Comparing Association Rules and Decision Trees for Disease Prediction

Carlos Ordonez
University of Houston
Houston, TX, USA

ABSTRACT

Association rules represent a promising technique to find hidden patterns in a medical data set. The main issue about mining association rules in a medical data set is the large number of rules that are discovered, most of which are irrelevant. Such number of rules makes search slow and interpretation by the domain expert difficult. In this work, search constraints are introduced to find only medically significant association rules and make search more efficient. In medical terms, association rules relate heart perfusion measurements and patient risk factors to the degree of stenosis in four specific arteries. Association rule medical significance is evaluated with the usual support and confidence metrics, but also lift. Association rules are compared to predictive rules mined with decision trees, a well-known machine learning technique. Decision trees are shown to be not as adequate for artery disease prediction as association rules. Experiments show decision trees tend to find few simple rules, most rules have somewhat low reliability, most attribute splits are different from medically common splits, and most rules refer to very small sets of patients. In contrast, association rules generally include simpler predictive rules, they work well with user-binned attributes, rule reliability is higher and rules generally refer to larger sets of patients.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; J.3 [**Computer Applications**]: Life and Medical Sciences —*Health*

General Terms

Algorithms, Experimentation

Keywords

Association rule, decision tree, medical data

1. INTRODUCTION

One of the most popular techniques in data mining is association rules [1, 2]. Association rules have been successfully applied with basket, census and financial data [17]. On the other hand, medical data is generally analyzed with classifier trees, clustering [17], regression [18] or statistical tests [18], but rarely with association rules. This work studies association rule discovery in medical records to improve disease diagnosis when there are multiple target attributes.

Association rules exhaustively look for hidden patterns, making them suitable for discovering predictive rules involving subsets of the medical data set attributes [26, 25]. Nevertheless, there exist three main issues. First, in general, in a medical data set a significant fraction of association rules is irrelevant. Second, most relevant rules with high quality metrics appear only at low support (frequency) values. Third and most importantly, the number of discovered rules becomes extremely large at low support. With these issues in mind, we introduce search constraints to reduce the number of association rules and accelerate search. On the other hand, decision trees represent a well-known machine learning technique used to find predictive rules combining numeric and categorical attributes, which raises the question of how association rules compare to induced rules by a decision tree. With that motivation in mind, we compare association rules and decision trees with respect to accuracy, interpretability and applicability in the context of heart disease prediction.

The article is organized as follows. Section 2 introduces definitions for association rules and decision trees. Section 3 explains how to transform a medical data set into a binary format suitable for association rule mining, discusses the main problems encountered using association rules, and introduces search constraints to accelerate the discovery process. Section 4 presents experiments with a medical data set. Association rules are compared with predictive rules discovered by a decision tree algorithm. Section 5 discusses related research work. Section 6 presents conclusions and directions for future work.

2. DEFINITIONS

2.1 Association Rules

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 X or Y, is called an itemset. In this work, a transaction is a patient record transformed into a binary format where only

http://doi.acm.org/10.1145/1183568.1183573

positive binary values are included as items. This is done for efficiency purposes because transactions represent sparse binary vectors.

Let $P(X)$ be the probability of appearance of itemset X in D and let $P(Y|X)$ be the conditional probability of appearance of itemset Y given itemset X appears. For an itemset $X \subseteq \mathcal{I}$, $support(X)$ is defined as the fraction of transactions $T_i \in D$ such that $X \subseteq T_i$. That is, $P(X) = support(X)$. The support of a rule $X \Rightarrow Y$ is defined as $support(X \Rightarrow Y) = P(X \cup Y)$. An association rule $X \Rightarrow Y$ has a measure of reliability called $confidence(X \Rightarrow Y)$ defined as $P(Y|X) = P(X \cup Y)/P(X) = support(X \cup Y)/support(X)$. The standard problem of mining association rules [1] is to find all rules whose metrics are equal to or greater than some specified minimum support and minimum confidence thresholds. A k-itemset with support above the minimum threshold is called frequent. We use a third significance metric for association rules called *lift* [25]: $lift(X \Rightarrow Y) = P(Y|X)/P(Y) = confidence(X \Rightarrow Y)/support(Y)$. Lift quantifies the predictive power of $X \Rightarrow Y$; we are interested in rules such that $lift(X \Rightarrow Y) > 1$.

2.2 Decision Trees

In decision trees [14] the input data set has one attribute called class $\mathcal{C}$ that takes a value from K discrete values $1, \ldots, K$, and a set of numeric and categorical attributes $A_1, \ldots, A_p$. The goal is to predict $\mathcal{C}$ given $A_1, \ldots, A_p$. Decision tree algorithms automatically split numeric attributes A_i into two ranges and they split categorical attributes A_j into two subsets at each node. The basic goal is to maximize class prediction accuracy $P(\mathcal{C} = c)$ at a terminal node (also called node purity) where most points are in class c and $c \in \{1, \ldots, K\}$. Splitting is generally based on the information gain ratio (an entropy-based measure) or the gini index [14]. The splitting process is recursively repeated until no improvement in prediction accuracy is achieved with a new split. The final step involves pruning nodes to make the tree smaller and to avoid model overfit. The output is a set of rules that go from the root to each terminal node consisting of a conjunction of inequalities for numeric variables ($A_i <= x, A_i > x$) and set containment for categorical variables ($A_j \in \{x, y, z\}$) and a predicted value c for class $\mathcal{C}$. In general decision trees have reasonable accuracy and are easy to interpret if the tree has a few nodes. Detailed discussion on decision trees can be found in [17, 18].

