Improving associative classification by incorporating novel interestingness measures

DOI: 10.1109/ICEBE.2005.76
Handle: http://hdl.handle.net/1942/1514

Improving associative classification by incorporating novel interestingness measures

Yu Lan[a], **Davy Janssens**[b], **Guoqing Chen**[a] **and Geert Wets**[b]

[a]School of Economics and Management, Tsinghua University, Beijing 100084, China
[b]Universiteit Hasselt, Campus Diepenbeek, Agoralaan-Gebouw D, B-3590 Diepenbeek, Belgium

Available online 3 October 2005.

Abstract

Associative classification has aroused significant attention in recent years and proved to generate good results in previous research efforts. This paper aims to contribute to this line of research by the development of more effective associative classifiers. Our goal is to achieve this by the incorporation of two novel interesting measures, i.e. intensity of implication and dilated chi-square, into an existing associative classification algorithm, respectively. The former interesting measure was merely proposed with the purpose of mining meaningful association rules, while the latter was designed to reveal the interdependence between condition and class variables. Each of these two measures is applied as the primary sorting criterion within the context of the well-known CBA algorithm in an attempt to organize the composition of the rule sets in a more reasonable sequence. Benchmarking experiments on 16 popular UCI datasets revealed that our algorithms could empirically generate accurate and significantly more compact decision lists. In addition to this, the algorithm was validated on a separate credit scoring dataset, which contained 7190 credit scoring samples.

Keywords: Associative classification; Intensity of implication; Dilated chi-square; Credit scoring

Article Outline

1. Introduction

Classification and association-rule discovery are two of the most important tasks addressed in the data mining literature. In recent years, extensive research has been carried out to integrate both approaches. By focusing on a limited subset of association rules, i.e. those rules where the consequent of the rule is restricted to the class variables, it is possible to build more accurate classifiers. Several publications have shown that associative classification is intuitive and effective in many cases (Dong et al., 1999, Liu et al., 1998, Liu et al., 2001, Wang and Zhou, 2000 and Yin and Han, 2003). Normally, association rules search globally for all rules that satisfy minimum support and minimum confidence thresholds. The richness of the rules gives this technique the potential of reflecting the true classification structure in the data.

Associative classification is first proposed in CBA (Liu et al., 1998), in which the popular Apriori algorithm has been applied in order to extract a limited number of association rules with their consequents limited to class labels. These rules are then sorted by descending confidence and are pruned in order to get a minimal number of rules that are necessary to cover training data and achieve satisfying accuracy. Another associative classifier ADT (Wang & Zhou, 2000) organizes the rule sets in the tree structure according to its defined relations. The decision tree pruning techniques is then applied to remove rules that are too specific. CPAR, CMAR and CAEP are three of the latest associative classification algorithms (Dong et al., 1999, Liu et al., 2001 and Yin and Han, 2003). They, respectively, propose expected accuracy, weighted chi-square and growth rate as rule interestingness measures, and all perform classification based on multiple rules that the new sample fires. The aim of this paper is to improve the CBA algorithm in order to generate a more accurate and compact decision list, which is convenient for decision makers to understand and adopt. Two novel interestingness measures, intensity of implication and dilated chi-square, are applied as the primary sorting criteria instead of the currently adopted confidence measure. Intensity of implication and dilated chi-square statistically reveal the interdependence between the antecedence and consequence of rules and empirically allocate rules in a more reasonable order.

The remainder of this paper is arranged as follows. Section 2 introduces the basic concepts of associative classification, along with the sorting mechanism applied by CBA. Section 3 elaborates on the weakness of conditional probability (confidence), as well as on the design of intensity of implication and dilated chi-square to overcome it. The results of the empirical evaluation are shown in Section 4. Section 5 gives some concluding remarks.

2. Associative classification

A comprehensive overview of the original CBA algorithm is necessary before we propose improvements to it. First, we will give an introduction to class association rules. Hereafter, the ranking and pruning mechanisms in CBA are described.

2.1. Class association rules

Let $I=\{i_1, i_2,\ldots i_k\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T\subseteq I$. We say that a transaction T contains X, a set of items in I, if $X\subseteq T$. An association rule is an implication of the form $X\Rightarrow Y$, where $X\subset I$, $Y\subset I$ and $X\cap Y=\varnothing$. The rule $X\Rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y. The rule $X\Rightarrow Y$ has support s in the transaction set D if s% of transactions in D contain $X\cup Y$. Given a set of transactions D, the problem of mining association rules is to generate all association rules that have support and confidence greater than a user-specified minimum support (minsup) and minimum confidence (minconf) (Agrawal et al., 1993 and Agrawal and Srikant, 1994).

