called the antecedent and Y is called the consequent. In our work, patient records consist of numeric and categorical values that are transformed into transactions, where each item corresponds to one numeric range or one categorical value. This transformation process will be discussed in Section III-A.

We proceed to define association rule metrics. Given an itemset X, support $s(X)$ is defined as the fraction of transactions $T_i \in D$ such that $X \subseteq T_i$. Consider $P(X)$ the probability of appearance of X in D, and $P(Y|X)$ the conditional probability of appearance of Y given X. $P(X)$ can be estimated as $P(X) = s(X)$. The support of a rule $X \Rightarrow Y$ is defined as $s(X \Rightarrow Y) = s(X \cup Y)$. An association rule $X \Rightarrow Y$ has a measure of reliability called confidence, defined as $c(X \Rightarrow Y) = s(X \Rightarrow Y)/s(X)$. Confidence can be used to estimate $P(Y|X)$: $P(Y|X) = P(X \cup Y)/P(X) = c(X \Rightarrow Y)$. A probabilistic interpretation of support and confidence is discussed in [10]. We use a third metric called lift [2], defined as $l(X \Rightarrow Y) = P(X \cup Y)/(P(X)P(Y)) = c(X \Rightarrow Y)/s(Y)$. Lift quantifies the relationship between X and Y. In general, a lift value greater than 1 provides strong evidence that X and Y depend on each other. A lift value below 1 states X depends on the absence of Y or vice versa. A lift value close to 1 indicates X and Y are independent.

The problem of mining association rules is defined as finding the set of all rules $\{X \Rightarrow Y\}$ such that $s(X \Rightarrow Y) \geq \psi$ and $c(X \Rightarrow Y) \geq \alpha$, given a support threshold ψ and a confidence threshold α. A k-itemset X such that $s(X) \geq \psi$ is called frequent.

III. Using the Train and Test Approach With Constrained Association Rules

This section presents an algorithm to find predictive association rules in a medical data set. The algorithm has three major steps. First, a medical data set with categorical and numeric attributes is transformed into a transaction data set, as defined in Section II. Second, four constraints are incorporated into the search process to find predictive association rules with medically relevant attribute combinations. Third, a train and test approach is used to validate association rules.

A. Medical Data Set Transformation

Consider a medical data set containing n patient records $S = \{s_1, s_2, \ldots, s_n\}$ with categorical, numeric, time, and image attributes. All attributes are uniformly treated as categorical or numeric to make the problem simpler. That is, if S has attributes $A_1, A_2, \ldots, A_p$, then A_l is either a categorical or numeric type. In order to use association rules, S is transformed into a transaction data set $D = \{T_1, T_2, \ldots, T_n\}$. Numeric attributes are binned (discretized) and each bin becomes an item. Categorical attributes are transformed into items by assigning an item to each categorical value. When an attribute is negated, additional items are created for each negated categorical value or each negated numeric range. Summarizing, each transaction T_i is a set of items that indicate the presence (or absence under negation) of one categorical value, or one numeric interval for one patient record s_i.

B. Constraints for Association Rules

This section introduces search constraints to find only predictive association rules and to reduce the number of patterns. Let $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$ be the set of items to be processed, obtained by the transformation process from the attributes $\mathcal{A} = \{A_1, \ldots, A_p\}$. Let attribute : $\mathcal{I} \rightarrow \mathcal{A}$ be a function that returns the attribute corresponding to one item.

The problem of discovering association rules is decomposed into two basic subproblems [19]: 1) finding all frequent k-itemsets X such that support $s(X) \geq \psi$ and 2) finding all rules $X \Rightarrow Y$ such that $c(X \Rightarrow Y) \geq \alpha$. This is motivated by the fact that the first subproblem is generally more difficult. But in Section IV we shall see that the second subproblem is actually harder for a medical data set if constraints are not used. In our algorithm we call the first subproblem Phase 1 and the second one Phase 2. We define four constraints, three of which are used in Phase 1 and one on Phase 2. Our work extends previous research on constraining association rules [14], [15], [16], [17], where most constraints work on Phase 1.

1) Constraints for Frequent Itemset Search on Phase 1: We define an item filtering constraint based on the predictive goal P. For heart disease prediction there will be a set of items for predicting existence of disease ($P =$ "Y") and another set of items to predict absence of disease ($P =$ "N"). Some items will be used for both predictive goals. This constraint will eliminate any combination containing the filtered items. In our case, item filtering is applied before itemset generation during preprocessing, instead of applying it during the search phase [20]. Item filtering is a succinct constraint [14], [16]. This constraint is manually

defined by the medical doctor (domain expert) to include interesting items given the prediction goal, but at the same time eliminating many item combinations that are not currently being analyzed. It should be clarified that some filtered items may be potentially valuable to help in understanding unknown relationships among risk factors/perfusion and arteries. That is, any rule with high reliability metrics has the potential of being medically significant.

The second constraint is called the rule size constraint, which is given by a maximum size κ. Frequent itemsets are generated up to size κ, eliminating complete branches of the search tree having frequent itemsets of size $\kappa+1, \kappa+2$ and so on. This constraint is simple, yet it produces simpler and fewer rules. Most approaches find all rules above thresholds in an exhaustive fashion [19], [21], but in the case of medical data such an approach is not practical. In general, search constraints should be used as early as possible to improve efficiency [15]. The rule size constraint is applied during itemset generation in Phase 1, but it could be applied in Phase 2.

We now introduce the third and fourth constraints. Input attributes are extended with two constraints: the group constraint and the antecedent-consequent constraint. We will now explain the group constraint that is used on Phase 1. The group constraint is used to avoid combinations of perfusion measurements and trivial combinations of risk factors. Let $\mathcal{G} = \{g_1, g_2, \ldots, g_p\}$ be a set of p group constraints, one for each attribute A_j; g_j is a positive integer if A_j is constrained to belong to some group or 0 if A_j is not group-constrained. We define the function group : $\mathcal{A} \rightarrow \mathcal{G}$ as group$(A_j) = g_j$. Since each attribute belongs to one group, then the group numbers induce a partition on the attributes. If $g_j > 0$ then there should be two or more attributes with the same group value of g_j. Otherwise, that would be equivalent to having $g_j = 0$. Itemset X is said to be group-interesting if, for every pair $\{i_j, i_{j'}\}$, it holds that group(attribute$(i_j)) \neq$ group(attribute$(i_{j'})$). The group constraint can be considered antimonotonic [14], meaning that if an itemset is not interesting, then any superset will not be interesting either. Our group constraint differs from item constraints [16] in the sense that it induces a partition on attributes and their corresponding items.

