Redundancy Reduction: Does It Help Associative Classifiers?

Luiza Antonie
University of Guelph
Guelph, Canada
lantonie@uoguelph.ca

Osmar R. Zaïane
University of Alberta
Edmonton, Canada
zaiane@cs.ualberta.ca

Robert C. Holte
University of Alberta
Edmonton, Canada
holte@cs.ualberta.ca

ABSTRACT

The number of classification rules discovered in associative classification is typically quite large. In addition, these rules contain redundant information since classification rules are obtained from mined frequent itemsets and the latter are known to be repetitive. In this paper we investigate through an empirical study the performance of associative classifiers when the classification rules are generated from frequent, closed and maximal itemsets. We show that maximal itemsets substantially reduce the number of classification rules without jeopardizing the accuracy of the classifier. Our extensive analysis demonstrates that the performance remains stable and even improves in some cases. Our analysis using cost curves also provides recommendations on when it is appropriate to remove redundancy in frequent itemsets.

1. INTRODUCTION

Classification is an important task in many applications. Association rule-based classifiers [1, 2] are classification systems that consist of a set of rules, with each rule predicting that an object belongs to a specific class if it has certain properties. The rules are discovered using an association rule mining algorithm [3]. The association rule mining problem has been thoroughly studied in the data mining community [4, 5, 6, 7], thus there are several fast algorithms for discovering these types of rules. An attractive characteristic that associative classifiers possess is their readability. However, the number of classification rules they generate is quite large.

Typically, associative classifiers generate classification rules from frequent patterns (i.e., all patterns that are seen frequently in the training data). Closed [8] and maximal [9] patterns are compressed representations of all the frequent patterns. They have been proposed to substantially reduce the number of frequent patterns. This reduction is achieved by eliminating redundancy present in the frequent patterns set. Closed patterns are a lossless form of compression, as the frequent patterns and their respective supports can be reproduced from this representation. On the other hand, maximal patterns represent a lossy compression since the support measures of the frequent patterns have to be recomputed.

Several research studies demonstrate the usefulness of closed and maximal itemsets in different applications. In the case of classification the use of these patterns has not been thoroughly explored. Closed and maximal frequent patterns reduce the number of association rules, but it is not clear that they reduce the number of classification rules as well. We hypothesize that they do and probably even improve the performance of the classification.

In this paper we investigate the performance of associative classifiers when the classification rules are generated from closed and maximal itemsets. Through our analysis we answer the following critical questions:

- How substantial is the reduction in the number of classification rules? It is known that closed and maximal representations reduce the number of patterns, but how does this translate to classification rules?
- What is the effect of rules extracted from closed and maximal itemsets on the classification performance? Are closed and maximal itemsets a good substitute for frequent itemsets in associative classifiers?

In the framework proposed in this paper, we integrate several associative classifiers with classification rules generated from frequent, closed and maximal patterns. Our hypothesis is that the use of closed and maximal is advantageous to associative classifiers. The benefit is twofold: first, the set of classification rules generated from closed and maximal itemsets is smaller; second, the performance level of the classifier stays the same or it improves. We test our hypothesis with an extensive experimental study and we show that this hypothesis holds over a large range of applications.

The contributions of this paper are as follows:

- We integrate associative classifiers with classification rules generated from closed and maximal itemsets;
- The classification stage is what distinguishes the existing associative classifiers. We study the potential of classification rules generated from closed and maximal patterns with several scoring schemes used in the classification stage;
- We carry out an extensive experimental study on well-known UCI datasets and on microarray data

SAC'16, April 4-8, 2016, Pisa, Italy
Copyright 2016 ACM 978-1-4503-3739-7/16/04…$15.00
http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

2. PREREQUISITES

Association rule mining is a data mining task that discovers relationships between items in a transactional database. The efficient discovery of such rules has been a major focus in the data mining research community, given their popularity in market basket analysis and other applications. From the original *apriori* algorithm [3] there have been a remarkable number of variants and improvements [4, 5, 6, 7].

Formally, frequent pattern mining is defined as follows. Let $\mathscr{I} = \{i_1, i_2, ...i_m\}$ be a set of items. Let $\mathscr{D}$ be a set of transactions, where each transaction T is a set of items such that $T \subseteq \mathscr{I}$. A transaction T is said to contain X, a set of items in $\mathscr{I}$, if $X \subseteq T$. We define in the following the types of patterns that we study in this paper.

Definition 1. Frequent itemset: An itemset $f \subseteq I$ is said to be frequent if its support s (i.e., the percentage of transaction in $\mathscr{D}$ that contain f) is greater than or equal to a given minimum support threshold.

Definition 2. Frequent closed itemset: A frequent itemset $c \subseteq I$ is said to be frequent closed if and only if there is no frequent itemset c' such that $c \subseteq c'$ and the support of c equals the support of c'.

Definition 3. Maximal frequent itemset: A frequent itemset $m \subseteq I$ is said to be maximal frequent if there is no other frequent itemset that is a superset of m.