To make association rules suitable for the classification task, the associative classification method focuses on a special subset of association rules, i.e. those rules with a consequent limited to class variables only, the so-called class association rules (CARs). Thus, only rules of the form $A\Rightarrow c_i$, where c_i is a possible class, are generated.

2.2. Ranking and pruning of CARs in CBA

Building a classifier in CBA is largely based on a database coverage pruning method, which is applied after all the CARs have been generated. At the first step of the pruning, the algorithm ranks all the CARs and sorts them in the descending sequence. As we will show in the next section, this rank will be subject to one of the modifications that were implemented. The ranking is as follows: given two rules r_i and r_j, $r_i>r_j$ (or r_i is said having higher rank than r_j), if (1) conf (r_i)>conf (r_j); or (2) conf (r_i)=conf (r_j), but sup (r_i)>sup (r_j); or (3) conf (r_i)=conf (r_j) and sup (r_i)=sup (r_j), but r_i is generated before r_j. Each training sample is classified by the rule that covers it and has the highest ranking. The pruning algorithm tries to select a minimal number of rule sets, each of which correctly classifies at least one training sample, to cover the training dataset and to achieve the lowest error rate. The default class is set as the majority class among the remaining samples that are not covered by any rule in the final classifier.

3. Novel interestingness measures

3.1. Limits of confidence

A profound examination of the algorithm identified a potential weakness in the way the rules are sorted. Since rules are inserted in the classifier primarily according to its confidence, this will determine to a large extent the accuracy of the final classifier. Confidence is a good measure for the quality of (class) association rules but it also suffers from certain weaknesses (Guillaume et al., 1998 and Janssens et al., 2005).

Firstly, the conditional probability of a rule $X\Rightarrow Y$ is invariable when the $s(Y)$ or $|D|$ varies, where $s(Y)$ denotes the subset of samples that contain Y and D is the whole database. Let $A=s(X)$, $B=s(Y)$, $n=|D|$, $n_a=|A|$, $n_b=|B|$, and $n_{ab}=|A\cap B|$. The confidence of rule $X\Rightarrow Y$ is calculated as n_{ab}/n_a. Keeping the numerator and denominator fixed, the confidence is stable when the size of $s(Y)$ or D changes. Nevertheless, as shown in Fig. 1, the rule $X\Rightarrow Y$ is more likely to happen when the size of $s(Y)$ increases or when the size of D decreases. It is not surprising that, when $s(Y)$ is close to the size of D, the observations which are covered by the

antecedent X of the rule, are also included in $s(Y)$. Furthermore, the implication will be more meaningful when the size of all the sets grows in the same proportion.

 (6K)

Fig. 1. Tree cases with constant confidence.

The second drawback for the use of conditional probability is that when for a particular class, the minsup parameter is set to 1% or even lower, it might very well happen that some rules have a high confidence parameter but meanwhile they might be confirmed by a very limited number of instances. As a result, those rules may stem from noise only. This is why it is always dangerous to look for implications with small support even though these rules might look very 'interesting'. As a result, choosing the most confident rules may not always be the best selection criterion.

Therefore, two novel interestingness measures that take both drawbacks into account, i.e. intensity of implication and dilated chi-square, were designed to adjust the ranking mechanism in CBA algorithm. The next sections elaborate on this.

3.2. Intensity of implication

Intensity of implication, introduced by Gras and Lahrer (1993) measures the distance to random choices of small, even non-statistically significant, subsets. In other words, it measures the statistical surprise of having so few negative examples on a rule as compared with a random draw. Now, let U and V be two sets randomly chosen from D with the same cardinality as $s(X)$ and $s(Y)$, respectively, i.e. $|U|=n_a$ and $|V|=n_b$. The comparison is illustrated in Fig. 2.

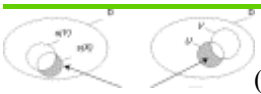 (12K)

Fig. 2. Comparison with random case.

