An ICA-Based Multivariate Discretization Algorithm*

Ye Kang[1,2], Shanshan Wang [1,2], Xiaoyan Liu[1], Hokyin Lai[1],
Huaiqing Wang[1], and Baiqi Miao[2]

[1] Department of Information Systems, City University of Hong Kong
[2] Management School, University of Science and Technology of China,
HeFei, AnHui Province
kye@mail.ustc.edu.cn, sswang@ustc.edu

Abstract. Discretization is an important preprocessing technique in data mining tasks. Univariate Discretization is the most commonly used method. It discretizes only one single attribute of a dataset at a time, without considering the interaction information with other attributes. Since it is multi-attribute rather than one single attribute determines the targeted class attribute, the result of Univariate Discretization is not optimal. In this paper, a new Multivariate Discretization algorithm is proposed. It uses ICA (Independent Component Analysis) to transform the original attributes into an independent attribute space, and then apply Univariate Discretization to each attribute in the new space. Data mining tasks can be conducted in the new discretized dataset with independent attributes. The numerical experiment results show that our method improves the discretization performance, especially for the nongaussian datasets, and it is competent compared to PCA-based multivariate method.

Keywords: Data mining, Multivariate Discretization, Independent Component Analysis, Nongaussian.

1 Introduction

Discretization is one of preprocessing technique used frequently in many data warehousing and data mining applications. It is a process of converting the continuous attributes of a data set into discrete ones. In most of databases, data is usually stored in mixed format: the attribute can be nominal, discrete or continuous. In practice, continuous attribute needs to be transformed discrete one so that some machine learning methods can operate on it. Furthermore, discrete values are more concise to represent and specify and easier to process and comprehend, because they are closer to knowledge-level representation. Therefore, discretization can highlight classification tasks and improve predictive accuracy in most cases[1].

Univariate Discretization is one of commonly used discretization strategy. It aims to find a partition of a single continuous explanatory attribute of a dataset at one time. But attributes in multivariate datasets are usually correlated with each other, discretizing them without considering the interaction between them can not get a global optimal result. Thus, Multivariate Discretization, which means discretizing attributes

* Supported by a SRG Grant (7001805) from the City University of Hong Kong.

J. Lang, F. Lin, and J. Wang (Eds.): KSEM 2006, LNAI 4092, pp. 556 – 562, 2006.

simultaneously or considering multiple attributes at the same time, draws more and more attention in recent years. However, few effective algorithms of Multivariate Discretization have been provided until now.

As the dataset usually comes with correlated attributes, to make them independent but the attribute information is not lost is a possible way for Multivariate Discretization. In this paper, a new Multivariate Discretization Algorithm using ICA (Independent Components Analysis) is presented. In this algorithm, we transform the original attributes into a new attributes space with ICA, and then conduct Univariate Discretization on the attributes of new space one by one as they are independent of each other after transform. Finally, a global optimal discretization results can be obtained. ICA is a statistical method, which can extract independent features from database, and then the database can be newly reformed approximately by the independent features as attributes. The accuracy of classification on this discretization results with the Multivariate Discretization algorithm proposed in this paper shows that this algorithm is competent to the published Multivariate Discretization approaches such as PCA-based Multivariate Discretization [2], especailly for nongaussian data.

The remainder of the paper is organized as follows: Section 2 gives an overview of related work, in Section 3 we discuss our transformation algorithm, and in Section 4 we report our experimental results. Finally we give our conclusion in Section 5.

2 Related Work

A large number of discretization algorithms have been proposed in past decades. Most of them are the Univariate Discretization methods. Univariate Discretization can be categorized in several dimensions: supervised or unsupervised, global or local, dynamic or static, merging(bottom-up) or splitting(top-bottom)[3]. For example, Chimerge[4] is a supervised, bottom-up algorithm, Zeta[5]is a supervised splitting algorithm, and Binning is a unsupervised splitting one. In discretization algorithms, stop criteria is an important factor. Most commonly used discretization criteria are Entropy measure, Binning, Dependency, Accuracy and so on [1]. Except that, recently Liu [6] provided Heterogeneity as another new criteria to evaluate a discretization scheme.

