Neurocomputing 145 (2014) 30–36

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Biology-constrained gene expression discretization for cancer classification

Hong-Qiang Wang [a,*], Gao-Jian Jing [b], Chunhou Zheng [c]

[a] Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Science, P.O. Box 1130, Hefei, Anhui 230031, China
[b] School of Mechanical and Automotive Engineering, Hefei University of Technology, Hefei, China
[c] College of Electrical Engineering and Automation, Anhui University, Hefei, China

ARTICLE INFO

Article history:
Received 22 December 2013
Received in revised form
31 March 2014
Accepted 7 April 2014
Available online 2 July 2014

Keywords:
Data discretization
Gene expression
Gene regulation
Cancer classification
High-throughput technology

ABSTRACT

In this paper, we propose a biology-constrained gene expression discretization method based on class distribution diversity. Inspired by the intrinsic relationship between gene expression and gene regulation, we constrain gene expression discretization to be of at most three discrete states and locate cut points using a regulatory states-guided mechanism. To take advantage of class label information, we define class distribution diversity (CDD) for an interval and devise three supervised discretization rules. The proposed method is very cost-efficient and simple to implement in practice. In the experiments, we evaluated the proposed method using four publicly available gene expression datasets involving four types of cancer: leukemia, prostate, lymphoma and liver cancer, and compared with two previous methods, Fayyad and Irani's (FI) and EBD. The experimental results show the effectiveness and efficiency of the proposed method.

1. Introduction

With the advent of high-throughput biological technology, an increasing amount of OMICs data are being generated [1,2]. Although the data are rich with information of biological system and potentially useful for deciphering cancer pathology, they are typically high-dimensional and noisy, thus posing an unprecedent knowledge discovery challenge [3]. Gene expression profiles, for example, have been proven to be more efficient to diagnose and classify cancer than traditional histological data, provided that they are properly preprocessed. Among existing data preprocessing methods, discretization transforms continuous data to be in a discrete form by reductionism and tends to yield more concise and more accurate decision rules [4–7]. On the other hand, useful information may be wrongly discarded during discretization, and it is challenging to develop an efficient gene expression discretization method that minimizes loss of cancer-related information.

Generally, data discretization can perform in a supervised or unsupervised manner. They differ in whether or not class membership information is used in forming discrete intervals. An unsupervised method does not use such information, and its two representative examples are equal-width (EW) and equal-frequency (EF) methods [8]. EW partitions the range of variables' values based on a prefixed interval width while EF based on sample fraction quantity. Although unsupervised methods are simple and take a relatively low computational cost, they are vulnerable to outliers and the results obtained are often unsatisfactory in practice. In contrast, supervised methods tend to be more sophisticated by incorporating class membership information and usually yield classifiers that have superior performance [8–10]. A supervised discretization method generally consists of two key steps: 1) scoring the goodness of a set of intervals and 2) searching for a good-scoring set of intervals in the discretization solution space. The scoring functions can be derived from statistics or informatics, such as χ^2-based measures [6,11] and entropy-based scores [12,13]. Besides the dichotomy of superivised or unsupervised, discretization methods can also be categorized into dynamic vs static, global vs local, splitting (top-down) vs merging (bottom-up) or direct vs incremental. Readers can refer to literatures [4,5] for more details.

FI, developed by Fayyad and Irani [10], is one of most commonly used discretization methods in practice. The method is supervised, in which a discretization solution is scored by using the entropy of the target variable that is induced by the solution and a recursive partitioning strategy based on minimum description length (MDL) is employed to find optimal discrete intervals in a greedy manner. The searching greediness often causes FI to trap at a local minimum and it is not guaranteed to find a globally

* Corresponding author. Tel./fax: +86 55165592751.
E-mail address: hqwang126@126.com (H.-Q. Wang).

http://dx.doi.org/10.1016/j.neucom.2014.04.064
0925-2312/

optimal discretization solution. Recently, Boulle [14] introduced Bayesian theory and developed a Bayesian score to assess the goodness of a discretization solution. In contrast to the entropy-based FI score, the Baysian score (BS) incorporates domain knowledge on the predictor variable to assess a discretization solution. Based on BS, the authors devised a new discretization method named MODL. However, MODL suffers from the forced assumption of uniform prior probability distribution over discretization solutions, and is not applicable in many practical cases. To overcome the assumption limitation, Lustgarten et al. introduced two priors, structure and parameter priors, to have a flexible calculation of BS. Specifically, the parameter prior is used to control the multi-normal distribution of the target variable in each interval and the structure prior to guide the selection of the number of intervals and the location of the cut points in a discretization solution. The improved MODL was named efficient Bayesian discretization (EBD). In addition to the improved calculation of BS, EBD also has a lower time complexity of $O(n^2)$, where n is the number of instances, than MODL ($O(n^3)$), which makes EBD more applicable in practice.

