Two Approaches to Gene Expression Analysis: Coring Clusterization versus SVM

Fatima, Adilova
Institute of Mathematics
Tashkent, Uzbekistan
e-mail:
fatima_adilova@rambler.ru

Rifkat, Davronov
Institute of Mathematics
Tashkent, Uzbekistan
e-mail: rifqat@rambler.ru

Anna, Tomskova
Institute of Mathematics
Tashkent, Uzbekistan
e-mail:
tomskovaanna@gmail.com

ABSTRACT
Recent progress in data mining has led to the development of numerous efficient and scalable methods for retrieval patterns in large biological databases. The question is how to bridge these two fields, data min-ing and bioinformatics for successful data processing. This paper presents the use of core clustering and modified SVM methods. Results of both methods are compared in gene expression task solving.

Keywords
Core Clusterization, Classification, Support Vector Machine.

1. INTRODUCTION
A critical problem in biodata analysis is to classify biosequences or structures based on their features and functions. The interaction among attributes of biological object could be very complicated and very often has graph representation. The clusterization is one of the popular tools for understanding the relationships among various conditions and the features of various objects. Typical methods include Bayesian classification, neural networks, SOM, support vector machines (SVMs), the k nearest neighbor (KNN), associative classification, etc. In [4] a new clustering method was proposed applicable to either weighted or unweighted graphs in which each cluster consists of a highly dense core region surrounded by a region with lower density. The nodes belonging to dense cores of cluster then divided into groups, each of which is the representative of one cluster. These groups are finally expanded into complete clusters covering all the nodes of the graph. The support vector machine (SVM) has been one of the most popular classification tools in bioinformatics [5]. The main idea of SVM is that the points of the two classes cannot be separated by a hyper-plane in the original space. These points may be trans-formed to a higher dimensional space so that they can be separated by a hyperplane. In SVM, the kernel is in-troduced so that computing the separation hyperplane becomes very fast. Saddle point search algorithm re-quires finding projections on intersection of cube and plane. The goal of our study was to compare these two approaches, improving them in some modifications, de-scribed below and testing both on task of gene expres-sion problem. The paper consists of 3 sections. In sections 1 and 2 we remind setting of coring clusteriza-tion and SVMs problems. In section 3 we show results of comparative computations on benchmark data, and the section 4 presents the conclusion.

2. PRELIMINARIES
2.1 Coring clusterization
Let us consider an undirected proximity graph

$$G = (V, E, W),$$

where V is a set of nodes, E is a set of edges, W is a matrix with entry w_{ij} being the weight of the edge between nodes i and j. In proximity graphs, V represents a set of data objects, $w_{ij} \geq 0$ represents the similarity of the objects i and j. A higher value of w_{ij} reflects a higher degree of similarity. Thus, applying a graph clustering method proximity graph will produce a set of subgraphs, such that each subgraph corresponds to a group of similar objects, which are dissimilar to objects of groups corresponding to other subgraphs. We assume that every cluster of the input graph has a region of high density called a 'cluster core', surrounded by sparser regions (non-core). The nodes in cluster cores are denoted as 'core nodes', the set of core nodes as the 'core set', and the subgraph consisting of core nodes as the 'core graph'.

For each node i of $H \subseteq V$, the local density at i is defined as

$$d(i, H) := \frac{\sum_{j \in H} w_{ij}}{|H|}. \quad (1)$$

The node with the minimum local density in H is referred to as the weakest node of H :

$$arg\, min_{i \in H}\, d(i, H).$$

We define the minimum density of H as

$$D(H) = \min_{i \in H} d(i, H)$$

to measure the local density of the weakest node of H. Let us consider the greedy procedure that iteratively com-putes $D(G)$ and removes the weakest nodes from G. By ana-lyzing the variation of the minimum density value D, we can identify core nodes located in the dense cores of clus-ters. That is, if the weakest node is in a sparse region, the D value will increase when this node is removed. On the other hand, if the removal of the weakest node causes a significant drop in D value, then this node is highly connected with a set of stronger nodes in a high density region. It is potentially a core node because its removal greatly reduces the density of nodes around it.

Our contribution to this method is to change the function in (1) to

$$d(i, H) = \frac{\max_j w_{ij}}{|H|}.$$

This is correct, because the function's property of monotony remains valid.

2.2 SVM

The standard Support Vector Machine (SVM) problem in learning classification is as follows. We denote by $\langle x_1, x_2 \rangle$ the inner product of given vectors x_1 and x_2. Suppose that we have a learning sample:

$$\{x_i, y_i\},\ x_i \in \mathbb{R}^n,\ y_i \in \{-1, 1\},\ i = 1, ..., l,$$

where $\mathbb{R}$ is the set of real numbers, $l \in \mathbb{N}$ and $\mathbb{N}$ is the set of natural numbers.

The standard formulation of SVM problem is:

$$\begin{array}{c} \min\limits_{w,b,\delta_i} \left(\frac{1}{2}\|w\|^2 + C \sum\limits_{i=1}^{l} \delta_i \right) \\ y_i \left(\langle w, x \rangle + b \right) \geq 1 - \delta_i \\ \delta_i \geq 0,\ C > 0,\ i = 1, ..., l. \end{array} \qquad (2)$$

The solution w^*, b^*, δ^* of (2) gives an optimal hyperplane $\langle w^*, x \rangle + b^* = 0$. Our contribution to the SVM method is that we preliminary calculated a significance of all variables based on Kullback-Leibler divergence [3].