3. CONSTRAINED ASSOCIATION RULES

We introduce a transformation process of a data set with categorical and numerical attributes to transaction (sparse binary) format. We then discuss search constraints to get medically relevant association rules and accelerate search. Search constraints for association rules to analyze medical data are explained in more detail in [26, 25].

3.1 Transforming Medical Data Set

A medical data set with numeric and categorical attributes must be transformed to binary dimensions, in order to use association rules. Numeric attributes are binned into intervals and each interval is mapped to an item. Categorical attributes are transformed by mapping each categorical value to one item. Our first constraint is the negation of an attribute, which makes search more exhaustive. If an attribute has negation then additional items are created, corresponding to each negated categorical value or each negated interval. Missing values are assigned to additional items, but they are not used. In short, each transaction is a set of items and each item corresponds to the presence or absence of one categorical value or one numeric interval.

3.2 Search Constraints

Our discussion is based on the standard association rule search algorithm [2], which has two phases. Phase 1 finds all itemsets having minimum support, proceeding bottom-up, generating frequent 1-itemsets, 2-itemsets and so on, until there are no frequent itemsets. Phase 2 produces all rules whose support and confidence are above user-specified thresholds. Two of our constraints work on Phase 1 and the other one works on Phase 2.

The first constraint is κ, the user-specified maximum itemset size. This constraint prunes the search space for k-itemsets of size such that $k > \kappa$. This constraint reduces the combinatorial explosion of large itemsets and helps finding simple rules. Each predictive rule will have at most κ attributes (items).

Let $\mathcal{I} = \{i_1, i_2, \ldots i_m\}$ be the set of items to be mined, obtained by the transformation process from the attributes $\mathcal{A} = \{A_1, \ldots, A_p\}$. Constraints are specified on attributes and not on items. Let $attribute()$ be a function that returns the mapping between one attribute and one item.

Let $\mathcal{C} = \{c_1, c_2, \ldots c_p\}$ be a set antecedent and consequent constraints for each attribute A_j. Each c_j can take two values: 1 if attribute A_j can only appear in the antecedent of a rule and 2 if A_j can only appear in the consequent. We define the function antecedent/consequent $ac : \mathcal{A} \rightarrow \mathcal{C}$ as $ac(A_j) = c_j$ to make reference to one such constraint. Let X be a k-itemset; X is said to satisfy the antecedent constraint if for all $i_j \in X$ then $ac(attribute(i_j)) = 1$; X satisfies the consequent constraint if for all $i_j \in X$ then $ac(attribute(i_j)) = 2$. This constraint ensures we only find predictive rules with disease attributes in the consequent.

Let $\mathcal{G} = \{g_1, g_2, \ldots g_p\}$ be a set of p group constraints corresponding to 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 at all. 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. Note that 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 $g_j = 0$. The itemset X satisfies the group constraint if for each item pair $\{a, b\}$ s.t. $a, b \in \mathcal{I}$ it is true $group(attribute(a)) \neq group(attribute(b))$. The group constraint avoids finding trivial or redundant rules.

3.3 Constrained Association Rule Algorithm

We join the transformation algorithm and search constraints from into an algorithm that goes from transforming medical records into transaction to getting predictive rules. The transformation process using the given cutoffs for numeric attributes and desired negated attributes, produces the input data set for Phase 1. Each patient record becomes a transaction T_i (see Section 2). After the medical data set is transformed, items are further filtered out depending on the prediction goal: predicting absence or existence of heart disease. Items can only be filtered after attributes are transformed because they depend on the numeric cutoffs and negation. That is, it is not possible to filter items based on raw attributes. This is explained in more detail in Section 4. In Phase 1 we use the $group()$ constraint to avoid searching for trivial itemsets. Phase 1 finds all frequent itemsets from size 1 up to size κ. Phase 2 builds only predictive rules satisfying the $ac()$ constraint. The algorithm main input parameters are κ, minimum support and minimum confidence.

4. EXPERIMENTS

Our experiments focus on comparing the medical significance, accuracy and usefulness of predictive rules obtained by the constrained association rule algorithm and decision trees. Further experiments that measure the impact of constraints in the number of rules and reducing running time can be found in [25]. Our experiments were run on a computer running at 1.2 GHz with 256 MB of main memory and 100 GB of disk space. The association rule and the decision tree algorithms were implemented in the C++ language.

4.1 Medical Data Set Description

There are three basic elements for analysis: perfusion defect, risk factors and coronary stenosis. The medical data set contains the profiles of $n = 655$ patients and has $p = 25$ medical attributes corresponding to the numeric and categorical attributes listed in Table 1. The data set has personal information such as age, race, gender and smoking habits. There are medical measurements such as weight, heart rate, blood pressure and pre-existence of related diseases. Finally, the data set contains the degree of artery narrowing (stenosis) for the four heart arteries.

4.2 Default Parameter Settings

This section explains default settings for algorithm parameters, that were based on the domain expert opinion and previous research work [25]. Table 1 contains a summary of medical attributes and search constraints.