2) Constraint for Filtering Rules on Phase 2: There is only one constraint used in Phase 2, applied to rules but not to associations, called the antecedent–consequent (AC) constraint. Intuitively, we can think of it as a template of predictive rules. We define a set of AC constraints on $\{A_1, \ldots, A_p\}$ by $\mathcal{C} = \{c_1, \ldots, c_p\}$. The constraint c_j takes one out of three possible values: $c_j = 1$ if attribute A_j can only appear in the antecedent of a rule; $c_j = 2$ if A_j can only appear in the consequent; and $c_j = 0$ if it can appear in either place. We define the function antecedent/consequent ac : $\mathcal{A} \rightarrow \mathcal{C}$ as $\mathrm{ac}(A_j) = c_j$. This constraint is specified over attributes, but not on items like [15] and [16]. Let X be a k-itemset; X is said to be antecedent-compliant if $\forall i_j \in X$: ac(attribute$(i_j)) \neq 2$; X is said to be consequent-compliant if $\forall i_j \in X$: ac(attribute$(i_j)) \neq 1$. The ac constraint cannot be applied in Phase 1 because it is neither succinct nor antimonotonic [15].

C. Training and Test Data Sets

In machine learning, it is customary to collect disjoint (independent) samples from a base data set to build and tune predictive (supervised) models [18]. The most common approach is called "train and test." The basic idea is to build a predictive model with a training sample and then validate the model using an independent test sample. This process has the goals of reducing model overfit, providing a realistic estimate of model accuracy and improving generalization when the model is used on new data.

We apply this idea by partitioning D into two disjoint subsets. We call the first subset the training set D_{train} and the second subset the test set D_{test}. We introduce a training fraction τ to control the size of the training set. Therefore

$$|D_{\mathrm{train}}| = \tau n$$
$$D = D_{\mathrm{train}} \cup D_{\mathrm{test}}$$

and

$$D_{\mathrm{train}} \cap D_{\mathrm{test}} = \emptyset.$$

Given the small size of our medical data set, we cannot apply a more reliable train/validate/test approach that requires three independent samples [10]. We extend the definition of association rules, given in Section II, to have two sets of metrics per rule based on a training and a test set. That is, each rule has six metrics in total. In general, metrics on the training set are used only for search purposes, and metrics on the test set are used to validate rules and are taken as the actual rule metrics.

We can informally think of the set of all association rules as a "global" predictive model. We compute two sets of rules, R_{train} on D_{train} and $R_{\mathrm{test}} \subseteq R_{\mathrm{train}}$, such that R_{test} also has metrics above ψ, α, λ on D_{test}. The computation of R_{test} is as follows. Each association has two sets of metrics, one for D_{train} and one for D_{test}. We search association rules based on D_{train} to get R_{train} based on thresholds ψ, α, λ. We set $R_{\mathrm{test}} = R_{\mathrm{train}}$. We then compute support, confidence, and lift for each rule in R_{test} based on D_{test}. Rules whose test metrics on D_{test} are below ψ, α, or λ are filtered out from R_{test}. This process is repeated a number of times (t) to achieve basic cross validation and to eliminate rules that cannot be generalized.

D. Algorithm to Find Predictive Rules

The constraints from Section III-B and the training/validation approach from Section III-C are assembled together in one algorithm that goes from transforming medical records to repeating validation on several independent test samples to get a set of accurate predictive association rules. The algorithm is shown in Fig. 1.

This is a summary of the input and output of the algorithm. The main input parameters are the predictive goal P ($P =$ "Y" for existence of disease and $P =$ "N" for absence of disease), κ (maximum rules size), t (number of times to train/test), as well as the ψ, α, λ thresholds. In general, the training sample fraction $\tau = 50\%$. The output is a set of rules (R_Y or R_N) that are valid on all t test sets.

Input: S, P, ψ, α, λ.
Parameters: τ, t (changing infrequently)
Output: R_N, R_Y

1) Data transformation:
   Transform $S = \{s_1, s_2, \ldots s_n\}$ into $D = \{T_1, \ldots, T_n\}$.
2) FOR $I = 1$ TO t DO
   - Partition D into D_{train} and D_{test}, based on τ.
   - Phase 1:
     Filter items depending on predictive goal P ("Y" or "N") before generating 1-itemsets. Search for frequent k-itemsets on D_{train} for $k \in \{1, \ldots, \kappa\}$, using ψ and the $group()$ constraint.
   - Phase 2:
     Generate rules using the $ac()$ constraint, minimum confidence α and minimum lift λ. Let the rule set be R_{train}.
   - Phase 3:
     Validate rules R_{train} on D_{test}. Set $R_{test} = R_{train}$. For each frequent itemset X compute test support $s(X, D_{test})$. For each rule $X \Rightarrow Y \in R_{test}$ compute test support $s(X \Rightarrow Y)$, compute test confidence $c(X \Rightarrow Y)$ and test lift $l(X \Rightarrow Y)$ on D_{test}. Eliminate rules from R_{test} s.t. $s(X \Rightarrow Y) < \psi$, or $c(X \Rightarrow Y) < \alpha$, or $l(X \Rightarrow Y) < \lambda$. Finally, set $R_I = R_{test}$.

   END
3) Get intersection of t rule sets and compute average rule metrics with (1) – (3)
   If P="Y" then $R_Y = R_1 \cap R_2 \cap \ldots \cap R_t$
   else
   If P="N" then $R_N = R_1 \cap R_2 \cap \ldots \cap R_t$

Fig. 1. Algorithm to find association rules with a train/test approach.

We run the transformation process once. Transformation requires the user to specify numeric cutoffs and negation. The transformation process binning numeric attributes and incorporating negation, creates the input data set for Phase 1, where each medical record s_i is transformed into T_i. Thereafter, medical records are manipulated as itemsets as defined in Section II.

Building the training and test samples is repeated several times. The constrained association rule algorithm produces different sets of rules with different training/test samples, where each set of rules has several different rules and common rules have slightly lower or higher metrics. We want to find rules that are valid on both D_{train} and D_{test} in general. This motivates repeating the training/test process t times to achieve basic cross validation and compute averages for rule metrics. Association rules do not represent a model. Therefore, they do not have a goodness of fit statistic; this is because each association rule simply represents an individual pattern.

Phase 1 eliminates many items depending on the predictive task, uses the group$(\cdot)$ constraint to discard irrelevant item combinations, and generates all associations up to size κ. Items are filtered depending on the prediction goal ($P =$ "N" for absence or $P =$ "Y" for existence of disease). Phase 2 filters predictive rules with the ac$(\cdot)$ constraint. Phase 3 eliminates unreliable and particular (not general) rules by computing metrics on D_{test}, producing a subset of rules R_I that remains valid on the test data set I. The process to create train/test samples and to discover/validate association rules is repeated t times, with t being a user-specified parameter. This process generates t independent training sets and t independent test sets. These t sets will produce different sets of rules that will have rules in common, but also different rules.