Let $N_{u\bar{v}} = |U \cap \bar{V}|$ be the random variable that measures the expected number of random negative examples under the assumption that U and V are independent, and $n_{a\bar{b}}$ the number of negative samples observed on the rule. Now, if $n_{a\bar{b}}$ is unusually small compared with $N_{u\bar{v}}$, the one we would expect at random, then we say that the rule $X \Rightarrow Y$ has a strong statistical implication. In other words, the intensity of implication for a rule $X \Rightarrow Y$ is stronger, if the quantity $\Pr[N_{u\bar{v}} \leq n_{a\bar{b}}]$ is smaller. Intensity of implication is then defined as $1 - \Pr[N_{u\bar{v}} \leq n_{a\bar{b}}]$. The random variable $N_{u\bar{v}}$ follows the hypergeometric law, which means $\Pr[N_{u\bar{v}} = k] = \Pr[\text{of } |U| \text{ examples selected at random, exactly } k \text{ are not in } V]$. Let $n_u=|U|$, $n_v=|V|$, $n_{\bar{v}} = |\bar{V}|$. It equals

$$\frac{C_{n_{ab}}^{k} C_{n-n_v}^{n_v-k}}{C_n^{n_v}}$$

Taking into account that $n_v=n_a$ and $n_v=n_b$, the intensity of implication can be written as:

$$1-\sum_{k=\max(0,n_a-n_b)}^{n_{ab}} \frac{C_{n_b}^{k} C_{n-n_b}^{n_a-k}}{C_n^{n_a}}$$

This formula for intensity of implication is suitable as long as the number of samples in the database, i.e. $|D|$, is reasonably small. Otherwise, the combination numbers in the above formula explode very quickly. Therefore, Suzuki and Kodratoff (1998) came up with an approximation of this formula for big datasets. They argue that if n_{ab} is small, which is often the case in rule discovery, then Poisson approximations can be applied. In that case, the above formula for intensity of implication reduces to a much simpler version that is easier to compute:

$$1-\sum_{k=\max(0,n_a-n_b)}^{n_{ab}} \frac{C_{n_b}^{k} C_{n-n_b}^{n_a-k}}{C_n^{n_a}} \approx 1-\sum_{k=0}^{n_{ab}} \frac{\lambda^k}{k!} e^{-\lambda}$$

With

$$\lambda = \frac{n_a(n-n_b)}{n}$$

Keeping the confidence of rule $X \Rightarrow Y$ constant, the intensity of implication varies with the size of $s(Y)$, with the size of D, and by dilation of n when n_a/n, n_b/n and n_{ab}/n stay constant, as Fig. 3 shows.

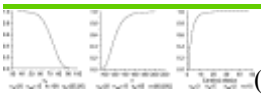
(8K)

Fig. 3. Sensitivity analysis of intensity of implication.

3.3. Dilated chi-square

Traditional chi-square test statistics (χ^2) is a widely used method for testing independence and/or correlation (Mills, 1955). Essentially, it is based on the comparison of observed frequency with the corresponding expected frequencies. Let f_0 be an observed frequency, and f be an expected frequency. The χ^2 value is defined as

$$\chi^2 = \sum \frac{(f_0 - f)^2}{f}$$

to test the significance of the deviation from the expected values. For each rule X⇒Y and the training dataset *D*, a 2*2 contingency table can be derived as Fig. 4:


(4K)

Fig. 4. A 2*2 contingency table for rule X⇒Y and dataset *D*.

The χ^2 value for rule X⇒Y can be calculated as

$$\chi^2 = \frac{(m_{11}m_{22} - m_{12}m_{21})^2 |D|}{(m_{11} + m_{12})(m_{21} + m_{22})(m_{11} + m_{21})(m_{12} + m_{22})}$$

However, simply using the traditional χ^2 value will be favourable to the situation where the distribution of row total is close to that of column total distribution. We, therefore, proposed dilated chi-square to conquer this shortcoming. The definitions of local maximum χ^2 and global maximum χ^2 are given first, along with their related properties. An example is then demonstrated to validate our opinion.

Definition 1

Given a dataset *D* and class label *Y*, the local maximum χ^2, denoted as $l\max(\chi^2)$, is the maximum χ^2 value for a fixed support count of *X*.

Definition 2

Given a dataset *D* and class label *Y*, the global maximum χ^2, denoted as $g\max(\chi^2)$, is the maximum χ^2 value for any possible support count of *X*.

Considering Fig. 4, |*D*| and the support count of *Y* are settled when given a dataset *D* and class label *Y*. $g\max(\chi^2)$ is the maximum χ^2 value that one rule $X \Rightarrow Y$ may obtain, while $l\max(\chi^2)$ is the maximum χ^2 value under the condition that the support count of *X* is fixed.