An **association rule** is an implication of the form "$X \Rightarrow Y$", where $X \subseteq \mathscr{I}, Y \subseteq \mathscr{I}$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ has a *support* s in the transaction set $\mathscr{D}$ if $s\%$ of the transactions in $\mathscr{D}$ contain $X \cup Y$. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases. It is said that the rule $X \Rightarrow Y$ holds in the transaction set $\mathscr{D}$ with *confidence c* if $c\%$ of transactions in $\mathscr{D}$ that contain X also contain Y. In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X. The problem of discovering all association rules from a set of transactions $\mathscr{D}$ consists of generating the rules that have a *support* and *confidence* greater than given thresholds. These rules are called *strong rules* and represent interesting patterns in data. In this context, a **classification rule** is a particular association rule that associates frequent features with a *class label*.

3. ASSOCIATIVE CLASSIFIERS

The use of association rule mining for building classification models is relatively new. These classification systems discover the strongest classification rules in the dataset and use them to classify new objects.

The first reference to using association rules as classification rules is credited to [10], while the first classifier using these association rules was CBA introduced by [2] and later improved in CMAR [1] and 2SARC [11]. The idea is relatively simple. Given a training set modelled with transactions where each transaction contains all features of an object in addition to the class label of the object, we can constrain the mining process to generate association rules that always have as consequent a class label. In other words, the problem consists of finding the subset of strong association rules of the form $X \Rightarrow C$ where C is a class label and X is a conjunction of features.

Associative classifiers have the advantage over other rule-based classification systems that they guarantee to find all interesting rules (in the support-confidence framework). However, this property also guarantees that the number of classification rules will be quite large. In this paper, we focus our efforts on lowering the number of rules by mining closed and maximal itemsets. In addition to reducing the size of the system (i.e., the number of classification rules), redundant information is also eliminated. The next section gives a detailed description of our framework.

4. INTEGRATING ASSOCIATIVE CLASSIFIERS WITH CLOSED AND MAXIMAL PATTERNS

Our hypothesis is that the use of closed and maximal patterns is beneficial to associative classifier. The benefit is twofold: first, the set of classification rules generated from closed and maximal itemsets is smaller; second, the performance level of the classifier stays the same or it improves. In our framework we integrate several associative classifiers with classification rules generated from frequent, closed and maximal patterns to investigate our hypothesis.

The classification stage plays an important role in building any associative classifier. That is why we integrate in our framework several classification schemes. We investigate these systems when classification rules are generated from frequent, closed and maximal itemsets. The goal of our study is to find out the effect that closed and maximal patterns produce when integrated with associative classifiers.

The modules of our framework are as follows:

- *Discover frequent, closed and maximal itemsets.* Let F be the set of all frequent itemsets, F=$\{\bigcup$ f such that f $\subseteq$ I, and support(f) $\geq$ *minsupp*$\}$. Let C be the set of all closed itemsets, C=$\{\bigcup$ c such that c $\subseteq$ I, it is closed and support(c) $\geq$ *minsupp*$\}$. Let M be the set of all maximal itemsets, M=$\{\bigcup$ m such that m $\subseteq$ I, it is maximal and support(c) $\geq$ *minsupp*$\}$. Since we are interested in building a classifier, only itemsets that have a class label are considered.

- *Generate classification rules from mined itemsets.* From one set of patterns (F, C or M) find all rules such that the consequence of the rule is a class label. Let R be the set of all classification rules, R=$\{\bigcup$ r such that r is of the form $X \rightarrow Y$, where Y is a class label and confidence(r) $\geq$ *minconf*$\}$. Let us consider that R_f, R_c and R_m are the rule sets generated from frequent, closed and maximal itemsets. Order classification rules in R_f, R_c and R_m by confidence and break ties by support.

- *Classify a new object using the set of classification rules (R_f, R_c or R_m).* The classification decision is made according to a scoring scheme. The scoring schemes that we investigate are as follows:

  1. Classify a new object according to the highest ranked rule that applies. This scheme is used in CBA [2]. We denote this method as ***FR*** (first rule)

  2. Classify a new object based on average of the confidences. Let us assume that k rules apply to a new object. Let S be the set of k rules. Divide S in subsets by class label: $S_1, S_2...S_n$. For each subset S_i sum the confidences of rules and divide by the number of rules in

S_i, this is the score that is associated to class i. The object is classified in the class with the highest score [12]. We denote this method as ***AvR*** (average the confidences of the rules that apply).

3. Classify a new object using a two stage classification approach [11].
   - for each instance in the training set, use R_f, R_c or R_m to collect a set of features (class-features or rule-features as defined in [11]); Rule features are generated as follows. Given a rule set (R_f, R_c or R_m) and a training set, the characteristics of the rule R_i in R with respect to each instance I in the training set become features for the second learning method. For each instance I, a rule either applies or not. This information along with the rule's confidence is used into the model for the second stage. Similarly class features are generated, but this time class aggregates are considered as features.
   - apply a learning method in this new feature space to learn how to use the rules in the prediction process; this scheme is used in *2SARC*;
   - classify the objects in the testing set using R_f, R_c or R_m and *2SARC* combined;

   2SARC-CF is the system that uses class features in the two stage classification approach; ***2SARC-RF*** is the system that uses rule features in the two stage classification approach.