At the end, the algorithm computes a rule set (R_Y for existence of disease, or R_N for absence of disease) that is the intersection of the t rule sets, further eliminating rules that may be particular to one run, or rules that are not valid in general. Recall R_Y corresponds to $P =$ "Y" and R_N refers to rules when $P =$ "N." The metrics of each rule are computed as averages of the test metrics on the t test sets. Let D_I be the Ith *test* set. Let $X \Rightarrow Y$ be a valid rule appearing on all t sets. Then

$$s(X \Rightarrow Y) = \frac{1}{t} \sum_{I=1}^{t} s(X \Rightarrow Y, D_I) \quad (1)$$

$$c(X \Rightarrow Y) = \frac{1}{t} \sum_{I=1}^{t} c(X \Rightarrow Y, D_I) \quad (2)$$

$$l(X \Rightarrow Y) = \frac{1}{t} \sum_{I=1}^{t} l(X \Rightarrow Y, D_I). \quad (3)$$

IV. Experiments

Our experiments were conducted on a computer with an 800-MHz CPU, 256 MB of memory, and 40 GB on disk. Our algorithm was programmed in the C++ language.

A. Medical Data Set Description

We used a medical data set obtained from a hospital. The data set contains 655 patient records with 113 attributes, combining numeric, categorical, and image data. There are risk factor attributes such as age, race, gender, and smoking habits. There are measurements on the patient such as weight, heart rate, blood pressure, and information regarding the pre-existence of other diseases like diabetes. Diagnostic procedures made by a clinician are also included. The data set has an important 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, the remaining attributes store imaging (perfusion) information from nine regions of the myocardium (heart muscle). Table I shows the 25 attributes we selected for our experiments that were the most important. Therefore, $p = 25$ and $n = 655$.

The attributes not used in our experiments involved heart image data for the patient performing exercise or resting, and attributes whose value distributions were skewed. To give an example, race was not used because 82% of patients were of white race, and more than 10% had missing race information. The selected attributes shown in Table I provide an accurate profile of each patient based on perfusion measurements for specific regions of the heart, known risk factors for heart disease, and the degree of disease in four arteries (percentage of artery narrowing). Perfusion measurements are obtained with a medical procedure that digitizes the flow pattern in the heart of a colored substance swallowed by the patient.

Attributes in Table I are abbreviated as follows: arteries use an acronym indicating their location in the heart whose value represents artery narrowing (LM, LAD, LCX, RCA); the nine

TABLE I
MEDICAL DATA SET ATTRIBUTES (CONSTRAINTS: 0 = OFF, 1 = ON)

Attribute name	Medical meaning	Neg	Constraints itemFilter	group	ac
AGE	Patient age	0	0	0	1
LM	Left Main	1	1	0	2
LAD	Left Anter Desc	1	1	0	2
LCX	Left CircumfleX	1	1	0	2
RCA	Right Coronary	1	1	0	2
AL	Antero-Lateral	0	1	1	1
AS	Antero-Septal	0	1	1	1
SA	Septo-Anterior	0	1	1	1
SI	Septo-Inferior	0	1	1	1
IS	Infero-Septal	0	1	1	1
IL	Infero-Lateral	0	1	1	1
LI	Latero-Inferior	0	1	1	1
LA	Latero-Anterior	0	1	1	1
AP	Apical	0	1	1	1
SEX	Gender	0	0	0	1
HTA	Hypertension Y/N	0	1	0	1
DIAB	Diabetes Y/N	0	1	0	1
HYPLPD	Hyperloip Y/N	0	1	0	1
FHCAD	Faml hist dis Y/N	0	1	0	1
SMOKE	Smokes Y/N	0	1	0	1
CLAUDI	Claudication Y/N	0	1	0	1
PANGIO	Prev angina Y/N	0	1	0	1
PSTROKE	Prior stroke Y/N	0	1	0	1
PCARSUR	Prior surgery Y/N	0	1	0	1
CHOL	Cholesterol	0	0	0	1

heart muscle regions (AL, IS, SA, AP, AS, SI, LI, IL, LA) are abbreviated with two letters that can be interpreted as coordinates of their specific location on a two-dimensional (2-D) map of the heart, whose attributes contain a *perfusion* measurement in $[-1, 1]$; a value close to 1 indicates a severe defect and a value close to -1 indicates no defect; finally, risk factors (AGE, CHOL, HTA, FHCAD, ...) are binary variables to allow their easy manipulation as items. The last four columns in Table I indicate how each constraint was set for each attribute, with 1 being "on" and 0 being "off." Constraints are explained in more detail as follows. Experiments have the goal of finding rules with perfusion measurements and risk factors in the antecedent and artery disease measurements in the consequent of a rule.

B. Parameter Settings

We explain settings for program parameters that were based both on medical opinion and experimental validation.

1) Transformation Parameters: Refer to Table I to interpret attribute names. The LAD, RCA, LCX, and LM attributes contain the percentage of vessel narrowing (or blockage) compared to a healthy artery. Attributes LAD, LCX, and RCA were binned at 50% and 70%. In cardiology, 70% or higher indicates significant coronary disease, and 50% indicates borderline disease. A value lower than 50% means the patient is healthy. The most common cutoff value used by the cardiology community to distinguish healthy from sick patients is 50%. The LM artery is treated different because it poses higher risk than the other three arteries. LM was binned at 30% and 50%. The reason behind it is that both the LAD and the LCX arteries branch from the LM artery, and thus a defect in LM is more likely to cause a larger diseased heart region. That is, narrowing (blockage) in the LM artery is likely to produce more disease than blockages on the other arteries. That is why its cutoff values are set 20% lower than the other vessels. The nine heart regions (AL, IL, IS, AS, SI, SA, LI, LA, AP) were partitioned into two ranges at a cutoff point of 0.2, meaning a perfusion measurement greater or equal than 0.2 indicated a severe defect. CHOL was partitioned with cutoff points 200 (warning) and 250 (high). These values correspond to known medical settings. The training fraction was set at $\tau = 50\%$. Every time the algorithm is run, new samples are created. Finally, only the four artery measurements had negation to find rules referring to healthy patients and sick patients. The rest of attributes did not require negation. Since most risk factors were binary and perfusion measurements were divided into two ranges, this eliminated the need to use negation on them. Negation was not considered useful for age and cholesterol level. After transforming S into D, we ended with 68 items. Therefore, $n = 655$ and $m = 68$ according to our definitions from Section II.