To our knowledge, there exists no method that can exploit a priori biological knowledge for discretizing gene expression data. In this paper, we propose a biology-constrained gene expression discretization method motivated by the intrinsic relationship of gene's expression and regulation. Biologically, the expression levels of a gene are often regulated in response to the endogenous or exogenous stimuli of cells. For simplicity, complex regulatory activity is often categorized into three states, down-regulated, non-regulated and up-regulated [15]. In light of the taxonomy of regulation activity, we argue that gene expression can be discretized to at most three basic intervals that associate with the three regulatory states. We incorporate this as a biological constraint into gene expression discretization to not only simplify the discretization but also make the discretization biologically understandable. On the other hand, we follow the supervised line described above to increase the efficiency of discretization. As a result, class distribution diversity (CDD) is defined to measure the discriminative power of an interval and three CDD-based discretization criteria devised. The use of the criteria make the proposed method free to iterative searches as in most previous methods and lead to a low computational cost.

The rest of the paper is organized as follows. In Section 2, we first review related biological knowledge on gene regulation and expression, and then present our method in detail. In Section 3, we evaluate the proposed method using four real-world gene expression data sets and compare it with two previous methods, FI and EBD. The influences of the parameters on the proposed method are also discussed in this section. Finally, we conclude the paper.

2. Methods

In order to adapt to and survive in variable environments, cells often actively regulate their gene expression to maintain a physiological balance. Therefore, regulatory states largely influence expression levels in a cell. In genetics, a gene can be in one of three regulatory states, *i.e.*, down-regulated (DR), non-regulated (NR) and up-regulated (UR) in a particular cellular status [16,17]. So, we reason that the whole expression range of a gene can be divided into three natural segments closely related to three regulatory states, possibly named DR-related, NR-related and UR-related, and these segments can be located from left to right along the expression range. Because differential regulatory patterns of a gene in a cell are deemed to be responsible for different cellular phenotypes, these segments can guide the seeking for discrete intervals that are responsible for the distinction of different phenotypes of interest. The two boundaries between two immediate neighbor segments would be potential candidates cut points for a discriminative discretization. Based on the logics, we devise our biology-constraint gene expression discretization method in the following sections.

2.1. Definition of class distribution diversity for a half-open interval

For convenience, we consider a binary cancer classification problem. Note that a multi-class classification problem can be handled by converting it into multiple binary problems. Let N_1 and N_2 represent the sample sizes of the two classes, class 1 and class 2. Given a left-side-half-open interval of a gene predictor, $v=(-\infty, l]$, we define its class distribution diversity (CDD), denoted by $D(v)$, w.r.t the two classes as

$$D(v)=\frac{n_1(v)}{N_1}-\frac{n_2(v)}{N_2} \tag{1}$$

where $n_1(v)$ and $n_2(v)$ represent the numbers of samples belonging to class 1 and 2 in the interval v, respectively. Note that a CDD can be positive or negative value, of which the positive indicate that class 1 dominates the interval and class 2 does otherwise. When l slides along the range from left to right, a series of intervals v_i can be obtained with the corresponding CDDs D_i calculated by Eq. (1).

2.2. The property of CDD

It can be imagined that for a binary problem, there could be three representative situations of regulatory state distribution between the two classes: i) The two classes share a same regulatory pattern, as shown in Fig. 1A; ii) The two classes have completely different regulatory patterns, as shown in Fig. 1B; iii) One class is in non-regulatory state while the other is both in down-regulatory state and in up-regulatory state, as shown in Fig. 1C. Assuming that the expression of a gene is normally distributed under a regulatory state, we simulated the gene expression distributions in the three regulation situations above. In each case, we uniformly divided the whole expression range into $m=50$ segments to form 50 left-side-half-open intervals upper-bounded by the right end of m segments. The CDDs for the intervals were calculated by Eq. (1) and are plotted in Fig. 1D.

First, for case i, all the intervals have a very small absolute value of CDDs due to the non-significant difference of class distributions, as shown in Fig. 1D. One can reason that genes with such kind of CDD distribution patterns would be non-informative to the class distinction and the expression range should be discretized into one state. Second in case ii, in sharp contrast, the CDD curve has a remarkable peak between the two regulatory states, as shown in Fig. 1D. It can be reasoned that such genes would be closely relevant to the class distinction and the expression range can be discretized into two parts that are separated by the peak. Third, compared with cases i and ii, case iii has a little complex CDD distribution, where two turning points appear at the boundaries of two adjacent regulatory states, as shown in Fig. 1D. The two turning points correspond to the maximum and minimum values of CDDs, respectively. We reason that in this case, the gene is also relevant but not as much as case ii to the class distinction, and the expression range can be discretized to three parts around the two turning points.