5. EXPERIMENTAL STUDY

To test our hypothesis and to study our framework we performed an extensive experimental study in which we evaluated all four classification systems in our framework on several datasets. The details of the datasets and the evaluation techniques used are described in the following sections.

5.1 Datasets and Experimental Setup

We evaluated our framework on several UCI datasets [13]. In addition, we studied the performance of associative classifiers in the classification of several challenging microarray datasets. In our evaluation we used the following experimental setup. We used an apriori-like algorithm to mine frequent, closed and maximal itemsets [14]. We generated classification rules from these patterns and we integrated them with the associative classifiers discussed in Section 4. A k-nearest neighbour algorithm is used in the second stage of the 2SARC system.

For UCI datasets we set the support threshold to 1%, 5% and 10%. The confidence threshold was set to 50%. On each UCI dataset we performed C4.5's shuffle utility [15] for shuffling the datasets. A 10-fold cross validation was performed on each dataset and the reported results are averages over the 10 folds. The continuous attributes were discretized using the entropy method in [16], with the code taken from MLC++ machine learning library [17]. It is the same discretization method for continuous attributes as used in [2]. The support threshold was harder to set for the microarray datasets. Across the set of datasets the support threshold ranged from 10% for *Prostate Cancer* dataset to 75% for *Lung Cancer* dataset. The confidence threshold was set to 50%. For each dataset we report the best result obtained under this parameter setting.

We evaluate the performance of the classifiers based on accuracy, Brier score, and cost curves [18]. Accuracy represents the percentage of the correct classifications out of the total number of classifications performed. The Brier score is the mean squared difference between the predicted probability of an instance, and one or zero depending if the actual label of the instance matches the prediction or not. The best possible score for Brier score is zero. Cost curves [18] are evaluation tools for classification systems that have been proposed as an alternative to ROC curves. Each classifier is represented by a straight line in the cost space. The y-axis is the normalized expected cost (NEC) of a classifier and is between 0 and 1. The x-axis (PC(+)) is the fraction of the total cost of using a classifier that is due to positive examples.

All the associative classifiers proposed in the literature so far have been evaluated under the assuption that they predict only classes. No score or probability estimation is given for the predictions. However, one can turn such a classifier into a probability estimator if a probability can be associated with the class prediction. We did just that for FR and AvR methods, where the probability of the prediction is the rule confidence and the average of the confidences respectively. Brier score and the area under the curve are reported for FR and AvR methods. Both these measures require a probability reported along with the prediction.

5.2 Results

This section presents the results that we obtained in our study. Our hypothesis has two components: first, we anticipate that the number of classification rules is significantly reduced when rules are obtained from closed and maximal itemsets; second, we expect to keep the same level or improve the performance of associative classifiers when the classification rules generated from closed and maximal patterns are used. The findings of our empirical studies clearly support our first hypothesis for the case of maximal frequent itemsets and in most cases of closed frequent itemsets. Indeed, in some cases the closed itemsets generated almost the same rules as those obtained from frequent patterns, but when maximal itemsets were used the number of classification rules always decreased. However, for our second hypothesis, the findings were not conclusive and further analysis (statistical and visual) with cost curves was necessary. In the remainder we will highlight these findings using different datasets and show how the rules were effectively reduced without putting the classification performance in jeopardy.

5.2.1 UCI Datasets

Table 1 shows the number of classification rules generated from frequent itemsets for different supports. Next to the number of rules, the percentage of rules reduced when they are obtained from closed and maximal is given. It can be observed from this table that our first hypothesis is correct. The number of classification rules is substantially reduced when closed and maximal patterns are employed. In addition, it can be observed that the use of maximal patterns is the most beneficial.

When closed patterns mined at support 1% are used, the set of classification rules is reduced up to 57.19% (*iris* dataset). Under the same support condition, the set of classification rules generated from maximal patterns is up to 87.03% (*iris* dataset) smaller than the rules obtained from frequent patterns. Similar trends are observed for supports of 5% and 10%. For three of the datasets (*led7*, *tic-tac-toe* and *waveform*) the set of classification rules generated from closed patterns is less than 0.5% smaller than the set obtained

Table 1: Number of classification rules generated from frequent itemsets and the percentage of rules dropped when closed and maximal itemsets are used; itemsets mined at 1%, 5% and 10% support