2) Constraints: The first parameter is the maximum association size κ. We used $\kappa \in \{2, \ldots, 4\}$ to study the individual impact of constraints. In the second set of experiments, to get simple rules, $\kappa = 4$. A lower κ produces fewer and simpler rules. A higher κ significantly increases the number of rules, and they become more complex.

The training sample fraction was $\tau = 50\%$. Association rule mining had the following thresholds for metrics. The minimum training support was fixed at $\psi = 1\% \approx 3$. That is, rules involving two or one patient(s) were filtered out. From a medical point of view, rules with high confidence are desirable but, unfortunately, they are infrequent. Based on previous experiments [4], [5] and the domain expert opinion, the minimum confidence was set at $\alpha = 70\%$. Rules with confidence lower than 70% are not medically reliable. In general, the lift threshold was $\lambda = 1$.

We will now explain attribute constraints. Table I summarizes attribute constraints with "1" indicating the constraint was on and "0" indicating the constraint was off. Item filtering, group constraints, and antecedent/consequent constraints were set as follows. AGE, SEX, and CHOL did not have item filtering, meaning that any AGE and CHOL range or gender could appear in any predictive rule. The remaining risk factors, all perfusion measurements, and arteries had item filtering. To predict absence of disease ($P =$"N") risk factors ("Y"), high perfusion measurements, and high artery blockage were filtered out. To predict existence of disease, ($P =$"Y") risk factors ("N"), low perfusion measurements, and low artery blockage were filtered out. AGE, arteries, SEX, SMOKE, and CHOL were not group-constrained. The nine perfusion measurements of the heart were constrained to be in the same group (group 1) to predict absence or existence of disease. When predicting no disease, ($P =$"N") HTA, DIAB, HYPLPD, FHCAD, CLAUDI were in group 2, and PANGIO, PSTROKE, PCARSUR were in group 3; these attributes were not constrained to predict existence of disease. Risk factors and perfusion measurements were constrained to appear in the antecedent $(\text{ac}(A_l) = 1)$. Arteries were constrained to appear in the consequent $(\text{ac}(A_l) = 2)$.

TABLE II
ITEM FILTERING CONSTRAINT: NUMBER OF ASSOCIATIONS
($0 =$ OFF, $1 =$ ON)

	P="N" (absence) itemFilter		P="Y" (existence) itemFilter	
κ	0	1	0	1
2	2003	351	2051	451
3	30100	2373	32844	3014
4	258906	9930	309146	12073

TABLE III
GROUP CONSTRAINT: NUMBER OF ASSOCIATIONS
($0 =$ OFF, $1 =$ ON)

κ	both group=0	P="N" (absence) group= 1	P="Y" (existence) group= 1
2	485	351	453
3	3940	2365	3095
4	19643	9818	11782

TABLE IV
ANTECEDENT/CONSEQUENT CONSTRAINT: NUMBER OF ASSOCIATION RULES ($0 =$ OFF, $1 =$ ON)

	P="N" (absence)		P="Y" (existence)	
κ	ac=0	ac= 1	ac=0	ac=1
2	144	24	52	7
3	2359	306	1222	110
4	16733	2218	9051	957

TABLE V
NUMBER OF ASSOCIATIONS AND RULES IN D_{train} AND D_{test} VARYING κ TO PREDICT EXISTENCE OF DISEASE ($P =$ "Y")

		no. of associations		no. of rules		
κ	ψ	train	test	train	test	time
2	0.01	493	467	8	5	7
3	0.01	3286	2948	145	77	51
4	0.01	11610	9327	1222	342	228

C. Number of Association Rules With and Without Constraints

We study each constraint individually to measure its relative importance with respect to number of patterns and time. Since the number of patterns is large, we turn all constraints on by default, turning one off at a time to see the increase in output size. We concentrate on studying the impact on the training set D_{train} since it is the one used to build the predictive association rules, and it is the most time demanding. Each table indicates when the constraint is turned on or off. Recall that D_{train} is built as a random sample from D every time. Each run generates a different set of rules. We repeated experiments five times and averages are reported.

Table II summarizes the number of associations at different itemset sizes, turning item filtering on and off (i.e., itemFilter $= 0$ and itemFilter $= 1$, respectively). Recall from Section IV-B that the prediction of absence and existence of heart disease have different settings. We can see that there is an important reduction in the number of patterns at $\kappa = 2$ and $\kappa = 3$, but there is an outstanding reduction in size at $\kappa = 4$. It is clear that more than 200 000 patterns are impossible to interpret. We can see that the reduction in size is similar for predicting the absence or existence of disease. We do not show the number of rules, but without constraining there is potentially an exponential number of rules that can be derived from associations. These experimental results indicated it was necessary to split experiments into two sets, one for predicting absence of disease ($P =$"N") and another one to predict its existence ($P =$"Y"). Item filtering is a simple, but important constraint.

We now study the impact of the *group* constraint. Results are shown in Table III, where "group $= 0$" indicates the constraint was off, and "group $= 1$" indicates the constraint was on. There is a different setting for this constraint depending on the prediction goal. For absence of disease ($P =$"N") there are three groups. For existence of disease ($P =$"Y") only heart region perfusion measurements are group-constrained. For $\kappa = 2$, the reduction in size is marginal. At $\kappa = 3$, the reduction in size becomes better, especially for predicting no disease. Finally, for $\kappa = 4$, the reduction in size is about 50% in both cases. At all itemset sizes, when the constraint is turned on, the reduction in size is bigger for predicting no disease, which is due to using three attribute groups instead of one.

Table IV shows the impact of applying the ac constraint. If "ac $= 0$," the constraint was off; otherwise "ac $= 1$." The ac settings are the same for both prediction goals, with the only difference being that item filtering and group constraints are different, as explained in Section IV-B. The impact of this constraint is significant. In every case, there is an order of magnitude reduction in size. The number of rules goes beyond 1000 only in the case of trying to predict absence of disease at $\kappa = 4$.

The number of rules shown when the constraint is on will not be the actual number, because rules will be filtered validating them on D_{test}, which is explained in the following section.

We conclude that item filtering and the antecedent/consequent constraint are essential to get a manageable number of rules. Given the number of rules and running time, it is necessary to have two sets of constraints, depending on the prediction goal. The group constraint has an important impact, but not as significant.

D. Validating Rules Based on Training and Test Sets

In this section, we show the effect of filtering association rules based on the test set. Recall that association rules are tested for generality and validity by partitioning the input data set into a training set and a test set. A valid rule must have minimum metrics on both sets. We would like to understand the extent to which the number of rules is reduced by varying κ or ψ. In the following experiments, we focus on predicting existence of disease ($P =$"Y"). As in the previous section, we repeated each experiment five times and averages are reported.