Property 1

:

$$l\max(\chi^2) = \frac{(n_1 n_2)^2 |D|}{(m_{11} + m_{12})(m_{21} + m_{22})(m_{11} + m_{21})(m_{12} + m_{22})}$$

where

$$n_1 = \min(\min(m_{11}+m_{12}, m_{21}+m_{22}), \min(m_{11}+m_{21}, m_{12}+m_{22}))$$

$$n_2=\min(\max(m_{11}+m_{12},m_{21}+m_{22}),\max(m_{11}+m_{21},m_{12}+m_{22}))$$

That is to say, the local max χ^2 value is arrived at the most deviation from the expected frequency when the support count of X is given.

Property 2

:

$$g\max(\chi^2)=|D|$$

Proof: Without lose of generality, We suppose $m_{11}+m_{21}\geq m_{12}+m_{22}$ and $m_{11}+m_{12}\geq m_{21}+m_{22}$, then

$$n_1^2=(\min(m_{21}+m_{22},m_{12}+m_{22}))^2\leq(m_{21}+m_{22})(m_{12}+m_{22})$$

$$n_2^2=(\min(m_{11}+m_{12},m_{11}+m_{21}))^2\leq(m_{11}+m_{12})(m_{11}+m_{21})$$

Therefore

$$l\max(\chi^2)\leq|D|=g\max(\chi^2)$$

The equation is arrived when $m_{21}+m_{22}=m_{12}+m_{22}$ and $m_{11}+m_{12}=m_{11}+m_{21}$, i.e. the distribution of row total equals that of column total.

We modified the example in (Liu et al., 2001) to illustrate the problem when simply choosing χ^2 value as the interestingness measure for associative classification, which is our motivation to design the novel measure, i.e. dilated χ^2.

Example 1: In a credit card application approval case, three rules are generated:

r_1

job=no⇒rejected (support count of rule=30, confidence=60%)

r_2

education=university⇒approved (support count of rule=199, confidence=99.5%)

r_3

number of children >4⇒rejected (support count of rule=2, confidence=100%).

The contingency tables for these three rules are displayed in Fig. 5:


(11K)

Fig. 5. Contingency tables for rules.

The χ^2 values of the three rules are, respectively, 88.7, 33.4 and 18.1, and the local maximum χ^2 values 287.2, 37.0 and 18.1. It is evident that the χ^2 values are favourable to the situation where the distribution of row total is close to that of column total. For a customer having no job and with university education, her application will be rejected according to r_1, if the choice of rules is based on only χ^2 values. However, r_2 is intuitively much better than r_1 since r_2 has much higher support and confidence. Moreover, although the support of r_3 is very low, r_3 has a 100% confidence. The interestingness of r_3 seems a bit underestimated by its χ^2 value.

Since the χ^2 value has a bias to different row total distributions, we adjust it to a more uniform and fare situation and get a novel interestingness measure called dilated χ^2 value, denoted as dia(χ^2) (Yu, Chen, Janssens, & Wets, 2004). More concretely, we heuristically dilate the χ^2 value according to the relationship between the local and global maximum χ^2 values for current rule and database. The dilation procedure is nonlinear and empirically achieved excellent results, as demonstrated in next section.

$$\frac{\mathrm{dia}(\chi^2)}{\chi^2} = \left(\frac{g\max(\chi^2)}{l\max(\chi^2)}\right)^{\alpha} = \left(\frac{|D|}{l\max(\chi^2)}\right)^{\alpha}, \text{where } 0 \le \alpha \le 1$$

Therefore

$$\mathrm{dia}(\chi^2) = \left(\frac{|D|}{l\max(\chi^2)}\right)^{\alpha}\chi^2$$

The parameter α is used to control the impact of global and local maximum χ^2 values and tuned for different classification problems. Although α can be any positive real number, it is restricted empirically between 0 and 1. The dilated χ^2 values for the three rules are, respectively, 117.0, 136.1 and 95.1 if α=0.5, which is much more reasonable to our intuition. For a given training dataset, the size of D is fixed and irrelevant to the ranking of interestingness of rules.

It can be seen that the dilated χ^2 value is sensitive when the size of $s(Y)$ or D varies. Furthermore, for these rules with high confidence and very low support, dilate χ^2 values estimate their interestingness in a more cautious way.