	support=1%			support=5%			support=10%		
dataset	**Frequent** (# rules)	**Closed** (% reduced)	**Maximal** (% reduced)	**Frequent** (# rules)	**Closed** (% reduced)	**Maximal** (% reduced)	**Frequent** (# rules)	**Closed** (% reduced)	**Maximal** (% reduced)
anneal	10500.10	27.92	29.61	2783.20	33.28	38.88	1259.10	33.40	43.47
australian	19457.20	15.30	23.83	4765.20	13.76	34.49	1942.90	13.10	42.92
breast-w	3153.80	9.14	41.66	894.90	2.56	49.46	484.80	1.98	51.20
cleve	7898.80	11.11	26.80	2730.90	5.31	38.18	1161.30	3.01	48.28
crx	24353.30	16.84	22.31	6180.80	19.07	31.89	2597.10	22.53	40.04
diabetes	635.60	13.09	51.64	233.90	9.79	66.61	94.90	6.74	67.44
german	31688.40	14.71	24.38	5328.60	8.54	37.01	1746.00	5.54	47.15
glass	958.60	44.12	48.26	321.00	48.69	58.60	199.20	50.10	62.00
heart	2242.10	7.33	32.74	1074.20	3.84	47.77	547.10	2.07	52.90
hepatitis	25074.90	14.96	18.51	10493.30	17.91	26.19	5208.40	16.69	29.81
horse	31909.10	20.66	27.52	2706.50	19.38	52.32	482.90	11.14	63.88
iris	92.50	57.19	87.03	65.50	54.81	88.55	50.20	53.19	91.04
labor	4790.90	34.25	37.36	684.90	49.06	65.02	137.30	42.32	72.10
led7	258.10	0.46	38.09	237.60	0.00	39.69	NA	NA	NA
pima	577.20	13.60	53.59	232.70	10.10	67.98	94.90	7.80	68.81
tic-tac-toe	6362.80	5.04	37.07	411.10	0.02	35.13	104.40	0.00	21.26
vehicle	48465.50	16.26	17.71	9854.00	20.51	25.24	1781.70	23.72	36.53
waveform	34886.30	0.01	27.67	610.20	0.00	37.30	47.80	0.00	21.97
wine	17860.30	20.77	23.81	5586.50	27.39	34.16	2490.30	32.29	40.06
zoo	25084.20	17.67	17.67	10963.30	22.38	22.38	5817.60	25.50	25.50
Average	14812.49	18.02	34.36	3307.92	18.32	44.84	1381.47	17.56	46.32

from frequent patterns. However, the use of maximal patterns for these datasets is still advantageous as it reduces the set of rules between 21.26% (*tic-tac-toe* dataset) and 39.69% (*led7* dataset). Note that for *led7* dataset no classification rules are generated at 10% support. Although the reduction in number of rules is sometimes small for closed patterns, the maximal patterns produce a substantial reduction in most cases. For instance, when *heart* dataset is mined at support 10% closed patterns reduce the number of rules by only 2.07%, while the use of maximal itemsets lowers the number of rules by 52.9%. The average provided in Table 1 shows that closed patterns reduce the number of rules by around 18% for all support thresholds, while the use of maximal patterns lowers the number of rules between 34.36% and 46.32% for the range of supports. Thus our hypothesis that the number of classification rules is substantially reduced by the use of closed and maximal patterns holds.

The second component of our hypothesis is about the level of performance of the classification systems. The results for the four associative classifiers investigated in our study are shown in Tables 2 to 5: Table 2 presents the accuracies of the ***FR*** method; Table 3 shows the accuracies of the ***AvR*** method; Table 4 presents the accuracies of the ***2SARC-CF*** method; and Table 5 presents the accuracies of the ***2SARC-RF*** method.

The results for ***FR*** method are presented in Table 2. When first rule scoring scheme ***FR*** is employed, the variation in the performance of the systems, measured by accuracy, using closed and maximal patterns compared to the system built with frequent patterns is as follows: for closed patterns it ranges from -0.27% to 0.93%; for maximal itemsets it ranges from -5.36% to 5.32%. It can be observed that the biggest variation in accuracy occurs for *iris* for which the accuracy went down due to the fact the number of rules discovered was very small to start with, and *labor* for which the accuracy went up substantially. When *iris* and *labor* datasets are discarded, the range is much smaller (-1.66% to 1.18%). On aver-

Table 2: Evaluation of FR scoring scheme

dataset	**accF**	**bsF**	**accC**	**bsC**	**accM**	**bsM**
anneal (1%)	93.74	0.05	93.74	0.05	93.96	0.05
australian (10%)	85.53	0.13	85.53	0.13	85.53	0.13
breast-w (1%)	95.41	0.04	95.4	0.04	95.41	0.04
cleve (10%)	84.52	0.14	84.52	0.14	83.85	0.15
crx (10%)	85.37	0.13	85.37	0.13	85.37	0.13
diabetes (5%)	74.32	0.19	74.32	0.19	75.5	0.17
german (1%)	71.1	0.27	71.1	0.27	71.1	0.27
glass (1%)	72.02	0.23	72.95	0.23	72.95	0.23
heart (1%)	81.87	0.17	81.87	0.17	82.24	0.17
hepatitis (5%)	80.56	0.19	80.56	0.19	80.56	0.19
horse (5%)	84.25	0.15	83.97	0.15	85.05	0.14
iris (1%)	94.01	0.05	94.67	0.04	89.33	0.05
labor (1%)	86.34	0.13	86.34	0.13	91.66	0.09
led7 (5%)	71.81	0.20	71.81	0.20	71.81	0.20
pima (5%)	74.2	0.19	74.2	0.19	74.99	0.18
tic-tac-toe (1%)	97.92	0.01	97.92	0.01	98.02	0.01
vehicle (1%)	59.72	0.34	59.72	0.34	59.6	0.34
waveform (1%)	81.86	0.16	81.86	0.16	82.0	0.16
wine (10%)	91.03	0.08	91.59	0.08	89.37	0.10
zoo (1%)	86.18	0.13	86.18	0.13	86.18	0.13
Average	82.58	0.15	82.68	0.15	82.72	0.15

age, both closed and maximal patterns improve by a small margin the classification performance. This trend is confirmed by the Brier score and by the area under the curve.