The first set of experiments shows the importance of filtering rules on the test by varying κ. Table V summarizes results. The reduction in the number of associations is small, with a reduction of about 10%–15%. The reduction becomes much more important for the number of rules. For $\kappa = 2$, the impact is small in most cases, which indicates most rules can be generalized. For $\kappa = 3$, the reduction is more than 50%, providing evidence that many rules are particular to the training sample. At $\kappa = 4$, the number of rules in the test set is about 30% of the total, with a reduction of about 70%, providing evidence that most rules may be particular to the training set. The trend indicates there will be a combinatorial explosion of rules that are valid only on the training set. Time grows fast as rule size κ grows.

TABLE VI
NUMBER OF RULES IN D_{train} AND D_{test} SETS VARYING MINIMUM SUPPORT ψ TO PREDICT EXISTENCE OF DISEASE (P ="Y")

κ	ψ	train	test	time
4	0.100	62	33	24
4	0.050	163	114	62
4	0.010	1222	342	228
4	0.005	2022	497	248

The difference in the relative number of patterns for associations and rules can be explained by the fact that associations are filtered on D_{train} and D_{test} based only on support, but rules require support, confidence, and lift to be greater than or equal to the respective thresholds in D_{train} and D_{test}. Therefore, validation is stricter for rules than for associations. The last column in Table V contains elapsed times in seconds. These times include all steps in the algorithm, starting with transformation and repeating the train/test cycle only once. As can be seen, time grows fast as k increases, even with all the number, of patterns eliminated by constraints and the validation phase.

The second set of experiments studies the growth of the number of patterns varying ψ, which is the main parameter to control the number of rules. Table VI contains a summary of results. At high support levels, the reduction in the number of rules is about 40%. For low support levels, the number of rules goes down to less than 35%. This indicates that as support is lowered, more rules become particular to the training set because they do not meet the minimum metrics in the test set. The last column in Table VI contains total elapsed times in seconds. Time growth is not as fast compared to varying κ, because constraints and test set validation significantly reduce the number of patterns.

E. Predicting Heart Disease With Association Rules

This section presents medically significant rules that predict artery disease based on perfusion measurements and risk factors. Rules are presented in two groups: 1) those which express that if there is no risk factor and a low perfusion measurement, then there is absence of heart disease and 2) those that state that if there exists a risk factor or a high perfusion measurement then there exists heart disease. We sought rules involving at most $\kappa = 4$ attributes; the reason behind this setting was that we wanted to get fewer and simpler rules. The default program parameter settings are described in Section IV-B. The rule metrics thresholds were $\psi = 1\%$, $\alpha = 70\%$, and $\lambda = 1$. These metrics allowed getting rules to appear in at least three patients, having reasonable sensitivity, and filtering out rules with unacceptable lift. We ran $t = 5$ train/test cycles to get five independent sets of rules to compute the intersection of rule sets, where each rule support, confidence, and lift are computed as averages of rule metrics from all test rule sets [by (1)–(3)]. Given the small size of our data set, $t = 5$ seemed reasonable to validate rules. Refer to Table I to understand the abbreviations of attribute names.

Medical doctors use sensitivity and specificity as two basic statistics to validate results. Sensitivity is defined as the probability of correctly identifying sick patients, whereas specificity is defined as the probability of correctly identifying healthy individuals. In general, it is difficult to get high sensitivity and high specificity with the same predictive model. That is, increasing sensitivity decreases specificity, and vice versa. In general, our experiments were optimized to increase sensitivity and decrease specificity. That is, it was preferred to increase the number of patients predicted to have heart disease at the risk of including other patients with borderline heart disease, that were perhaps healthy. Lift was used together with confidence to understand sensitivity and specificity.

TABLE VII
SUMMARY OF RULES IN R_N AND R_Y

Rule subset	R_N		R_Y	
	No. of rules	%	No. of rules	%
All	850	100	197	100
$c = 1$ (+)	241	28	10	5
$c \geq 0.9$ (+)	519	61	39	20
$l \geq 2$ (+)	0	0	20	10
$l \geq 1.5$	76	9	180	91
$l \geq 1.2$	346	40	197	100
$s \geq 0.3$	60	7	0	0
$s \geq 0.2$	135	16	2	1
$s \geq 0.1$	351	41	35	18
Two items in conseq. (+)	23	3	0	0
Two items (+)	7	1	4	2

1) Rule Summary: Table VII shows a summary of rules that remain valid on all test sets. The most significant or interesting rules are marked with a + symbol, meaning rules with very high confidence or high lift. The first general observation is there are many more rules predicting absence of disease; unfortunately, most such rules had low lift. These results state that finding interesting rules to predict the existence of disease are likely to appear mainly at low support levels. There were more rules predicting existence of disease with high lift (above 2). The proportion of high confidence ($\geq 90\%$) rules was much higher for rules predicting absence of disease; in fact, more than half of the total number of rules. More than 60% of rules predicting absence of disease had lift close to 1 ($l < 1.2$), stressing the difficulty in isolating important risk factors or low perfusion measurements in predicting whether a specific artery is healthy. There were many more rules with high support or high confidence predicting absence of disease compared to their counterpart.

Any rule with confidence greater or equal than 90%, with high lift (above 1.5), or with two arteries in the consequent can enrich medical knowledge. In particular, rules with very high confidence and high lift that are not currently part of an expert system knowledge base, or which do not correspond to common diagnostic patterns, represent valuable new medical knowledge. Getting rules with combinations of arteries in the consequent was one of the reasons to use association rules. Unfortunately, there were no rules predicting the existence of disease having two arteries in the consequent, but there were 23 for healthy arteries. This is an important finding, because even though it is possible to identify a link between perfusion measurements/risk factors and one diseased artery, there was not a single rule predicting two diseased arteries. Manual inspection revealed there were some rules predicting the existence of disease in two arteries on the training set, but unfortunately none passed validation on all test sets. Simple rules, which are preferred because they

