A Modified Chi2 Algorithm Based on the Significance of Attribute

Hao Zhang, Duoqian Miao, Ruizhi Wang
Department of Computer Science and Technology, Tongji University
Shanghai, 201804, P.R.China
vivianzhang_hao@hotmail.com, Miaoduoqian@163.com, wrz977@sohu.com

Abstract

Discretization is one of the important components of the data preprocessing. Discretization can turn numeric attributes into discrete ones. There are many different kinds of discretization methods. This paper describes the Chi2 algorithm which is a simple and general discretization algorithm. In this algorithm, the χ^2 statistic value is used as an evaluative standard to discretize the numeric attributes. However, the Chi2 algorithm dose not consider the sequence of discretization for each attribute in the second phase. And the inconsistency rate cannot fully reflect the characteristic of dataset. These drawbacks will affect the result of discretization finally. In this paper, some concepts of the rough set are introduced to improve the Chi2 algorithm.

1. Introduction

Discretization is an effective method in dealing with continuous attributes for rule generating. Through discretization, the data pattern can meet the need of the classification algorithm; in the meantime, the quality of the knowledge we gain can be improved. This demands the studies on the discretization methods. There are three different axes by which discretization methods can be classified: local vs. global, supervised vs. unsupervised and static vs. dynamic. Local discretization method uses the localized regions of the instance space to discretize, while the global method [1] uses the entire instance space to discretize. Unsupervised discretization method doesn't utilize instance class labels in the discretization process, such as equal width interval [2] and equal frequency interval [2] methods. By contrast, those methods utilizing the class labels are called supervised method [2]. Many discretization methods require a parameter p, which indicates the maximum number of intervals to produce in the discretization process for an attribute. Static methods perform the discretization on each attribute and determine the value of p for each attribute independent of other attributes. Conversely, dynamic methods search through the space of possible p values for all attributes simultaneously, thereby capturing interdependencies in attribute discretization.

The ChiMerge algorithm [3] proposed by Kerber is a supervised global discretization method. The ChiMerge algorithm consists of an initialization step and a bottom-up merging process. The merging process is repeated until a stopping criterion is met. It uses the χ^2 test to determine whether adjacent intervals should be merged.

However, the ChiMerge algorithm requires significance level α to be specified. Nevertheless, too big or too small, the α will underdiscretize or overdiscretize an attribute. Then, Liu and Setiono proposed the Chi2 algorithm [4][5] using the ChiMerge algorithm as a basis. The Chi2 algorithm improved the ChiMerge algorithm in calculating the significance level α based on the training data itself. Then the value of α differed from attribute to attribute, so the merging process would only continue on those attributes that needed it.

The Chi2 algorithm has some drawbacks; in order to deal with them, Tay and Shen proposed the modified Chi2 algorithm [6]. They use the quality of approximation to replace the inconsistency rate, and the degree of freedom of each two adjacent intervals is under consideration. Then the two adjacent intervals having a maximal difference in the calculated χ^2 value and the threshold should be merged first. Although the modified Chi2 algorithm considers the effect of the degrees of freedom, it ignores the effect of variance in the two merging intervals. Chao-Ton Su and Jyh-Hwa Hsu proposed the extended Chi2 algorithm [7] to deal with the problem. In their algorithm, they presented a method to determine the predefined inconsistency rate based on the least upper bound of data misclassification error.

In this paper, we use the level of consistency to

Proceedings of the 2006 IEEE/WIC/ACM International Conference on
Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops)(WI-IATW'06)
0-7695-2749-3/06 $20.00 © 2006 IEEE

replace the inconsistency rate. In addition, the sequence of discretization for each attribute is taken into consideration. These two remedies can overcome the drawbacks of the Chi2 algorithm. The effectiveness of our proposed method is demonstrated by three data sets. Comparing the implementation results using See5, the algorithm after modification performs better than the original Chi2 algorithm.

2. Modification

2.1. Chi2 Algorithm

The Chi2 algorithm consists of two phases:

```
Phase 1：set siglevel =0.5;
        do while(Inconsistency(data)< δ )
        {  for each attribute
            {    Sort(attribute,data);
                 chi-sq-initialization(attribute,data);
                 do{ chi-sq-calculation(attribute,data)
                   }while(Merge(data))
            }
            siglevel0=siglevel;
            siglevel=decreSiglevel(siglevel);
        }
Phase 2：set all siglevel[i]=siglevel0 for attribute i;
        do until no-attribute-can-be-merged
        {  for each attribute i that can be merged
             {  Sort(attribute,data);
                chi-sq-initilization(attribute,data);
                do{
                   chi-sq-calculation(attribute,data)
                   }while(Merge(data))
                if(Inconsistency(data)< δ )
                siglevel[i]=decreSiglevel(siglevel[i]);
                else
                attribute i can not be merged;
             }
        }
```

The first phase can be regarded as a generalization of the ChiMerge algorithm. The goal of it is to determine a proper threshold while keeping the fidelity of the original data. The second phase is a finer process of the first phase, which uses separate significance levels for each attribute.

