Approximate Equal Frequency Discretization Method

Sheng-yi Jiang, Xia Li, Qi Zheng, Lian-xi Wang
School of Informatics Guangdong University of Foreign Studies Guangzhou 510006 China
jiangshengyi@163.com

Abstract

Many algorithms for data mining and machine learning can only process discrete attributes. In order to use these algorithms when some attributes are numeric, the numeric attributes must be discretized. Because of the prevalent of normal distribution, an approximate equal frequency discretization method based on normal distribution is presented. The method is simple to implement. Computing complexity of this method is nearly linear with the size of dataset and can be applied to large size dataset. We compare this method with some other discretization methods on the UCI datasets. The experiment result shows that this unsupervised discretization method is effective and practicable.

1. Introduction

Discretization is to divide the range of the continuous attribute into intervals. Every interval is labeled a discrete value, and then the original data will be mapped to the discrete values. Discretization of the continuous attributes is an important preprocessing approach for data mining and machine learning algorithm. An effective discretization method not only can reduce the demand of system memory and improve the efficiency of data mining and machine learning algorithm, but also make the knowledge extracted from the discretized dataset more compact, easy to be understand and used. Research shows that picking the best split points is a NP-complete problem[1]. The result of discretization is related not only with the discretization algorithm itself but also with the data distribution and the number of split points. When the same discretization algorithm is applied to different dataset, we may get different result. We can only know the effectiveness of the discretization method by the result of post processing. So whether the discretization method is good or not is also related with the induction algorithm adopted later.

There are many classic methods to discretize continuous attribute, including equal width method(EW),equal frequency method(EF),statistic test method[2-6],information entropy method[7-11] and clustering-based method[12][13] etc. In general all these methods use the selected split points to divide the range of the continuous attribute into finite intervals and label every interval. These algorithms are categorized as supervised discretization method and unsupervised discretization method based on whether it uses class information or not. This paper studies the method of unsupervised discretization.

This article is organized as follows. Section 2 introduces equal frequency method and approximate equal frequency method. Section 3 presents experimental evaluation with real and synthetic data sets. Section 4 is the conclusions and directions for future work.

2.Approximate Equal Frequency Discretization Method

2.1 Method Description

Equal frequency method is an unsupervised discretization algorithm, which tries to put the same number of values into each interval. If there are n points in the whole range of the attribute value and we divide it into k intervals, then every interval will have n/k points. The equal frequency discretization algorithm is simple to implement, but due to ignoring the distribution information of the dataset, the algorithm can not set interval boundary on the right position in some situation, so that it doesn't perform well. Normal distribution has good characteristic, and it is also a theoretical basis of many statistic method. There are many statistic variables in nature science and behavior science are approximately submission to normal distribution in large sample. Thus, in this paper we present an approximate equal frequency discretization method (AEFD) which based on normal distribution.

978-0-7695-3571-5/09 $25.00 © 2009 IEEE
DOI 10.1109/GCIS.2009.131

The idea of AEFD is that if a variable is normally distributed, then the frequency that the observation value is located in an interval is equal to the probability that the variable's value is located in the interval. The normally distributed variable's quantile is used to divide the range of the variable's value into several intervals that the variable's value has the same probability of locating in each interval.

Suppose dividing the range of attribute value into k intervals: $(b_i, b_{i+1}](i = 0,1,\cdots,k-1)$, $b_0 = -\infty, b_k = \infty$, the normally distributed variable has the equal probability 1/k falling into each of the intervals. After the discretization, we map every attribute value in $(b_i, b_{i+1}](i = 0,1,\cdots,k-1)$ to a same symbol. AEFD contains three steps. Details are as follows:

Step1: calculate split points, decide initial dividing intervals.

Calculate split points according to $b_i = \bar{x} + Z_{\alpha_i} \cdot \sigma(i = 1,2,\cdots,k-1)$, here $\bar{x}$ and σ is mean and standard deviation of the attribute value, Z_{α_i} is the α_i quantile of standard normal distribution $\xi \sim N(0,1)$, $P(\xi \le Z_{\alpha_i}) = \alpha_i = \frac{i}{k}(i = 1,2,\cdots,k-1)$.

Via split points we can get the initial k intervals: $(b_i, b_{i+1}](i = 0,1,\cdots,k-1)$.

