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