DILATED CHI-SQUARE: A NOVEL INTERESTINGNESS MEASURE TO BUILD ACCURATE AND COMPACT DECISION LIST

Yu Lan*, Guoqing Chen*, Davy Janssens** and Geert Wets**
* *School of Economics and Management, Tsinghua University, Beijing 100084, China*
Email: {yul1; chengq} @em.tsinghua.edu.cn
***Limburg University Centre, Universitaire Campus, gebouw D,B-3590 Diepenbeek, Beigium*
Email: {davy.janssens; geert.wets} @luc.ac.be

Abstract: Associative classification has aroused significant attention in recent years. This paper proposed a novel interestingness measure, named dilated chi-square, to statistically reveal the interdependence between the antecedents and the consequent of classification rules. Using dilated chi-square, instead of confidence, as the primary ranking criterion for rules under the framework of popular CBA algorithm, the adapted algorithm presented in this paper can empirically generate more accurate and much more compact decision lists.

Key words: dilated chi-square, associative classification, CBA

1. INTRODUCTION

In recent years, extensive research has been carried out to integrate classification and association rules [1-5]. By focusing on a limited subset of association rules, i.e. those rules where the consequent of the rule is restricted to the class attribute, it is possible to build more accurate classifiers.

Associative classification is first proposed in CBA system [1], which uses a slightly adapted version of the well known Apriori algorithm [6] in order to extract meaningful association rules with their consequents limited to class labels. These rules are then primarily sorted by descending confidence and pruned in a way to get a minimal number of rules that are necessary to cover

the training data and to achieve satisfactory accuracy. The aim of this paper is to improve CBA algorithm and generate a more accurate and compact decision list. Instead of confidence, a novel interestingness measure called dilated chi-square is applied as the primary sorting criterion.

The remaining of the paper is arranged as follows. Section 2 elaborates on the weakness of confidence. Section 3 describes in detail the design of dilated chi-square to overcome it. The empirical research is presented in section 4. Section 5 gives our concluding remarks.

2. LIMITS OF CONFIDENCE

The rules in CBA are sorted primarily by descending confidence, which 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 [7].

First, the confidence of a rule X => Y is invariable when the size of s(Y) or D varies. s(Y) is the subset of the samples which are covered by the consequent of the rule, while D is the total samples in the dataset. The confidence of rule X=>Y is Supp(X ∪ Y)/Supp(X). Keeping the numerator and denominator fixed, the confidence is stable when the size of s(Y) or D changes. Nevertheless, the rule X =>Y is more likely to happen when the size of s(Y) increases or when the size of D decreases.

Second, the minimal support is always set to 1% or even lower in practice. It might very well happen that some rules have a high confidence but on the other hand they might be confirmed by a very limited number of instances, and that those rules stem from noise only.

Therefore, a novel interestingness, i.e. dilated chi-square was designed to overcome these two drawbacks. The next section elaborates on this.

3. DILATED CHI-SQUARE

Traditional Chi-square test statistics (χ^2) is a widely used method for testing independence or correlation. This statistic is based on the comparison between the observed and the corresponding expected frequencies.

For each rule X=>Y generated from a training dataset D, a 2*2 contingency table can be derived as Figure 1:

	Y	¬Y	Row Total
Satisfy X	m_{11}	m_{12}	Support count of X
Not Satisfy X	m_{21}	m_{22}	|D|-Support count of X
Column Total:	Support count of Y	|D|-Support count of Y	|D|

Figure 1. A 2*2 contingency table for rule X=>Y and dataset D

The chi-square value for rule X=>Y can be calculated as

$$\chi^2 = \frac{(m_{11}m_{22} - m_{12}m_{21})^2|D|}{(m_{11}+m_{12})(m_{21}+m_{22})(m_{11}+m_{21})(m_{12}+m_{22})} \quad (1)$$

However, simply using the traditional chi-square value will be favorable in a situation where the distribution of the row total is close to that of the column total distribution. We then adjust it according to local and global maximum chi-square that we define.

*Definition 1:*Given a dataset D, the local maximum chi-square, denoted as $lmax(\chi^2)$, is the maximum chi-square value for a fixed support count of X.

*Definition 2:*Given a dataset D, the global maximum chi-square, denoted as $gmax(\chi^2)$, is the maximum chi-square value for any possible support count of X.

Property 1: $lmax(\chi^2)=(n_1n_2)^2|D|/[(m_{11}+m_{12})(m_{21}+m_{22})(m_{11}+m_{21})(m_{12}+m_{22})]$, where $n_1 = \min(\min(m_{11}+m_{12},m_{21}+m_{22}),\min(m_{11}+m_{21},m_{12}+m_{22}))$ and $n_2 = \min(\max(m_{11}+m_{12},m_{21}+m_{22}),\max(m_{11}+m_{21},m_{12}+m_{22}))$. The local maximum chi-square value is arrived at the largest deviation from the expected frequency, assumed that the support count of X is given.

Property 2: $gmax(\chi^2)=|D|$. The equation is arrived when $m_{21}+m_{22} = m_{12}+m_{22}$ and $m_{11}+m_{12} = m_{11}+m_{22}$, i.e. the distribution of row total equals that of column total.