Step2: merge the initial divided intervals. That is to merge the intervals, which contains few records, into the nearest interval.

Calculate frequency , maximum and minimum of records contained in every interval $(b_i, b_{i+1}](i = 0,1,2,\cdots,k-1)$ ..Searching from right to left, when the frequency of records contained in interval $(b_i, b_{i+1}](i = 0,1,2,\cdots,k-1)$ is less than MinF which is a fraction of frequency (1/k), merge the interval into its nearest interval and update frequency, maximum and minimum of records contained in the interval. The whole process will not stop until no interval will be merged.

Step3: map values of the continuous attribute into discrete values with respect to the divided intervals.

2.2 Time Complexity of the Algorithm

In Step1, operation is fixed, time complexity is $O(1)$; In step2 and step3, we only need to scan the dataset once to know the objects contained in each intervals, so time complexity of these two step is $O(n*m)$, n is the size of the dataset and m is the number of the attribute to be discretized. Hence the whole algorithm need to scan the dataset two times, the total time complexity is $O(n*m)$ .AEFD can be applied to large size dataset, the time complexity of it is less than that of equal frequency method, PKID algorithm[17] .

2.3 Further Discussion of the Algorithm AEFD

(1)About the number of the initial divided intervals k

The number of the initial divided intervals k is between with $\lceil \log(n) \rceil$ and $\lceil \log(n) - \log(\log(n)) \rceil + 1$, if $\lceil \log(n) - \log(\log(n)) \rceil > 19$, then let k=20. i.e. k is no more than 20. In latter experiment, we set $k = \min\{\lceil \log(n) - \log(\log(n)) \rceil + 1, 20\}$.

(2) Condition to merge two intervals

When a interval contains fewer records than a fraction of the frequency (1/k),MinF, the interval need to be merged because of its less frequency. MinF can be set in the range [0.3,0.5]. In our experiment, we set MinF=1/3.

(3) Strategy to merge two intervals

There are two different strategies to merge the objects in the low-frequency interval. The first is to see the interval as a unit to be merged into the nearest interval. Second is to merge every object in the interval one by one into the nearest interval. Considering the efficiency of the algorithm, we prefer to see the interval as a whole. Also there are two methods to measure how close the two intervals are, one is the distance between the two centroids of the two intervals, the another one is the distance between the border points of the two intervals. The experiment shows that the two measure of the distance of two intervals has almost the same result. In this paper we use the distance between two border points of the two intervals.

(4) Distribution of the data

If we know the distribution of the data, step1 can be changed to decide the split points based on the given distribution, and the result will be better.

3. Experimental Results

In order to test the performance of AEFD algorithm, we select twenty-nine datasets from UCI[15] and a real salary dataset to be discretized. The character of the selected datasets is as table 1. Algorithms in Weka software[16] ,including C4.5, Ripper, Naïve-Bayse classifer, equal width discretization method (EW), PKID and supervised discretization method MDL[7], are used here.

Table 1 summary of datasets in our experimental

Dataset	Nominal/ Continuous attributes	Instance size	Number of Class
Adult	8/6	32561	2
Aneal	32/6	898	6
Austra	8/6	690	2
Breast	0/9	699	2
Credit	9/6	690	2
Dermatology	33/1	366	6
Diabetes	0/8	768	2
Ecoli	1/8	336	8
Flag	10/18	194	6
German	7/13	1000	2
Glass	0/9	214	6
Haberman	0/3	306	2
Heart	7/6	270	2
Hepatitis	6/13	155	2
Horse-colic	7/19	368	2
Hypothyroid	7/18	3163	2
Ionosphere	0/34	351	2
Iris	0/4	150	3
Labor	0/8	57	2
Letter-recognition	0/16	20000	26
Liver	0/6	345	2
Musk_clean	0/188	6598	2
Pendigits	0/16	10992	10
Pima	0/8	768	2
Satimage	0/36	6435	6
Sonar	0/60	208	2
Vehicle	0/18	846	4
Vowel-context	1/10	990	11
Wine	0/13	178	3
Salary	0/1	80	4

Salary dataset contains after-tax salary data of eighty employees in a department of a university, it contains two attributes, title and after-tax salary, the distribution of the whole data is shown as table 2.

