Discretization Based on Chi2 Algorithm and Visualize Technique for Association Rule Mining

Nuntawut Kaoungku*, Tippaya Thinsungnoen, Pongsakorn Durongdumrongchai,
Kittisak Kerdprasop, Nittaya Kerdprasop

School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, Thailand.

*Corresponding Author: b5111299@gmail.com

Abstract

This research aims at studying the discretization based on Chi2 algorithm and visualize technique for association rule mining. Numeric attributes with large distinct values normally do not appear in the association rules. We thus study the discretization method for numeric attributes with the integrated Chi2 algorithm and visualize technique to handle numeric attributes prior to the association analysis phase. We comparatively experiment with our proposed method against existing techniques. The comparative metrics are accuracy and number of rules.

Keywords: Discretization, Association Rule Mining, Chi2, Visualize.

1. Introduction

Currently, many organizations and merchants do not use paper to record data, but they use computers for recording. When the data are in the digital form, we can use computer to analyze these data in order to obtain the model for future use such as to predict patient's disease, to understand customer purchase behavior, or to assess learning behavior of student. The automatic induction of model from electronic data is known as data mining⁽¹⁾.

Association rule mining is one well-known technique in data mining. It is the induction of relationships of events or objects and generate these relationship as association rules to understand the current data or to predict the occurrence of an event or object in the future. There are many researches about the increase in efficiency in association rule mining, such as increase the speed or the accuracy of the association rule mining.

The data currently available are in a variety of types, such as numeric, text or character. But, the result of relationship induction from the numerical data in association rule mining was not good enough because the numerical data have a wide range of values. Thus, the solution of numerical data handling for association rule mining is discretization technique. There exist have many techniques for discretizing numerical data⁽²⁾, such as Chi2 algorithm and Extend-Chi2 algorithm⁽³⁾.

This research aims at proposing the efficient discretization technique based on Chi2 algorithm and visualized cut-point analysis for association rule mining to handle numerical data. The problem caused by the numerical data in association rule mining is that they make association rules disappear due to the sparseness of each numeric value. We, therefore, propose an efficient algorithm to solve this problem.

2. Related Work

Related researches can be divided into several types: research for proposing the new idea, research to improve the original algorithm, research to discretize numerical data for classification, and research to discretize numerical data for association rule mining. Thus, we study related work along these research themes with the details as follows:

Gyenesei⁽⁴⁾ has proposed an algorithm to discretize numerical data by using fuzzy sets technique to reduce runtime and to increase number of effective rules in the association rule mining. The experimental result has been compared between discretization by non-normal distributed data and normal distributed data. The result of the discretization by normal distributed data gives more effective association rules than non-normal distributed data,

which is measured by the minimum support, minimum confidence and runtime in association rule mining.

Tong et al.⁽⁵⁾ has proposed a method for association rule mining with numerical data using k-means clustering algorithm and Euclidean-based distance calculation. They used synthetic data to test the performance of the algorithm. Their algorithm results in less number of association rules, but higher in the values of support and confidence compared to association rule mining without any handling method for numerical data.

Ke et al.⁽⁶⁾ has proposed a method of the association rule mining with numerical data using mutual information and clique (MIC). This technique has divided the work into 3 parts. First, applying discretization to numerical data. Second, using the data obtained from the first step to create the MI graph. And finally, the result of the second step was used in the finding of frequent itemsets. The experimental result used 6 datasets and compared the runtime, number of rules and other measures.

Wei⁽⁷⁾ has proposed discretization technique to handle numerical data for association rule mining with clustering and genetic algorithm. His proposed technique is multivariate discretization based on density-based clustering and genetic algorithm (MVD-CG), which is an algorithm that improves the multivariate discretization algorithm (MVD). The experiment used real data and compared between MVD-CG algorithm and MVD algorithm. Measurement metrics are number of rules and effective of rules. The result is that MVD-CG algorithm has higher confident than MVD algorithm.

Sug⁽⁸⁾ has proposed multi-dimensional association rule mining, which is different from original association rule mining in term of the difference of column. This method can reduce the data size and runtime in association rule mining. The experiment used real data from UCI. The result is that the algorithm can create small multi-dimensional table and reduce the number of rules.

From the related work, it can be seen that there exist many techniques for discretization. However, most researches do not take into consideration the distribution of the data in each rang of the divided value. We notice that it may be possible that some of the data that was in the range with very small amount might be unnecessary to be used in the association rule mining. We thus propose the visualize technique to detect ranges of values with limited amount of data in order to consider a cut point during discretization process.