Table 3: Evaluation of AvR scoring scheme

dataset	accF	bsF	accC	bsC	accM	bsM
anneal (1%)	94.86	0.04	94.97	0.03	94.64	0.04
australian (10%)	85.97	0.11	85.53	0.11	85.38	0.12
breast-w (1%)	93.87	0.05	93.57	0.05	95.55	0.04
cleve (10%)	82.51	0.14	82.84	0.14	79.87	0.17
crx (5%)	85.39	0.11	85.66	0.11	85.39	0.12
diabetes (5%)	75.24	0.17	75.24	0.17	75.23	0.17
german (1%)	71.2	0.21	71.2	0.21	71.2	0.21
glass (1%)	70.62	0.24	71.08	0.24	71.54	0.24
heart (10%)	80.02	0.16	81.13	0.15	82.61	0.14
hepatitis (1%)	84.59	0.13	82.67	0.15	84.59	0.13
horse (10%)	84.22	0.13	84.49	0.13	82.6	0.14
iris (1%)	94.01	0.05	94.68	0.04	89.33	0.05
labor (5%)	88.0	0.11	86.33	0.12	89.66	0.11
led7 (1%)	70.81	0.20	70.78	0.20	70.76	0.20
pima (1%)	74.59	0.17	74.59	0.17	75.51	0.17
tic-tac-toe (1%)	92.82	0.06	93.22	0.06	93.65	0.06
vehicle (1%)	57.45	0.28	57.69	0.28	57.21	0.29
waveform (1%)	75.54	0.18	75.54	0.18	75.54	0.18
wine (10%)	89.93	0.09	92.74	0.06	89.95	0.09
zoo (5%)	87.18	0.10	87.18	0.10	87.18	0.10
Average	81.94	0.14	82.05	0.14	81.86	0.14

The accuracy results for average rules scoring scheme ***AvR*** is shown in Table 3. The difference in performance for closed patterns when compared to frequent patterns ranges from -2.29% to 0.66%; for maximal itemsets it ranges from -5.33% to 1.11%. On average, the closed patterns are beneficial to the classifier and slightly increase its performance in terms of accuracy, while the maximal patterns slightly decrease the classification accuracy. The performance of the classifier is the same on average for both closed and maximal, when the classifier is evaluated with the Brier score, while the performance slightly decreases for both closed and maximal when area under the curve is considered.

The performance of ***2SARC-CF*** system is presented in Table 4. The use of closed patterns decreases the performance by up to 1% and increases it by up to 1.66%; the variation in performance for maximal itemsets ranges from -3.34% to 3.33%. Note that, again, for maximal itemsets the range is much smaller (-1% to 0.58%) when the difference is computed without *iris* and *labor* datasets. On average, the performance increases slightly when closed patterns are used and decreases slightly when classification rules are generated from maximal patterns.

When ***2SARC-RF*** scoring scheme is considered (Table 5), the variation in the performance of the system for closed patterns ranges from -1.33% to 1.8%; for maximal itemsets it ranges from -3.99% to 3.67%. Note that for maximal itemsets the range is much smaller (-1.24% to 0.48%) when *iris* and *labor* datasets are discarded. On average, both closed and maximal patterns decrease insignificantly the classification performance.

These very small increases or decreases in accuracy do not seem to be significant, but a verification is required to support our hypothesis. We performed a series of Wilcoxon signed ranked tests

Table 4: Accuracy for 2SARC-CF scoring scheme

dataset	Frequent	Closed	Maximal
anneal	98.45	98.01	98.01
australian	87.27	87.41	87.85
breast-w	97.28	97.29	97.28
cleve	84.84	84.5	84.18
crx	86.97	86.84	86.83
diabetes	77.05	77.18	76.14
german	74.9	74.7	74
glass	70.99	70.68	71.13
heart	85.2	84.83	84.46
hepatitis	87.08	86.42	86.41
horse	84.78	84.53	84.25
iris	95.34	95.33	92
labor	93.33	94.99	96.66
led7	73.85	73.85	73.87
pima	75.51	75.11	74.86
tic-tac-toe	99.69	100	98.85
vehicle	66.91	68.21	67.38
waveform	79.7	79.8	79.06
wine	96.64	97.73	96.06
zoo	95.09	94.09	94.09
Average	85.54	85.58	85.17

Table 5: Accuracy for 2SARC-RF scoring scheme

dataset	Frequent	Closed	Maximal
anneal	99.01	99.01	99.12
australian	88.28	87.86	87.84
breast-w	97.43	97.29	97.42
cleve	84.19	83.52	84.21
crx	86.97	86.39	86.96
diabetes	75.49	75.36	74.33
german	73.6	73.8	73.8
glass	71.97	71.52	71.57
heart	85.94	85.94	85.94
hepatitis	85.84	87.64	85.8
horse	83.98	85.04	83.42
iris	95.99	96.65	92
labor	92.99	91.66	96.66
led7	74.28	74.09	74.08
pima	74.86	74.33	73.82
tic-tac-toe	100	100	100
vehicle	71.99	71.15	70.92
waveform	80.8	80.88	81.28
wine	97.77	97.22	97.74
zoo	98.33	97.09	97.09
Average	85.98	85.82	85.7