2.3. Three discretization criteria based on class distribution diversity

Assume that the expression range of a gene is uniformly divided into m ($m \geq 50$) segments. Let l_i, $i=1,2,\ldots,m$, denote the upper-boundaries of these segments, we can have m half open

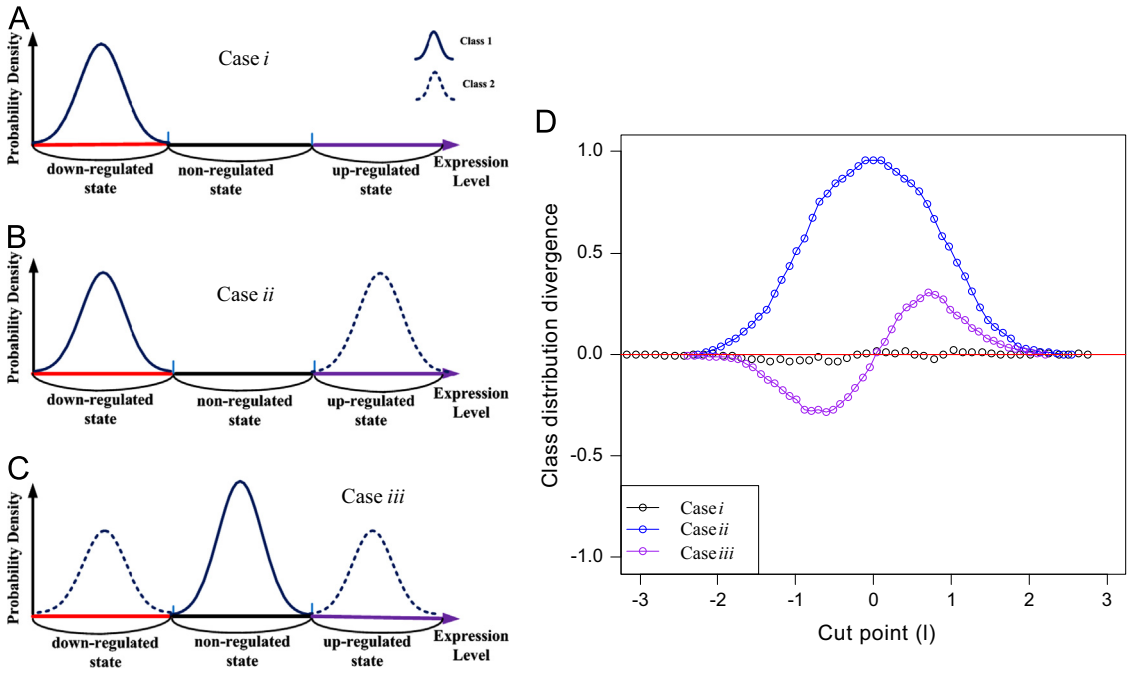


Fig. 1. Illustration of the property of CDD. In A, the gene is in a same regulatory state for the two classes and the expression distributions in both classes are $N(0,0.5)$. In B, the gene is in down-regulatory state for class 1 and is in non-regulatory state for class 2, and the expression distributions in both classes are $N(-1,0.5)$ and $N(1,0.5)$, respectively. In C, the gene is in non-regulatory state for class 1 and is both in down-regulatory and up-regulatory state for class 2, and the expression distribution is $N(0,0.5)$ in class 1 and a mixture distribution of $N(-1,0.5)$ and $N(1,0.5)$ in class 2.

intervals, $V_i=(-\infty, l_i)$, each having a CDD D_i calculated by Eq. (1). Without loss of generality, let $V_{\max}$ and $V_{\min}$, upper-bounded by $L_{\max}$ and $L_{\min}$, respectively, be the two intervals with the maximum and minimum CDDs, $D_{\max}$ and $D_{\min}$, respectively. Given a discriminative gene, the values of $D_{\max}$ and $D_{\min}$ of it will be in one of the following three cases:

1) $D_{\max}>0$, $D_{\min}=0$
2) $D_{\max}=0$, $D_{\min}<0$
3) $D_{\max}>0$, $D_{\min}<0$

and the two boundaries, $L_{\max}$ and $L_{\min}$, could be potentially the cut points for a reasonable discretization solution. To characterize the overall discriminative power of the gene, we further define a global CDD Δ as the difference between $D_{\max}$ and $D_{\min}$, i.e.

$$\Delta=|D_{\max}-D_{\min}| \tag{2}$$

The global CDD is actually the CDD for the interval of $(L_{\max}, L_{\min}]$ for $L_{\max}\le L_{\min}$ or that of $(L_{\min}, L_{\max})$ for $L_{\max}>L_{\min}$. The larger the Δ is, the more discriminative to the class distinction the gene is. Given a proper constant $0<\alpha<1$, if $\Delta<\alpha$, the gene can be said to be non-informative to the classification and consequently, its expression range can be discretized into one state. Otherwise, the gene can be said to be discriminative and its expression range would be discretized into two or three states. Note that the number of discrete states is biology-constrained to be no more than three. In this case, if $|D_{\max}|>|D_{\min}|$, $L_{\max}$ would be naturally a cut point and $L_{\min}$ can be another cut point if the absolute value of $D_{\min}$ is large enough, and vice versa. So, given a proper constant $0<\lambda<\alpha$ and let

$$d=min(|D_{\max}|,|D_{\min}|), \tag{3}$$

where min means to take a minimum value, we can reason that if the following equation holds

$$d\ge\lambda \tag{4}$$

$L_{\max}$ and $L_{\min}$ will be two cut points for a reasonable discretization solution.

