Application of biclustering algorithm to extract rules from labeled data

Zhang Yanjie and Sun Hongbo
School of Computer and Control Engineering, Yantai University, Yantai, China

Received 25 January 2018
Revised 10 April 2018
Accepted 12 April 2018

Abstract

Purpose – For many pattern recognition problems, the relation between the sample vectors and the class labels are known during the data acquisition procedure. However, how to find the useful rules or knowledge hidden in the data is very important and challengeable. Rule extraction methods are very useful in mining the important and heuristic knowledge hidden in the original high-dimensional data. It can help us to construct predictive models with few attributes of the data so as to provide valuable model interpretability and less training times.

Design/methodology/approach – In this paper, a novel rule extraction method with the application of biclustering algorithm is proposed.

Findings – To choose the most significant biclusters from the huge number of detected biclusters, a specially modified information entropy calculation method is also provided. It will be shown that all of the important knowledge is in practice hidden in these biclusters.

Originality/value – The novelty of the new method lies in the detected biclusters can be conveniently translated into if-then rules. It provides an intuitively explainable and comprehensive approach to extract rules from high-dimensional data while keeping high classification accuracy.

Keywords Biclustering algorithm, Crowdsourced big data and analytics, Rule extraction

Paper type Research paper

1. Introduction

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as ribosomal RNA, transfer RNA or small nuclear RNA genes, the product is a functional RNA. Gene expression data is a kind of data matrix used to represent the expression level of different genes under specific conditions simultaneously. Each element is a real number which is often the logarithm of the relative abundance of the mRNA of the gene (Madeira and Oliveira, 2004). Usually, in the data matrix, genes are arranged in the row direction, while the column direction represents different time or different environmental conditions.

This work was partially supported by National Key Research and Development Program of China (Grant No.2017YFB1400105), Shandong Province Young Scientists Award Fund No.BS2012SW023 and Shandong Province Young Backbone University Teachers Civil Visiting Scholars Project.

International Journal of Crowd Science
Vol. 2 No. 2, 2018
pp. 86-98
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-01-2018-0002

As an array contains tens of thousands of probes, a microarray experiment can accomplish many genetic tests in parallel. The acquired gene expression data are a typical kind of high-dimensional data. The huge number of features or attributes adds great difficulties to the prediction and interpretability capabilities of all of the models applied to analyze it because of redundant features and noises. How to find the useful rules or knowledge hidden in the data is very important and challengeable. Feature selection is usually a necessary step to facilitate further processing which is especially true for high dimensional data (Gorzałczany and Rudziński, 2017). In machine learning and statistics, attribute selection is the process of selecting a subset of the fewest number of informative attributes for classification, rule extraction and the other applications (Shrivastava and Barua, 2015).

Rule-based expert systems are often applied to classification problems in various application fields, like fault detection, biology and medicine (Dahal *et al.*, 2015; Shrivastava and Barua, 2015). In Roubos *et al.*'s (2003) study, the authors show a compact, accurate and interpretable fuzzy rule-based classifiers obtained from labeled observation data. To implement it, an iterative approach for developing fuzzy classifiers was proposed. The initial model was derived from the data and subsequently, feature selection and rule-base simplification were applied to reduce the model, while a genetic algorithm was used for parameter optimization. Moreover, the researchers proposed different optimization-based methods such as ant colony optimization and particle swarm optimization to extract rules (Chen *et al.*, 2015; Indira and Kanmani, 2015).

Support vector machines (SVMs) are learning systems based on the statistical learning theory and exhibit good generalization ability on different kinds of real data sets (Han *et al.*, 2015). Companying with the study on SVMs, it has gradually turned into a leading machine learning technique and has been applied in a wide range of areas such as bioinformatics, pattern recognition, text classification and so on (Shi *et al.*, 2015). Researchers interested in this topic can easily access to a lot of free software or toolbox. However, the results given by SVMs are usually difficult to explain. In safety-critical or medical applications, an explanation capability is an absolute requirement. A rule extraction method based on SVMs was proposed in Núñez *et al.*'s (2002) study. The authors introduced a SVM plus prototypes procedure for rule extraction. This method allows giving explanation ability to SVMs. Once determined the decision function by means of a SVM, a clustering algorithm was used to determine prototype vectors for each class. These points were combined with the support vectors using geometric methods to define ellipsoids in the input space with minimum overlapping between classes, which were later transferred to if-then rules.

One important analysis task of microarray data concerns the simultaneous identification of groups of genes that show similar expression patterns across specific groups of experimental conditions (Wang *et al.*, 2014; Maulik *et al.*, 2015). Most of time, it is not the sample vectors as integrity shows the strong coherence with each other, but the elements at some specific positions among different sample vectors show the local similarity (Valarmathi *et al.*, 2015). Besides classical clustering methods such as hierarchical clustering, in recent years, biclustering has become a popular approach to analyze biological data sets and a wide variety of algorithms, and analysis methods have been published (Czibula *et al.*, 2015; Shinde and Kulkarni, 2016; Indira and Kanmani, 2015).