Chi-square 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 chi-square value, denoted as $dia(\chi^2)$. More concretely, we heuristically use formula 3 to dilate the chi-square value according to the relationship between the local and global maximum chi-square values for current rule and database. The dilation procedure is nonlinear and empirically achieved excellent results, as shown in the next section.

$$\frac{dia(\chi^2)}{\chi^2} = \left(\frac{gmax(\chi^2)}{lmax(\chi^2)}\right)^{\alpha} = \left(\frac{|D|}{lmax(\chi^2)}\right)^{\alpha}, where \quad 0 \le \alpha \le 1 \quad (2)$$

Therefore

$$dia(\chi^2) = \left(\frac{|D|}{lmax(\chi^2)}\right)^{\alpha} \chi^2 \quad (3)$$

The parameter α is used to control the impact of global and local maximum chi-square values and can be tuned for different classification problems. It is visible that the dilated chi-square value is sensitive when the size of s(Y) or D varies. Furthermore, for these rules with high confidence and very low support, dilate chi-square values estimate their interestingness in a more cautious way.

We now adapted CBA by taking dilated chi-square as the primary criteria to sort the class association rules. Rule r_i has a higher rank than *rule* r_j if it has a larger value of dilated chi-square. When two rules have the same values of dilated chi-square, they are ranked according to the ranking mechanism of the original CBA.

4. EMPIRICAL SECTION

This part is to validate our adapted CBA algorithm on 16 binary classification datasets from UCI [8]. The average results of 10-fold cross validation are described in table 1:

Table 1. Results on UCI datasets

Datasets	Adapted CBA(α=best)		Original CBA		C4.5 [9]	NB
	error rate	no. of rules	error rate	no. of rules	error rate	error rate
austra	13.04%	12.4	14.35%	130.5	13.48%	18.70%
breast	3.58%	28.3	3.86%	42.2	4.43%	2.58%
cleve	16.13%	9.6	17.16%	63.8	20.79%	16.17%
crx	13.04%	12.4	14.93%	138.2	12.75%	18.99%
diabetes	21.74%	10.7	22.26%	38.5	22.92%	24.22%
german	26.80%	19.7	26.70%	134	27.60%	25.30%
heart	16.67%	7.4	17.78%	37.6	18.89%	14.81%
hepati	16.83%	11.3	16.21%	25.2	16.77%	15.48%
horse	14.12%	1	19.03%	87.9	15.22%	20.92%
hypo	0.85%	10.9	1.64%	30	0.85%	1.90%
iono	6.55%	18.5	8.25%	44.8	9.69%	8.26%
labor	8.33%	4.4	10.00%	12.5	15.79%	8.77%
pima	22.00%	10.7	23.43%	38.3	22.66%	25%
sick	3.25%	1	2.64%	47.4	2.07%	4.32%
sonar	18.74%	21.8	22.60%	41	18.75%	25.48%
ti-tac	3.34%	9	0.00%	8	14.20%	29.65%
average	12.81%	11.82	13.80%	49.34	14.80%	16.28%

As shown in Table 1, adapted CBA has a lowest average error rate on these benchmarking datasets if the best parameter α is selected. The average number of rules that adapted CBA generated on these datasets is nearly one fourth of the original CBA! We also run the adapted CBA with α set at 0.8 for all datasets. The average error rate and number of rules are respectively 14.02% and 12.7.

Wilcoxon signed-rank test was used to give statistical comparisons between adapted CBA (for best α) and each of other classifiers considered in this paper. The results are depicted in Table 2.

Table 2. P-values of the Adapted CBA algorithm versus other classifiers

p-values for one tail test	Original CBA	C4.5	Naïve Bayes
Adapted CBA (α =best)	0.0107	0.0035	0.0125

5. CONCLUSION

A novel interestingness measure name dilated chi-square is proposed in this paper. We adapt CBA algorithm, which can be used to build classifiers based on class association rules, by coupling it with dilated chi-square. More concretely, dilated chi-square is adopted as the primary criteria to rank the class association rules at the first step of the database coverage pruning procedure in the CBA algorithm. Experiments on wide-range datasets proved that this adapted CBA, compared with original CBA, C4.5 decision tree and Naive Bayes, achieves significantly better performance and generates classifiers much more compact than CBA.

ACKNOWLEDGEMENT

The work was partly supported by the National Natural Science Foundation of China (79925001/70231010), the MOE Funds for Doctoral Programs (20020003095), and the Bilateral Scientific and Technological Cooperation Between China and Flanders/Czech.

REFERENCES

1. B.Liu, W.Hsu, and Y.Ma. Integrating Classification and Association Rule Mining. in the 4th International Conference on Discovery and Data Mining. 1998. New York,U.S.: pp. 80-86.
2. G.Dong, et al. CAEP:Classification by aggregating emerging patterns. in 2nd International Conference on Discovery Science,(DS'99),volume 1721 of Lecture Notes in Artificial Intelligence. 1999. Tokyo,Japan: Springer-Verlag: pp. 30-42.
3. W.Liu, J.Han, and J.Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. in ICDM'01. 2001. San Jose, CA: pp. 369-376.
4. X.Yin and J.Han. CPAR:Classification based on predictive association rules. in 2003 SIAM International Conference on Data Mining (SDM'03). 2003. San Fransisco,CA: pp. 331-335.
5. K.Wang and S.Zhou. Growing decision trees on support-less association rules. in KDD'00. 2000. Boston,MA: pp. 265-269.
6. R.Agrawal and R.Srikant. Fast algorithm for mining association rules. in the 20th International Conference on Very Large Data Bases. 1994. Santiago,Chile: pp. 487-499.
7. Janssens, D., et al. Adapting the CBA-algorithm by means of intensity of implication. in the First International Conference on Fuzzy Information Processing Theories and Applications. 2003. Beijing, China: pp. 397-403.
8. C.L.Blake and C.J.Merz, UCI repository of machine learning databases.1998, Irvine,CA:University of California, Dept. of Information and Computer Science. http://www.ics.uci.edu/~mlearn/mlrepository.htm.
9. J.R.Quinlan, C4.5 programs for machine learning. 1993: Morgan Kaufmann.