行政院國家科學委員會專題研究計畫　成果報告

變數精準粗略集之理論與應用(I)

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-034-003-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：中國文化大學應用數學系

計畫主持人：許志華

計畫參與人員：陳隆昇、蕭宇翔

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 10 月 11 日

行政院國家科學委員會補助專題研究計畫 ■ 成 果 報 告 □期中進度報告

變數精準粗略集之理論與應用(I)

計畫類別：■ 個別型計畫　　□ 整合型計畫
計畫編號：NSC 94－2213－E－034－003－
執行期間：94 年 8 月 1 日至 95 年 7 月 31 日

計畫主持人：許志華　中國文化大學應用數學系
共同主持人：
計畫參與人員：陳隆昇　國立交通大學工業工程與管理學系
蕭宇翔　國立清華大學工業工程與工程管理學系

成果報告類型(依經費核定清單規定繳交)：■精簡報告　□完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：中國文化大學應用數學系

中　華　民　國　95　年　10　月　8　日

一、中文摘要

變數精準粗略集理論是資料探勘的重要工具之一，已廣泛應用於不同領域的知識獲取。然而，變數精準粗略集卻無法應用於資料含有連續型屬性的問題，它需要一個將屬性資料離散化的方法來進行資料的前置處理。就資料探勘而言，資料離散化是處理連續型資料非常有效的方法，尤其是處理分類問題。此外，變數精準粗略集理論缺乏一個適當的方法來決定精準參數值以確定其最簡化屬性集合。然而，目前相關的文獻研究卻很少見。

本計劃第一年，我們將首先提出一利用 Chi2 演算法為基礎，所發展的資料離散化演算法(extended Chi2 algorithm)。利用五個例子對所提的演算法和三個資料離散化演算法(Chi2 演算法, 修正的 Chi2 演算法和布林推論演算法)進行比較，最後並給予扼要討論。

關鍵詞：變數精準粗略集理論、資料探勘、資料離散化、最簡化屬性集合

Abstract

The Variable Precision Rough Sets (VPRS) model is a powerful tool for data mining, as it has been widely applied to acquire knowledge. Despite its diverse applications in many domains, the VPRS model unfortunately cannot be applied to real world classification tasks involving continuous attributes. This requires a discretization method to pre-process the data. Discretization is an effective technique to deal with continuous attributes for data mining, especially for the classification problem. The modified Chi2 algorithm is one of the modifications to the Chi2 algorithm, replacing the inconsistency check in the Chi2 algorithm by using the quality of approximation, coined from the Rough Sets Theory (RST), in which it takes into account the effect of degrees of freedom. However, the classification with a controlled degree of uncertainty, or a misclassification error, is outside the realm of RST. This algorithm also ignores the effect of variance in the two merged intervals. In this study we propose a new algorithm, named the extended Chi2 algorithm, to overcome these two drawbacks. By running the software of See5, our proposed algorithm possesses a better performance than the original and modified Chi2 algorithms.

Keywords: VPRS model, data mining, discretization, β-reducts

二、緣由與目的

A number of methods based on the entropy measure establish a strong group of works in the discretization domain. This concept uses class entropy as a criterion to evaluate a list of best cuts, which together with the attribute domain induce the desired intervals (Nguyen, 1998).

The ChiMerge algorithm introduced by Kerber (1992) is a supervised global discertization method. The user has to provide several parameters such as the significance level α, and the maximal intervals and minimal intervals during the application of this algorithm. ChiMerge requires α to be specified. Nevertheless, too big or too small a α will over-discretize or under-discretize an attribute. Liu, et al. (1997) proposed a Chi2 algorithm that uses a ChiMerge algorithm as a basis, whereby the Chi2 algorithm improves the ChiMerge algorithm in that the value of α is calculated based on the training data itself.