3. COMPUTING EXPERIMENTS

In order to compare the results of the methods described above we used the problem of gene expression analysis [4]. The problem of tissue clustering aims to find connections between gene expressions and condition of tissues to predict the condition of a tissue based on its gene expressions. The database used in the experiment is publicly available at: www.microarray.princeton.edu/oncology/affydata/index.html. This data contains 62 samples including 40 tumor and 22 normal colon tissues. Each sample consists of a vector of 2000 gene expressions. We will set aside the sample labels (tumor/normal) and cluster the samples based on the similarities between their gene expressions. Ideally, the task was to partition the sample set into two clusters such that one contains only tumor tissues and the other contains only normal tissues.

3.1 Coring clusterization

The proximity graph constructed from the gene expression vectors is a complete graph of 62 nodes. Because relative values are more important than absolute values in gene expressions computing, edge weights are computed based on the Pearson correlation coefficient. Specifically, the weight function is defined by:

$$w_{ij} = \frac{1}{2000} \sum_{k=1}^{2000} \frac{1}{s_i s_j} (i_k - m_i)(j_k - m_j),$$

where i_k and j_k are kth gene expressions of samples i and j, m_i, m_j, s_i, s_j are means and standard deviations of i_ks and j_ks. Initially, the coring method identified 12 core nodes. The dendrogram of these core nodes exposes two well-separated groups, one contains 10 nodes and the other has 2 nodes. Expanding these cluster cores yields two clusters. One has 40 samples consisting of 37 tumor and 3 normal tissues. The other contains 22 samples consisting of 3 tumor and 19 normal tissues.

Figure 1 shows the comparison of clustering results by the coring method, and results from [1] and [2]. The result of [2] consists of 6 clusters, but joining clusters 1, 4 and 5 into one group of normal tissues and 2, 3 and 6 into another group of tumor tissues yields a clustering similar to the result of [1].

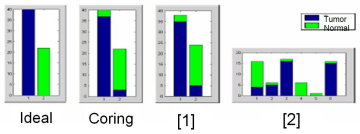

Figure 1: Comparison of the coring method with methods from [1] and [2]

3.2 SVM method

Suppose parent matrix $X = \{x_{ij}\}_{i,j=1}^{n,m}$, $n, m \in \mathbb{N}$, where element x_{ij} means availability of jth feature at ith sample. First, we centralize and normalize this matrix according to the following formulas:

$$a_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

$$b_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - a_j)^2},\ j = 1, ..., m$$

$$y_{ij} = \frac{x_{ij} - a_j}{b_j}.$$

Then received matrix $Y = \{y_{ij}\}_{i,j=1}^{n,m}$, $n, m \in \mathbb{N}$, will pos-sess the following properties:

$$\frac{1}{n} \sum_{i=1}^{n} y_{ij} = 0,$$

$$\frac{1}{n} \sum_{i=1}^{n} y_{ij}^2 = 1,$$

for every $j = 1, ..., m$.

Then matrix Y was processed using a standard SVM method [5]. On base of training set (32 samples) Δ-margin plane was developed. This computing experiment resulted the six errors, particularly 2 samples from the first class, and 4 samples of the second class. Therefore the quality of partition can be estimated as 93, 75 % on train-ing sample and 86, 67 % on testing sample. Applying Kullback-Leibler divergence formula, we designed the new matrix with weighted variables for which over were used standard SVM. But the results we received were different: only four errors, 2 errors from the first class and 2 of the second. Naturally, the accuracy of partition became equal to 93,75 % on training set and 93,33 % on test-ing samples.

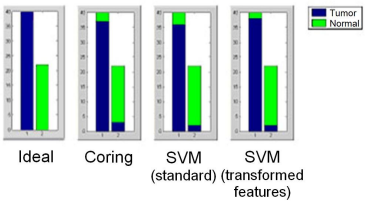

Figure 2: Comparison of coring method with SVM standard method and modified procedures

4. CONCLUSION

We tested core clustering method in two variants and SVM method in standard and modified form. Experiments with coring clusterization demonstrated good clustering results; the method is simple and fast, but definition the values of two free parameters needs in future research. Core nodes can represent informative data objects and also make the method robust to noise.

Standard SVM method gives the same results as coring clusterization, but after transformation of initial features space based on Kullback-Leibler divergence, the accuracy of partition improved on 7,2%.

Thus, we can conclude that coring clusterization gives more possibilities for interpretation, it is more robust to noise, but SVM used in transformed space of initial vari-ables is more accurate.

REFERENCES

[1] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array", *Proc. Natl. Acad. Sci. USA*, 96(12):6745-6750, 1999.

[2] A. Ben-Dor, R. Shamir and Z. Yakhini, "Clustering gene expression patterns", *Journal of Computational Biology*, 1999.

[3] S. Kullback, R. A. Leibler, "On Information and Sufficiency", *Annals of Mathematical Statistics* 22 (1), 79-86, 1951.

[4] V. Thang, A. Casimir, I. Kulikowski, B. Muchnik, "Coring Method for Clustering a Graph", *DIMACS Technical Report*, 2008.

[5] V. N. Vapnik, "The Nature of Statistical Learning Theory", *New York*, 1995.