The sensitivity analysis in Section 3.2 is also applied to dilated chi-square when α is set at 0.3 and 0.8. As shown in Fig. 6, in the first case where n_b increases while n_a, n_{ab} and n remain

stable, the dilated χ^2 first gradually declines to zero, when n_b equals 75. This is the situation when the confidence of the rule is equal to the proportion of class Y in the whole dataset D, i.e. 0.75. Dilated χ^2 then climbs up sharply if n_b continues to increase, which indicates the negative relationship between X and Y. Therefore, those rules whose confidence is less than their corresponding class proportion are not expected to exist. The similar mechanism occurs in the second case and dilate χ^2 is close to zero when size of D equals 113. The third case show that dilated χ^2 increases linearly if all subsets are cardinally dilated.

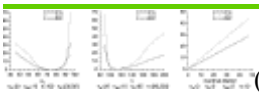 (11K)

Fig. 6. Sensitivity analysis of dilated chi-square.

We now adapted CBA by taking intensity of implication and dilated χ^2, respectively, as the primary criteria to sort the rule set. Rule r_i has a higher rank than rule r_j if it has a larger value of intensity of implication (or dilated χ^2). When two rules have the same values of intensity of implication (or dilated χ^2), they are ranked according to the ranking mechanism of the original CBA, which is mentioned in Section 2.2.

4. Empirical section

Before the adapted CBA algorithms are validated on our 16 benchmarking datasets from UCI, we first introduce the adaptations by means of a toy example to observe the behaviour of the new algorithms.

4.1. Toy example

The toy example, as shown in Fig. 7, has eight attributes and 20 examples. Attribute h is set as the class attribute. Each of these attributes has two possible values. Original CBA was run on this toy example using its default parameter sets. It generated 34 class association rules in total. These rules were then sorted and pruned to build the classifier.

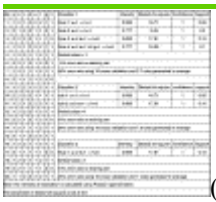 (48K)

Fig. 7. Toy example and three classifiers.

The details of the final classifiers are listed in Fig. 7. Classifier 1, which is generated by original CBA, consists of four rules and has an error rate of 5% on the training dataset. But the original CBA will get a 40% error rate using 10-fold cross validation. Classifier 2 and 3 are generated by adapted CBA that incorporates intensity of implication and dilate χ^2 as the

primary sorting criteria, respectively. Although their error rates on training dataset are both 10%, which is a bit higher than classifier 1, these two algorithms achieve 20 and 25% error rate in average using 10-fold cross validation, which gives an initial indication about the better generalization abilities of the adapted algorithms on this toy example. In addition, Classifier 2 and 3 are more compact (less rules) than classifier 1 that is generated by original CBA, which favours Occam's Razor theory.

4.2. Benchmarking on real life datasets

In order to get a more comprehensive evaluation, 16 UCI datasets (Blake & Merz, 1998) are classified by original CBA, the classical decision tree technique C4.5 (Quinlan, 1993) and Naïve Bayes. All of these selected classifiers are white-box techniques. C4.5 and Naïve Bayes are implemented by the software package of WEKA (Witten & Frank, 2000). The continuous attributes are discretized based on entropy (Fayyad & Irani, 1993) if needed. 10-fold cross validation is used to test the performance of these classifiers in an attempt to reduce the fluctuation that stems from random sampling.

As shown in Table 1, adapted CBA^1 and CBA^2, which, respectively, correspond to the new algorithms that incorporate intensity of implication and dilated χ^2, perform better than any of the other classifiers in comparison to the average error rate. The average error rate of adapted CBA^1 on these 16 datasets is 13.21%, and that of adapted CBA^2 is only 12.81% if the best parameter α is selected for each dataset. Furthermore, adapted CBA^1 generates 17.925 rules in average, which is almost one third of those rules generated by original CBA. The classifiers built by adapted CBA^2 are more compact and averagely contain 11.82 rules. The original CBA also has a better performance than C4.5 and Naïve Bayes. Although Naïve Bayes performs excellent on several datasets such as breast, heart and labor, its behaviour is unstable since it assumes attributes are independent, which is a very fragile assumption in real life datasets. The performance of C4.5 on discretized datasets is better than on original datasets, so only the former result is presented.

Table 1.