$c = 1$:
LI[-1,0.2) SMOKE=n CHOL[200,250) ⇒ LCX[0,50) s=0.03,c=1,l=1.62
HTA=n SMOKE=n CHOL[0,200) ⇒ LCX[0,50) s=0.02, c=1,l=1.62
AGE[40,60) IS[-1,0.2) CHOL[0,200) ⇒ LM[0,30) s=0.05,c=1,l=1.07
SEX=F HTA=n CHOL[0,200) ⇒ RCA[0,50) s=0.02,c=1,l=1.76
$c \geq 0.9$:
AP[-1,0.2) CHOL[250,500) ⇒ LM[0,30) s=0.06,c=0.99,l=1.06
AGE[60,100) LA[-1,0.2) DIAB=n ⇒ LM[0,30) s=0.22,c=0.95,l=1.01
AS[-1,0.2) SEX=F FHCAD=n ⇒ LM[0,30) s=0.15,c=0.98,l=1.06
LI[-1,0.2) SMOKE=n PCARSUR=n ⇒ LM[0,30) s=0.06,c=0.99,l=1.06
AGE[0,40) AL[-1,0.2) DIAB=n ⇒ LAD[0,50) s=0.03,c=0.91,l=1.86
AGE[40,60) IL[-1,0.2) SEX=F ⇒ LCX[0,50) s=0.10,c=0.92,l=1.49
AGE[40,60) IL[-1,0.2) SMOKE=n ⇒ RCA[0,50) s=0.10,c=0.91,l=1.6
AGE[40,60) DIAB=n ⇒ LM[0,30) s=0.33,c=0.95,l=1.02
$l \geq 1.5$:
AGE[0,40) AL[-1,0.2) ⇒ LAD[0,50) s=0.03,c=0.91,l=1.86
SEX=F HTA=n CHOL[0,200) ⇒ RCA[0,50) s=0.02,c=1.00,l=1.76
AGE[40,60) LI[-1,0.2) CHOL[200,250) ⇒ LCX[0,50) s=.02,c=.95,l=1.5
Two arteries in the consequent:
AGE[0,40) AL[-1,0.2) ⇒ LM[0,30) LAD[0,50) s=0.03,c=0.91,l=1.96
SEX=F CHOL[0,200) ⇒ LM[0,30) RCA[0,50) s=0.03,c=0.80,l=1.45
IL[-1,0.2) SEX=F ⇒ LM[0,30) LCX[0,50) s=0.18,c=0.77,l=1.31
Two items (simple):
CLAUDI=n ⇒ LM[0,30) s=0.71,c=0.94,l=1.01
AGE[40,60) ⇒ LM[0,30) s=0.37,c=0.95,l=1.01
AS[-1,0.2) ⇒ LM[0,30) s=0.73,c=0.94,l=1.01
CHOL[250,500) ⇒ LM[0,30) s=0.07,c=0.99,l=1.06

Fig. 2. Association rules predicting absence of heart disease (P ="N").

$c = 1$ and $l >= 2$:
IS[0.2,1.0) CLAUDI=y PSTROKE=y ⇒not RCA[0,50) s=0.02,c=1,l=2.3
$c = 1$ and $l < 2$:
SA[0.2,1.1) CHOL[200,250) ⇒ not LAD[0,50) s=.03,c=1,l=1.98
SA[0.2,1.1) SMOK=y CHOL[200,250)⇒not LAD[0,50) s=.02,c=1,l=1.9
SI[0.2,1.1) SEX=M CHOL[200,250) ⇒ not LAD[0,50) s=.02,c=1,l=1.9
$c \geq 0.9$ and $l \geq 2$:
SA[0.2,1.1) CHOL[200,250) ⇒ LAD[70,100) s=0.03,c=0.95,l=3
SI[0.2,1.1) SEX=M CHOL[200,250) ⇒LAD[70,100) s=0.01,c=.91,l=2.9
SA[0.2,1.1) PANGIO=y ⇒ LAD[70,100) s=0.02,c=0.91,l=2.9
$c >= 0.9$ and $l < 2$:
DIAB=y SMOKE=y 0.CHOL[0,200) ⇒not LAD[0,50) s=.02,c=.93,l=1.8
SA[0.2,1.1) CLAUDI=y PSTROK=y ⇒not LAD[0,50) s=.02,c=.95,l=1.8
IL[0.2,1.1) SEX=M CHOL[200,250) ⇒not LAD[0,50) s=.03,c=.93,l=1.8
AP[0.2,1.1) HYPLPD=y CHOL[0,200) ⇒not LAD[0,50) s=.02,c=.96
AGE[60,100) AP[0.2,1.1) SEX=F ⇒ not LAD[0,50) s=.05,c=.94,l=1.8
$c < 0.9$ and $l >= 2$:
AGE[60,100) SA[0.2,1.1) PANGIO=y ⇒LAD[70,100) s=.02,c=.88,l=2.8
AGE[40,60) SEX=M CHOL[250,500) ⇒not LCX[0,50) s=.02,c=.85,l=2
SA[0.2,1.1) CLAUDI=y ⇒ LAD[70,100) s=0.03,c=0.76,l=2.46
AGE[60,100) IL[0.2,1.1) CHOL[250,500) ⇒ RCA [70,100) s=.02,c=.89
Two items (simple):
AP[0.2,1.1) ⇒ not(LAD[0,50)) s=0.24,c=0.78,l=1.54
SA[0.2,1.1) ⇒ not(LAD[0,50)) s=0.17,c=0.80,l=1.58

Fig. 3. Association rules predicting existence of heart disease (P ="Y").

tend to be easier to interpret by medical doctors and have higher support, were scarce; there were only seven and four rules predicting the absence and existence of disease, respectively.

2) Predicting Absence of Heart Disease: Fig. 2 shows some selected rules with high quality metrics. For each numeric attribute X, we use the notation $X[a, b)$, meaning that $a \leq X$ and $X < b$; this provides a more concise notation. We have a group of rules with 100% confidence; the important fact was that there was only one rule for RCA involving only risk factors and no perfusion, and there was only one rule for LCX with a low perfusion measurement. Rules for LM are abundant, combining low perfusion measurements in most regions and risk factors equal to "N." When looking at the set with confidence above 0.9 and below 1, there are a few more LCX rules involving low perfusion measurements, but only one RCA rule involves perfusion and risk factors. Then we have the rule set with high lift ($l \geq 1.5$) that involved LAD, LCX, and RCA alone in the consequent and 11 rules combining LAD and LM. In general, the support for those rules was low, and confidence was below 90%.

Many rules predicting no disease represent valuable new medical knowledge. In particular, rules with borderline risk factors reveal interesting relationships among risk factors, normal (low) perfusion measurements, and healthy arteries. Those rules with warning cholesterol levels (200–250) show relationships to the LCX and RCA arteries. A rule with high cholesterol levels (250–500) implies a still-healthy LM artery. Rules with adult age (40–60) imply a healthy RCA artery. A rule about males shows a healthy LM artery, but its lift is low.

Rules with high support in [0.2, 0.3] involved only the LM, LCX, and RCA arteries. The rules with $s \geq 0.4$ involved only LM. This confirms the higher likelihood of having a healthy LM artery, given the fact that the other arteries branch from it. Another observation is that LM rules have higher confidence than rules with any of the other arteries. The rules with two items in the consequent are valuable since they were one of the main reasons to use association rules to predict multiple combinations of target variables. They involved all combination between LM and the other three arteries, but there was not any combination among the branching arteries. This means it is unlikely two branching arteries are healthy in the same person. Finally, we show some rules with only one item in the antecedent; these rules have high confidence and high support, but low lift.