Attribute	Description	Constraints		
		neg	group H D	ac
AGE	Age of patient	N	0 0	1
LM	Left Main	Y	0 0	2
LAD	Left Anterior Desc.	Y	0 0	2
LCX	Left Circumflex	Y	0 0	2
RCA	Right Coronary	Y	0 0	2
AL	Antero-Lateral	N	1 1	1
AS	Antero-Septal	N	1 1	1
SA	Septo-Anterior	N	1 1	1
SI	Septo-Inferior	N	1 1	1
IS	Infero-Septal	N	1 1	1
IL	Infero-Lateral	N	1 1	1
LI	Latero-Inferior	N	1 1	1
LA	Latero-Anterior	N	1 1	1
AP	Apical	N	1 1	1
SEX	Gender	N	0 0	1
HTA	Hyper-tension Y/N	N	2 0	1
DIAB	Diabetes Y/N	N	2 0	1
HYPLD	Hyperloipidemia Y/N	N	2 0	1
FHCAD	Family hist. of disease	N	2 0	1
SMOKE	Patient smokes Y/N	N	0 0	1
CLAUDI	Claudication Y/N	N	2 0	1
PANGIO	Previous angina Y/N	N	3 0	1
PSTROKE	Prior stroke Y/N	N	3 0	1
PCARSUR	Prior carot surg Y/N	N	3 0	1
CHOL	Cholesterol level	N	0 0	1

Table 1: Attributes of medical data set.

Transformation parameters

To set the transformation parameters default values we must discuss attributes corresponding to heart vessels. The LAD, RCA, LCX and LM numbers represent the percentage of vessel narrowing (stenosis) compared to a healthy artery. Attributes LAD, LCX and RCA were binned at 50% and 70%. In cardiology a 70% value or higher indicates significant coronary disease and a 50% value indicates borderline disease. Stenosis below 50% indicates the patient is considered healthy. The LM artery has a different cutoff because it poses higher risk than the other three arteries. LAD and LCX arteries branch from LM. Therefore, a defect in LM is likely to trigger more severe disease. Attribute LM was binned at 30% and 50%. 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, meaning a perfusion measurement greater or equal than 0.2 indicated a severe defect. CHOL was binned at 200 (warning) and 250 (high). AGE was binned at 40 (adult) and 60 (old). Finally, only the four artery attributes (LAD, RCA, LCX, LM) had negation to find rules referring to healthy patients and sick patients. The other attributes did not have negation.

Search and filtering constraints

The maximum itemset size was set at $\kappa = 4$. Association rule mining had the following thresholds for metrics. The minimum support was fixed at $1\% \approx 7$. That is, rules referring to 6 or less patients were eliminated. Such threshold eliminated rules that were probably particular for our data set. From a medical point of view, rules with high confidence are desirable, but unfortunately, they are infrequent. Based on the domain expert opinion, the minimum confidence was set at 70%, which provides a balance between sensitivity (identifying sick patients) and specificity (identifying healthy patients) [26, 25]. Minimum lift was set slightly higher than 1 to filter out rules where X and Y are very likely to be independent. Finally, we use a high lift threshold (1.2) to get rules where there is a stronger implication dependence between X and Y.

The group constraint and the antecedent/consequent constraint had the following settings. Since we are trying to predict likelihood of heart disease, the 4 main coronary arteries LM, LAD, LCX and RCA are 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, risk factors (medical history and measurements) and perfusion measurements (9 heart regions) appear in the antecedent, whereas the four artery measurements appear in the consequent of a rule. From a medical perspective, determining the likelihood of presenting a risk factor based on artery disease is irrelevant. The 9 regions of the heart (AL, IS, SA, AP, AS, SI, LI, IL, LA) were constrained to be in the same group (group 1). The group settings for risk factors varied depending on the type of rules being mined (predicting existence or absence of disease). Combinations of items in the same group are not considered interesting and are eliminated from further analysis. The 9 heart regions were constrained to be on the same group because doctors are interested in finding their interaction with risk factors, but not among them. The default constraints are summarized in Table 1. Under column "group", the H subcolumn presents the group constraint to predict healthy arteries and the D subcolumn has the group constraint to predict diseased arteries.

4.3 Predictive Association Rules

The goal is to link perfusion measurements and risk factors to artery disease. Some rules were expected, confirming valid medical knowledge, and some rules were surprising, having the potential to enrich medical knowledge. We show some of the most important discovered rules. Predictive rules were grouped in two sets: (1) if there is a low perfusion measurement or no risk factor then the arteries are healthy; (2) if there exists a risk factor or a high perfusion measurement then the arteries are diseased. The maximum association size κ was 4.

Minimum support, confidence and lift were used as the main filtering parameters. Minimum lift in this case was 1.2. Support was used to discard low probability patterns. Confidence was used to look for reliable prediction rules. Lift was used to compare similar rules with the same consequent and to select rules with higher predictive power. Confidence, combined with lift, was used to evaluate the significance of each rule. Rules with confidence $\geq$ 90%, with lift $>= 2$, and with two or more items in the consequent were considered medically significant. Rules with high support, only risk factors, low lift or borderline confidence were considered interesting, but not significant. Rules with artery figures in wide intervals (more than 70% of the attribute range) were not considered interesting, such as rules having a measurement in the 30-100 range for the LM artery.

Rules predicting healthy arteries

The default program parameter settings are described in Section 4.2. Perfusion measurements for the 9 regions were in the same group (group 1). Rules relating no risk factors (equal to "n") with healthy arteries were considered medically important. Risk factors HTA, DIAB, HYPLD, FHCAD, CLAUDI were in the same group (group 2). Risk factors describing previous conditions for disease (PANGIO, PSTROKE, PCARSUR) were in the same group (group 3). The rest of the risk factor attributes did not have any group constraints. Since we were after rules relating negative risk factors and low perfusion measurements to healthy arteries, several items were filtered out to reduce the number of patterns. The discarded items involved arteries with values in the higher (not healthy) ranges (e.g. $[30, 100]$, $[50, 100]$, $[70, 100]$), perfusion measurements in $[0.2, 1]$ (no perfusion defect), and risk factors equal to "y" for the patient (person presenting risk factor). Minimum support was 1% and minimum confidence was 70%.