Following the description above, three gene expression discretization criteria based on CDD can be summarized as follows.

Criterion 1. *Given two constants,* $0<\lambda<\alpha<1$, if $\Delta<\alpha$ or $\max(|D_{\max}|,|D_{\min}|)<\lambda$, the expression levels of gene g will have one discrete state in the whole range. Otherwise, the expression levels will have two or more discrete states.

Criterion 1 determines whether the expression levels are discretized into one discrete state or not. A gene being with one discrete state means that the gene follows a same regulatory mechanism in both classes. On the other hand, Criterion 1 implies that genes with enough large Δ are relevant to a class classification and should be discretized into more than one discrete state. For such kind of genes, the following two criteria should be applied for further discretization:

Criterion 2. *Given two constants,* $0<\lambda<\alpha<1$, if $\Delta\ge\alpha$, $\max(|D_{\max}|,|D_{\min}|)\ge\lambda$ and $\min(|D_{\max}|,|D_{\min}|)<\lambda$, the expression levels of gene g will be discretized into two states in the whole range. The two corresponding discrete intervals are $(-\infty, a)$ and $[a, +\infty)$, respectively, where

$$a=\begin{cases} L_{\max} & if \quad |D_{\max}|>|D_{\min}| \\ L_{\min} & if \quad |D_{\max}|\le|D_{\min}| \end{cases} \tag{5}$$

Criterion 3. *Given two constants,* $0<\lambda<\alpha<1$, if $\Delta\ge\alpha$ and $\min(|D_{\max}|,|D_{\min}|)\ge\lambda$, the expression levels of gene g will be discretized into three states in the whole range. The three corresponding discrete intervals are $(-\infty, a)$, $[a, b]$, and $(b, +\infty)$, respectively, where $a=\min(L_{\min}, L_{\max})$, $b=\max(L_{\min}, L_{\max})$.

Remark 1. Following the three criteria, the expression range of a gene can be discretized so that the gene is empirically most discriminative to the class distinction of interest.

Remark 2. In the three criteria, the parameter α plays a critical role in the discretization for determining whether the gene is discretized into one state or more. Too small values of α will over-discretize a non-informative gene while too large values will under-discretize a truly differentially expressed gene. Another parameter λ is used to dichotomize the discretization into two or three states. The selection of the two parameters is likely data-dependent and our numeric experiments have showed that value of α around 0.5 and value of λ around 0.1 can be a reasonable choice for many practical cases.

In summary, the proposed method can be shown in Fig. 2

3. Experimental results

To evaluate the proposed method, we collected four expression data sets of different types of cancer, Leukemia [18], Prostate [19], Lymphoma [20] and Liver [21]. These data sets were widely used as benchmark data sets for algorithm evaluation in bioinformatics area. In the Leukemia data, all the samples (72) are categorized into two classes: 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML). In the Prostate data, all the samples (102) are categorized into 50 normal and 52 prostate cancer. In the Lymphoma data, all the samples (77) are grouped into 58 DLBCL and 19 FL. In the Liver data, all the samples (60) are grouped into 20 non-recurrent and 40 recurrent liver cancer. The numbers of gene variables in the four data sets are 7129, 7129, 12600 and 7129. In the experiments, we performed 10-fold cross validation on the whole data 10 times for each data set. In each run, we learned a discretization model for each gene using 9 folds as a training set and applied the learned discretization model to discretize both the training set and the rest fold (test set). For the proposed method, we set the two parameters to be $\alpha=0.5$ and $\lambda=0.1$ (the settings will be discussed in Section 3.3). For comparison, two previous methods, FI and EBD, were simultaneously applied to the four data sets. FI is commonly used as a standard algorithmic benchmark for data discretization in practice while EBD is a recently developed method which has been proved to work well for gene expression data discretization [6].

Algorithm

Input: a set of n tissue samples and the expression values of a gene in the n samples

Output: Discrete intervals

BEGIN

Step.1 Initializing m,α and λ

Step.2 Finding the maximum (max) and minimum (min) expression values across the n samples;

Step.3 Uniformly dividing the expression range [min, max] into m parts and forming m left-side-half-open intervals;

Step.4 Calculating *the* CDDs for the m intervals by Eq.(1) and locating the intervals with the maximum and minimum CDDs;

Step.5 Calculating the global CDD by Eq.(2);

Step.6 Discretizing the expression range by Criterion 1. If the condition for two or more states holds, go to Step 7, and Step.9 otherwise;

Step.7 Discretizing the expression range by Criterion 2. If the condition for 3 states holds, go to Step 8, and Step.9 otherwise;

Step.8 Discretizing the expression range by Criterion 3;

Step.9 Outputting the discrete intervals obtained in Step. 6, 7 or 8.

END

Fig. 2. Algorithm of the proposed method.