Table 6: Maximal versus frequent on UCI datasets

M vs. F	wins	losses	ties
FR	9	4	7
AvR	8	9	3
2SARC-CF	5	14	1
2SARC-RF	5	13	2

Table 7: Number of classification rules generated from frequent itemsets and the percentage of rules dropped when closed and maximal itemsets are used

dataset	**Frequent**	**Closed**	**Maximal**
	(# rules)	(% saved)	(% saved)
Breast Cancer	11304	20.21	20.25
Lung Cancer	10139	22.6	22.6
AML-ALL	40441.4	15.23	15.23
Prostate Cancer	95307.8	14.28	14.28

between the classifier built from frequent patterns and the model built from maximal patterns. In addition, we tested the classifier built from frequent patterns versus the model built from closed patterns. There was indeed no statistically significant difference in the performance. The Wilcoxon test is a non-parametric test [19], it does not make any assumptions about the distributions of the values. Based on average performance we can conclude that using maximal patterns is advantageous because the reduction in the number of rules is significant, while the improvement or the drop in the performance level is not statistically significant.

Table 6 shows on how many datasets the use of maximal patterns performs better (wins), as well as (ties) or worse (losses) than when frequent patterns are used.

The use of maximal patterns is more beneficial to ***FR*** and ***AvR*** systems. This is due to their naïve scoring schemes and thus reducing the redundancy in the rule set is beneficial. ***2SARC-CF*** and ***2SARC-RF*** use scoring schemes that are learned automatically from data and thus making the system less sensitive to the redundancy in the set of rules.

This empirical study found no significant effect on the accuracy of classifiers with different rule selection schemes when closed or maximal frequent patterns are used instead of all frequent itemsets. The gain, however, is in the reduction of the number of classification rules. It remains to see whether this observation is still true with more challenging datasets such as microarray data which contain a relatively small set of samples.

5.2.2 Microarray Datasets

Microarray data contains measurements of a large number of genes for a particular sample. Due to high acquisition costs, generally there is a small number of samples in microarray datasets. The large feature space (given by the number of genes) and the small number of samples make the construction of a good classifier difficult. We have access to microarray data for breast cancer, lung cancer, leukemia and prostate cancer. A method for reducing the dimensionality of the feature space for these microarray data has been proposed in [20]: first, biclusters are found in data (a bicluster represents a subset of genes that are similar for a subset of samples); second, transform the original data based on bicluster membership. This transformation reduces dramatically the feature space. In addition, the new features are binary, representing the membership in a bicluster. This new representation is highly suitable for association rule mining. Thus we investigate our framework on these microarray datasets and the results are presented in Tables 7 and 8. It is relevant to notice that while the authors of [20] claim to have reached the best known classification results on these microarray datasets, our results using 2SARC with maximal patterns outperformed their classification results on Lung Cancer and AML-ALL datasets.

Table 8: Accuracy on microarray datasets

dataset	**Frequent**	**Closed**	**Maximal**
FR scoring scheme			
Breast Cancer	58.0	58.0	58.0
Lung Cancer	96.1	96.1	96.1
AML-ALL	75.72	75.72	75.72
Prostate Cancer	79.82	79.82	79.82
AvR scoring scheme			
Breast Cancer	58.0	58.0	58.0
Lung Cancer	96.1	96.1	96.1
AML-ALL	77.16	77.16	77.16
Prostate Cancer	84.34	83.6	83.64
2SARC-CF scoring scheme			
Breast Cancer	81.58	81.58	81.58
Lung Cancer	97.24	96.7	96.7
AML-ALL	92.88	92.88	92.88
Prostate Cancer	86.6	86.6	86.6
2SARC-RF scoring scheme			
Breast Cancer	85.5	85.5	85.5
Lung Cancer	96.72	96.72	96.72
AML-ALL	91.44	91.44	91.44
Prostate Cancer	86.6	86.6	86.6

Table 7 shows the number of classification rules and their reduction when rules are generated from closed and maximal patterns. For all the microarray datasets the reduction in the number of rules is almost the same for closed and maximal, indicating yet again that using maximal patterns is indeed a winning strategy. The accuracy results for ***FR***, ***AvR*** and ***2SARC-RF*** methods (shown in Table 8) remain the same when closed and maximal patterns are used instead of frequent ones. The only variation occurs for ***2SARC-CF*** for *Lung Cancer* dataset: the performance insignificantly decreases for the approaches using closed and maximal patterns.

The results on microarray data and UCI datasets confirm our hypothesis, that the use of closed and maximal patterns maintains the level of performance while reducing the number of classification rules. Thus, based on our results so far, we can conclude that the use of maximal patterns is advantageous. However, we want to study when exactly is the use of maximal patterns more advantageous than closed patterns and vice-versa. We use cost curve analysis for this purpose.