The program produced a total of 9,595 associations and 771 rules in about one minute. Although most of these rules provided valuable knowledge, we only describe some of the most surprising ones, according to medical opinion. Figure 1 shows rules predicting healthy arteries in groups. These rules have the potential to improve the expert system. The group with confidence=1 shows some of the few rules that had 100% confidence. It was surprising that some rules referred to young patients, but not older patients. The rules involving LAD had high lift with localized perfusion defects. The rules with LM had low lift confirming other risk factors may imply a healthy artery. The group with two items shows the only rules predicting absence of disease in two arteries. They include combinations of all the arteries and have high lift. These rules highlight low cholesterol level, female gender and young patients. It turned out all of them refer to the same patients. The 90% confidence group shows fairly reliable rules. Unfortunately, their lift is not high. The group with only risk factors shows rules that do not involve any perfusion measurements. These rules highlight the importance of smoking habits, diabetes, low cholesterol, gender and age in having no heart disease. The last group describes rules with high support. Most of them involve the LCX artery, the IL region and some risk factors. These rules had low lift stressing the importance of many other factors to have healthy arteries. Summarizing, these experiments show LCX is more likely to be healthy given absence of risk factors and low perfusion measurements. Lower perfusion measurements appeared in heart regions IL and LI. Some risk factors have less importance because they appear less frequently in the rules. But age, sex, diabetes and cholesterol level appear frequently stressing their importance.

```
Con↓dence = 1:
 IF   0 <= AGE < 40.0   - 1.0 <= AL < 0.2  PCARSUR = n
  THEN   0 <= LAD < 50, s=0.01 c=1.00 l=2.1
 IF   0 <= AGE < 40.0   - 1.0 <= AS < 0.2  PCARSUR = n
  THEN   0 <= LAD < 50, s=0.01 c=1.00 l=2.1
 IF   40.0 <= AGE < 60.0  SEX = F  0 <= CHOL < 200
  THEN   0 <= LCX < 50, s=0.02 c=1.00 l=1.6
 IF   SEX = F  HTA = n  0 <= CHOL < 200
  THEN   0 <= RCA < 50, s=0.02 c=1.00 l=1.8
Two items in the consequent:
 IF   0 <= AGE < 40.0   - 1.0 <= AL < 0.2
  THEN   0 <= LM < 30  0 <= LAD < 50, s=0.02 c=0.89 l=1.9
 IF   SEX = F  0 <= CHOL < 200
  THEN   0 <= LAD < 50  0 <= RCA < 50, s=0.02 c=0.73 l=2.1
 IF   SEX = F  0 <= CHOL < 200
  THEN   0 <= LCX < 50  0 <= RCA < 50, s=0.02 c=0.73 l=1.8
Con↓dence >= 0.9:
 IF   40.0 <= AGE < 60.0   - 1.0 <= LI < 0.2  0 <= CHOL < 200
  THEN   0 <= LCX < 50, s=0.03 c=0.90 l=1.5
 IF   40.0 <= AGE < 60.0   - 1.0 <= IL < 0.2  0 <= CHOL < 200
  THEN   0 <= LCX < 50, s=0.03 c=0.92 l=1.5
 IF   40.0 <= AGE < 60.0   - 1.0 <= IL < 0.2  SMOKE = n
  THEN   0 <= LCX < 50, s=0.01 c=0.90 l=1.5
 IF   40.0 <= AGE < 60.0  SEX = F  DIAB = n
  THEN   0 <= LCX < 50]), s=0.08 c=0.92 l=1.5
 IF   HTA = n  SMOKE = n  0 <= CHOL < 200
  THEN   0 <= LCX < 50, s=0.02 c=0.92 l=1.5
Only risk factors:
 IF   0 <= AGE < 40.0
  THEN   0 <= LAD < 50, s=0.03 c=0.82 l=1.7
 IF   0 <= AGE < 40.0  DIAB = n
  THEN   0 <= LAD < 50, s=0.03 c=0.82 l=1.7
 IF   40.0 <= AGE < 60.0  SEX = F  DIAB = n
  THEN   0 <= LAD < 50, s=0.07 c=0.72 l=1.5
 IF   40.0 <= AGE < 60.0  SMOKE = n
  THEN   0 <= LCX < 50, s=0.11 c=0.75 l=1.2
 IF   40.0 <= AGE < 60.0  SMOKE = n
  THEN   0 <= RCA < 50, s=0.11 c=0.76 l=1.3
Support >= 0.2:
 IF   - 1.0 <= IL < 0.2  DIAB = n
  THEN   0 <= LCX < 50, s=0.41 c=0.72 l=1.2
 IF   - 1.0 <= LA < 0.2
  THEN   0 <= LCX < 50, s=0.39 c=0.72 l=1.2
 IF   SEX = F
  THEN   0 <= LCX < 50, s=0.23 c=0.73 l=1.2
 IF   40.0 <= AGE < 60.0   - 1.0 <= IL < 0.2
  THEN   0 <= RCA < 50, s=0.21 c=0.73 l=1.3
```

Figure 1: Association rules for healthy arteries.

Rules predicting diseased arteries

The default program parameter settings are described in Section 4.2. Refer to Table 1 to understand the meaning of abbreviations for attribute names. The four arteries (LAD, LCX, RCA, LM) had negation. Rules relating presence of risk factors (equal to "y") with diseased arteries were considered interesting. There were no group constraints for any of the attributes, except for the 9 regions of the heart (group 1). This allowed finding rules combining any risk factors with any perfusion defects. Since we were after rules relating risk factors and high perfusion measurements indicating heart defect to diseased arteries, several unneeded items were filtered out to reduce the number of patterns. Filtered items involved arteries with values in the lower (healthy) ranges (e.g. $[0, 30)$, $[0, 50)$, $[0, 70)$), perfusion measurements in $[-1, 0.2)$ (no perfusion defect), and risk factors having "n" for the patient (person not presenting risk factor). Minimum support was 1% and minimum confidence was 70%.