In the Univariate Discretization algorithms, each attribute is viewed as independently determining the class attribute. Therefore, it can not generate global optimal intervals by discretizing all the involved attributes in a multivariate dataset one by one. A solution to this problem is to discretize attributes simultaneously, that is to consider multiple attribute at a time, which is known as Multivariate Discretization. Several approaches about this have been presented.

Ferrandiz′ [7] discussed the multivariate notion of neighborhood, which is extending the univariate notion of interval. They proposed Bipartitions based on the Minimum Description Length (MDL) principle, and apply it recursively. This method is thus able to exploit correlations between continuous attributes. However it reduces predictive accuracy as it makes only local optimization.

Bay [8] provided an approach to the Multivariate Discretization problem considering interaction among attributes. His approach is to finely partition each continuous attribute into n basic regions and iteratively merge adjacent intervals with similar multivariate distribution. However, this approach is not very effective because of its high computational complexity.

Sameep [2] proposed a PCA-based Multivariate Discretization method. His method first computes a positive semi-defined correlation matrix from the dataset. Suppose its eigenvalues are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ of which the corresponding eigenvectors are e_1, e_2, …, e_d. Only the first k ($k<d$) eigenvectors with greater variance from the data are retained. Then all the data in original space are projected to the eigenspace which is spanned by the retained eigenvectors. Since each dimension in eigenspace is not correlated, the new attributes can be discretized separately by simple Distance-based Clustering or the Frequent Item Sets method. Once cut points are obtained, they are projected to the original data set which correlated most closely with this corresponding eigenspace dimension. This approach considers the correlation information among attributes through PCA transform. But PCA which relies on second-order statistics of the data often fails where the data are nongaussian [9].

In this paper, a new ICA-based multivariate discretization algorithm is proposed. The original attributes are transformed into a new attributes space with ICA, and then conduct Univariate Discretization on the attributes of new space one by one. The numerical experiment results show that this method impoves discretization performance, especially for the nongaussian datasets, and it is competent to other Multivariate Discretization method, such as PCA-based method.

3 ICA-Based Multivariate Discretization Algorithm

This section gives a detailed description of our algorithm and the background of it will also be introduced.

3.1 ICA (Independent Component Analysis)

ICA is on the base of Central Limit Theorem which tells that a sum of independent variables tends to follow a Gaussian distribution. Assuming there are n features, we denote x_j as the j-th feature and X as the random vector composed of x_1, …, x_n. The objective of ICA is to find n independent components s_i of X:

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \ldots + a_{jn}s_n \text{, for all } j \tag{1}$$

Let $\boldsymbol{A}$ be the matrix with element a_{ij} and $\boldsymbol{S}$ be the vector (s_1, …, s_n), then the above equation can be rewritten as follows:

$$\boldsymbol{X=AS} \tag{2}$$

$$\mathrm{S=A^{-1}X=WX} \tag{3}$$

where $\boldsymbol{W}$ denotes the weighed matrix of $\boldsymbol{X}$ which is the inverse of $\boldsymbol{A}$. All we observe is the random vector X, but we must estimate both $\boldsymbol{A}$ and $\boldsymbol{S}$ from it. The final aim of this estimation process is to obtain the values of $\boldsymbol{W}$ that can make $\boldsymbol{S}$ maximally nongaussianity, and they are just the independent components of $\boldsymbol{X}$.

Since there are many ICA algorithms provided by researchers such as Kernel [10], Hyvärinen and Oja [11], and Comon [12], in this paper we adopted FastICA algorithm which was introduced by Hyvärinen and Oja[11] due to low linear computational complexity.

From Central Limit Theorem it is known that the sum of two independent variables is more Gaussian than themselves. So $\boldsymbol{X}$ is more Gaussian than $\mathbf{S}$. In other words, $\boldsymbol{S}$ is more nongaussian than $\boldsymbol{X}$. As the linear sum of Gaussian distribution components still follows Gaussian distribution, ICA is only fit for nongaussian datasets (i.e., the more nongaussian the attribute variable of a data set is, the better). One of the classical measures of nongaussianity is kurtosis. The kurtosis of a variable y is defined by the following equation:

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2 \tag{4}$$

where y is zero-mean and of unit variance. Kurt will be zero for Gaussian variable. So the absolute value of kurtosis $|kurt(y)|$ is usually used as measure of non-gaussianity. And ICA is more suited to the variable with larger value of $|kurt(y)|$.