Such applications can be addressed by a biclustering process whose aim is to discover biclusters (Cheng and Church, 2000). The so called bicluster is a subset of genes and conditions of the original expression matrix where the selected genes present a coherent

behavior under all the experimental conditions contained in the bicluster. In other words, the data in the same bicluster show a high degree of local similarity. The difference between a bicluster and a submatrix is that all the biclusters are definitely submatrices, but only those submatrices whose row or column vectors satisfying some kind of linear relations will be treated as biclusters. Biclustering algorithms are just a kind of data processing algorithms to find those submatrices lying in the original data matrix showing the local similarity. This technology has found numerous applications in research and applied areas like biology, drug discovery, toxicological study and diseases diagnosis (Alon *et al.*, 1999; Alizadeh *et al.*, 2000; Golub *et al.*, 1999; Pomeroy *et al.*, 2002).

However, the number of biclusters lying in the data, the size and the spatial positioning relations among these biclusters is completely unknown and strongly data dependent (Rabia *et al.*, 2016). In Kaiser and Leisch's (2008) study, the authors introduced the R package which contains a collection of biclustering algorithms, preprocessing methods for two way data and validation and visualization techniques for bicluster results. In Amela *et al.*'s (2006) study, the authors provided the Biclustering Analysis Toolbox, BicAT, as a software platform for clustering-based data analysis that integrates various biclustering and clustering techniques in terms of a common graphical user interface. Furthermore, BicAT provides different facilities for data preparation, inspection and post processing such as discretization, filtering of biclusters according to specific criteria or gene pair analysis for constructing gene interconnection graphs. The toolbox is described in the context of gene expression analysis but is also applicable to other types of data. The authors compared different biclustering techniques with each other with respect to the biological relevance of the clusters as well as with other characteristics such as robustness and sensitivity to noise (Shi *et al.*, 2015; Maulik *et al.*, 2015).

When the biclusters have been detected by applying the biclustering algorithm, the problem is how to translate the biclusters into the corresponding rules. In fact, it can be easily implemented combining with the data discretization schemes. As each bicluster is a submatrix, the line and column numbers that it covers are known. As each experiment condition can be treated as an attribute and all the column numbers of the bicluster can be used as a prerequisite for a rule, a bicluster detection procedure can also be thought as an attribute's selection processing. This kind of rule extraction provides a comprehensive interpretable way compared with the other methods while keeping high classification accuracy.

2. The proposed method

For many pattern recognition problems, the relation between the samples and the classes are known during the acquisition procedure of the data. And this kind of data is called labeled data. Suppose a bicluster B is composed of the row numbers set $\{i_1, i_2,\ldots,i_m\}$ and column numbers set $\{j_1,j_2,\ldots,j_n\}$ of a labeled data matrix D, then the function $\Phi(B) = \{i_1, i_2,\ldots, i_m\}$ is defined to determine the set of those row numbers that the elements of B lies in D and $|\Phi(B)| = m$, so is the definition of the function $\Psi(B) = \{j_1, j_2, \ldots, j_n\}$. The position of a bicluster B lying in the original data matrix D can be determined by $\Phi(B)$ together with $\Psi(B)$.

Usually, the data in the same class show some kind of behavior similarity is called a rule. A rule is applicable only for the sample vectors in the same class. The similarity of sample vectors spanning over the class boundary cannot be thought of the knowledge to distinguish data among different classes. That means biclustering processing results with labeled data are meaningful depending on the class labels. Directly biclustering with D without

considering the labels of the sample vectors does not help to find the biclusters which will be translated into the rules eventually.

2.1 Data discretization

The flowchart of the new rule extraction method is illustrated in Figure 1, which is mainly composed of four sequential processing procedures: data discretization, biclustering processing, bicluster significance evaluation and rule translation based on the discretization schemes. Data discretization is a technique to partition continuous attributes into a finite set of adjacent intervals to generate attributes with a small number of distinct values (Kurgan and Cios, 2004). Discretization algorithms have played an important role in data mining and knowledge discovery (Tsai *et al.*, 2008). They not only produce a concise summarization of continuous attributes to help the experts understand the data more easily but also make learning more accurate and faster (Oliveira, 1999).

Assuming that a dataset consists M examples and S target classes, a discretization algorithm would discretize the continuous attribute a in this dataset into n discrete intervals $\{[d_0, d_1],(d_1, d_2],\ldots,(d_{n-1}, d_n]\}$, where d_0 is the minimal value and d_n is the maximal value of attribute a. Such a discrete result is called a discretization scheme on attribute a. This discretization scheme should keep the high interdependency between the discrete attribute and the target class to carefully avoid changing the distribution of the original data.