Benchmark experiments

Dataset		**Original CBA**		**Adapted CBA^1**		**Adapted CBA^2**		**C45**	**NB**
		Error rate (%)	**No. of rules**	**Error rate (%)**	**Num. of rules**	**Error rate (%)**	**No. of rules**	**Error rate (%)**	**Error rate (%)**
1	Austra	14.35	130.5	13.48	26.4	13.04	12.4	13.48	18.70
2	Breast	3.86	42.2	4.72	28.4	3.58	28.3	4.43	2.58
3	Cleve	17.16	63.8	15.47	16.9	16.13	9.6	20.79	16.17
4	Crx	14.93	138.2	12.90	34.2	13.04	12.4	12.75	18.99
5	Diabetes	22.26	38.5	24.21	10.4	21.74	10.7	22.92	24.22
6	German	26.70	134	25.60	56.5	26.80	19.7	27.60	25.30
7	Heart	17.78	37.6	16.30	13.6	16.67	7.4	18.89	14.81

Dataset		Original CBA		Adapted CBA[1]		Adapted CBA[2]		C45	NB
		Error rate (%)	No. of rules	Error rate (%)	Num. of rules	Error rate (%)	No. of rules	Error rate (%)	Error rate (%)
8	Hepati	16.21	25.2	18.67	18.4	16.83	11.3	16.77	15.48
9	Horse	19.03	87.9	14.12	1	14.12	1	15.22	20.92
10	Hypo	1.64	30	1.23	24.4	0.85	10.9	0.85	1.90
11	Iono	8.25	44.8	9.10	21.7	6.55	18.5	9.69	8.26
12	Labor	10.00	12.5	11.67	4.2	8.33	4.4	15.79	8.77
13	Pima	23.43	38.3	23.17	11	22.00	10.7	22.66	25
14	Sick	2.64	47.4	2.43	10.7	3.25	1	2.07	4.32
15	Sonar	22.60	41	18.31	27.4	18.74	21.8	18.75	25.48
16	Ti-tac	0.00	8	0.00	8	3.34	9	14.20	29.65
Average		13.80	49.34	13.21	17.925	12.81	11.82	14.80	16.28

A deeper insight into the comparisons between original CBA, adapted CBA[1] and CBA[2] reveals that the adapted algorithms are more suitable for the application of credit scoring. Tests on three related datasets, austral, crx and german, show that adapted CBA[1] and CBA[2] have satisfactory accuracy improvements and achieve much more compact classifiers at the same time. Another interesting test can be seen on the horse dataset: While original CBA gets a 19.03% 10-fold cross validation error rate with an average of 87.9 rules, the adapted CBA[1] and CBA[2] both get only one rule and suffer from a 14.12% error rate. More concretely, all ten-fold classification loops implemented by these two adapted algorithms result in the same rule: if surgery=2 & outcome=1 then surgical lesion=2 (default class=1). This simple decision rule achieved excellent performance.

Although it is difficult to compare two classifiers based on datasets from different domains, wilcoxon signed-rank test is applied to give a rough statistical comparison.

As shown in Table 2, significant improvement is achieved by adapted CBA[2] at a 5% confidence interval. Although the performance of adapted CBA[1] could not generate the same good result as CBA[2], it performs best in several cases and requires no parameter selection. We, therefore, conclude that intensity of implication and dilated χ^2 are both appropriate measures for associative classification. Benchmarking tests of classification algorithms on UCI datasets are sometimes criticised for their incapability of representing the often more complex relationships which are present in larger real-world datasets. For this reason, an additional benchmarking test (in the context of credit scoring) is discussed in the next section.

Table 2.

Performance comparison

p-values for one tail test	Original CBA	C4.5	Naïve Bayes
Adapted CBA[1]	0.1652	0.0844	0.1057
Adapted CBA[2]	0.0107	0.0035	0.0125

4.3. Tests on a credit scoring dataset

Credit scoring is a qualified assessment and formal evaluation procedure of a particular company's credit history. In addition to this, it offers a capability of repaying obligations by credit bureaus. It measures the default probability of the borrower, and its ability to repay fully and timely its financial debt obligations (Guo, 2003). However, it is not possible to score all companies (or individuals) due to its enormous cost. It cannot be performed frequently either. As a result, statistical models and artificial intelligent technologies have been applied to help banks to identify these high-risk companies effectively and efficiently during past years. As mentioned above, the adapted CBA algorithms revealed their potential in generating accurate and compact decision lists on credit scoring datasets. We carried out them on an additional dataset, which is adopted from one major financial institution in the Benelux (Belgium, The Netherlands and Luxembourg) and contains 7190 credit scoring samples. 2/3 of its samples were taken as a training set and 1/3 of them were presented as a test set for the learning algorithms. Given other numerous empirical findings that were achieved in this domain (Altman, 1968, Kim, 1993, Moody and Utans, 1995 and Pinches and Mingo, 1973) by means of applied traditional statistical methods (such as discriminant analysis), neural networks and inductive learning algorithms (such as decision trees), the adapted CBA algorithms were also benchmarked against these learning algorithms on this dataset. Back-propagation algorithm and three-layer architecture were employed for neural network. The number of neurons in hidden layer was selected in order to achieve lowest error rate. The experiment results are summarized in Table 3.