The program produced a total of 10,218 associations and 552 rules in less than one minute. Most of these rules were considered important and about one third were medically significant. Most rules refer to patients with localized perfusion defects in specific heart regions and particular risk factors with the LAD and RCA arteries. It was surprising there were no rules involving LM and only 9 with LCX. Tomography or coronary catheterization are the most common ways to detect heart disease. 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.

Figure 2 shows groups of rules predicting diseased arteries. Hypertension, diabetes, previous cardiac surgery and male sex constitute high risk factors. The 100% confidence group shows some of the only 22 rules with 100% confidence. They show a clear relationship of perfusion defects in the IS, SA regions, certain risk factors and both the RCA and LAD arteries. The rules with RCA have very high lift pointing to specific relationships between this artery and cholesterol level and the IS region. It was interesting the rule with LAD>= 70 also had high lift, but referred to different risk factors and region SA. The group of rules with two items in the consequent shows the only rules involving two arteries. They show a clear link between LAD and RCA. It is interesting these rules only involve a previous surgery as a risk factor. These four rules are surprising and extremely valuable. This is confirmed by the fact that two of these rules had the highest lift among all discovered rules (above 4). The 90% confidence group shows some outstanding rules out of the 35 rules that had confidence 90-99%. All of these rules have very high lift with a narrow range for LAD and RCA. These rules show that older patients of male gender, high cholesterol levels and localized perfusion measurements, are likely to have disease on the LAD and RCA arteries. The group involving only risk factors in the antecedent shows several risk factors and disease on three arteries. Unfortunately their support is relatively low, but they are valuable as they confirm medical knowledge. The rule with lift=2.2 confirms that gender and high cholesterol levels may lead to disease in the LCX artery. The group with support above 0.15 shows the rules with highest support. All of them involved LAD and combinations of risk factors. Their lift was low-medium, confirming more risk factors are needed to get a more accurate prediction. There were no high-support rules involving LCX, RCA or LM arteries, confirming they have a lower probability of being diseased.

```
con↓dence = 1:
 IF   0.2 <= SA < 1.0  HYPLPD = y  PANGIO = y
  THEN   70 <= LAD < 100, s=0.01 c=1.00 l=3.2
 IF   60 <= AGE < 100  0.2 <= SA < 1.0  FHCAD = y
  THEN not(0 <= LAD < 50, s=0.02 c=1.00 l=1.9
 IF   0.2 <= IS < 1.0  CLAUDI = y  PSTROKE = y
  THEN   not(0 <= RCA < 50), s=0.02 c=1.00 l=2.3
 IF   60 <= AGE < 100.0  0.2 <= IS < 1.0  250 <= CHOL < 500
  THEN   70 <= RCA < 100, s=0.02 c=1.00 l=3.2
 IF   0.2 <= IS < 1.0  SEX = F  250 <= CHOL < 500
  THEN   70 <= RCA < 100, s=0.01 c=1.00 l=3.2
 IF   0.2 <= IS < 1.0  HTA = y  250 <= CHOL < 500])
  THEN   70 <= RCA < 100, s=0.011 c=1.00 l= 3.2
Two items in the consequent:
 IF   0.2 <= AL < 1.1  PCARSUR = y
  THEN   70 <= LAD < 100  not(0 <= RCA < 50), s=0.01 c=0.70 l=3.9
 IF   0.2 <= AS < 1.1  PCARSUR = y
  THEN   70 <= LAD < 100  not(0 <= RCA < 50), s=0.01 c=0.78 l=4.4
 IF   0.2 <= AP < 1.1  PCARSUR = y
  THEN   70 <= LAD < 100  not(0 <= RCA < 50), s=0.01 c=0.80 l=4.5
 IF   0.2 <= AP < 1.1  PCARSUR = y
  THEN   not(0 <= LAD < 50)  not(0 <= RCA < 50), s=0.01 c=0.80 l=2.8
con↓dence >= 0.9:
 IF   0.2 <= SA < 1.1  PANGIO = y])
  THEN   70 <= LAD < 100, s=0.023 c=0.938 l= 3.0
 IF   0.2 <= SA < 1.0  SEX = M  PANGIO = y
  THEN   70 <= LAD < 100, s=0.02 c=0.92 l=2.9
 IF   60 <= AGE < 100.0  0.2 <= IL < 1.1  250 <= CHOL < 500
  THEN   70 <= RCA < 100, s=0.02 c=0.92 l=2.9
 IF   0.2 <= IS < 1.0  SMOKE = y  250 <= CHOL < 500
  THEN   70 <= RCA < 100, s=0.02 c=0.91 l=2.9
Only risk factors:
 IF   SEX = M  PSTROKE = y  250 <= CHOL < 500
  THEN   not(0 <= LAD < 50), s=0.01 c=0.73 l=1.4
 IF   40.0 <= AGE < 60.0  SEX = M  250 <= CHOL < 500
  THEN   not(0 <= LCX < 50), s=0.02 c=0.83 l=2.2
 IF   SMOKE = y  PANGIO = y  250 <= CHOL < 500
  THEN   not(0 <= RCA < 50), s=0.01 c=0.80 l=1.9
Support >= 0.15:
 IF   0.2 <= IL < 1.1
  THEN   not(0 <= LAD < 50), s=0.25 c=0.71 l=1.4
 IF   0.2 <= AP < 1.1
  THEN   not(0 <= LAD < 50), s=0.24 c=0.78 l=1.5
 IF   0.2 <= IL < 1.1  SEX = M
  THEN   not(0 <= LAD < 50), s=0.19 c=0.72 l=1.4
 IF   0.2 <= AP < 1.1  SEX = M
  THEN   not(0 <= LAD < 50), s=0.18 c=0.75 l=1.5
 IF   60 <= AGE < 100.0  0.2 <= AP < 1.1
  THEN   not(0 <= LAD < 50), s=0.18 c=0.87 l=1.7
```

Figure 2: Association rules for diseased arteries.