As having been introduced before, each column of D can be considered as an attribute or feature no matter what real physical meaning it has. If a sample vector $V \in D$ has value $V(a)$ with respect to the attribute a, then the discretized value $V_D(a)$ of V on a is determined by the discretization scheme. For an example, if it is known that $V(a) \in (d_i, d_{i+1}]$, $i = 0, 1,\ldots,n-1$, then after the discretization processing, the value of $V_D(a)$ will be $i + 1$.

The wine data contains the chemical analysis of 178 wines produced in the same region in Italy but derived from three different cultivars. The problem is to distinguish the three different types based on 13 continuous attributes derived from chemical analysis: alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoids phenols, proanthocyaninsm color intensity, hue, OD280/OD315 of diluted wines and proline (Roubos *et al.*, 2003). In Figure 2, the original wine data with all of its 13 attributes are shown. The wine data are also a kind of high-dimensional data even though their dimensionality is less than the real gene expression data. It has been widely applied in the research studies on machine learning such as attribute selection, pattern recognition and rule extraction. The discretized wine data with all of their 13 attributes are shown in Figure 3. The data discretization schemes are listed in Table I where the method proposed in Tsai *et al.*'s (2008) study is applied.

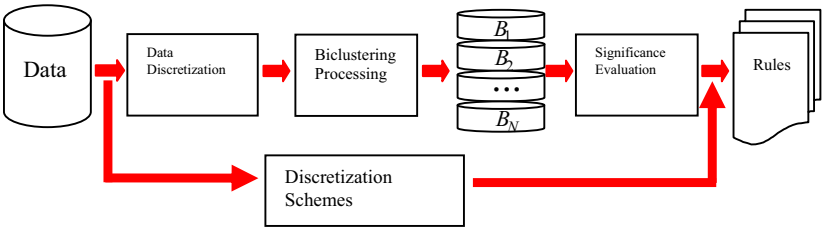

Figure 1.
The whole rule extraction flowchart

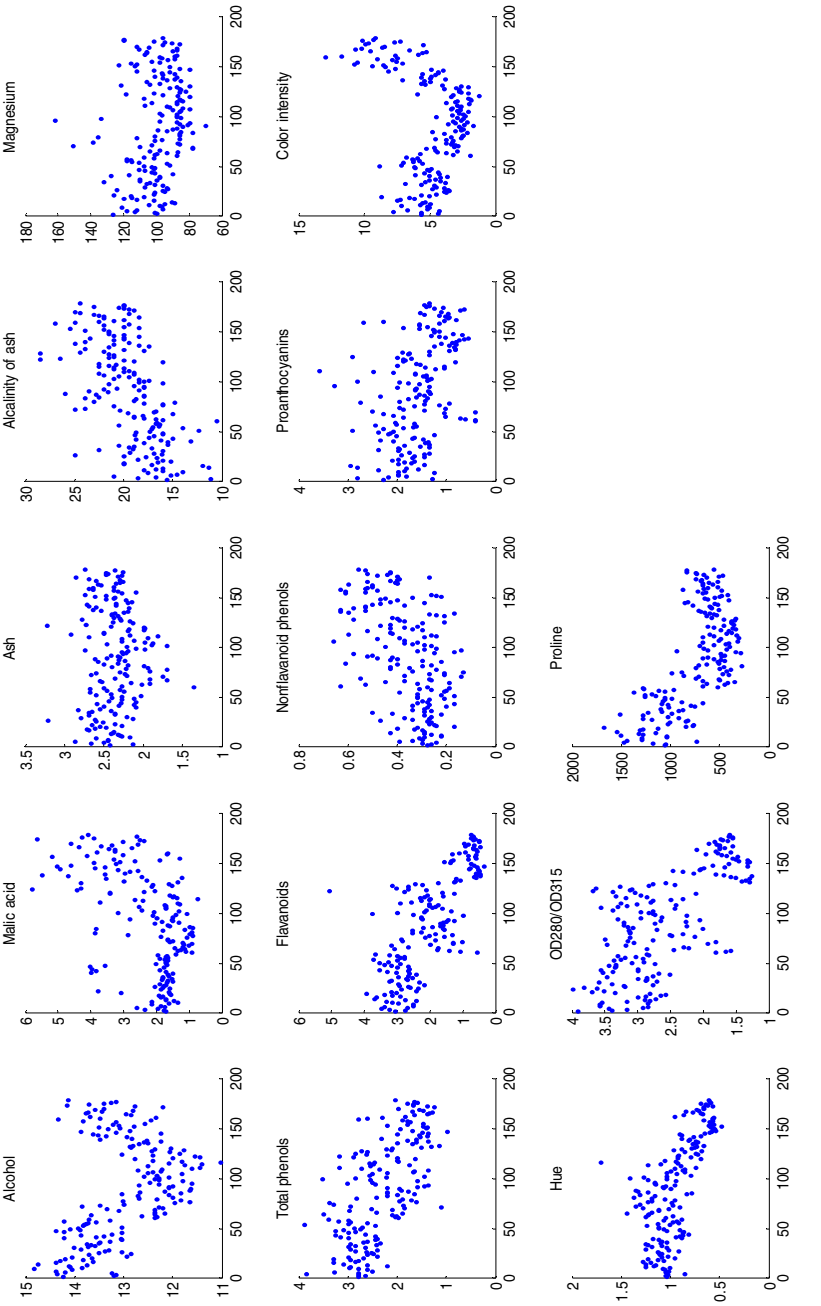

Figure 2.
The original wine data with all of the 13 attributes

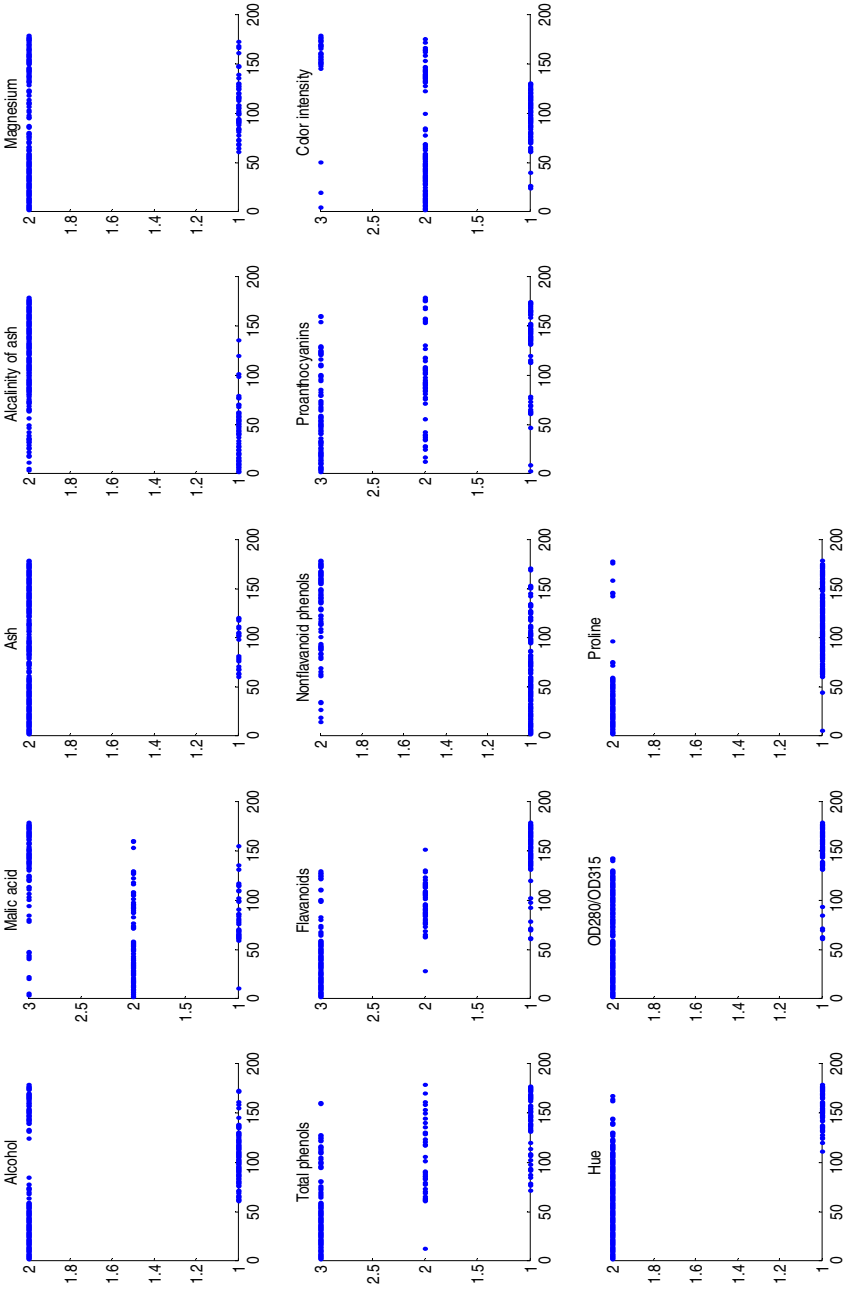

Figure 3.
The discretized wine data with all of the 13 attributes

Table I.
Discretization schemes of all the attributes of the wine data

No.	Attribute name	d_0	d_1	d_2	d_3
1	Alcohol	11.030	12.780	14.830	
2	Malic acid	0.740	1.475	2.235	5.800
3	Ash	1.360	2.030	3.230	
4	Alcalinity of ash	10.600	17.900	30	
5	Magnesium	70	88.500	162	
6	Total phenols	0.980	1.840	2.335	3.880
7	Flavanoids	0.340	1.400	2.310	5.080
8	Nonflavanoid phenols	0.130	0.395	0.660	
9	Proanthocyanins	0.410	1.305	1.655	3.580
10	Color intensity	1.280	3.820	7.550	13
11	Hue	0.480	0.785	1.710	
12	OD280/OD315	1.270	2.115	4	
13	Proline	278	755	1680	