Table 2 Distribution of the Salary dataset

Title	number of person	range of after-tax salary
professor	8	5266.08-4858.24
associated processor	22	4272.34-3644.98
lecturer	40	3438.55-2885.22
assistant	10	2702.28-2420.35

We use AEFD algorithm with k specified 9 to discretize the after-tax salary and get 6 split points which is 2466.44, 2797.07, 3035.43, 3239.07, 3619.90, 4392.53, and 7 intervals. It is apparent that after discretization, professor and associate professor is corresponding to 1 interval respectively, and lecture is corresponding to 3 different intervals, assistant is corresponding to 2 different intervals.

We use PKID algorithm to discretize the after-tax salary, get 7 split points which is 2793.75, 2913.11, 2974.46, 3099.82, 3541.77, 3990.89, 4155.65, and 8 intervals. The interval (4155.65, 5266.08) contains two types of object, which are professor and associate professor. So the effect of PKID algorithm is not better than AEFD algorithm.

We use MDL algorithm to discretize the after-tax salary, get 3 split points which is 2793.75, 3541.77, 4565.29 and 4 intervals. Different title is corresponding to different discretization interval respectively. It's a perfect result.

3.1.Performance Comparison on C4.5 classifier

We apply AEFD, EW, MDL method, PKID method[17] method to discretize the selected 29 datasets respectively, and use C4.5 with10-fold cross-validation to classify the datasets before and after discretization. We compare the classification precision with the result in related literature; the test result is as table 3.

Table 3 Accuracy of C4.5

Datasets	Before discretization	AEFD	EW	PKID	MDL
Adult	86.24%	84.53%	84.25%	85.95%	86.55%
Aneal	91.53%	85.69%	88.75%	92.36%	91.62%
Austra	84.93%	87.39%	84.93%	85.14%	85.22%
Breast	94.56%	96.09%	94.99%	94.41%	95.71%
Credit	85.22%	86.78%	84.64%	85.10%	87.10%
Dermatology	93.99%	93.96%	93.99%	93.85%	93.99%
Diabetes	72.40%	76.56%	74.61%	74.04%	78.13%
Ecoli	82.65%	78.87%	72.35%	65.77%	83.10%
Flag	59.33%	61.65%	61.65%	62.11%	62.63%
German	70.50%	73.07%	71.70%	70.80%	72.10%
Glass	72.90%	70.28%	51.40%	51.21%	74.77%
Haberman	71.57%	73.20%	71.34%	73.53%	71.34%
Heart	76.67%	78.74%	73.33%	77.78%	81.48%
Hepatitis	64.19%	65.81%	65.24%	65.48%	65.24%
Horse-colic	67.93%	65.90%	67.12%	67.69%	66.30%
hypothyroid	99.28%	99%	97.51%	95.23%	99.27%
Ionosphere	91.45%	89.91%	87.75%	89.12%	89.17%
Iris	94.00%	96.53%	92%	91.53%	93.33%
Labor	73.68%	75.61%	64.91%	69.82%	80.70%
Letter recognition	88.20%	80.57%	81.30%	77.92%	78.76%
Liver	66.67%	70.32%	56.81%	57.45%	63.19%
Musk_clean	96.89%	96.08%	96.17%	93.86%	96.83%
Pendigits	94.92%	87.86%	85.10%	60.62%	89.43%
Pima	73.96%	75.01%	74.35%	73.93%	77.73%
Satimage	85.94%	83.73%	84.10%	79.42%	83.23%

Sonar	74.18%	73.03%	69.09%	68.89%	80.05%
Vehicle	72.72%	68.18%	70.60%	62.25%	70.60%
Vowel-context	80.11%	74.51%	75.05%	49.33%	79.23%
Wine	93.26%	91.29%	78.65%	79.72%	94.38%
Average	81.37%	80.69%	77.71%	75.67%	81.76%

3.2.Performance Comparison on RIPPER classifier

We randomly disorder the selected 26 datasets ten times, and apply AEFD, EW, PKID method, MDL method to discretize them respectively. We use RIPPER with10-fold cross-validation to classify the datasets before and after discretization, and compare the classification precision with the result in related literature; the test result is as table 4.