3) Predicting Existence of Heart Disease: Fig. 3 shows some selected rules with high quality metrics. Rules with high confidence and high lift are extremely valuable. For numeric attributes X, we use the notation $X(a, b)$, meaning that $a \leq X$ and $X < b$. Most rules with 100% confidence referred to the LAD, and only one referred to the RCA artery. We show the only rule that had $c = 1$ and $l \geq 2$ involving the RCA artery; this rule has significant medical value because it combines two risk factors and a specific perfusion defect with high quality metrics. The next group involves rules with $c = 1$ and l slightly below 2; all those rules involve the LAD artery and none of the other arteries; most rules involved warning cholesterol levels and older patients (above 60). The next rule group involves rules with $c \geq 0.9$ and $l \geq 2$; all seven rules involved LAD with warning cholesterol levels; most involved sex = M and perfusion defects in SA. The next rule group involves rules with $c \geq 0.9$ and $l < 2$; there were only 22 rules involving only LAD. Curiously, some rules had sex = F and low cholesterol levels; HTA and PSTROKE come out as important factors. Then we have 12 interesting rules with lower confidence but high lift ($c < 9$ and $l \geq 2$); these rules involve only LAD and RCA, with older age, high cholesterol, and perfusion defects in SA and AP. It is interesting that in rules with high confidence ($c \geq 0.9$) or high lift ($l \geq 2$), high cholesterol levels, older age, hypertension (HTA), and smoking appear as risk factors, but they do not appear combined without a perfusion defect.

There were only two rules with support above 0.2 involving the LAD artery. There were only three rules with only risk factors involving low cholesterol levels, sex = M or DIAB/SMOKE; all these rules had LAD in the consequent. We found a few rules combining smoking with high cholesterol, a well-known risk combination that included perfusion defects.

Finally, there were only four simple rules with only two items (one item in the antecedent), and we show two of them. These rules have fairly high support, borderline confidence, and small lift. These rules confirm it is not possible to find a single attribute predicting heart disease with high confidence and high lift. Several rules predicting disease represent valuable new medical knowledge. In particular, rules with highly narrowed arteries (narrowing above 70%), with high cholesterol levels (250–500), less significant risk factors (like DIAB or FHCAD), or with sex = F show surprising relationships among well-known risk factors.

V. Related Work

We start by discussing research on data mining techniques used with medical and biological data. Important issues [13] when using machine learning or data mining techniques in the medical domain include fragmented data collection, strict privacy regulations, rich data types (image, numeric, categorical, missing information), complex taxonomies classifying attributes, and an already rich and complex knowledge base. The HDP system to aid diagnosis of heart disease is presented in [22], [23], and [24]. This body of work provides evidence that a computer program can improve differential diagnosis for cardiovascular disease made by medical doctors. Bayesian networks are combined with AI inference mechanisms [23]. More general research using association rules in the medical community includes the following. Infection detection and monitoring has been automated by association rules [25], [26]. Association rules describe what drugs are frequently coprescribed with antacids [27]. Frequent patterns in gene data have been discovered with association rules [28], [29]. Protein interaction within protein groups is an important topic [30]. Common risk factors in pediatric diseases are another important medical application [31]. Fuzzy sets have been used to extend association rules [32]. To our knowledge, reducing the number of association rules with constraints and validating/filtering rules on an independent set has not previously been studied in the medical domain.

Association rules were introduced in a well-known article [19]. In [33], there is a study of I/O complexity of early association rule algorithms as well as some statistical metrics. Both [12] and [34] use different approaches to automatically bin numeric attributes. Domain experts preferred to use standard cutoff numbers for binning numeric attributes, to improve result interpretation and validation with previous work. There is previous work on using constraints to reduce the number of rules. Our constraints exhibit similarities to those proposed in [14], [15], [20]. Reference [20] proposes algorithms that can incorporate constraints to include or exclude certain items in the frequent itemset generation phase. Association rule constraints are studied in depth in [15], where constraints are defined as item boolean expressions between two variables. Subsequently, [14] extended that approach by allowing more general expressions on items. It is well known that simple constraints on support can be used for pruning the search space in Phase 1 [17]. Several search constraints for association rules are studied in [16].

VI. Conclusion

We used association rules to predict the degree of narrowing in four arteries based on heart perfusion measurements and risk factors. We studied two complementary tasks: predicting the absence and predicting the existence of heart disease. We focused on two main research issues. The first issue is the large number of rules that are obtained by the standard association rule algorithm. The second issue is the validation of rules on an independent set, which is required to eliminate unreliable rules, or rules that cannot be generalized. Four constraints were proposed in this work to reduce the number of rules: item filtering, attribute grouping, maximum itemset size, and antecedent/consequent rule filtering. Contrasting with previous work, our constraints are specified on raw attributes instead of items, item filtering is applied earlier before generating frequent itemsets, and the group constraint induces a partition on attributes allowing easier manipulation. In order to validate rules, we used the train and test approach that uses two disjoint samples from a data set to search and validate rules. All features are assembled together in one algorithm that combines search constraints and train/test validation. The algorithm performs several train and test cycles to achieve basic cross-validation and reduce the number of rules with poor generalization potential. Experiments on a real data set studied the impact of constraints and the elimination of unreliable rules with validation on the test set. The reduction in output size provided by constraints and validation is significant. In particular, the reduction provided by item filtering and antecedent/consequent rule filtering was about an order of magnitude. We presented medically significant rules discovered on our medical data set that remain valid in several independent train/test cycles. Rules predicting absence of disease are abundant, but have low lift metrics and therefore poor predictive quality. The set of rules predicting the existence of disease is smaller and Tends to have higher lift, which is consistent with our medical goal. Rules predicting heart disease provide accurate profiles of patients with localized heart problems, specific risk factors, and the degree of disease in one artery.

There are many interesting aspects for future research. Constraints for association search may be relaxed, given the high number of rules that are filtered out in the validation phase. Some statistic or quality metric is needed to compare different sets of rules. We would like to know which attributes make rules fail the validation phase and which attributes tend to appear more frequently after validation. Support, confidence, and lift have different importance to filter rules on the test set; we conjecture that confidence is the most important metric to validate rules. Given the small size of our data set, we could not apply more rigorous techniques like train/validate/test or ten-fold cross validation; such techniques may be applied on larger data sets. Artery attributes may be helped with fuzzy discretization in order to get rules with higher confidence or higher support. A hierarchy of perfusion measurements is required to control the rule discovery process, in order to increase or decrease sensitivity to detect sick patients with high accuracy without significantly losing sensitivity.