3.2 ICA-Based Discretization Algorithm

Our method is composed of the following four steps:

(1) Centering and whitening
Given a multivariate dataset, let x_i denotes (i=1, …, n) it's the i-th attribute which consists of m records, then the most basic and necessary step is to center x_{i} before the application of the ICA algorithm. Centering x_{i} is to subtract its mean value so as to make x_{i} zero-mean. This preprocessing can simplify ICA algorithm.

After centering, whitening as another important preprocessing should be taken, which transform the observed random vector $\boldsymbol{X}$ into white vector (i.e., the components are uncorrelated and their variances equal unity) denoted by $\boldsymbol{Y}$. One popular whitening method is adopted here, which is using the eigenvalue decomposition (EVD) of the covariance matrix $\mathrm{E}\{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\} = \boldsymbol{EDE}^{\mathrm{T}}$, where $\boldsymbol{E}$ is the orthogonal matrix of eigenvectors of $\mathrm{E}\{\boldsymbol{XX}^{\mathrm{T}}\}$ and $\boldsymbol{D}$ is the diagonal matrix of its eigenvalue, $\boldsymbol{D} = \mathrm{Diag}(d_{1,} d_{2,} ..., d_m)$. Thus,

$$\boldsymbol{Y} = \boldsymbol{ED}^{-1/2}\mathrm{E}^{\mathrm{T}}\boldsymbol{X} \tag{5}$$

Where the matrix $\boldsymbol{D}^{-1/2}$ is computed by a simple component-wise operation as $\boldsymbol{D}^{-1/2} = \mathrm{Diag}(d_1^{-1/2}, \ldots, d_n^{-1/2})$. It is easy to check that now $\mathrm{E}\{\boldsymbol{YY}^{\mathrm{T}}\} = \boldsymbol{I}$.

(2) Transforming attributes space by FastICA into new attributes space
After centering and whitening, FastICA is used to transform the original multi-attribute space into new independent multi-attribute space. Let z_i (i=1,…, n) denotes a new attribute which contains m data points, and each of them is independent of others. Finally, the class attribute is appended to the new space accordingly, and each instance

in the new dataset has the same class label as before. During transform, attribute information contained in the original dataset is preserved maximally in the new dataset.

(3) Using Univariated Discretization
After the new attributes z_i, (i=1, …, n) are obtained, we apply Unviariated Discretization method to each of them, and finally get the discretized intervals of the new attributes.

So far, many Univarivated Discretization have been proposed, in our experiment, we use the MDL method of Fayyad and Zrani [13] which is the only supervised discretization method provided in Weka and is also known for its good performance.

4 Experiments

In this section, we validate the Multivariate Discretization method proposed in our paper in terms of the quality of the discretization results. Here we use the results of Classification tasks on our discretization data to test the performance of our algorithm.

4.1 Experiment Setting

All the datasets used in this experiment are from UCI repository[1]. In order to simplify our experiment, those datasets with only continuous attributes were chosen. Table 1 gives a description of the chosen datasets. We used WEKA[2], software which contains Classification and Discretization tool packages to evaluate our discretization results.

Table 1. Data Sets Used in Evaluation

Dataset	Records	Attributes	Num of Class labels
Iris	150	4	3
Waveform	300	21	3
Glass	214	9	7
Cancer	683	8	2

Having chosen the datasets, we first took away the class attribute of each dataset, then centered and whitened the remaining continuous attributes, and transformed the original datasets by FastICA algorithm into new attributes space. A Matlab implementation of the FastICA algorithm is available on the World Wide Web free of charge[3]. At last we obtained a new dataset with independent attributes carrying the

[1] http://www.ics.uci.edu/~mlearn/MLRepository.html

[2] http://www.cs.waikato.ac.nz/~ml/

[3] http://www.cis.hut.fi/projects/ica/fastica/

information of the original dataset attributes. After transformat, the class attribute taken away before it was appended, then a new dataset was completed.

Discretization tool package of WEKA includes both supervised and unsupervised Univariate Discretization methods. The supervised discretization method is based on MDL [13]. As the attributes in the new transformed space are independent, they can be discretized separately. The discretized datasets was then processed by four classification algorithms of WEKA, respectively, C4.5, IBK, PART, NaiveBayes, and the error rates of classification using 10-fold cross-validation are reported in Table 2.

