Computers in Biology and Medicine 38 (2008) 886–893

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

Incorporating PCA and fuzzy-ART techniques into achieve organism classification based on codon usage consideration

Kun-Lin Hsieh[a,*], I-Ching Yang[b]

[a]Department of Information Management, National Taitung University, 684, Sec. 1, Chung - Hua Road, Taitung, Taiwan
[b]Department of Natural Science Education, National Taitung University, 684, Sec. 1, Chung-Hua Road, Taitung, Taiwan

Received 6 November 2006; accepted 19 May 2008

Abstract

To recognize the DNA sequence and mine the hidden information to achieve the classification of organisms are viewed as a difficult work to biologists. As we know, the amino acids are the basic elements to construct DNA. Hence, if the codon usage of amino acids can be analyzed well, the useful information about classification of organisms may be obtained. However, if we choose too many amino acids to perform the clustering analysis, the high dimensions also lead the clustering analysis to be a complicated structure. Hence, in this study, we will incorporate the principle component analysis and fuzzy-ART clustering techniques into constructing an integrated approach. The useful information about organisms classification based on the codon usage can be mined by using the proposed approach. Finally, we also employ a case including 18 bacteria to demonstrate the rationality and feasibility of our proposed approach.

Keywords: Principle component analysis (PCA); Fuzzy adaptive resonance theory (fuzzy-ART); Codon usage

1. Introduction

The classification of organisms can make us to understand the origin of lives. Until now, many studies and researches had focused to address such issue [1–8]. As we know, the coding structure of DNA sequence was frequently used to discuss or to study since meeting the issue of classifying different organisms. Hence, the similarity analysis or clustering analysis of the DNA sequence will be a worthy study to address the classification problem. However, most techniques with the quantitative characteristics cannot be directly employed to DNA sequence. Restated, the transformation for DNA sequence will be necessary action for the subsequent analysis. According to the philosophy of organism evolution, the useful messages can be transferred from DNA to mRNA, and then it will be transferred from mRNA to protein. Next, such useful messages can also be transferred from mRNA to protein via codon types. Generally, each amino acid can match to codon-one at least or it can match to codon-six at most. The codon to encode the same amino acid will be called as the synonymity codon. The frequency of using synonymity codon during the encoding process of protein may be different [9–13], and the particular organism or gene generally focus on one or several specific synonymity codon. Although the codon will be recognized to be a complicated case, it still hidden the important meanings [14,15], e.g. the information providing the recommendation to the classification problem.

In this study, we initially intend to transfer DNA sequence into a quantitative structure based on codon usage. Then, we will apply the transferred form into performing the subsequent clustering analysis. As we know, if we choose too many amino acids to make analysis, it will lead the clustering analysis to meet the case with the multip dimensions [16]. It will cause the clustering analysis to be a complicated operation. Hence, we will also intend to combine the techniques with the dimension reduction characteristic. Hence, we will propose an integrated approach based on soft computing concept, which will incorporate the dimension reduction and clustering technique to resolve the organism classification. Finally, an illustrative data including 18 bacteria will be applied to demonstrating the rationality and feasibility of our proposed approach.

[*] Tel.: +886 89 318855x1250; fax: +886 89 321981.
E-mail address: klhsieh2644@mail2000.com.tw (K.-L. Hsieh).

0010-4825/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2008.05.007

2. Background review

2.1. Principle component analysis (PCA)

The philosophy of principle component analysis (PCA) can be denoted as that summarizing all parameters to make the necessary analysis since facing such problem [16–19]. PCA will be frequently viewed as one technique to reduce the dimension of problem. Restated, the practitioner can apply PCA to transfer those parameters with the high correlation into the few independent parameters (or it will be called as the principle component term) and the variation of the original data can be still explained well. Those few principle component terms will be the index to explain the summarization of parameters. The equation of PCA can be given as follows:

$$\begin{aligned} \mathrm{PC}(1) &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ \mathrm{PC}(2) &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ \mathrm{PC}(m) &= a_{m1}X_1 + a_{m2}X_2 + \cdots + a_{mp}X_p \end{aligned} \tag{1}$$

where PC(1), PC(2), ..., PC(m) will denote the first principle component term, the second principle component term, ..., the n-th principle component term. The summarization characteristics will be represented according to the coefficients $a_{11}, \ldots, a_{1p}$ in the linear equation. As for the philosophy to determine those principle terms will be recommended as "The capability of variation explanation for the first principle component term will be the largest one and the capability of the remainder variation for the second principle component term will be the largest one." If we take such philosophy into performing analysis, we will get m ($m \leqslant p$) principle component terms and the generalized form will be given as follows:

$$\mathrm{PC}(m) = am1X1 + am2X2 + \cdots + ampXp \tag{2}$$

where X_j, $j = 1, 2, \ldots, P$, and we can re-write it into the following equation:

$$Y = \beta 1X1 + \beta 2X2 + \cdots + \beta pXp \tag{3}$$

2.2. Fuzzy adaptive resonance theory (fuzzy-ART)

The behavior in fuzzy adaptive resonance theory (fuzzy ART) lends itself well to simple geometrical interpretation owing to an internal representation of category prototypes as hyperrectangles in the input space. As for the category choice process, by which fuzzy ART always responds the same way to a familiar input: it recalls the smallest hyperrectangle containing this input [20]. Hyperrectangle overlaps have been argued to be an inconvenience if categories are mutually exclusive [21]. In order to learn intersecting and overlapping categories, a neural network must be capable of repressing previously known categories while it forms new ones. In other words, it must be able to make temporary abstraction of previous knowledge. The generalization would allow the learning of the tulips category first, and the flowers category next, whereas discrimination would allow the reverse. In the case of fuzzy ART, increasing the value of a network parameter called vigilance allows formation of new, more specific categories intersecting broad ones that are already known. The network is thus capable of discrimination. However, reducing the same parameter value does not yield generalization. This is due to the predilection of fuzzy ART for the smallest hyperrectangle containing the input. To avoid a category proliferation problem that could otherwise occur [22,23] recommend input normalization by a procedure called complement coding. Let a be an M-dimensional vector $(a_1, a_2, \ldots, a_M)$, where $0 \leqslant a_i \leqslant 1$. The complement coded input I is obtained as $I = (a_1, a_2, \ldots, a_M, 1 - a_1, 1 - a_2, \ldots, 1 - a_M) = (a, a_c)$. Assign to each category j a vector $w_j = (w_{j1}, w_{j2}, \ldots, w_{j2M})$ of adaptive weights. Each category is initially uncommitted, and its weights are initialized to one. The functionality of fuzzy ART may be described as a three-step algorithm [24]:

Step 1. *Category choice*: Upon presentation of an input I, a choice function T_j is computed for each category j.

Step 1. The norm operator $|\bullet|$ is defined as $|x| = \sum_{i=1}^{2M} |x_i|$, the symbol ^ denotes the fuzzy AND operator, that is, $x \wedge y = (\min(x_1, y_1), \ldots, \min(x_{2M}, y_{2M}))$, and α is a user-defined parameter, $\alpha > 0$. The category J for which the choice function is maximal, that is, $T_j = \max\{T_j,\ j = 1, 2, 3, \ldots\}$ is chosen for the vigilance test.

Step 2. *Vigilance test*: The similarity between w_J and I is compared to a parameter ρ called vigilance, $0 \leqslant \rho \leqslant 1$, in the following test:

$$T_j = \frac{|\boldsymbol{I} \wedge \boldsymbol{w}_j|}{\alpha + |\boldsymbol{w}_j|} \tag{1}$$

If the test is passed, then resonance occurs and learning takes place. If the test is failed, then mismatch reset occurs: the value of T_j is set to -1 for the duration of the current input presentation, another category is chosen in Step 1, and the vigilance test is repeated. Categories are searched, that is, chosen and then tested, until one that meets (1) is found. This category is said to be selected for I. It is either already committed or uncommitted, in which case it becomes committed during resonance.

Step 3: *Resonance*: Resonance makes reference to the internal dynamics of the neural network as it pays attention to the vector $(I \wedge w_J)$. During resonance, the weight vector w_J of the selected category I is updated according to the equation

$$w_J^{(\mathrm{new})} = \beta(I \wedge w_J^{(\mathrm{old})}) + (1 - \beta) w_J^{(\mathrm{old})} \tag{2}$$

where β is a learning rate parameter, $0 \leqslant \beta \leqslant 1$. What is learned is not the input I itself, but rather an attended weight vector $(I \wedge w_J^{(\mathrm{old})})$: fuzzy ART thus learns prototypes, rather than exemplars. The special case $\beta = 1$ is called fast learning and is assumed throughout this