Tay, et al. (2002) indicated that although the Chi2 algorithm automates the ChiMerge algorithm by calculating a significance value α based on the training data set, it still has two drawbacks: (1) the Chi2 algorithm requires the user to provide an inconsistency rate to stop the merging procedure. This is unreasonable since an inappropriate threshold will result in over-merging. (2) This merging criterion does not consider the degrees of freedom, but rather only the fixed degrees of freedom (the classes' number minus one). According to the statistical point of view, this is inaccurate (Montgomery, et al. 1999), since the power of a statistical test is affected by the degrees of freedom of a test. They utilize the quality of approximation to replace the inconsistency checking of the Chi2 algorithm and consider the degrees of freedom of each two adjacent intervals, in which the two adjacent intervals when it has a maximal difference in the calculated χ^2 value and the threshold should be merged first.

The modified Chi2 algorithm introduced by Shen et al. (2001) can be sectioned into two phases: The first phase of the modified Chi2 algorithm can be regarded as a generalization version of the ChiMerge algorithm. Instead of specifying a χ^2 threshold, the modified Chi2 algorithm provides a wrapping that automatically increments the χ^2 threshold (decreasing the significant level α). A consistency check is used as a stopping criterion to make sure that the modified Chi2 algorithm automatically determines a proper χ^2 threshold while still keeping the fidelity of the original data.

The second phase is a finer process of the first phase, beginning with the significant level α_0 determined in the first phase, where each attribute i is associated with a sigLvl[i] and they take turns for merging. A consistency check is conducted after each attribute's merging. If the inconsistency rate does not exceed the pre-defined inconsistency rate (δ), then sigLvl[i] is decreased for attribute i's next round of merging. Otherwise, the attribute i will not be involved in further merging. This process is repeated until no attribute's value can be merged.

In the modified Chi2 algorithm, inconsistency checking (InConCheck (data) $<\delta$) of the original Chi2 algorithm is replaced by the quality of approximation L_c after each step of discretization ($L_{c-discretized} \le L_{c-original}$). This inconsistency rate is utilized as the termination criterion. The quality of approximation coined from the Rough Sets Theory is defined as follows:

$$L_c = \frac{\sum card(\underline{B}X_i)}{card(U)}, \qquad (2.1)$$

where U is the set of all objects of the data set:

X can be any subset of U;

$\underline{B}X$ is the lower approximation of X in B ($B \subseteq A$);

A is the set of attributes.

The card denotes set cardinality.

The merge criterion of the original Chi2 algorithm does not consider the degrees of freedom, as it only used the fixed degrees of freedom (the classes' number minus one). The original Chi2 algorithm merges the pair of adjacent intervals with the lowest x^2 value being the critical value. The merge criterion of modified Chi2 considers the degrees of freedom of each of the two adjacent intervals. When two adjacent intervals have a maximal difference in the calculated χ^2 value, the threshold should be merged first.

三、結果與討論

The extended Chi2 algorithm

Step 1. Initialize:

Set the significant level as α=0.5;

calculate the pre-defined

inconsistency rate ξ.

Step 2. Calculate the chi-square value:
For each numeric attribute, sort data on the attribute and use formula (3.2) to compute the x^2 value.

Step 3. Merge:
For a comparison, compute the x^2 value and corresponding threshold; merge the adjacent two intervals which have the maximal normalized difference and the computed x^2 value is smaller than the corresponding threshold. If no two adjacent intervals satisfy this condition, then go to Step 5.

Step 4. Check inconsistency rate for merger:
Check the merged inconsistency rate, and if the merged inconsistency rate exceeds the pre-defined inconsistency rate, then discard the merger. Go to step 5. Otherwise, go to step 2.

Step 5. Decrease the significance level:
Decrease $\alpha \rightarrow \alpha_0$.

Step 6. Calculate finer the chi-square value:
For each numeric attribute, sort data on the attribute and use formula (3.2) to compute the x^2 value.