2.2 Criterion for rule extraction

The symbol $B_j|\hat{\omega}, j = 1, 2, \ldots, m$ is used to represent all of the biclusters detected within the sample vectors belonging to the same class $\hat{\omega}$. Then $|\Phi(B_j|\hat{\omega})|$ and $|\Psi(B_j|\hat{\omega})|$ are two very important factors which can be used to evaluate the significance of the bicluster $B_j|\hat{\omega}$. As we expect the number of rules should be as small as possible, the requirement has the meaning in twofold. First, the smallest number of rules means the nature of the data has been well grasped by the rules. Second, the rules should have restriction on all of the vectors within the same class $\hat{\omega}$. In other words, the rules should be adapted to all of the vectors in the same class. The above analysis tells us the first rule selection criterion which is:

$$|\Phi(B_{j_1}|\hat{\omega})| + |\Phi(B_{j_2}|\hat{\omega})| + \cdots + |\Phi(B_{jm}|\hat{\omega})| \geq |\Phi(\hat{\omega})| \tag{1}$$

In equation (1), the subscript variable m means the number of biclusters within the same class $\hat{\omega}$ has been used to extract the corresponding rules. Undoubtedly, the value of m should be as small as possible.

Usually, there are a huge number of biclusters which can be detected. The significance of each detected bicluster must be evaluated by calculating its information entropy. Based on it, the significance of all these biclusters can be sorted in a decreased manner. Applying the first rule selection criteria, the minimum number m can be determined. It must be pointed out that since every bicluster is corresponded to a rule, then minimum number of rules to well express the knowledge hidden in the class $\hat{\omega}$ is also m. Here, we simultaneously draw an important conclusion which is the least number of rules providing 100 per cent of recognition accuracy is data-dependent. We also give the way to determine the exact value of it. However, when the number of rules is fixed, the new algorithm provides the most distinct and convenient way to find the rules while giving the assurance of the maximum accuracy.

Though the introduced above rule extraction method from bicluster is done within each of the class respectively. Whether the combination of rules from different classes can be applied to the whole data should be done with careful consideration. If each column of the data D is treated as an attribute, then it usually has different values. The information entropy corresponding to each attribute within the same class $\hat{\omega}$ can be calculated as

$H(a|\hat{\omega}) = \sum_{i=1}^{N(a|\hat{\omega})} -p_i \ln p_i$, the variable $N(a|\hat{\omega})$ means the number of different values corresponding to the attribute a in the class $\hat{\omega}$. Suppose that all of the sample vectors in D can be classified into n classes, if $H(a|\hat{\omega}) = \sum_{i=1}^{n} \frac{H(a|\omega_i)}{n}$ then the attribute a is of the least importance, ass all of the sample vectors have the same value on it.

2.3 Bicluster significance evaluation

How to select the significant bicluster among the huge number of detected biclusters is very important. Given a data matrix D, each row vector of D can be considered as a sample vector, assume the number of sample vectors lying in D is N, all these samples belong to different classes named ω_i, $i = 1, 2, \ldots, M$. Define N_i is the number of samples in class ω_i, then p_i which means the probability of one sample belonging to the class ω_i can be estimated by N_i/N. The expected information entropy provided by D is:

$$I(N_1, N_2, \cdots, N_M) = -\sum_{i=1}^{M} p_i \log_2 p_i \tag{2}$$

If an attribute A has a number of k values which are $\{a_1, a_2, \ldots a_k\}$, then the whole samples set can be classified into k different subset $S_1, S_2, \ldots, S_k$ by only using attribute A. Assume N_{ij} is the number of samples in the subset S_j which belongs to the class ω_i, then the information entropy of classified result by attribute A is defined as:

$$I(A) = \sum_{j=1}^{k} \left[\left(\frac{N_{1j} + N_{2j} + \cdots + N_{Mj}}{N} \right) I\left(N_{1j}, N_{2j}, \cdots, N_{Mj}\right) \right] \tag{3}$$

where $I\left(N_{1j}, N_{2j}, \cdots, N_{Mj}\right) = -\sum_{i=1}^{M} p_{ij} \log_2 p_{ij}$ and $p_{ij} = \frac{N_{ij}}{|S_j|}$ which means the probability of samples lying in subset S_j belonging to the class ω_i. The whole information gain acquired by attribute A is:

$$Gain(A) = I(N_1, N_2, \cdots, N_M) - I(A) \tag{4}$$

Assume that there is a bicluster B when doing biclustering with E(D), it is a fact that the more row numbers B covers and the less attributes B has, the more meritorious B is. Based on it, we define $\Delta(B) = \frac{|\Phi(\omega)|}{|\Phi(B)|}$ as a weight to indicate the importance of information provided by B. $|\Phi(\omega)|$ is the number of samples belonging to the class ω where the bicluster B is founded. Equation (3) only instructs how to calculate the information entropy with one attribute. As each bicluster B satisfies $|\Psi(B)| \geq 2$, which means we have to take a number of $|\Psi(B)|$ attributes' information entropy into account. For any two different attributes, A_1 and A_2, if $I(A_1) < I(A_2)$, then the values taken by A_1 are more regular than the values taken by A_2. When applying these two attributes to classify an unknown input sample, the classified result based on A_1 will be more accurate than that of A_2. Considering all of

the aforementioned analysis, we define the following formula as an index to evaluate the significance of the bicluster B:

$$IE(B) = \Delta(B) \max_{A \in |\Psi(B)|} Gain(A) \tag{5}$$

2.4 Translation of biclusters to rules

Assume there is a bicluster $B|\hat{\omega}$ lying in the class $\hat{\omega}$ where the number of sample vectors in $\hat{\omega}$ is $\Phi(\hat{\omega}) = N\hat{\omega}$ satisfying $\Phi(B|\hat{\omega}) = \{i_1, i_2, \cdots, i_m\}$, $\Psi(B|\hat{\omega}) = \{j_1, j_2, \cdots, j_n\}$, $m \leq N\hat{\omega}$, $n \leq \Psi(D)$, the bicluster $B|\hat{\omega}$ can be conveniently translated into the corresponding rule accompanied by the data discretization schemes. As there are a number of $|\Psi(B|\hat{\omega})| = n$ attributes in $B|\hat{\omega}$, the translated rule has n antecedents which are related with attributes $a_{j_1}, a_{j_2}, \cdots, a_{j_n}$, respectively. Here, the attribute a_{j_1} is used for explanation. If the data discretization scheme on attribute a_{j_1} is $\{[d_0, d_1], (d_1, d_2], \cdots, (d_{n-1}, d_n]\}_{a_{j_1}}$, as we have known how the data discretization works, the value j_1 on which the attribute a_{j_1} is means the original attributea_{j_1}'s value without doing data discretization belongs to the range $(d_{j_1-1}, d_{j_1}]$; by this way, the first antecedent of the rule is if $a_{j_1} \in (d_{j_1-1}, d_{j_1}]$. Keep on this kind of processing till to the attribute a_{j_n}, the full description of the rule is determined.

3. Computation example

The well-known wine data are applied as the experiment data to illustrate the feasibility and effectiveness of the proposed new method. There are 59 samples in class ω_1, 71 samples in class ω_2 and 48 samples in class ω_3. The number of all of the samples is 178. Each sample vector has 13 numerical attributes whose values are different from each other observably. Compared with the real gene expression data, the wine data have smaller dimensionality, while they are all numerical data with high dimensionality (Wang *et al.*, 2014). The application of wine data will help to save a lot of time to verify the feasibility of the proposed algorithm without destroying the nature of the research data object. And it also facilitates the comparison among the different research results.

Here, the data discretization method proposed in Tsai *et al.*'s (2008) study is applied and the discretization schemes are listed in Table I. The discretized wine data are illustrated in Figure 4. As there are three classes, the discretized data within each class is isolated as a single picture. According to the data discretization scheme, each of the 13 attributes is discretized into less than three intervals which means the discretized data are only composed of three different numbers 1, 2 and 3. Each number is represented by a colorful square for intuitive illustration. The whole processing is followed by the procedures shown in Figure 1.

Biclustering with the discretized wine data, a number of 5,716 biclusters are founded. Among all these biclusters, 217 biclusters are within ω_1, 3,794 biclusters are within ω_2 and 1,705 biclusters are within ω_3. Using equation (5) to calculate every biclusters' IE, biclusters B_1, B_2 and B_3 belonging to the class ω_1, ω_2 and ω_3 are selected respectively. Accompanying with the data discretization schemes listed in Table 1, the selected biclusters can be translated into three rules as follows:

(1) $\Psi(B_1) = \{1, 3, 5, 11, 12, \omega_1\}, |\Phi(B_1)| = 59, \frac{|\Phi(B_1)|}{\Phi(\omega_1)} = 100\%$, each row vector of B_1 is (2, 2, 2, 2, 2, 1), B_1 is translated into the rule as: for a sample vector $\vec{\alpha}$, if $a_1 \in$ (12.78, 14.83] and $a_3 \in$ (2.03, 3.23] and $a_5 \in$ (88.5,162] and $a_{11} \in$ (0.785, 1.71] and $a_{12} \in$ (2.115, 4] then $\vec{\alpha} \in \omega_1$.