3. Background

This research aims at studying the discretization and proposing a new method based on Chi2 algorithm and visualize technique for association rule mining. The related theories are divided into 4 parts, that is, association rule mining, discretization algorithms, the cut points in Chi2 algorithm, and visualization by cut points.

3.1 Association Rule Mining

Association rule mining is a popular analysis technique to automatically find the relationships between the data. There are many methods for association rule mining. In this papers we use Apriori⁽⁹⁾ algorithm for association rule mining. In the table 1 is a list of customer purchases that will be used as an example to explain the association rule mining process. The data are counted to find the frequent customer purchases on each itemset, and then take the frequent itemsets to generate the association rules, which is a rule in the form of “If condition Then result”. The measurement metrics used in the selection of frequent itemsets and rules are the following:

- Support is the frequency of the occurring event. Give the items A and B, the computation for support of A and B to be purchased is as follows:

$$Support(A \rightarrow B) = P(A \wedge B) \quad (1)$$

As an example, support (Coca cola $\rightarrow$ Bread) = $2/5 = 0.4$ or 40%.

- Confidence is the frequency of the incident with other events occurring together. The computation for confident is as follows:

$$Confidence(A \rightarrow B) = \frac{Support(A \rightarrow B)}{Support(A)} \quad (2)$$

For the same example as above, confidence (Coca cola $\rightarrow$ Bread) = $0.4/0.4 = 1.0$ or 100%.

Table 1. Purchase transactions of customers.

Order	Coca cola	Bread	Candy	Milk
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Table 2. Contingency table for cut point computing demonstration.

	Class 1	Class 2	Sum
Interval 1	A ₁₁	A ₁₂	R ₁
Interval 2	A ₂₁	A ₂₂	R ₂
Sum	C ₁	C ₂	N

Sample	K=1	K=2	
2	0	1	1
3	1	0	1
total	1	1	2

$$\begin{aligned} E_{11} &= (1/2)*1 = .05 \\ E_{12} &= (1/2)*1 = .05 \\ E_{21} &= (1/2)*1 = .05 \\ E_{22} &= (1/2)*1 = .05 \end{aligned}$$

$$X^2 = (0-.5)^2/.5 + (1-.5)^2/.5 + (1-.5)^2/.5 + (0-.5)^2/.5 = 2$$

Sample	K=1	K=2	
3	1	0	1
4	1	0	1
total	2	0	2

$$\begin{aligned} E_{11} &= (1/2)*2 = 1 \\ E_{12} &= (0/2)*2 = 0 \\ E_{21} &= (1/2)*2 = 1 \\ E_{22} &= (0/2)*2 = 0 \end{aligned}$$

$$X^2 = (1-1)^2/1 + (0-0)^2/0 + (1-1)^2/1 + (0-0)^2/0 = 0$$

Fig. 1. Example of calculation the Chi2 cut point between intervals.

3.2 Discretization algorithms

(a) Chi2 algorithm

Chi2 algorithm⁽¹⁰⁾ that is based on the X^2 statistics was used to perform discretization over the numerical data. The computation for x^2 is as follows:

$$X^2 = \sum_{i=1}^x \sum_{j=1}^k \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \quad (3)$$

where:

k = number of classes,

A_{ij} = number of patterns in the ith interval, jth class,

E_{ij} = expected frequency of A_{ij} = R_i * C_j / N,

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

C_j = number of patterns in the jth class = $\sum_{i=1}^x A_{ij}$,

N = total number of patterns = $\sum_{i=1}^x R_i$

Fig. 2. Example of interval integration by Chi2 considering value.

(b) CAIM algorithm

Class-Attribute Interdependence Maximization (CAIM)⁽¹¹⁾ is a discretization algorithm by supervised learning. Main idea of the CAIM algorithm is to use class-attributed interdependence of the class and the numeric column to create minimal interval. The computation for CIAM is as follows:

$$CAIM(C, D | F) = \frac{\sum_{r=1}^n \frac{\max_r^2}{M_{+r}}}{n} \quad (4)$$

where:

C = class,

D = discretization,

F = columns,

n = number of intervals,

max_r = maximum of the q_{ir},

M_{+r} = number of all continuous columns

(c) CACC algorithm

Class-Attribute Contingency Coefficient (CACC)⁽¹²⁾ is a discretization algorithm developed from CAIM algorithm for solving the overfitting problem. The computation for CACC is as follows:

$$CACC = \sqrt{\frac{y'}{y' + M'}} \quad (5)$$

where:

