Support Vector Clustering of Microarray Gene Expression Data

Biter Bilen

Bilkent University, Faculty of Science, Department of Molecular Biology and Genetics, 06800 Bilkent, Ankara, TURKEY

biter@bilkent.edu.tr

7, April 2005

ABSTRACT

In this paper Gaussian kernel approach will be explored for microarray gene expression data using support vector clustering (SVC). SVC uses the idea of support vector machines. The data points are mapped to a high dimensional feature space with a kernel function, and a minimal enclosing sphere in looked for. Cluster boundaries in data space are complex shapes and are formed from the sphere boundary points in the feature space. The performance of the algorithm and the biological implications will be demonstrated as a future work.
Key words: support vector clustering, gene expression.

INTRODUCTION

The genes in a living organism function collaboratively. However traditional methods in molecular biology generally work on a single gene in one experiment, hence the throughput is very limited and determination of the whole picture is hard. In the last several years, DNA microarray technology brought up the idea of monitoring the whole genome in a single chip to have a better global view simultaneously.

The DNA microarray technology has two variants in terms of the property of arrayed DNA sequence with known identity: In the first one probe cDNA (500~5,000 bases long) is immobilized to a solid surface such as glass using robot spotting and exposed to a set of targets either separately or in a mixture [1]. In the second one, an array of oligonucleotide (20~80-mer oligos) or peptide nucleic acid (PNA) probes is synthesized either in situ (on-chip) or by conventional synthesis followed by on-chip immobilization. The array is exposed to labeled sample DNA, hybridized, and the identity or abundance complementary sequences are determined.

This microarray technology accelerates the rate at which gene expression pattern information is accumulated. Hence there is an increasing need to reveal the patterns hidden in the data. However, the nature of studies of multiconditional gene expression patterns widely varies. Accordingly, we are interested in analysis tools that may be useful in all such contexts. Clustering techniques are applicable as they would cluster sets of genes that "behave similarly" under the set of given conditions.

The term cluster analysis aims for grouping objects of similar kind into respective categories in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise.

In this paper, the support vector clustering method will be applied on DNA microarray data, which is performed to determine the expression profile of p53 protein (codon 249) mutant and wild type HepG2 cells after the induction of the double stranded DNA break by Adriamycin treatment.

In the next section, support vector clustering algorithm is explained. The experiments and results and the discussion about them will be mentioned in the later versions of the paper.

METHOD

Support Vector Machines (SVM)

Support Vector Machines rely on separating the data when not linearly separable by transforming them to a higher dimensional feature space with an appropriate nonlinear mapping, in which, data from two categories can always be separated by a hyperplane.

An important property of SVM is that the construction of the classifier only needs to evaluate an inner product between two vectors of the kernel function. An explicit mapping into the high dimensional feature space is not necessary. This avoids the complex calculations in the feature space.

The separation process results in a classification function that depends only on a small number of input vectors [2].

The Support Vector Clustering Algorithm

In the Support Vector Clustering (SVC) algorithm data points are mapped from data space to a high dimensional feature space using a Gaussian kernel. In the feature space, the smallest sphere that encloses the image of the data is looked for. This sphere is mapped back to data space, where it forms a set of contours, which enclose the data points. These contours are interpreted as cluster boundaries. Points enclosed by each separate contour are associated with the same cluster.

The clustering level can be controlled by changes in the width parameter of the Gaussian kernel. As this parameter is increased, the number of disconnected contours in data space increases too, leading to an increasing number of clusters.

The SVC algorithm can also deal with outliers by employing a soft margin constant that allows the sphere in feature space not to enclose all points. Large values of this parameter, can also deal with overlapping clusters [3]. The algorithm has two major steps as described below:

Cluster Boundary Detection

Let $\{xi\}$, in the subspace of χ, be a data set of N points, with χ, in the subspace of IR^d, the data space. Using a nonlinear transformation Φ from χ to some high dimensional feature-space, the smallest enclosing sphere of radius R is looked for described by the constraints:

$$||\Phi(x_j) - a||^2 \leq R^2 \text{ for all } j ,$$

where $\| \cdot \|$ is the Euclidean norm and a is the center of the sphere. Soft constraints are incorporated by adding slack variables ξj:

$$||\Phi(x_j) - a||^2 \leq R^2 + \xi_j \qquad (1)$$

with $\xi j \geq 0$. To solve this problem we introduce the Lagrangian

$$L = R^2 - \sum_j (R^2 + \xi_j - ||\Phi(x_j) - a||^2)\beta_j - \sum_j \xi_j \mu_j + C\sum_j \xi_j , \qquad (2)$$