3.1. Distribution of numbers of discrete states

We first statistically summarized the discretized results by FI, EBD and our method for the four data sets, as shown in Table 1. When a gene is discretized into one discrete state, it means that the gene is indiscriminative to the cancer classification. The mean proportions of gene variables with one state by our method are 0.93, 0.99, 0,92 and 0.97 for leukemia, prostate, lymphoma and liver data sets, respectively, which are consistently larger than those by FI (0.89, 0.78, 0.87 and 0.98) and EBD (0.87,0.77,0.75 and 0.97), as shown in Table 1. For genes with >1 discrete states, our method obtained larger mean numbers of states per gene for all the four data sets (2.1 for Leukemia, 2.27 for Prostate, 2.13 for Lymphoma and 2.52 for Liver), as shown in Table 1, suggesting the better power of discovering more complex expression patterns. Finally, we compared the mean number of states per gene among the three discretization methods on each data set, as shown in Table 1. The number reflects the overall complexity of a discretization. From Table 1, it can be seen that our methods obtained the smallest mean numbers per gene almost for all 4 data set, confirming the superior performance of our method in distinguishing non-informative (one state) and informative (more than one states) genes.

We also examined the discretized results for individual genes by three methods. Take the leukemia data as an example. As we know, it has been previously reported that three genes, DEK, FUS and HOXA9, are closely related to AML but not ALL [22]. According to the fact, a reasonable discretization method should discretize these genes into more than one discrete states. Table 2 lists the mean numbers of discrete states obtained by our method, FI and EBD. From Table 2, it can be seen that our method discretized three genes into 1.91, 1.81 and 2.88 states on an average, respectively, while both previous methods, FI and EBD, discretized the three genes into <1.58 states for all the three genes. Obviously, our methods obtained more reasonable discretizations for the three truly discriminative genes. In addition to the three leukemia-relevant genes, Table 2 also reports the discretized results of

Table 1
Statistics of the discretized results by our method, FI and EBD for the leukemia, prostate, lymphoma and liver data sets.

Dataset	Mean fraction of genes with 1 discrete state			Mean # of discrete states per gene with >1 discrete states			Mean # of discrete states per gene		
	Our method	FI	EBD	Our method	FI	EBD	Our method	FI	EBD
Leukemia	0.93	0.89	0.87	2.1	2.01	2.01	1.07	1.11	1.13
Prostate	0.99	0.78	0.77	2.27	2.13	2.31	1.0	1.25	1.99
Lymphoma	0.92	0.87	0.75	2.13	2.01	2.01	1.08	1.13	1.12
Liver	0.97	0.98	0.97	2.52	2.02	2.02	1.03	1.02	1.01

Table 2
Numbers of discrete states of five individual genes for the leukemia data by our method, FI and EBD.

Probe sets	Gene symbol	Description	Our method	FI	EBD
X64229_at	DEK	DEK oncogene	1.91	1.5	1.40
X71428_at	FUS	Fused in sarcoma	1.81	1.5	1.56
U41813_at	HOXA9	Homeobox A9	2.88	1.16	1.08
M22898_at	TP53	Tumor protein p53	1.67	1.02	1.03
U15131_at	ST5	Suppression of tumorigenicity 5	2.00	1.98	2

another two generic cancer genes, TP53 and ST5 [23,24] by three methods. Although FI and EBD worked reasonably well for ST5, they improperly discretized TP53 to nearly one state on average. Taken together, these results confirm that our method is powerful in reliably finding discriminative genes and discretizing gene expression reasonably well.

3.2. Classification performance of the discretized results

Classification performance is another important index for evaluating a discretization method. To evaluate the classification performance of the proposed method, we employed two popular classifiers, C4.5 and naïve Bayes (NB), which were among the top 10 of data mining algorithms in [25]. In the experiment, we considered two measures of classification performance, correct classification rate (CCR) and discretization generalization (DG). CCR is defined as the proportion of correct class predictions on test samples by the classifier and DG is defined as the ratio of the CCR on a test set to that on the training set. Generally, DG evaluates discretized predictors directly, while CCR provides an indirect measure of discretization by considering the performance of classifiers learned from the discretized predictors.

Table 3 shows the mean CCRs for our method, FI and EBD on the four data sets. From this table, it can be seen that our method obtained higher mean CCRs for almost all the data sets except the liver data, regardless of which classifier is used. The liver data contain fewer samples relative to other three data sets, which is likely the reason for the relatively low accuracies for the data set. Table 3 also revealed that given the discretized results by either of FI and EBD, different classifiers led to a very large disparity of CCRs. For example, for FI, NB obtained a CCR that is 20.82% higher than that of C4.5 for the liver cancer data and for EBD, C4.5 obtained a CCR which is 14.35% higher than that of NB for the leukemia data, as shown in Table 3. On the contrary, irrespective of which classifier is used, our method led to similar CCRs for all the four data sets, suggesting that our method is favorably classification model-free. Table 4 compares the mean DGs among our method, FI and EBD on the four data sets. The larger the value of DG is, the better the generalization of discretization is. From Table 4, we see that our method obtained the highest DG among the three methods for all the four data sets, suggesting the better discretization generalization of our method. Overall, our method can obtain more favorable classification performances than the two previous methods, FI and EBD.