M' = number of sampling data

$$y' = M \left[\left(\sum_{i=1}^S \sum_{r=1}^n \frac{q_{ir}^2}{M_{i+} M_{+r}} \right) - 1 \right] / \log(n)$$

when

M = number of sampling data,

n = number of data,

q_{ir} = number of class i by sampling data
($i=1,2,\dots,S$ and $r=1,2,\dots,n$) in the
interval $(d_{r-1}, d_r]$,

M_{i+} = number of class i by sampling data

(d) AMEVA algorithm

AMEVA algorithm⁽¹³⁾ is discretization algorithm by supervised learning. This algorithm increase performance in terms of correlations between parameter and decrease number of intervals of the Chi2 algorithm. The computation for AMEVA is as follows:

$$Ameva(k) = \frac{x^2(k)}{k(l-1)} \quad (6)$$

where:

k = number of intervals,

$$x^2(k) = N \left(-1 + \sum_{i=1}^l \sum_{j=1}^k \frac{n_{ij}^2}{n_i n_j} \right)$$

3.3 Cut point in Chi2 Algorithm

Discretization by Chi2 algorithm is to find the cut points to divide the numerical data to interval data. The algorithm is based on the bottom-up division of numerical data in each row into intervals, and then gradually merging each interval based on independence, which can be computed by Chi2 value. We can demonstrate the cut point calculation from contingency table in table 2 and equation (3). Figure 1 shows an example of calculating the Chi2 in each interval, such as intervals 2, 3 and 4 has Chi2 0. Figure 2 shows an example of integration the intervals by considering minimal Chi2 value, because the minimal Chi2 value means less independent between intervals. If the intervals are less independent, they should be in the same interval. For instance, the intervals 3, 4 and 5 have the Chi2 values 0. They should be in the same interval. Repeat the interval merging until Chi2 values of all intervals are greater than threshold that the user has defined.

Fig. 3. Example of comparing the number of rules in each interval.

3.4 Visualization Cut Point Consideration Through

Discretization numerical data by various algorithms has to consider the cut points, which can be used for grouping the numerical data to intervals. But in some situation the data in the discretized interval has too small amount of data. This can lead to an inefficient association rule mining. Thus, we propose to use visualization technique⁽¹⁴⁾ in the post-processing of the discretization steps to see the distribution of data in each interval, and then adjust the cut points to fit the data distribution in each interval. Figure 3 shows an example chart comparing the number of data in each interval. It can be seen that the data in the intervals 1, 7 and 12 are minimal comparing to other intervals. The cut points should be adjusted to re-distribute data in the discretized intervals. Figure 4 example of interval integration by visualize technique.

Fig. 4. Example of interval integration by visualize technique

Table 3. Comparative results of accuracy and number of rules by five discretization algorithms.

Algorithms \ Datasets	Simulated Data		ECOLI		APD		CLE	
	#Rules	Acc.	#Rules	Acc.	#Rules	Acc.	#Rules	Acc.
AMEVA	81	88.19	107	97.57	987	95.29	111	45.31
CAIM	79	88.39	52	97.47	971	94.87	43	83.13
CACC	81	88.19	107	97.57	987	95.21	75	65.99
Chi2	41	88.17	47	97.67	40	81.85	239	81.37
Chi2+Visualize	37	95.41	47	97.68	36	87.08	220	81.58

Fig. 5. The accuracy comparison of the five discretization algorithms.

Fig. 6. The number of rules comparison of the five discretization algorithms.

Fig. 7. Conceptual framework of the research.

4. Proposed Discretization Method

This research has proposed a methodology to perform discretization based on the Chi2 algorithm and the visualize technique for association rule mining. Figure 7 sketches the proposed method to discretize numerical attributed for association rule mining. The pre-processing module can be divided into two parts: discretization step and visualization by cut point consideration step. Discretization in our method is based on the Chi2 algorithm because it is easy to understand. After the discretization step to visualization has been applied to see distribution of data in each interval. If amount of data in any interval is too small, that interval will be merged with the previous interval or the next interval, which is considered by the distance to the nearest interval. Finally, the result is data discretized by new cut points. These data will be further used in the association rule mining

5. Experimental Setting and Results

The proposed discretization method has been experimented with both synthetic data and real data from the UCI Machine Learning Repository. The UCI data are Appendicitis data (APD) with 106 records and 7 attributes. Ecoli data with 336 records and 8 attributes. Cleveland data (CLE) with 303 records and 13 attributes. Each of these datasets has been divided into training dataset (70%) and test dataset (30%). The discretization algorithm is encoded with the R language⁽¹⁵⁾. The algorithm proposed in this research is called Chi2+Visualize discretization algorithm. Its performance has been compared with the AMEVA,

CAIM, CACC, and Chi2 algorithms. The performance metrics are number of rules and accuracy of the algorithms.