5.2.3 Cost Curves Analysis

In the previous sections we presented and discussed the accuracy obtained for all the studied methods. To gain a better insight into the examined framework we perform an analysis based on cost curves. Cost curves [18] are evaluation tools for classification systems that have been proposed as an alternative to ROC curves.

Their advantage is that in their visualization one can easily see the performance of a classifier over the entire range of class frequencies and costs.

In our analysis we are interested to see under what conditions the use of closed and maximal patterns is advantageous to an associative classifier. Cost curves allow us to easily visualize and detect these conditions.

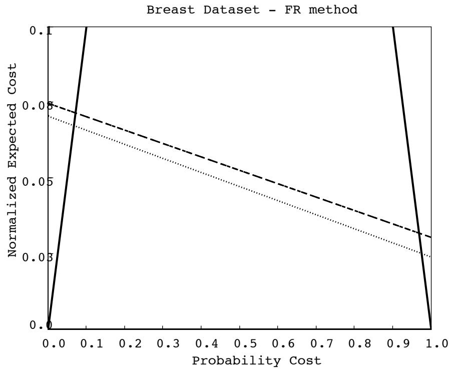

Figure 1: Cost curve performance for FR method on Breast-w dataset: frequent patterns - long-dashed line; closed patterns - dashed line; maximal patterns - dotted line; trivial classifiers - solid lines. Note that the dashed and long-dashed lines are overlapping.

Figures 1 to 4 show the performance of classification for several datasets. All the graphs show the performance of a classification method when classification rules are generated from frequent, closed and maximal patterns. Thus, three classifiers are compared in each graph. The classifier built from frequent patterns is represented by a long-dashed line. The one built from closed itemsets is shown with a dashed line. The dotted line corresponds to the classification system built from maximal patterns. The two solid lateral lines represent the trivial classifiers: the leftmost line represents the classifier that always predicts the negative class, while the rightmost line represents the classifier that always predicts positives. Note that cost curve evaluation can be done only for 2-class datasets.

Let us analyze the *Breast-w* dataset. As shown in Table 1 the use of closed patterns reduces only slightly (2%-9%) the set of classification rules, while the set of classification rules generated from maximal patterns is substantially smaller (40% to 50% smaller) than the one generated from all frequent itemsets. The accuracy for frequent and closed patterns is almost identical, while for maximal patterns it increases with 0.58%. Based on this information, one may conclude that the use of maximal patterns with ***FR*** method is the best choice. The cost curve shown in Figure 1 confirms this choice since the dotted line representing the maximal is indeed the lowest across the full range of possible $PC(+)$ values. Note that the dashed and long-dashed lines are overlapping, indicating that the classifier based on closed and frequent patterns have the same performance. Let us now look at the same dataset when ***2SARC-CF*** method is used. The reduction in number of rules is the same as with ***FR***. However, the performance in terms of accuracy is identical for frequent, closed or maximal. Thus one may assume again, that the use of maximal patterns (they lead to the smallest set of classification rules) would be the best choice. However, the analysis with cost curves sheds new light on this choice (see Figure 2). It can be observed from the graph that maximal patterns should be used only for a probability cost higher than 0.65, while for a smaller probability, closed patterns should be preferred. The use of closed patterns reduce only slightly the set of classification rules. This would be an indication that also the performance should not vary too much between frequent and closed. This assumption is validated by both Figure 1 and 2. In Figure 1 frequent and closed have exactly the same performance over the entire range of class frequencies, while in Figure 2 there is only a maximum difference of 0.0015 in the normalized expected cost.

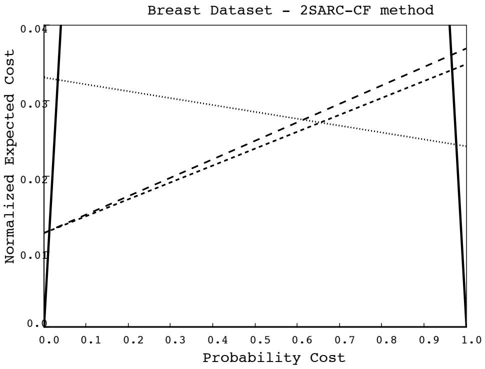

Figure 2: Cost curve performance for 2SARC-CF method on Breast-w dataset: frequent patterns - long-dashed line; closed patterns - dashed line; maximal patterns - dotted line; trivial classifiers - solid lines

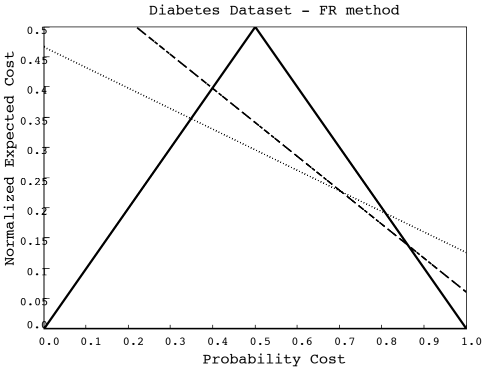

Figure 3: Cost curve performance for FR method on Diabetes dataset: frequent patterns - long-dashed line; closed patterns - dashed line; maximal patterns - dotted line; trivial classifiers - solid lines