Step 7. Finer merge:
For a comparison, compute the x^2 value and corresponding threshold; merge the adjacent two intervals which have a maximal normalize difference and the computed x^2 value is smaller than the corresponding threshold. If no two adjacent intervals satisfy this condition, then go to Step 9.

Step 8. Check the inconsistency rate much finer for a merger:
Check the merged inconsistency rate; if the merged inconsistency rate exceeds the pre-defined inconsistency rate, then discard the merger. Go to step 9. Otherwise, go to step 6.

Step 9. Decrease finer the significance level:
Decrease the significance level; then stop.

The formula for computing the χ^2 value is:

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})^2}{E_{ij}}, \qquad (3.2)$$

where $n = 2$;

$k =$ number of classes;

$A_{ij} =$ number of objects in the ith interval, jth class;

$R_i =$ number of objects in the ith interval $= \sum_{j=1}^{k} A_{ij}$;

$C_j =$ number of objects in the jth class $= \sum_{i=1}^{n} A_{ij}$;

$N =$ total number of objects $= \sum_{i=1}^{n} R_i$;

$E_{ij} =$ expected frequency of $A_{ij} = \frac{R_i * C_j}{N}$.

If either R_i or C_j is 0, then E_{ij} is set to 0.1. The degrees of freedom of the χ^2 statistic are one less than the number of classes

Five data sets are demonstrated to present the effectiveness of the proposed extended Chi2 algorithm. The five data sets are taken from the University of California, Irvine's repository of machine learning databases.

We ran See5 on both the original data sets and the discretized data sets. The parameters of See5 utilize its default setting. The ten-fold cross-validation test method is applied to all data sets. The data set is divided into 10 parts of which nine parts are used as training sets and the remaining one part as the testing set. The experiments were repeated 10 times. The final predictive accuracy is taken as the average of the 10 predictive accuracy values.

T The extended Chi2 algorithm is compared with the original Chi2 and modified Chi2 and Boolean reasoning algorithm with the predefined inconsistency rate (δ) value equal to 0 in the experiment.

The discretized data sets are sent into See5. The predictive accuracy and its standard deviation of these methods are listed in Table 1. From Table 1, we know that the predictive accuracy of the extended Chi2 algorithm outperforms other discretization algorithms.

Table1. The Predictive Accuracy Using See5 With the Discretization Algorithm

Data Set	See5				
	Continuous	Original Chi2 Algorithm	Modified Chi2 Algorithm	Extended Chi2 Algorithm	Boolean Reasoning Algorithm
Bupa	67.5 ± 2.4%	65.2 ± 3.2%	67.5 ± 1.9%	68.4 ± 2.7%	68.1 ± 2.3%
Glass	68.6 ± 2.5%	93.1 ± 2.1%	93.4 ± 2.3%	93.5 ± 1.3%	71.9 ± 2.8%
Iris	94.0 ± 2.1%	94.0 ± 2.1%	93.3 ± 2.2%	94.0 ± 2.1%	96.0 ± 1.8%
Breast Cancer	94.9 ± 0.8%	95.5 ± 1.0%	96.0 ± 0.9%	96.5 ± 0.8%	95.2 ± 0.8%
Heart dissease	51.9 ± 1.4%	52.5 ± 2.3%	53.2 ± 2.7%	54.2 ± 1.7%	55.9 ± 2.6%

四、計劃成果自評

Many classification algorithms developed in the data mining community can only acquire knowledge on the nominal attributes' data sets. However, many real world classification tasks exist that involve continuous attributes, such that these algorithms cannot be applied unless the continuous attributes are discretized. The VPRS model is a powerful mathematical tool for data analysis and knowledge discovery from inconsistent and ambiguous data. It cannot be applied to extract rules from the continuous attributes unless they are first discretized.