2.2. Problem analysis

In the original Chi2 algorithm, there are two drawbacks: (1) The Chi2 algorithm requires the user to provide the stopping criterion—inconsistency rate. However, this value is hard to confirm. It differs from dataset to dataset. Only through several experiments can we find a better value for a dataset. And the inconsistency rate cannot fully reflect the characteristic of dataset. (2) In the second phase of Chi2 algorithm, the attributes are discretized in random sequence. The sequence of discretization for each attribute directly decides whether the attribute could be further merged, which would affect the result of discretization finally. In order to solve these problems, some concepts of the rough set are introduced to improve the Chi2 algorithm.

2.3. Some concepts of rough set

In the theory of rough set, an information system consists of four parts, as:

$$S= (U, R, V, f)$$

Where:

U is a nonempty set of projects;

R is a nonempty set of attributes. We have $R = C \cup D, C \cap D = \varnothing$, where C is a nonempty set of condition attributes and D is a nonempty set of decision attributes;

V is the union of attributes domains;

f is an information function.

For every subset of attributes $P \subseteq R$, an indiscernibility relation ind(P) is defined as follows: $ind(P) = \{(x, y) \mid (x, y) \in U \times U, \forall r \in P, f(x,r) = f(y,r)\}$

Then $U / ind(P)$ denotes the set of all the equivalence class in relation ind(P).

The lower approximation of the set $X \subseteq U$ and $B \subseteq R$ is defined as:

$$B\_(X) = \cup\{Y_i \mid Y_i \in U / ind(B), Y_i \subseteq X\}.$$

The upper approximation of the set $X \subseteq U$ and $B \subseteq R$ is defined as:

$$B^{-}(X) = \cup\{Y_i \mid Y_i \in U / ind(B), Y_i \cap X \neq \varnothing\}.$$

In the theory of rough set, the significance of attribute is defined as follows: After one attribute is deleted, the decrement of the classification quality of the information system is this attribute's significance value. The larger the value, the more significant the attribute is.

Let C denote the condition attributes and d denote the decision attribute. Then the positive of the information system is defined as: $POS_C(d) = \bigcup_{X \in U/d} C\_(X)$. The classification quality, $r_C(d)$, is defined as $r_C(d) = card(POS_C(d)) / card(U)$. For every $c \in C$, the classification quality excluding attribute c is $r_{C\setminus\{c\}}(d) = card(POS_{C\setminus\{c\}}(d)) / card(U)$, so the significance of attribute c is defined as:

$$\sigma_{C,d}(c) = r_C(d) - r_{C\setminus\{c\}}(d) \qquad (1)$$

2.4. Modification of Chi2 Algorithm

Firstly, we use the level of consistency to replace the inconsistency rate. The level of consistency L, is defined

Proceedings of the 2006 IEEE/WIC/ACM International Conference on
Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops)(WI-IATW'06)
0-7695-2749-3/06 $20.00 © 2006 IEEE

as follows:

$$L = card(POS_C(d)) / card(U) \qquad (2)$$

In the Chi2 algorithm [4][5], inconsistency checking (InConCheck(data)< δ) is used. In the modified algorithm, after each step of discretization, we require the dataset to maintain the level of consistency ($L_{discretized} = L_{original}$). By using the level of consistency, the discretization process has been completely automatic. What is more important is that the fidelity of the dataset can be maintained to be the same after discretization.

Secondly, the significance of attribute is introduced into the modification of the Chi2 algorithm.

Before implementing the second phase, we compute the significance of each attribute according to the formula (1). Then we arrange the attributes in ascending order in accordance with the significance value. If some attributes have the same significance value, the number of the representative values of each attribute will be taken into consideration, by which these attributes will be arranged in descending order. Then the sequence of discretization for each attribute is confirmed in the second phase of the algorithm. Through this modification, attributes with less significance value can be merged as much as possible; meanwhile, those more significant ones can be avoided being further merged. Then the important information of the original dataset can be retained as much as possible after discretization.

These two remedies can overcome the drawbacks of the Chi2 algorithm. This modification not only makes the discretization process completely automatic, but also retains the valuable information of the dataset.

3. Experimental results

In order to compare the modified Chi2 algorithm with the original Chi2 algorithm, three data sets are chosen to be discretized. They are taken from the University of California, Irvine's repository of machine learning databases [8].

The three data sets used in the experiment are the Iris, the Bupa and the Breast Cancer. The three data sets are described below:

Iris: This data set contains 150 examples (50 examples of setosa, 50 examples of versicolor, and 50 examples of verginical). Each example is described by 4 attributes: sepal-length, sepal-width, petal-length and petal-width.