ACKNOWLEDGMENT

The author would like to thank the Emory University Radiology Department for providing the medical data set used in this article. The author would also like to thank Dr. C. Santana, from the Emory University Hospital, for sharing his expertise on cardiology.

REFERENCES

[1] 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," *J. Nucl. Med.*, vol. 40, no. 5, 1999.

[2] C. Ordonez and E. Omiecinski, "Discovering association rules based on image content," in *Proc. IEEE Advances in Digital Libraries Conf. (ADL'99)*, 1999, pp. 38–49.

[3] L. Braal, N. Ezquerra, E. Schwartz, and E. V. Garcia, "Analyzing and predicting images through a neural network approach," in *Proc. Visualization in Biomedical Computing*, 1996, pp. 253–258.

[4] C. Ordonez, E. Omiecinski, L. de Braal, C. Santana, and N. Ezquerra, "Mining constrained association rules to predict heart disease," in *Proc. IEEE ICDM Conf.*, 2001, pp. 433–440.

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

[6] N. Ezquerra and R. Mullick, "Perfex: An expert system for interpreting myocardial perfusion," *Expert Syst. Appl.*, vol. 6, pp. 455–468, 1993.

[7] D. Cooke, C. Santana, T. Morris, L. de Braal, C. Ordonez, E. Omiecinski, N. Ezquerra, and E. V. Garcia, "Validating expert system rule confidences using data mining of myocardial perfusion SPECT databases," in *Proc. Computers in Cardiology Conf.*, 2000, pp. 116–119.

[8] P. Clark and T. Nisblett, "The CN2 induction algorithm," *Mach. Learn.*, vol. 3, pp. 261–283, 1989.

[9] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[10] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, 1st ed., New York: Springer-Verlag, 2001.

[11] R. Bayardo and R. Agrawal, "Mining the most interesting rules," in *Proc. ACM KDD Conf.*, 1999, pp. 145–154.

[12] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," in *Proc. ACM SIGMOD Conf.*, 1996, pp. 1–12.

[13] J. F. Roddick, P. Fule, and W. J. Graco, "Exploratory medical knowledge discovery: Experiences and issues," *SIGKDD Explorations*, vol. 5, no. 1, pp. 94–99, 2003.

[14] L. V. S. Lakshmanan, R. Ng, J. Han, and A. Pang, "Optimization of constrained frequent set queries with 2-variable constraints," in *Proc. ACM SIGMOD Conf.*, 1999, pp. 157–168.

[15] R. Ng, L. Lakshmanan, and J. Han, "Exploratory mining and pruning optimizations of constrained association rules," in *Proc. ACM SIGMOD Conf.*, 1998, pp. 13–24.

[16] J. Pei and J. Han, "Constraints in data mining: Constrained frequent pattern mining: A pattern-growth view," *SIGKDD Explorations*, vol. 4, no. 1, pp. 31–39, 2002.

[17] K. Wang, Y. He, and J. Han, "Pushing support constraints into association rules mining," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 3, pp. 642–658, May/Jun. 2003.

[18] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[19] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Conf.*, 1993, pp. 207–216.

[20] R. Srikant, Q. Vu, and R. Agrawal, "Mining association rules with item constraints," in *Proc. ACM Knowledge Discovery and Data Mining Conf.*, 1997, pp. 67–73.

[21] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *Proc. ACM SIGMOD Conf.*, 2000, pp. 1–12.

[22] H. S. Fraser, W. J. Long, and S. Naimi, "Evaluation of a cardiac diagnostic program in a typical clinical setting," *J. Amer. Med. Inform. Assoc. (JAMIA)*, vol. 10, no. 4, pp. 373–381, 2003.

[23] W. J. Long, "Medical reasoning using a probabilistic network," *Appl. Artif. Intell.*, vol. 3, pp. 367–383, 1989.

[24] W. J. Long, H. S. Fraser, and S. Naimi, "Reasoning requirements for diagnosis of heart disease," *Artif. Intell. Med.*, vol. 10, no. 1, pp. 5–24, 1997.

[25] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, and S. A. Moser, "Association rules and data mining in hospital infection control and public health surveillance," *J. Amer. Med. Inform. Assoc. (JAMIA)*, vol. 5, no. 4, pp. 373–381, 1998.

[26] S. E. Brossette, A. P. Sprague, W. T. Jones, and S. A. Moser, "A data mining system for infection control surveillance," *Meth. Inf. Med.*, vol. 39, no. 4, pp. 303–310, 2000.

[27] T. J. Chen, L. F. Chou, and S. J. Hwang, "Application of a data mining technique to analyze coprescription patterns for antacids in Taiwan," *Clin. Ther.*, vol. 25, no. 9, pp. 2453–2463, 2003.

[28] C. Becquet, S. Blachon, B. Jeudy, J. F. Boulicaut, and O. Gandrillon, "Strong association-rule mining for large-scale gene-expression data analysis: A case study on human SAGE data," *Genom Biol.*, vol. 3, no. 12, 2002.

[29] C. Creighton and S. Hanash, "Mining gene expression databases for association rules," *Bioinformatics*, vol. 19, no. 1, pp. 79–86, 2003.

[30] T. Oyama, K. Kitano, T. Satou, and T. Ito, "Extraction of knowledge on protein-protein interaction by association rule discovery," *Bioinformatics*, vol. 18, no. 5, pp. 705–714, 2002.

[31] S. M. Down and M. Y. Wallace, "Mining association rules from a pediatric primary care decision support system," in *Proc. AMIA Symp.*, 2000, pp. 200–204.

[32] M. Delgado, D. Sanchez, M. J. Martin-Bautista, and M. A. Vila, "Mining association rules with improved semantics in medical databases," *Artif. Intell. Med.*, vol. 21, no. 1-3, pp. 241–245, 2001.

[33] J. L. Han, "Background for association rules and cost estimate of selected mining algorithms," in *Proc. ACM Information and Knowledge Management Conf.*, 1996, pp. 73–80.

[34] B. Lent, A. Swami, and J. Widom, "Clustering association rules," in *Proc. IEEE ICDE Conf.*, 1997, pp. 220–231.

Carlos Ordonez received the B.Sc. degree in applied mathematics and the M.S. degree in computer science from the UNAM University, Mexico City, Mexico, in 1992 and 1996 respectively, and the Ph.D. degree in computer science from the Georgia Institute of Technology, Atlanta, in 2000.

He is currently with Teradata (NCR), San Diego, CA, conducting research on databases and machine learning. He has published more than 20 scientific articles in international conferences and journals.