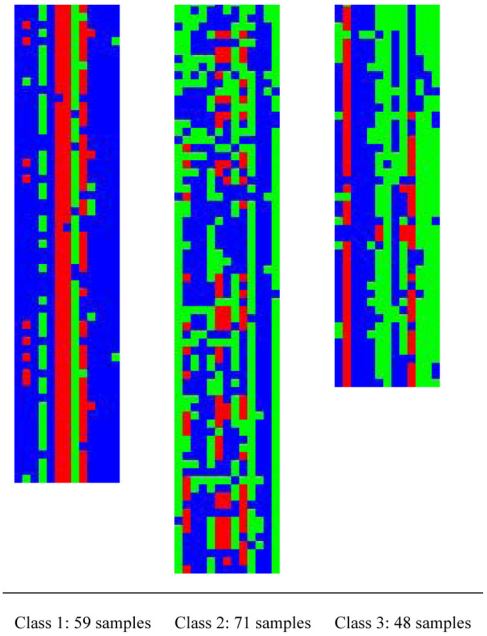

Figure 4. Illustration of discretized wine data (each color square corresponds to a discretized number)

(2) $\Psi(B_2) = \{13, \omega_2\}, |\Phi(B_2)| = 67, \frac{|\Phi(B_2)|}{\Phi(\omega_2)} = 94.37\%$, each row vector of B_2 is (13, 2), B_2 is translated into the rule as: for a sample vector $\vec{\alpha}$, if $a_{13} \in (278, 755]$ then $\vec{\alpha} \in \omega_2$.

(3) $\Psi(B_3) = \{3, 4, \omega_3\},\} |\Phi(B_3)| = 47, \frac{|\Phi(B_3)|}{\Phi(\omega_3)} = 97.92\%$, each row vector of B_3 is (2, 2, 3), B_3 is translated into the rule as: for a sample vector $\vec{\alpha}$, if $a_3 \in (2.03, 3.23]$ and $a_4 \in (17.9, 30]$ then $\vec{\alpha} \in \omega_3$.

As these three biclusters totally cover 173 sample vectors out of the whole 178 sample victors, the three extracted knowledge offer a recognition accuracy of 97.19 per cent.

4. Conclusions

Rule extraction methods as an approach which tries to find the useful knowledge hidden in the high-dimensional data are very useful. The so-called rules are in practice, and some sample vectors in the data show coherent similarity with each other. Because the data in the same bicluster are closely related to each other, the transition from bicluster to rule has a natural consistency. As the elements of a bicluster are just lying in the original data and can be conveniently translated into a corresponding rule, the results of the new method have good explanation ability. The difference of the new method with the other methods lies in it applies the biclustering algorithm to discover the local similar biclusters existing among the original data matrix.

In the large amount of bicluster results detected, a specially modified information entropy calculation method is provided to evaluate the significance of all the detected biclusters. Then, all of the biclusters can be sorted in a decreased manner according to their information entropy values. By this way, those most significant biclusters within the samples belonging to the same class individually can be selected to extract the rules. We will

try to deal with more different types of data and compare the results with the results of existing literature. Processing with real gene expression data is ongoing and will be presented in the future work.

References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Jr, Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M. (2000), "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", *Nature*, Vol. 403 No. 6769, pp. 503-511.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999), "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 96 No. 12, p. 6745.

Amela, P., Prelić, B., Philip, Z., Anja, W., Peter, B. and Wilhelm, G. (2006), "A systematic comparison and evaluation of biclustering methods for gene expression data", *Bioinformatics*, Vol. 22 No. 9, pp. 1122-1129.

Chen, L., Sun, Y. and Zhu, Y. (2015), "Extraction methods for uncertain inference rules by ant colony optimization", *Journal of Uncertainty Analysis and Applications*, Vol. 3 No. 1, pp. 1-19.

Cheng, Y. and Church, G.M. (2000), "Biclustering of expression data", *8th International Conference on Intelligent Systems for Molecular Biology 2000*, Vol. 8, pp. 93-103.

Czibula, G., Czibula, I.G., Sîrbu, A.M. and Mircea, I.G. (2015), "A novel approach to adaptive relational association rule mining", *Applied Soft Computing*, Vol. 36, pp. 519-533.

Dahal, K., Almejalli, K., Hossain, M.A. and Chen, W. (2015), "Ga-based learning for rule identification in fuzzy neural networks", *Applied Soft Computing*, Vol. 35, pp. 605-617.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999), "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, Vol. 286 No. 5439, pp. 205-214.

Gorzałczany, M.B. and Rudziński, F. (2017), "Interpretable and accurate medical data classification – a multi-objective genetic-fuzzy optimization approach", *Expert Systems with Applications*, Vol. 71, pp. 26-39.

Han, L., Luo, S., Yu, J., Pan, L. and Chen, S. (2015), "Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes", *IEEE Journal of Biomedical and Health Informatics*, Vol. 19 No. 2, pp. 728-734.

Indira, K. and Kanmani, S. (2015), "Association rule mining through adaptive parameter control in particle swarm optimization", *Computational Statistics*, Vol. 30 No. 1, pp. 251-277.