4.4 Predictive Rules from Decision Trees

In this section we explain experiments using decision trees. We used the CN4.5 decision tree [14] algorithm using gain ratio for splitting and pruning nodes. Due to lack of space we do not discuss experiments with CART decision trees [18], but results are similar. In some experiments the height of trees had a threshold to produce simpler rules. We show some classification rules with the percentage of patients (ls) they involve and their confidence factor (cf). The confidence factor has a similar interpretation to association rule confidence, but the percentage refers to the fraction of patients where the antecedent appears (i.e. support of antecedent itemset). For instance, if cf is less than 100% and $ls = 10\%$ then the actual support of the rule is less than 10%. These experiments focused on predicting LAD disease using its binary version LAD$\geq$ 50 as the target class. This artery was recommended for analysis by the domain expert because in general it is the most common to be diseased. Then it should be easier to find rules involving it. Due to lack of space we do not show experiments using RCA, LCX or LM as the dependent variable, but results are similar to the ones described below.

The first set of experiments used all risk factors and perfusion measurements without binning as independent variables. That is, the decision tree automatically splits numerical variables and chooses subsets of categorical values to perform binary splits. The first experiment did not have a threshold for the tree height. This produced a large tree with 181 nodes and 90% accuracy. The tree had height 14 with most classification rules involving more than 5 attributes (plus one for the predicted LAD disease). With the exception of five rules all rules involved less than 2% of the patients. More than 80% of rules referred to less than 1% of patients. Many rules involved attributes with missing information. Many rules had the same variable being split several times. A positive point was a few rules had $cf = 1.0$, but with splits for perfusion measurements and artery disease including borderline cases and involving a few patients. Therefore, even though this decision tree had all our variables and was 90% accurate it was not medically useful. In the second experiment we decided to set a threshold for height of the tree equal to 10. The resulting tree had 83 nodes out of which 43 were terminal nodes and accuracy went down to 77%. Most decision rules predicting diseased arteries had repeated attributes (splits on same variable twice), more than 5 attributes, perfusion cutoffs higher than 0.50, low cf and involved less than 1% of the population. Therefore, this tree was not useful either. This motivated getting smaller trees with simple rules involving larger sets of patients at the risk of getting lower confidence factors. This affects accuracy, of course, but it provides more control on the type of rules we want.

We constrained the decision tree to have maximum height equal to 3 to obtain simpler classification rules comparable to association rules. The resulting tree had low accuracy (65% accuracy) and only 6 terminal nodes. Figure 3 shows the classification rules letting the decision tree split variables automatically. Fortunately these rules are simpler than the previous ones. We discuss rules predicting healthy vessels. Rule 1 covers a wide group of patients, but it is too imprecise about patient's age since the range for AGE is too wide. Also, the split for AP leaves a big gap between it and 0.2 leaving potentially many patients with defects in AP incorrectly included. Then rule 1 cannot be medically used to predict no heart disease. Rule 2 goes against medical knowledge since it implies that two perfusion defects on young patients imply no disease. It is no coincidence this rule has such low support. We now explain rules predicting diseased LAD. Rule 1 is interesting since it involves 10% of patients and has decent confidence, but it combines almost absence of perfusion defect with existence of perfusion defect giving a "mixed" profile of such patients. Rule 2 is of little value since it includes absence of perfusion defects (range [-1,0.2]). We are rather interested in knowing the fraction of patients between the given splits for perfusion figures and 0.2. The only interesting aspect is that it refers to very old patients. Rule 3 combines absence and borderline perfusion defects with low support and then it is not medically useful. Rule 4 is the best rule found by the decision tree since it involves a perfusion defect on adult patients and has remarkable high confidence. As a note, a very similar rule was found by association rules. In short, discovered classification rules were very few, had split points that affected medical interpretation and did not include most risk factors.

In the last set of experiments we used items (binary variables) as independent variables like association rules to obtain similar rules with a tree height limited to 3. That is, we used the variable LAD>= 50 as the dependent variable and binned numerical variables (perfusion measurements, AGE and CHOL) and categorical variables as independent variables. Most of the rules were much closer to the prediction requirements. The tree had 10 nodes out of which 3 involved rules predicting diseased arteries and 3 involved rules predicting no disease. Figure 4 shows the discovered rules classified in two groups. We discuss rules predicting healthy arteries. Rule 1 has low confidence factor, relates absence of two perfusion defects (something not interesting in this case) and has low confidence. Therefore, it is not useful. Rule 2 and 3 might be useful because they involve a risk factor combined with perfusion defects, but they have low confidence and combine a perfusion defect with an absence of perfusion defect (something not medically meaningful). We now discuss rules predicting diseased arteries. Rule 1 is not useful because it involves a perfusion with no defect and its confidence is low. Rule 2 might be useful and was not found with constrained association rules. However, we stress this rule was not found because AGE did not have negation. Rule 3 is the only rule found by the decision tree that is one of the many rules found with constrained association rules with LAD>= 50.

```
Predicting healthy arteries:
  IF (  SA <= 0.37  AP <= 0.66  Age <= 78)
   THEN not(LAD >= 50) ls=76% cf=0.58
  IF (  SA > 0.37  Age <= 53  AS > 0.67)
   THEN not(LAD >= 50) ls=0.3% cf=1.00
Predicting diseased arteries:
  IF (  SA <= 0.37  AP > 0.66)
   THEN   LAD >= 50 ls=10% cf=0.80
  IF (  SA <= 0.37  AP <= 0.66  Age > 78)
   THEN   LAD >= 50 ls=4% cf=0.74
  IF (  SA > 0.37  Age <= 53  AS <= 0.67)
   THEN   LAD >= 50 ls=1% cf=0.85
  IF (  SA > 0.37  Age > 53)
   THEN   LAD >= 50 ls=8% cf=0.98
```