3.3. Influences of the parameters α and λ on discretization performance

The proposed method has two important tunable parameters, α and λ, which can influence the discretization performance. We experimentally examined the influence of the two parameters using the four gene expression data sets, as shown in Fig. 3. As described above, the proposed method first uses α to dichotomize the expression range into one or more than one discrete states. We varied α among {0.1, 0.2, 0.3, 0.4, 0.5, 0.7 and 0.9} and observed the changing curves of CCR (NB classifier), as shown in Fig. 3A. Note that in order to diminish the effect of another parameter λ, we tried λ={0.01, 0.05,0.1,0.15,0.2,0.3,0.4 and 0.5} for each value of α and averaged the obtained CCRs. From Fig. 3A, it can be found that CCRs significantly increase with α for $\alpha<0.5$ for all the data sets and almost fix for $\alpha>0.5$. Note that no CCR values are available when all the genes are discretized with one state. The decreasing CCRs for small values of α (e.g. $\alpha<0.5$) should be because non-discriminative genes were overdiscretized according to Criterion 1. The observations suggest that a value of α around 0.5 could be an acceptable choice in practice.

We next observed the changing curves of CCR with λ by fixing α=0.5, as shown in Fig. 3B. We varied λ={0.01,0.05,0.1,0.15, 0.2,0.3,0.4 and 0.5}. From Fig. 3B, it can be seen that CCRs did not change as much as that in Fig. 3A, and the value of λ around 0.1 could be generally proper in practice.

3.4. Time complexity

Time complexity is often used for evaluating a discretization method. In particular, fast discretization is especially necessary to analyze high-dimensional gene expression data in practice. In the proposed method, the most intensively computational part is searching for the maximum/minimum gene expression values across samples, as shown in Fig. 2. The time complexity is right that of a maximum/minimum searching algorithm, i.e., $O(n)$. As we know, the time complexities of FI and EBD are $O(n\log n)$ and $O(n^2)$, where n is the number of instances. So, the linear complexity suggests that our method is far more cost-efficient in computation

Table 3
Mean CCRs and standard error of the mean (% ± SEM) of our method, FI and EBD for the leukemia, prostate, lymphoma and liver data sets.

Classification model	C45			NB		
Dataset	Our method	FI	EBD	Our method	FI	EBD
Leukemia	88.00 ± 0.91	95.67 ± 0.82	**96.67** ± 1.10	**96.10** ± 0.80	80.28 ± 1.51	82.32 ± 1.17
Prostate	**84.11** ± 0.80	83.78 ± 0.68	81.21 ± 0.58	**85.08** ± 1.10	89.76 ± 0.75	83.76 ± 0.91
Lymphoma	**88.63** ± 1.80	71.25 ± 1.45	72.43 ± 1.32	**87.15** ± 1.3	85.45 ± 1.22	86.22 ± 1.41
Liver	**61.76** ± 2.7	50.00 ± 2.03	60.00 ± 2.08	67.50 ± 1.4	70.82 ± 1.49	**72.33** ± 1.42

Table 4
Mean discretization generalizations and standard error of the mean (% ± SEM) of our method, FI and EBD for the leukemia, prostate, lymphoma and liver data.

Classification model	C45			NB		
Dataset	Our method	FI	EBD	Our method	FI	EBD
Leukemia	**86.53** ± 1.02	80.58 ± 1.42	83.58 ± 1.34	**97.76** ± 0.79	95.89 ± 0.92	**97.76** ± 1.12
Prostate	**89.47** ± 1.4	82.65 ± 0.57	79.38 ± 0.57	**97.72** ± 1.70	91.81 ± 0.83	88.91 ± 0.72
Lymphoma	**89.85** ± 2.0	77.17 ± 1.79	73.17 ± 1.66	**94.94** ± 2.4	84.11 ± 1.38	92.57 ± 1.39
Liver	**62.15** ± 2.1	55.11 ± 2.06	55.50 ± 2.16	68.10 ± 3.4	71.67 ± 1.43	**70.94** ± 1.48