Table 3 shows the number of rules and accuracy of algorithms with different datasets. It can be seen that the Chi2+Visualize algorithm with the synthetic data and Ecoli data shows higher accuracy with less number of rules when compared to other algorithms. But the performance of our algorithm on the CLE and APD data shows lower accuracy than some algorithms.

Figure 5 shows a chart comparing the accuracy of algorithms with different datasets. It can be seen that the Chi2+Visualize with synthetic data and Ecoli data show higher accuracy when compared to other algorithms. Figure 6 shows a chart comparing the number of rules with different datasets. It can be seen that the Chi2+Visualize with synthetic, Ecoli and APD data has less number of rules when compared with other algorithms.

6. Conclusions

This research aims at studying the discretization method based on Chi2 algorithm and visualize technique for association rule mining. The problem of association rule mining with numerical data is that there will be large number of rules and the obtained association rules are not effective enough to predict the future data. Thus, we propose to use the cut point from discretization by Chi2 algorithm to see the distributed data in each interval, and then adjust the cut points to fit the distribution in each interval. The experimental results reveal that the proposed algorithm can reduce the number of rules and increase accuracy in predicting the future data. However, the application of our method over some data show low accuracy, but can be traded-off by small number of rules. But in some data our method shows low accuracy and large number of rules. We hypothesize that this dataset may be non-normal distribution and this kind of distribution has strong effect to our method. However, this hypothesis needs theoretical and experimental proofs further.

References

- (1) Berry, Michael J., and Gordon Linoff. : "Data mining techniques: for marketing, sales, and customer support.", John Wiley & Sons, Inc., 1997.
- (2) Liu, Huan, et al. : "Discretization: An enabling technique.", Data mining and knowledge discovery, vol. 6, No. 4, pp. 393-423, 2002

- (3) Su, Chao-Ton, and Jyh-Hwa Hsu. : "An extended chi2 algorithm for discretization of real value attributes.", Knowledge and Data Engineering, IEEE Transactions on, vol. 17, No. 3, pp. 437-441, 2005
- (4) Gyenesei, Attila : "A Fuzzy Approach for Mining Quantitative Association Rules.", Acta Cybern, Vol. 15, No. 2, pp. 305-320, 2001
- (5) Tong, Qiang, et al. : "A method for mining quantitative association rules.", Jisuanji Gongcheng/ Computer Engineering, vol. 33, No. 10, pp. 34-35, 2007
- (6) Ke, Yiping, James Cheng, and Wilfred Ng. : "MIC framework: an information-theoretic approach to quantitative association rule mining.", Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. IEEE, pp. 112-112, 2006.
- (7) Wei, Hantian. : "A novel multivariate discretization method for mining association rules.", Information Processing, 2009. APCIP 2009. Asia-Pacific Conference on, Vol. 1, pp. 378-381, 2009
- (8) Sug, Hyontai. : "Discovery of multidimensional association rules focusing on instances in specific class.", International Journal of mathematics and Computers in Simulation, vol. 5, No. 3, pp. 250-257, 2011
- (9) Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami : "Mining association rules between sets of items in large databases.", ACM SIGMOD Record, Vol. 22, No. 2, ACM, 1993.
- (10) Liu, Huan, and Rudy Setiono. : "Chi2: Feature selection and discretization of numeric attributes.", 2012 IEEE 24th International Conference on Tools with Artificial Intelligence. IEEE Computer Society, pp. 388-388, 1995
- (11) Kurgan, Lukasz A., and Krzysztof J. Cios. : "CAIM discretization algorithm.", Knowledge and Data Engineering, IEEE Transactions on, vol. 16, No. 2, pp. 145-153, 2004
- (12) Tsai, Cheng-Jung, Chien-I. Lee, and Wei-Pang Yang. : "A discretization algorithm based on class-attribute contingency coefficient.", Information Sciences, vol. 178, No. 3, pp. 714-731, 2008
- (13) Gonzalez-Abril, L., et al. : "Ameva: An autonomous discretization algorithm.", Expert Systems with Applications, Vol. 36, No. 3, pp. 5327-5332, 2009
- (14) Keim, Daniel A. : "Information visualization and visual data mining.", Visualization and Computer Graphics, IEEE Transactions on, vol. 8, No. 1, pp. 1-8, 2002
- (15) Ihaka, Ross, and Robert Gentleman : "R: a language for data analysis and graphics.", Journal of computational and graphical statistics, Vol. 5, No. 3, pp. 299-314, 1996