Table 4 Accuracy of RIPPER

Data sets	Before discretization	AEFD	EW	PKID	MDL
Aneal	93.58%	88.87%	92.19%	94.50%	94.55%
Austra	84.93%	85.64%	85.25%	84.99%	86.01%
Breast	94.56%	95.99%	93.98%	93.83%	95.26%
Credit	85.22%	65.47%	84.83%	84.99%	86.64%
Dermatology	88.58%	89.75%	89.51%	89.73%	88.33%
Diabetes	72.40%	74.66%	73.11%	74.74%	77.51%
Ecoli	82.23%	79.52%	74.85%	78.69%	82.26%
Flag	59.90%	62.89%	61.65%	62.58%	61.70%
German	70.50%	71.24%	69.86%	70.15%	71.91%
Glass	72.90%	63.22%	50.19%	60.42%	70.65%
Haberman	72.81%	75.20%	73.33%	72.25%	72.12%
Heart	76.67%	80.04%	78.26%	78.89&	82.33%
Hepatitis	62.97%	70.06%	69.10%	69.74%	68.07%
Horse-colic	67.93%	84.59%	84.62%	72.20%	86.09%
Hypothyroid	99.16%	98.96%	97.19%	97.72%	99.14%
Ionosphere	91.45%	90.10%	87.44%	88.69%	91.74%
Iris	94.00%	95.47%	92.53%	92.60%	94.87%
Labor	73.68%	83.33%	83.51%	80.70%	87.02%
Liver	66.67%	63.80%	62.09%	54.84%	63.19%
Pendigits	94.27%	90.37%	89.44%	82.46%	90.46%
Pima	73.96%	74.79%	74.01%	74.05%	77.30%
Satimage	85.79%	82.32%	83.40%	76.30%	83.56%
Sonar	74.71%	70.91%	66.88%	57.50%	79.18%
Vehicle	68.91%	64.11%	63.42%	58.39%	67.74%
vowel-context	70.64%	65.29%	66.65%	43.41%	73.62%
Wine	93.26%	89.66%	83.99%	90.00%	94.66%
Average	79.68%	79.09%	78.13%	76.22%	81.77%

3.3. Performance Comparison on Naïve-Bayes Classifier

We randomly disorder the selected 29 datasets ten times, and apply AEFD, EW, PKID method, MDL method to discretize them. We use Naïve-Bayes with 3-fold cross-validation to classify the datasets before and after discretization, and compare the classification precision with the result in related literature; the test result is as table 5.

Table 5 Accuracy of Naïve-Bayes

Data sets	Before discretization	AEFD	EW	PKID	MDL
Adult	83.38%	82.58%	81.86%	83.75%	83.88%
Aneal	79.91%	90.38%	92.17%	95.81%	95.79%
Austra	77.19%	86.06%	85.29%	86.42%	85.65%
Breast	96.01%	97.32%	97.25%	97.25%	97.22%
Credit	77.74%	85.78%	84.81%	85.67%	86.23%
dermatology	97.46%	97.90%	97.68%	97.46%	97.89%
Diabetes	75.42%	75.47%	75.36%	73.54%	77.92%
Ecoli	84.88%	83.78%	80.51%	80.71%	85.77%
Flag	43.35%	59.59%	59.38%	59.95%	59.23%
German	74.52%	75.12%	75.43%	74.78%	74.67%
Glass	46.54%	64.02%	57.01%	65.00%	72.34%
Haberman	74.97%	75.95%	75.26%	72.06%	72.55%
Heart	84.33%	84.07%	83.30%	81.81%	83.04%
Hepatitis	69.74%	68.58%	68.28%	67.35%	71.10%
Horse-colic	67.28%	68.91%	68.97%	69.84%	66.44%
hypothyroid	97.51%	98.24%	96.87%	97.62%	98.60%
Ionosphere	82.85%	89.15%	90.51%	88.15%	90.94%
Iris	95.40%	94.73%	94.13%	92.60%	94.47%
Labor	91.93%	92.11%	92.81%	91.93%	93.51%
Letter recognition	64.14%	72.33%	69.78%	73.21%	73.72%
Liver	55.77%	64.96%	64.12%	62.26%	63.19%
Musk_clean	83.76%	79.68%	79.81%	90.96%	91.56%
Pendigits	80.37%	85.98%	85.63%	85.45%	87.18%
Pima	73.45%	74.49%	74.26%	72.04%	77.28%
Satimage	79.55%	81.61%	80.59%	82.18%	82.23%
Sonar	68.17%	75.96%	75.91%	72.55%	85.10%
Vehicle	44.53%	58.44%	60.07%	61.02%	62.07%
Vowel-context	61.89%	62.55%	62.37%	55.72%	64.19%
Wine	96.85%	97.36%	96.35%	95.06%	98.88%
Average	76.17%	80.11%	79.51%	79.73%	81.82%

The experiment result shows that the classification precision after AEFD discretize preprocessing is higher than that after the preprocessing by EW, PKID method in the most situation. In close to fifty percent situations, the classification precision after discretize preprocessing by AEFD is better than thatbefore discretize preprocessing and close to the classic supervised discretization algorithm MDL.