Diabetes is another interesting dataset to be studied. The trend in rule reduction is similar to the *Breast-w* dataset, but the performance of the classifiers on this dataset is much poorer. The performance of ***FR*** method for *Diabetes* dataset is shown in Figure 3. Contrary to the classification systems in Figure 2, the classifiers do not perform well on the entire range of class frequencies. Indeed, the trivial classifier can outperform those classifiers. They should be used only on a smaller range of $PC(+)$ ([0.35-0.85]). Outside this interval trivial classifiers perform better. The use of maximal patterns is advantageous in the [0.35-0.7] range, while any of the frequent or closed patterns should be used in the [0.7-0.85] interval. Again, the classifiers built from frequent and closed patterns

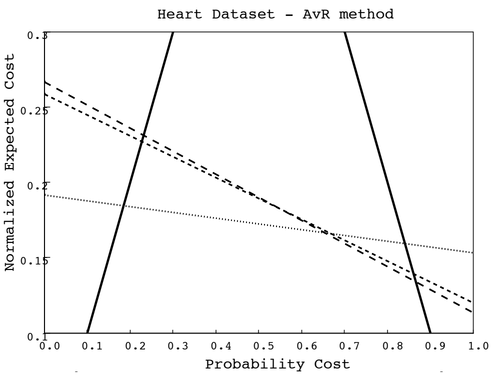

Figure 4: Cost curve performance for AvR method on Heart dataset: frequent patterns - long-dashed line; closed patterns - dashed line; maximal patterns - dotted line; trivial classifiers - solid lines

overlap in this graph.

Figure 4 presents the performance of ***AvR*** method for *Heart* dataset. ***AvR*** method performs better than the trivial classifiers in the [0.18-0.86] interval. Maximal patterns are more beneficial than frequent or closed patterns in the $PC(+)$ range [0.18-0.64]. Frequent patterns or even closed itemsets should be favored on the remaining interval of $PC(+)$.

In this section we have analyzed with cost curves several interesting cases. In general, our investigations using cost curves suggest that the use of maximal patterns leads to the best classification performance over most of the probability ranges. In applications were the class distribution changes, one may want to use the cost curves to determine the best classifier.

6. CONCLUSIONS

In this paper we investigated the performance of associative classifiers when the classification rules are generated from frequent, closed and maximal itemsets. We showed that maximal itemsets substantially reduce the number of classification rules without jeopardizing the accuracy of the classifier. Our extensive analysis demonstrates that the performance remains stable and even improves in some cases. Our analysis using cost curves also provides recommendations on when it is appropriate to remove redundancy in frequent itemsets. Based on our thorough analysis we are confident that any investigation of associative classifiers should consider first and foremost classification rules generated from maximal patterns.

7. REFERENCES

[1] W. Li, J. Han, and J. Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules," in *Proc. of ICDM*, 2001, pp. 369–376.

[2] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in *Proc. of SIGKDD*, 1998, pp. 80–86.

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

[4] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," in *Proc. of SIGMOD*, 1997, pp. 255–264.

[5] B. Goethals and M. Zaki, Eds., *FIMI'03: Workshop on Frequent Itemset Mining Implementations*, ser. CEUR Workshop Proceedings series, vol. 90, 2003.

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

[7] O. R. Zaïane and M. El-Hajj, "Pattern lattice traversal by selective jumps," in *Proc. of SIGKDD*, 2005, pp. 729–735.

[8] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules," *Lecture Notes in Computer Science*, vol. 1540, pp. 398–416, 1999.

[9] R. J. Bayardo, "Efficiently mining long patterns from databases," in *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, 1998, pp. 85–93.

[10] R. Bayardo, "Brute-force mining of high-confidence classification rules," in *Proc. of SIGKDD*, 1997, pp. 123–126.

[11] M.-L. Antonie, O. R. Zaïane, and R. Holte, "Learning to use a learned model: A two-stage approach to classification," in *Proc. of ICDM*, 2006.

[12] M.-L. Antonie and O. R. Zaïane, "Text document categorization by term association," in *Proc. of ICDM*, 2002, pp. 19–26.

[13] C. Blake and C. Merz, "UCI repository of machine learning databases," http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

[14] C. Borgelt, "Apriori software," http://fuzzy.cs.uni-magdeburg.de/∼borgelt/software.html, 2008.

[15] J. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[16] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *IJCAI*, 1993, pp. 1022–1029.

[17] R. Kohavi, G. John, R. Long, D. Manley, and K. Pfleger, "Mlc++: a machine learning library in c++," in *IN TOOLS WITH ARTIFICIAL INTELLIGENCE*. IEEE Computer Society Press, 1994, pp. 740–743.

[18] C. Drummond and R. C. Holte, "Cost curves: An improved method for visualizing classifier performance," *Machine Learning*, vol. 65(1), pp. 95–130, 2006.

[19] J. Demsar, "Statistical comparisons of classifiers over multiple datasets," *Journal of Machine Learning Research*, vol. 7, no. 1, pp. 1–30, 2006.

[20] N. Asgarian and R. Greiner, University of Alberta, Tech. Rep., 2007, http://www.cs.ualberta.ca/ greiner/R/RoBiC/.