where $\beta j \geq 0$ and $\mu_j \geq 0$ are Lagrange multipliers, C is a constant, and $C\sum_j \xi j$ is a penalty term. Setting the derivative of L to zero with respect to R, a and ξj, respectively, leads to

$$\sum_j \beta_j = 1 \quad (3)$$

$$a = \sum_j \beta_j \Phi(x_j) \quad (4)$$

$$\beta_j = C - \mu_j . \quad (5)$$

The KKT complementarity conditions of Fletcher [4] result in

$$\xi_j \mu_j = 0, \quad (6)$$

$$(R^2 + \xi_j - \|\Phi(x_j) - a\|^2)\beta_j = 0. \quad (7)$$

It follows from Eq. (7) that the image of a point x_i with $\xi_i > 0$ and $\beta_i > 0$ lies outside the feature-space sphere. Eq. (6) states that such a point has $\mu_i = 0$, hence from Eq. (5) that $\beta_i = C$ is concluded. This will be called a bounded support vector or BSV. A point x_i with $\xi_i = 0$ is mapped to the inside or to the surface of the feature space sphere. If its $0 < \beta_j < C$ then Eq. (7) implies that its image $\Phi(x_i)$ lies on the surface of the feature space sphere. Such a point will be referred to as a support vector or SV. SVs lie on cluster boundaries, BSVs lie outside the boundaries, and all other points lie inside them. When $C \geq 1$ no BSVs exist because of the constraint (3).

Using these relations we may eliminate the variables R, a and μ_j , turning the Lagrangian into the Wolfe dual form that is a function of the variables β_j:

$$W = \sum_j \Phi(xj)^2 \beta_j - \sum_{i,j} \beta_i \beta_j \Phi(xi) \cdot \Phi(x_j). \quad (8)$$

Since the variables μ_j don't appear in the Lagrangian they may be replaced with the constraints:

$$0 \leq \beta_j \leq C, j = 1, \ldots, N. \quad (9)$$

We follow the SV method and represent the dot products $\Phi(x_i) \cdot \Phi(x_j)$ by an appropriate Mercer kernel $K(x_i, x_j)$.

The Lagrangian W is written as:

$$W = \sum_j K(x_j , x_j)\beta_j - \sum_{i,j} \beta_i \beta_j K(x_i, x_j). \quad (10)$$

At each point x the distance of its image in feature space from the center of the sphere is defined as:

$$R^2(x) = \|\Phi(x) - a\|^2 . \quad (11)$$

In view of (4) and the definition of the kernel the following is got:

$$R^2(x) = K(x, x) - 2\sum_j \beta_j K(x_j , x) + \sum_{i,j} \beta_i \beta_j K(x_i, xj) . \quad (12)$$

The radius of the sphere is:

$$R = \{R(x_i) \mid x_i \text{ is a support vector } \} . \quad (13)$$

The contours that enclose the points in data space are defined by the set

$$\{x \mid R(x) = R\} . \quad (14)$$

They are interpreted by us as forming cluster boundaries. In view of equation (14), SVs lie on cluster boundaries, BSVs are outside, and all other points lie inside the clusters.

Cluster Assignment

The cluster boundary detection algorithm does not find points that belong to different clusters. To do so, a geometric approach is used: given a pair of data points that belong to different clusters, any path that connects them must exit from the sphere in feature space. Therefore, such a path contains a segment of points y such that $R(y) > R$. This leads to the definition of the adjacency matrix A_{ij} between a pair of points xi and xj whose images lie in or on the sphere in feature space:

$A_{ij} =$ 1, if, $R(y) \leq R$, for all y on the line segment connecting xi and xj, $R(y) \leq R$
0, otherwise

Clusters are defined as the connected components of the graph induced by A. Checking the line segment is implemented by sampling a number of points. BSVs are unclassified by this procedure since their feature space images lie outside the enclosing sphere. One may decide either to leave them unclassified, or to assign them to the cluster that they are closest to.

EXPERIMENTS

A Matlab implementation was unsuccessful in giving results, arising virtual memory problems due to the high dimensionality of the data. The code will be imported to C language to make memory and speed optimizations.

RESULTS

The results will be mentioned after the completion of the experiments.

DISCUSSION

The discussion will be mentioned after the completion of the experiments.

REFERENCES

[1] R. Ekins and F.W. Chu (1999): Microarrays: their origins and applications. — Trends in Biotechnology, Vol.17, pp.217-218.

[2] Vapnik, V.N. (1998): Statistical Learning Theory. — Wiley, New York.

[3] A. Ben-Hur, D. Horn, H. T. Siegelmann, V. Vapnik. (2001) Support Vector Clustering: — Journal of Machine Learning Research Vol.2: pp.125-137.

[4] R. Fletcher. (1987): Practical Methods of Optimization. — Wiley-Interscience, Chichester.