Table 3.

Experiment results on credit scoring dataset

	Logit	Adapted CBA[2]	LDA	Original CBA	NN	Adapted CBA[1]	C45	QDA
Error rate	25.04%	26.49%	26.90%	27.08%	27.37%	27.70%	29.79%	62.09%
No. of rules	–	51	–	393	–	186	–	–

McNemar test (Dietterich, 1998) is applied to examine whether the predictive performance of these algorithms are significantly different. The McNemar test results (*p* values) are listed in Table 4.

Table 4.

McNemar test on credit scoring dataset

	Adapted CBA^2	**LDA**	**Original CBA**	**NN**	**Adapted CBA^1**	**C45**	**QDA**
Adapted CBA^2	1	0.7184	0.5541	0.3771	0.2207	0.0009**	0.000**
LDA	–	1	0.8415	0.5485	0.3375	0.0034**	0.000**
Original CBA	–	–	1	0.7518	0.3929	0.0051**	0.000**
NN	–	–	–	1	0.7184	0.0193*	0.000**
Adapted CBA^1	–	–	–	–	1	0.0383*	0.000**
C45	–	–	–	–	–	1	0.000**
QDA	–	–	–	–	–	–	1

*Significant at 5%, **significant at 1%.

For traditional statistical methods LDA (linear discriminant analysis) and QDA (quadratic discriminant analysis), specific structures are imposed and parameters are estimated in order to fit the training dataset. Although QDA shows terrible prediction ability, LDA performs well on our dataset. These rest algorithms are free from structural assumptions that underlie statistical methods, and can extract knowledge from data automatically. The *p*-values in Table 4 reveals that there are no significant differences among original CBA, adapted CBA and Neural Networks at a 5% confidence level, while they are all significantly better than C4.5 decision tree. Taking the interpretability of the classification model into account, our two adapted CBA algorithm seem to be appropriate choices for credit scoring because they generated much more compact decision lists (less number of rules) than original CBA. A deeper insight into the structure of the rules shows that original CBA and adapted CBA^1 both focus on choosing classification rules that predict good clients (with bad clients as the default class). But according to the use of intensity of implication, numerous rules that have high confidence and low support have a lower rank than in original CBA. These rules are finally discarded since they are not fired by any training samples, which are matched by these rules with higher intensity of implications, thus making the decision lists generated by adapted CBA^1 more compact. Taking dilated chi-square as the primary criteria for ranking, adapted CBA^2 pays much more attention to those classification rules for bad clients (with good clients as the default class) and as a result creates significantly more compact rule sets as well. In addition, decision makers in financial institutions probably prefer rules that predict bad clients, which will be extraordinary costly if they are regarded as good ones.

5. Conclusion

In this paper, intensity of implication is adopted as an interestingness measure for class association rules. Another novel interestingness measure, called dilated chi-square is designed to reveal the statistical interdependence between the antecedents and consequents of association rules.

In a next stage the CBA algorithm was adapted, by coupling it with intensity of implication and dilated chi-square, respectively. More concretely, intensity of implication (or dilated chi-square) was adopted as the primary criterion to rank class association rules at the first step of

the database coverage pruning procedure in the original CBA algorithm. Benchmarking experiments on wide-range datasets, especially in credit scoring domains, proved that these two adapted algorithms could build accurate decision lists and generate classifiers that are significantly more compact than CBA.

References

Agrawal and Srikant, 1994 R. Agrawal and R. Srikant, Fast algorithm for mining association rules, *Proceedings of 20th international conference on very large data bases, Santiago, Chile* (1994) (pp. 487–499).

Agrawal et al., 1993 R. Agrawal, T. Imielinski and A. Swami, Mining association rules between sets of items in large databases., *Proceedings of the 93' ACM SIGMOID conference on management of data, Washington, DC* (1993) (pp. 207–216).