In this study we propose an extended Chi2 algorithm that determines the pre-defined misclassification rate (δ) from the data itself. We also consider the effect of variance in the two adjacent intervals. With these modifications, the extended Chi2 algorithm not only handles misclassified or uncertain data, but also becomes a completely automated discretization method and its predictive accuracy is better than the original Chi2 algorithm.

Five real world data set experiments were conducted to demonstrate the feasibility of the proposed algorithm. The experimental results show that our proposed algorithm could acquire a higher predicted accuracy than the original and modified Chi2 algorithm. Furthermore, the tree size is significantly smaller than using the original data with See5.

For m attributes, the computational complexity of original Chi2 algorithm at phase 1 has $O(Kmn \log n)$, where n is the number of objects in the dataset, and K is the number of incremental steps. A similar complexity can be obtained for phase 2. Although our proposed algorithm adds one step (i.e., to select the merging intervals), it does not increase the computational complexity as compared to the original Chi2 algorithm. The computational complexities of the original Chi2 algorithm, modified Chi2 algorithm, and our proposed algorithm are the same.

The above research results have been accepted for publication in *IEEE Transactions on Knowledge and Data Engineering.*

五、參考文獻

1. An, A., Shan, N., Chan, C., Cercone, N. and Ziarko, W., 1996, "Discovering Rules for Water Demand Prediction: An Enhanced Rough-set Approach," *Engineering Applications in Artificial Intelligence*, Vol. 9, No.6, pp. 645-653
2. Beynon, M., 2001, "Reducts within the variable precision rough sets model: A further investigation," *European Journal of Operational Research*, Vol. 134, pp. 592-605.
3. Beynon, M., 2002, "The Identification of Low-Paying Workplaces: An Analysis Using the Variable Precision Rough Sets Model," *The Third International Conference on Rough Sets and Current Trend in Computing*, *Lecture Notes in Artificial Intelligence Series*, Springer-Verlag, pp. 530-537.
4. Chmielewski, R. and Grzymala-Busse, W., 1996, "Global Descretization of Continuous Attributes as Preprocessing for Machine Learning," *International Journal of Approximate Reasoning*, Vol. 15, No. 4, pp. 319-331.
5. Dougherty, J., Kohavi, R. and Sahami, M., 1995, "Supervised and Unsupervised Discretization of Continuous Features," *Machine Learning: Proceedings of the Twelfth International Conference*, San Francisco, pp. 194-202.
6. Kattan, M. W. and Cooper, R. B., 1998, "The Predictive Accuracy of Computer-based Classification Decision Techniques. A Review and Research Directions," *Omega-International Journal of Management Science*, Vol. 26, No. 4, pp. 467-482.
7. Kerber, R., 1992, "ChiMerge: Discretization of Numeric Attributes," Proceeding tenth International Artificial Intelligence, pp. 123-128.
8. Li, R. P. and Wang, Z. O., 2002, "An Entropy-Based Discretization Method for Classification Rules with Inconsistency Checking," *Proceedings of the First Conference on Machine Learning and Cybemetics*, Beijing, pp. 243-246.
9. Liu, H. and Setiono, R., 1997, "Feature Selection via Discretization," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 9, No. 4, pp. 642-645.
10. Montgomery, D. C. and Runger, G. C., 1999, *Applied Statistics and Probability for Engineers*, Jone Wiley & Sons.
11. Nguyen, H. S. and Nguyen, S. H., 1998, "Discretization Methods in Data Mining," *Rough Sets in Knowledge Discovery*, Physica-Verlag, Heidelberg, pp. 451-482.
12. Shen, L. and Tay, E. H., 2001, "A discretization method for Rough Sets Theory," Intelligent Data Analysis, Vol. 5, pp. 431-438.
13. Tay, E. H. and Shen, L., 2002, "A Modified Chi2 Algorithm for Discretization," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 14, No. 3, pp. 666-670.
14. Ziarko, W., 1993, "Variable Precision Rough Set Model," *Journal of Computer and System Science*, Vol.46, pp. 39-59.