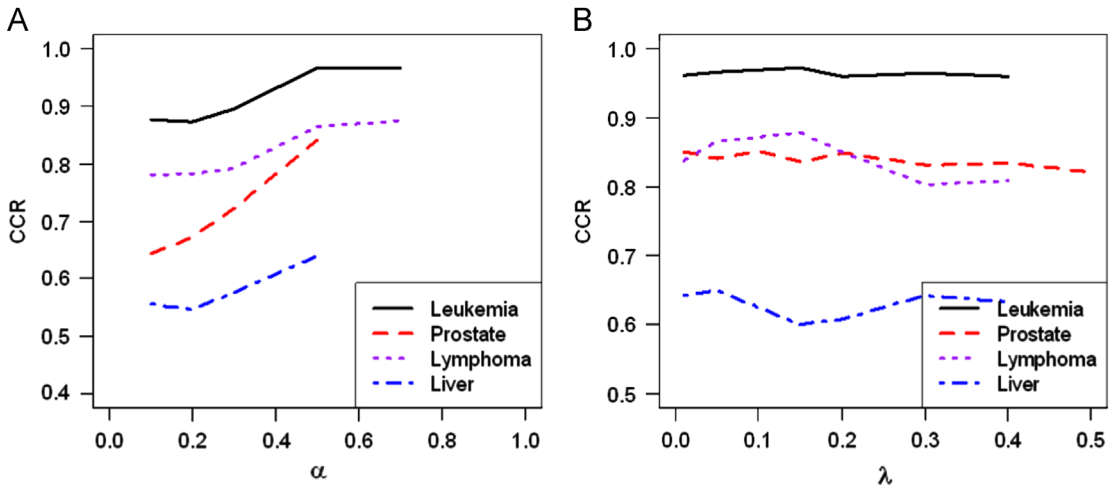


Fig. 3. Changing curves of CCR with α (A) and λ (B) on the leukemia, prostate, lymphoma and liver cancer data sets.

Table 5
Mean CPU times (second) per training fold of our method, FI and EBD on the leukemia, prostate, lymphoma and liver cancer data sets.

Method	Leukemia	Prostate	Lymphoma	Liver
Our method	**6.86**	**13.59**	**8.16**	**5.96**
FI	387.50	1248.31	517.12	352.50
EBD	886.23	3321.42	1278.41	947.41

than FI or EBD. On the other hand, our method does not employ any greedy-searching or dynamic programming process as in most of traditional discretization methods including FI and EBD. It is notorious that such iterative procedures can lead to a very heavy computational burden. So, the lack of them can also speed up our method. Table 5 reported the CPU times of our method, FI and EBD for four data sets, Leukemia, Prostate, Lymphoma and Liver. All the programs were in R codes and run on personal computer with window XP system with Intel (R) CPU @2.19 GHz and Ram 2.0 GB. Table 4 clearly reveals that our method can be faster by orders of magnitude than the two previous methods.

4. Conclusions

In this paper, we have proposed a biology-constrained gene expression discretization method based on class distribution diversity and evaluated it on four benchmark gene expression data sets, Leukemia, Prostate, Lymphoma and Liver cancer. There exists no method that can exploit a priori biological knowledge for discretizing gene expression data. Inspired by the tri-state paradigms of gene regulation, the proposed method constrains gene expression to be discretized into at most three states and locates discrete intervals through maximizing class distribution diversity in association with gene regulation property. Two adjustable parameters are provided to users, which make the proposed method flexible to various data scenarios. Another advantage of the proposed method is the linear time complexity of it, which makes the proposed method very cost-efficient and especially favorable for analyzing high-dimensional OMICs data in practice. The experimental results on four real-world data sets showed the effectiveness and efficiency of the proposed method.

No method always performs best in any data scenario. The proposed method obeys this law. Because the current calculation of CDD is only applicable to binary classification problems, our method is currently not applicable to multi-class data scenarios straightforward. On the other hand, the tri-state regulatory scheme adopted is a simplified assumption about gene regulatory activity. So, our method is not guaranteed to deal with more complex regulatory patterns.

It is well-known that a discretization process can reveal intrinsic properties of the variable and can lead to a knowledge-level representation. Because the proposed method naturally relates discrete intervals to the tri-states gene regulatory activity, the discretized results can provide potential clues about gene regulatory mechanisms underlying gene expression data. Therefore, the proposed method can help to recover gene regulatory networks from gene expression data. In future work, we will extend the proposed method to gene regulatory network analysis.

Acknowledgment

This work was supported by the grants of the National Natural Science Foundation of China, Nos. 61374181, 61272339, and 61300058, by Anhui Provincial Natural Science Foundation Grant no. 1408085MF133 and by K. C. Wong education foundation.

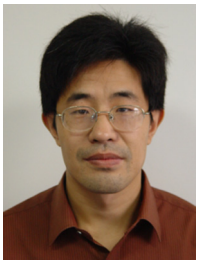
References

[1] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary microarray, Science 270 (1995) 467–470.
[2] M. Garber, M.G. Grabherr, M. Guttman, C. Trapnell, Computational methods for transcriptome annotation and quantification using RNA-seq, Nat. Methods 8 (2011) 469–477.
[3] C.L. Nutt, D.R. Mani, R.A. Betensky, P. Tamayo, J.G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M.E. McLaughlin, T.T. Batchelor, P.M. Black, A.v. Deimling, S. L. Pomeroy, T.R. Golub, D.N. Louis, Gene expression-based classification of malignant gliomas correlates better with survival than histological classification, Cancer Res. 63 (2003) 1602–1607.
[4] H. Liu, F. Hussain, C. Tan, M. Dash, Discretization: an enabling technique, Data Min. Knowl. Discov. 6 (2002) 393–423.
[5] S. Garcıa, J. Luengo, J.A. Saez, V. Lopez, F. Herrera, A survey of discretization techniques: taxonomy and empirical analysis in supervised learning, IEEE Trans. Knowl. Data Eng. 25 (2013) 734–749.
[6] J. Lustgarten, S. Visweswaran, V. Gopalakrishnan, G. Cooper, Application of an efficient Bayesian discretization method to biomedical data, BMC Bioinform. 12 (2011) 309.
[7] D. Johnstone, C. Riveros, M. Heidari, R. Graham, D. Trinder, R. Berretta, J. Olynyk, R. Scott, P. Moscato, E. Milward, Evaluation of different normalization and analysis procedures for illumina gene expression microarray data involving small changes, Microarrays 2 (2013) 131–152.
[8] J. Dougherty,R.Kohavi andM. Sahami, Supervised and unsupervised discretization of continuous features, in: A. Prieditis, S,J. Russell (Eds.), Proceedings of the International Conference on Machine Learning, Tahoe City, California, USA, (1995), pp. 194–202.
[9] R. Kohavi, M. Sahami, Error-based and entropy-based discretization of continuous features, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (1996), pp. 114–119.