Kaiser, S. and Leisch, F. (2008), "A toolbox for bicluster analysis in R", *Department of Statistics: Technical Reports*, available at: http://epub.ub.uni-muenchen.de/3293/

Kurgan, L.A. and Cios, K.J. (2004), "Caim discretization algorithm", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16 No. 2, pp. 145-153.

Madeira, S.C. and Oliveira, A.L. (2004), "Biclustering algorithms for biological data analysis: a survey", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 1 No. 1, pp. 24-45.

Maulik, U., Mallik, S., Mukhopadhyay, A. and Bandyopadhyay, S. (2015), "Analyzing large gene expression and methylation data profiles using StatBicRM: statistical biclustering-based rule mining", *Plos One*, Vol. 10 No. 4.

Núñez, H., Angulo, C. and Català, A. (2002), "Rule extraction from support vector machines", *Eurorean Symposium on Artificial Neural Networks, Bruges*, Vol. 80, pp. 107-112.

Oliveira, J.V.D. (1999), "Semantic constraints for membership function optimization", *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, Vol. 29 No. 1, pp. 128-138.

Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y.H., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S. and Golub, T.R. (2002), "Prediction of Central nervous system embryonal tumour outcome based on gene expression", *Nature*, Vol. 415 No. 6870, p. 436.

Rabia, A., Verma, C.K. and Namita, S. (2016), "A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data", *Genomics Data*, Vol. 8, pp. 4-15.

Roubos, H., Setnes, M. and Abonyi, J. (2003), "Learning fuzzy classification rules from labeled data", *Information Sciences*, Vol. 150 Nos 1/2, pp. 77-93.

Shi, Y., Zhang, L., Tian, Y. and Li, X. (2015), "Knowledge extraction from support vector machines", Intelligent Knowledge, pp. 101-111.

Shinde, S. and Kulkarni, U. (2016), "Extracting classification rules from modified fuzzy min–max neural network for data with mixed attributes", *Applied Soft Computing*, Vol. 40, pp. 364-378.

Shrivastava, A. and Barua, K. (2015), "An efficient tree based algorithm for association rule mining", *International Journal of Computer Applications*, Vol. 117 No. 11, pp. 31-32.

Tsai, C.J., Lee, C.I. and Yang, W.P. (2008), "A discretization algorithm based on class-attribute contingency coefficient", *Information Sciences*, Vol. 178 No. 3, pp. 714-731.

Valarmathi, M.L., Siji, P.D. and Mohana, S. (2015), "Efficient association rule mining based on correlation analysis", *International Journal of Applied Engineering Research*, Vol. 10 No. 11, pp. 29367-29384.

Wang, H.Q., Jing, G.J. and Zheng, C. (2014), "Biology-constrained gene expression discretization for cancer classification", *Neurocomputing*, Vol. 145 No. 18, pp. 30-36.

Further reading

Asadi, S. and Shahrabi, J. (2016), "ACORI: a novel ACO algorithm for rule induction", *Knowledge-Based Systems*, Vol. 97, pp. 175-187.

Hartigan, J.A. (1972), "Direct clustering of a data matrix", *Journal of the American Statistical Association*, Vol. 67 No. 337, pp. 12079-12084.

Mousavi, S., Esfahanipour, A. and Zarandi, M.H.F. (2014), "A novel approach to dynamic portfolio trading system using multitree genetic programming", *Knowledge-Based Systems*, Vol. 66 No. 1, pp. 68-81.

About the authors

Zhang Yanjie was born in Wendeng, Shandong Province, China on November 16th, 1975. In 2004, he got his doctor's degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. In 2001, he got his master's degree in information and control engineering, and in 1998, he got his bachelor's degree in mathematics and systematic science from Shandong University, Jinan, China. He is an Associative Professor of Yantai University, Shandong, China now. His research interests include pattern recognition, bioinformatics and information security. Zhang Yanjie is the corresponding author and can be contacted at: yanjie.zhang@126.com

Sun Hongbo became a Member (M) of IEEE in 2010. Hongbo was born in Fuping, Shannxi Province, China, on 13th February, 1977. In 2011, he got his doctor's degree in control science from Tsinghua University, Beijing, China. In 2005, he got his master's degree in software engineering of Tsinghua University, and in 1998, he got his bachelor's degree in information science from Beijing Institute of Technology, China. He is a Lecturer of Yantai University, Shandong, China now. He has been a Postdoctoral Researcher in Department of Automation, Tsinghua University, Beijing, China, from July 2011 to June 2014. From November 2009 to November 2010, he has worked in National Research Council Canada as an International Visiting Worker. Between 2005 and 2006, he served as a Software Engineer of National CIMS ERC, Tsinghua University, Beijing, China. And during 1998 to 2002, he was an Assistant Professor of Shenyang Institute of Technology, Liaoning, China. His research interests include system integration, artificial intelligence, algorithm, large-scale simulation and e-commerce.