Figure 3: Decision tree rules with numeric dimensions and automatic splits.

```
Predicting healthy arteries:
  IF (not([0.2 <= AP < 1.1])not([0.2 <= IL < 1.1)
   THEN not([LAD >= 50]) ls=54% cf=0.63
  IF (not([0.2 <= AP < 1.1])[0.2 <= IL < 1.1  HYPLPD = n])
   THEN not([LAD >= 50]) ls=5.5% cf=0.64
  IF (  0.2 <= AP < 1.1]not([60 <= Age < 100])not([0.2 <= IL < 1.1]))
   THEN not([LAD >= 50]) ls=3.8% cf=0.64
Predicting diseased arteries:
  IF (not([0.2 <= AP < 1.1])[0.2 <= IL < 1.1  HYPLPD = y])
   THEN   LAD >= 50 ls=7.6% cf=0.60
  IF (  0.2 <= AP < 1.1]not([60 <= Age < 100])[0.2 <= IL < 1.1)
   THEN   LAD >= 50 ls=7% cf=0.73
  IF (([0.2 <= AP < 1.1])[60 <= Age < 100)
   THEN   LAD >= 50 ls=20% cf=0.86
```

Figure 4: Decision tree rules with manually binned variables.

4.5 Discussion

Our experiments provide some evidence that decision trees are not as powerful as association rules to exploit a set of numeric attributes manually binned and categorical attributes and several related target attributes. Decision trees do not work well with combinations of several target variables (arteries), which requires defining one class attribute for each values combination. Decision trees fail to identify many medically relevant combinations of independent numeric variable ranges and categorical values (i.e. perfusion measurements and risk factors). When given the ability to build height-unrestricted trees decision trees tend to find complex and long rules, making rule applicability and interpretation difficult. Also, in such case decision trees find few predictive rules with reasonably sized (> 1%) sets of patients; this is a well-known drawback known as data set fragmentation [18]. To complicate matters, rules sometimes repeat the same attribute several times creating a long sequence of splits that needs to be simplified. However, it could be argued that we could build many decision trees with different independent attributes containing all different combinations of risk factors and perfusion variables for each target artery, following a similar approach to the constraints we introduced, but that would be error-prone, difficult to interpret and slow given the high number of attribute combinations. Another alternative is to create a family of small trees, where each tree has a weight, but each small tree becomes similar to a small set of association rules. We believe, for the purpose of predicting disease with several related target attributes, association rules are more effective. However, our constraints for association rules may be adapted to decision trees, but that is subject of future work. Decision trees do have advantages over association rules. A decision tree partitions the data set, whereas association rules on the same target attribute may refer to overlapping subsets; sometimes this makes result interpretation difficult. A decision tree represents a predictive model of the data set, whereas association rules are disconnected among themselves. In fact, the large number of discovered association rules may require rule summarization. A decision tree is guaranteed to have at least 50% prediction accuracy and generally above 80% accuracy for binary target variables, whereas association rules specifically require trial and error runs to find a good or acceptable threshold.

5. RELATED WORK

Important related work on using data mining and machine learning techniques in medical data includes the following. Some particular issues in medical data [29] include distributed and uncoordinated data collection, strong privacy concerns, diverse data types (image, numeric, categorical, missing information), complex hierarchies behind attributes and a comprehensive knowledge base. A well-known program to help heart disease diagnosis based on Bayesian networks is described in [15, 23, 22]. Association rules have been used to help infection detection and monitoring [7, 8], to understand what drugs are co-prescribed with antacids [10], to discover frequent patterns in gene data [5, 11], to understand interaction between proteins [27] and to detect common risk factors in pediatric diseases [13]. Fuzzy sets have been used to extend association rules [12]. In [26] we explore the idea of constraining association rules in binary data for the first time and report preliminary findings from a data mining perspective. Finally, [25] studies the impact of each constraint on the number of discovered rules and algorithm running time and also proposes a summarization of a large number of rules having the same consequent.

Association rules were proposed in the seminal paper [1]. Quantitative association rules are proposed in [31]; such technique automatically bins attributes, but such rules have not been shown to be more accurate than decision trees. Both [31] and [21] use different approaches to automatically bin numeric attributes. Instead, in our approach it was preferred to use well-known medical cutoffs for binning numeric attributes, to improve result interpretation and validation. Our search constraints share some similarities with [4, 24, 32]. In [32] the authors propose 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 [30] 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 relevant rules. A work which studies constraining association rules in more depth is [24], where constraints are item boolean expressions involving two variables. It is well-known that simple constraints on support can be used for pruning the search space in Phase 1 [34]. Association rules and prediction rules from decision trees are contrasted in [16]. The lift measure for association rules was introduced in [6]. Rule covers [19, 20] and basis [33, 3, 28, 9] are alternatives to get condensed representations of association rules.