[10] Usama M. Fayyad, Keki B. Irani, Multi-interval discretization of continuous-valued attributes forcClassification learning, in: Proceedings of the International Joint Conference on Uncertainty in AI (1993), pp. 1022–1027.
[11] G.V. Kass, An exploratory technique for investigating large quantities of categorical data, Appl. Stat. 29 (1980) 119–127.
[12] Y. Kodratoff, J. Catlett, On changing continuous attributes into ordered discrete attributes, Machine Learning-EWSL-91, Springer, Berlin Heidelberg (1991) 164–178.
[13] A. Quinlan, I. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (2010) 841–842.
[14] M. Boulle, MODL: a bayes optimal discretization method for continuous attributes, Mach. Learn. 65 (2006) 131–165.
[15] H.-S. Wong, H.-Q. Wang, Constructing the gene regulation-level representation of microarray data for cancer classification, J. Biomed. Inform. 41 (2008) 95–105.
[16] H.Q. Wang, D.S. Huang, Regulation probability method for gene selection, Pattern Recognit. Lett. (2006) 116-112.
[17] D. Li, R., Li, H.Q., Wang Novel Discretization Method for microarray-based cancer classification, in: Proceedings of the ICIC 2012, LNCS 7389, (2012) 327–333.
[18] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[19] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2002) 203–209.
[20] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, T.R. Golub, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nat. Med. 8 (2002) 68–74.
[21] N. Iizuka, M. Oka, H. Yamada-Okabe, M. Nishida, Y. Maeda, N. Mori, T. Takao, T. Tamesa, A. Tangoku, H. Tabuchi, K. Hamada, H. Nakayama, H. Ishitsuka, T. Miyamoto, A. Hirabayashi, S. Uchimura, Y. Hamamoto, Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection, Lancet 361 (2003) 923–929.
[22] ⟨http://www.bioinformatics.org/legend/leuk_db.htm⟩, (Dec. 16, 2013).
[23] P.A.J. Muller, K.H. Vousden, p53 mutations in cancer, Nat. Cell Biol. 15 (2013) 2–8.
[24] J.H. Lichy, W.S. Modi, H.N. Seuanez, P.M. Howley, Identification of a human chromosome 11 gene which is differentially regulated in tumorigenic and nontumorigenic somatic cell hybrids of HeLa cells, Cell Growth Differ. 3 (1992) 541–548.
[25] X. Wu, V. Kumar, Data Mining and Knowledge Discovery, The Top Ten Algorithms in Data Mining, Chapman & Hall/CRC Press, Boca Raton, 2009.

Hong-Qiang Wang is currently an associate professor in the Intelligent Computing Lab, Institute of Intelligent Machines, Chinese Academy of Science. He received his bachelor and master degrees of engineering from Hefei University of Technology, and his Ph.D. degree in the Department of Automation, University of Science & Technology of China (USTC), HeFei, China. He has also held research positions in City University of Hong Kong, the Hong Kong Polytechnic University, and University of Georgia. His research interests include machine learning, pattern recognition, neural networks and bioinformatics.

Gao-Jian Jing is currently a graduate student in School of Mechanistic and Automotive Engineering, Hefei University of Technology. He received the B.Eng. degree in School of Mechanistic and Automotive Engineering. His research interests include machine learning, pattern recognition, neural network and bioinformatics.



Chun-Hou Zheng received the B.Sc degree in Physics Education in 1995 and the M.Sc. degree in Control Theory & Control Engineering in 2001 from QuFu Normal University, and the Ph.D degree in Pattern Recognition & Intelligent System in 2006, from University of Science and Technology of China. From Feb. 2007 to Jun. 2009 he worked as a Postdoctoral Fellow in Hefei Institutes of Physical Science, Chinese Academy of Sciences. From Jul. 2009 to Jul. 2010 he worked
as a Postdoctoral Fellow in the Dept. of Computing, The Hong Kong Polytechnic University. He is currently a Professor in the College of Electrical Engineering and Automation, Anhui University, China. His research interests include Pattern Recognition and Bioinformatics.