Altman, 1968 E. Altman, Financial ratios, discriminant analysis and the prediction of corporate bancruptcy, *The Journal of Finance* **23** (1968), pp. 589–609.

Blake and Merz, 1998 C.L. Blake and C.J. Merz, *UCI repository of machine learning databases*, University of California, Department of Information and Computor Science, Irvine, CA. (1998)http://www.ics.uci.edu/~mlearn/mlrepository.htm.

Dietterich, 1998 T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, *Neural Computation* **10** (1998) (7), pp. 1895–1924.

Dong et al., 1999 G. Dong, X. zhang, L. Wong and J. Li, CAEP:Classification by aggregating emerging patterns., *Proceedings of the second international conference on discovery science.*, *Lecture notes in artificial intelligence 1999* **Vol. 1721**, Springer, Tokyo, Japan (1999).

Fayyad and Irani, 1993 U.M. Fayyad and K.B. Irani, Multi-interval discretization of continuous valued attributes for classification learning., *Proceedings of the 13th international joint conference on artificial intelligence, Chambery, France* (1993).

Gras and Lahrer, 1993 G. Gras and A. Lahrer, L'implication statistique: une nouvelle methode d'analysis de donnees., *Mathematiques, Informatique et Sciences Humaines No. 20* (1993).

Guillaume et al., 1998 S. Guillaume, F. Guillet and J. Philippe, Improving the discovery of association rules with intensity of implication, *Principles of data mining and knowledge discovery.*, *Lecture notes in artificial intelligence* **Vol. 1510** (1998) (pp. 318–327).

Guo, 2003 M.H. Guo, *Credit rating*, China Renmin University Press, Beijing (2003).

Janssens et al., 2005 Janssens, D., Wets, G., Brijs, T., & Vanhoof, K., (2005). Adapting the CBA-algorithm by means of intensity of implication. *Expert systems with Application, 28,* 105–117.

Kim, 1993 J.W. Kim, Expert systems for bond rating: A comparative analysis of statistical, rule-based and neural network systems, *Expert Systems* **10** (1993), pp. 167–171. Abstract-Compendex | Order Document

Liu et al., 1998 B. Liu, W. Hsu and Y. Ma, Integrating classification and association rule mining., *Proceedings of the fourth international conference on discovery and data mining, New York, US* (1998) (pp. 80–86).

Liu et al., 2001 W. Liu, J. Han and J. Pei, CMAR: Accurate and efficient classification based on multiple class-association rules., *Proceedings of ICDM'01, San Jose, CA* (2001) (pp. 369–376).

Mills, 1955 F. Mills, *Statistical methods*, Pitman, London (1955).

Moody and Utans, 1995 J. Moody and J. Utans, Architecture selection strategies for neural networks application to corporate bond ratings. In: A. Refenes, Editor, *Neural networks in the capital markets*, Wiley, Chichester (1995), pp. 277–300.

Pinches and Mingo, 1973 G.E. Pinches and K.A. Mingo, A multivariate analysis of industrial bond ratings, *Journal of Finance* **30** (1973) (1), pp. 1–18. Abstract-EconLit | Order Document

Quinlan, 1993 J.R. Quinlan, *C4.5 programs for machine learning*, Morgan Kaufmann, San Mateo, CA (1993).

Suzuki and Kodratoff, 1998 E. Suzuki and Y. Kodratoff, Discovery of surprising exception rules based on intensity of implication., *Proceedings of PKDD'98*, Springer, Berlin (1998) (pp. 10–18).

Wang and Zhou, 2000 K. Wang and S. Zhou, Growing decision trees on support-less association rules., *Proceedings of KDD'00, Boston, MA* (2000).

Witten and Frank, 2000 I.H. Witten and E. Frank, *Data mining: Practical machine learning tools and techniques with Java implementations*, Morgan Kaufmann, San Francisco, CA (2000).

Yin and Han, 2003 X. Yin and J. Han, CPAR: Classification based on predictive association rules., *Proceedings of 2003 SIAM international conference on data mining, San Fransisco, CA* (2003) (pp. 331–335).

Yu et al., 2004 L. Yu, G. Chen, D. Janssens and G. Wets, Dilated Chi-square: A novel interestingness measure to build accurate and compact decision list., *Proceedings of the international conference on intelligent information processing, Beijing, China* (2004) (pp. 233–237).

Corresponding author. Tel.: +32 11268960; fax: +32 11232217.