Bupa: This data set contains 345 examples (145 examples of normal, 200 examples of a liver malfunction). Each example is described by 6 attributes: MCV, ALKPHOS, SGPT, SGOT, GAMMAGT and DRINKS.

Breast Cancer: This data set contains 699 examples, where 16 examples have missing attributes values. Removing examples with missing attributes value, we use 683 examples (444 examples of benign, 239 examples of malignant). Each example is described by 9 attributes: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses.

Table 1 gives a summary of data sets used in this experiment.

Table 1. Data sets information

Name	Examples	Continuous attributes	Classes
Iris	150	4	3
Bupa	345	6	2
Breast Cancer	699	9	2

In this experiment, See5-demo is chosen as the benchmark for comparing the performance of the original Chi2 algorithm and the modified Chi2 algorithm. The parameters of See5 utilize its default setting. See5 works well for decision-making problems and it is a well-known method. That's why we choose it as the benchmark. The 10-fold cross-validation test method is applied to all the data sets. The data set is divided into ten parts, of which nine parts are used as the training sets and the remaining one part as the testing set. The experiments are repeated ten times. The final predictive accuracy is taken as the average of the ten predictive accuracy values. The final predictive accuracy and the tree size given by the See5 will be compared to analyze the performance of these two algorithms.

Because the modification on the Chi2 algorithm is made to maintain the fidelity of the dataset, then the modified algorithm is compared with the original Chi2 algorithm with the inconsistency rate value equal to 0 in the experiment.

From table 2, by comparing the predictive accuracy, we know that the modified algorithm outperforms the original one. In the modified algorithm, a higher predictive accuracy can be acquired. It also shows that the modified algorithm has no significant difference in tree size compared to the original one. Sometimes, the tree size is greater than using the original data set with See5. In a word, the modified algorithm performs better than the original algorithm.

4. Conclusion

In the data mining field, many classification algorithms can only acquire knowledge on the datasets with nominal attributes. However, in the real world, there are many datasets containing continuous attributes. In order to applying these classification algorithms, continuous attributes must be discretized.

The Chi2 algorithm is a simple and general discretization algorithm based on the χ^2 statistic value.

But it still has some drawbacks. In this paper, some methods are proposed to modify the Chi2 algorithm. The level of consistency is utilized to replace the inconsistency rate checking so that the fidelity of the dataset can be maintained to be the same after discretization. And the sequence of discretization for each attribute is taken into consideration in the modified Chi2 algorithm. The significance of attributes defined in the rough set theory is used to confirm the sequence. These modifications not only make the discretization process completely automatic, but also retain the valuable information of the dataset.

Through the experiment, the results show that the modified algorithm performs better in predictive accuracy than the original Chi2 algorithm. However, the modified algorithm has no significant difference in tree size compared to the original one, which remains one of the key problems for us to address in the future.

References

[1] M.Chmielewski and J.Grzrmala-Busse. Global Discretization of Continuous Attributes as Preprocessing for Machine Learning. Int'l J.Approximate Reasoning, vol.15, no.4, pp.319-331, Nov.1996

[2] J.Dougherty, R.Kohavi, and M.Sahami. Supervised and Unsupervised Discretization of Continuous Features. In: Proceedings of the 12th International Conference on Machine Learning, San Francisco, 1995, pp. 194-202

[3] R.Kerber. ChiMerge: Discretization of numeric attributes. In: Proceedings of the 10th National Conference on Artificial Intelligence, San Jose, AAAI Press/The MIT Press, July 1992, pp.123-128

[4] Huan Liu and Rudy Setiono. Chi2: Feature Selection and Discretization of Numeric Attributes. In: Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence, Washington D.C., IEEE CS Press, Nov. 1995, pp. 388 – 391

[5] Huan Liu and Rudy Stetiono. Feature Selection via Discretization. IEEE Transactions on Knowledge and Data Engineering, 1997, 9(4): 642-645

[6] Francis E.H. Tay. Member, IEEE. and Lixiang Shen. A Modified Chi2 Algorithm for Discretization. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(3): 666-670

[7] Chao-Ton Su and Jyh-Hwa Hsu. An Extended Chi2 Algorithm for Discretization of Real Value Attributes. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(3): 437-441

[8] C.J. Merz, P.M. Murphy. UCI Repository of machine learning database. http://www.ics.uci.edu/~mlearn/M-LRepository.html

Table 2. The predictive accuracy and the tree size comparison of the discretization algorithms

Data set	Continuous		Chi2 Algorithm (δ =0)		Chi2 Algorithm after modification	
	Tree size	Predictive accuracy	Tree size	Predictive accuracy	Tree size	Predictive accuracy
Iris	4	92%	3.3	94.66%	4	95.32%
Bupa	24.3	61.49%	25.4	64.92%	25	66.95%
Breast Cancer	9.5	94.6%	9	95.65%	9.5	95.75%