4. Conclusion

AEFD algorithm is based on normal distribution theory and notices that the large samples of real data is approximately submission to normal distribution. The algorithm is simple, without complicated theory, and easy to understand and implement. The algorithm has nearly linear time complexity and the experimental results show that the AEFD is better than the previous

unsupervised discretization algorithm. Similar to EF method, AEFD does not fully consider the data's distribution information. In some situations, it can not set the interval border on the properest position. The further work is to find method that can recognize intervals of one dimension data with different distribution density, and further to find the nature discrete intervals to improve discretization result.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.60673191), the Natural Science Research Programs of Guangdong Province's Institutes of Higher Education (No.06Z012) and Guangdong University of Foreign Studies Team Research Program of Innovations (No.GW2006-TA-005).

References

[1] H.S.Nguyen,A.Skowron.Quantization of Real Values Attributes Rough Set and Boolean Reasoning Approaches[A].Proc. of the 2th joint Annual Conf on Information Sci[C].USA Wrightsville Beach,NC,1995:34-37.

[2] Kerber R. Discretization of Numeric Attributes [A]. The 9th International Conference on Artificial Intelligence [C].1992.

[3] Liu H, Setiono R. Feature selection via discretization[J]. IEEE Transactions on Know ledge and Data Engineering, 1997, 9(4):642-645.

[4] Tay E H, Shen L. A modified Chi2 algorithm for discretization[J].IEEE Transactions on Knowledge and Data Engineering,2002,14(3):666-670.

[5] LI Gang,TONG Fu.An Unsupervised Discretization Algorithm Based on Mixture Probabilistic Model[j]. Chinese Journal of Computers, 2002,25(2):158- 164

[6]I. Kononenko. Inductive and bayesian learning in medical diagnosis[J].Applied Artificial Intelligence.1993,7:317–337.

[7] Fayyad U M,Irani K B.Multi-interval discretization of continuous valued attributes for classification learning[A]. Proc. of the 13th International Joint Conference on Artificial Intelligence[C],1993:1022-1029.

[8] Clarke EJ,Braton BA.Entropy and MDL discretization of continuous variables for Bayesian belief networks[J]. International Journal of Intelligence Systems, 2000,15:61-92.

[9] Chiu D K Y, Cheng B, Wong A K C. Information Synthesis Based on Hierarchical Maximum Entropy Discretization.Journal of Experimental and Theoretical Artificial Intelligence,1990,2:117-129.

[10] XIE Hong,CHENG Hao-Zhong, NIU dong-Xiao. Discretization of Continuous Attributes in Rough Set[J]. , Chinese Journal of Computers, 2005,28(9):1570-1574.

[11] Chang-Hwan Lee. A Hellinger-based discretization method for numeric attributes in classification learning. Knowledge-Based Systems. 2007,20(4): 419-425.

[12]LI Xing-sheng,LI De-yi.A New Method Based on Density Clustering for Discretization of Continuous Attributes[J].Acta Simulata Systematica Sinica. 2003.6:804-806.

[13]XI Jing,OUYANG Wei min.Clustering Based Algorithm for Best Discretizing Continuous Valued Attributes[J]. Mini-micro systems. 2000,21 (10):1025-1027

[14]Dougherty J R, Kohavi, Sahami M. Supervised and Unsupervised Discretization of Continuous Features. Machine Learning[A] .Proc of 12th International Conference, Morgan Kaufmann[C].1995:194-202

[15] Asuncion, A. & Newman, D.J. UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. 2007.

[16] weka. http://www.cs.waikato.ac.nz/ml/weka/

[17]Y Yang, GI Webb.A Comparative Study of Discretization Methods for Naive-Bayes Classifiers.Pacific Rim Knowledge Acquisition Workshop (PKAW'02), Tokyo, 2002: 159-173