Table 2. Classification Error Comparison

Dataset	Mean kurtosis		C4.5	IBK	PART	NB	PCA+C4.5
Waveform	2.9237	Original	24.56	25.34	23.6	18.24	N
		ICA	17.32	16.86	16.72	18.12	
Iris	2.8941	Original	6	6	4.67	6	4.9
		ICA	1.3	1.3	1.3	1.3	
Glass	2.2055	Original	26.19	21.50	24.30	25.23	29
		ICA	27.57	28.04	29.91	29.91	
Cancer	1.6425	Original	4.25	3.07	4.25	2.49	4.1
		ICA	6.59	6.88	6.30	5.41	

Two groups of datasets were used in this experiment, one was composed of the original datasets that was downloaded from UCI Repository, and the other was composed of the new independent attribute data sets that were transformed from the original data sets using ICA. The supervised Univariate Discretization based on MDL was conducted on both the group's data sets. And their classification errors are reported in Table 2. From Table 2, for the datasets Waveform and Iris, ICA-based discretization method improved the classification accuracy significantly. However, it did not work very well on the datasets Glass and Cancer. This is because ICA has its roots for datasets with nongaussian distribution. We have mentioned before that kurtosis is one classic measure of nongaussianity, so the kurtosis of each dataset is given in the second column of the table. As there are more than one attributes for each dataset, the given value is the mean kurtosis of all the attributes. It can be seen that Waveform and Iris are much more nongaussian than Glass and Cancer as they have larger mean kurtosis. This can explain why our method works better for the former two datasets. The last column lists the results of the PCA-based Multivariate Discretization method from[2]. And we can see that our method is competent.

5 Conclusions

In this paper, we proposed ICA-based Multivariate Discretization method. It uses ICA to transform original dataset to a new dataset in which the attributes are independent of each other, and then conducts the Univariate Discretization on the new dataset. The numerical experiment results show that the discretization results of this method could improve the classification accuracy, especially for the nongaussian datasets, and it is competent compared to other multivariate method, such as PCA-based method and so on.

References

1. LIU, H., *Discretization: An Enabling Technique.* Data Mining and Knowledge Discovery, 2002. **6**: p. 393-423.
2. Sameep Mehta, *Toward Unsupervised Correlation Preserving Discretization.* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2005. **17**(9): p. 1174-1185.
3. Dougherty, J., Kohavi, R., & Sahami, M. *Supervised and unsupervised discretization of continuous features*. in *Proceedings of the Twelfth International Conference on Machine Learning*. 1995.
4. Kerber, R. *Chimerge discretization of numeric attributes.* in *Proceedings of the 10. th. International. Conference on Artificial Intelligence*. 1991.
5. K.M.HO. *Zeta: A Global Method for Discretization of Continuous Variables*. in *The Third International Conference on Knowledge Discovery and Data Mining*. 1997.
6. Liu, X. and H. Wang, *A Discretization Algorithm Based on a Heterogeneity Criterion.* IEEE Transactions on Knowledge and Data Engineering, September 2005. **17**(9): p. 1166-1173.
7. Ferrandiz, S. *Multivariate Discretization by Recursive Supervised Bipartition of Graph.* in *4th International Conference, MLDM 2005,Leipzig, Germany,July 9-11. Proceedings*. 2005.
8. Bay, S.D., *Multivariate Discretization of Continuous Variables for Set Ming.* Knowledge and Information Systems, 2001. **3**(4): p. 491-512.
9. Yaping HUANG and S. LUO. *Genetic Algorithm Applied to ICA Feature Selection.* in *Proceedings of the International Joint Conference on Neural Networks*. 2003.
10. Bach, F.R. and M.I. Jordan, *Kernel Independent Component Analysis.* Journal of Machine Learning Research, 2002. **3**.
11. Hyvärinen, A., *Independent Component Analysis:Algorithms and Applications.* Neural Networks, 2000. **13**: p. 411-430.
12. Comon, P., *Independent component analysis, A new concept?* Signal Processing, 1994. **36**: p. 287-314.
13. Fayyad, U. and K.B. Irani. *Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning*. in *Proceeding of 13th International Joint Conference on Artificial Intelligence*. 1993.