6. CONCLUSIONS

In this work constrained association rules were used to predict multiple related target attributes, for heart disease diagnosis. The goal was to find association rules predicting healthy arteries or diseased arteries, given patient risk factors and medical measurements. This work presented three search constraints that had the following objectives: producing only medically useful rules, reducing the number of discovered rules and improving running time. First, data set attributes are constrained to belong to user-specified groups to eliminate uninteresting value combinations and to reduce the combinatorial explosion of rules. Second, attributes are constrained to appear either in the antecedent or in the consequent to discover only predictive rules. Third, rules are constrained to have a threshold on the number of attributes to produce fewer and simpler rules. Experiments with a medical data set compare predictive constrained association rules with rules induced by decision trees, using one of the best currently available decision tree algorithms. Rules are analyzed in two groups: those that predict healthy arteries and those that predict diseased arteries. Decision trees are built both on raw numeric and categorical attributes (original medical dataset) as well as using transformed attributes (binned numeric features and binary coded categorical features). Experimental results provide evidence that decision trees are less effective than constrained association rules to predict disease with several related target attributes, due to low confidence factors (i.e. low reliability), slight overfitting, rule complexity for unrestricted trees (i.e. long rules) and data set fragmentation (i.e. small data subsets). Therefore, constrained association rules can be an alternative to other statistical and machine learning techniques applied in medical problems where there is a requirement to predict several target attributes based on subsets of independent numeric and categorical attributes.

Our work suggests several directions to improve decision trees and association rules. We want to adapt search constraints to decision trees to predict several related target attributes. A hybrid set of attributes may be better, where some attributes may be automatically binned by the decision tree, while other attributes may be manually binned by the user. A family of small decision trees may be an alternative to using a large number of association rules. Decision trees may be used to pre-process a data set to partition it into focused subsets, where association rules may be applied in a second phase.

Acknowledgments

The author thanks Dr. Cesar Santana from the Emory University Hospital and Dr. Hiroshi Oyama from the University of Tokyo School of Medicine for many helpful discussions.

7. 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] Y. Bastide, N. Pasquier, R. Taouil, and G. L. Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In *Computational Logic*, pages 972–986, 2000.

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

[5] 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.*, 3(12), 2002.

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

[7] 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 Am Med Inform Assoc. (JAMIA)*, 5(4):373–381, 1998.

[8] S.E. Brossette, A.P. Sprague, W.T. Jones, and S.A. Moser. A data mining system for infection control surveillance. *Methods Inf Med.*, 39(4):303–310, 2000.

[9] A. Bykowski and C. Rigotti. DBC: a condensed representation of frequent patterns for efficient mining. *Information Systems*, 28(8):949–977, 2003.

[10] 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*, 25(9):2453–2463, 2003.

[11] C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79–86, 2003.

[12] M. Delgado, D. Sanchez, M.J. Martin-Bautista, and M.A. Vila. Mining association rules with improved semantics in medical databases. *Artificial Intelligence in Medicine*, 21(1-3):241–5, 2001.

[13] S.M. Down and M.Y. Wallace. Mining association rules from a pediatric primary care decision support system. In *Proc of AMIA Symp.*, pages 200–204, 2000.

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

[15] H.S. Fraser, W.J. Long, and S. Naimi. Evaluation of a cardiac diagnostic program in a typical clinical setting. *J Am Med Inform Assoc. (JAMIA)*, 10(4):373–381, 2003.

[16] A. Freitas. Understanding the crucial differences between classification and association rules - a position paper. *SIGKDD Explorations*, 2(1):65–69, 2000.

[17] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 1st edition, 2001.

[18] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer, New York, 1st edition, 2001.

[19] M. Kryszkiewicz. Concise representation of frequent patterns based on disjunction-free generators. In *Proc. IEEE ICDM Conference*, pages 305–312, 2001.

[20] M. Kryszkiewicz. Reducing borders of k-disjunction free representations of frequent patterns. In *ACM SAC Conference*, pages 559–563, 2004.

[21] B. Lent, A. Swami, and J. Widom. Clustering association rules. In *Proc. IEEE ICDE Conference*, pages 220–231, 1997.

[22] W.J. Long. Medical reasoning using a probabilistic network. *Applied Artifical Intelligence*, 3:367–383, 1989.

[23] W.J. Long, H.S. Fraser, and S. Naimi. Reasoning requirements for diagnosis of heart disease. *Artificial Intelligence in Medicine*, 10(1):5–24, 1997.

[24] 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.

[25] C. Ordonez, N. Ezquerra, and C.A. Santana. Constraining and summarizing association rules in medical data. *Knowledge and Information Systems (KAIS)*, 9(3):259–283, 2006.

[26] C. Ordonez, E. Omiecinski, Levien de Braal, Cesar Santana, and N. Ezquerra. Mining constrained association rules to predict heart disease. In *IEEE ICDM Conference*, pages 433–440, 2001.

[27] T. Oyama, K. Kitano, T. Satou, and T. Ito. Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics*, 18(5):705–714, 2002.

[28] V. Phan-Luong. The representative basis for association rules. In *IEEE ICDM Conference*, pages 639–640, 2001.

[29] J.F. Roddick, P. Fule, and W.J. Graco. Exploratory medical knowledge discovery: Experiences and issues. *SIGKDD Explorations*, 5(1):94–99, 2003.

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

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

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

[33] R. Taouil, N. Pasquier, Y. Bastide, and L. Lakhal. Mining bases for association rules using closed sets. In *IEEE ICDE Conference*, page 307, 2000.

[34] K. Wang, Y. He, and J. Han. Pushing support constraints into association rules mining. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 15(3):642–658, 2003.