

A Multiclass Classification of Cancer Data: Using a Kernel Based Clustering k-NN Support Vector Machine

Tripti Swarnkar¹, Chinmaya Dash²

¹Institute of Technical Education & Research/CA, Bhubaneswar, India
Email: tripti_sarap@yahoo.com

²Institute of Technical Education & Research/CSE, Bhubaneswar, India
Email: chinmaya_nitr_008@hotmail.com

Abstract: Support vector machines (SVM) have been promising methods for classification because of their solid mathematical foundations which convey several salient properties that other methods hardly provide. However, despite of the prominent properties of SVM, they are not as favored for large-scale data as complexity of SVM is highly dependent on the size of a data set. Microarray gene expression data that usually have a large number of dimensions, over thousands of genes, and a small number of samples, e.g., a few tens of patients.

This paper presents a noble and efficient approach, *Dimensionality Reduction* (T-test), followed by *Clustering kNN Support Vector machines* (CK-SVM), which is specially designed for handling very large data sets like Microarray gene expression data. The CK-SVM classifies by reflecting the degree of a training data point, as a support vector by using Gaussian function, with K-nearest neighbor (k-NN) and Euclidean distance measure. To add local control property a simple clustering scheme is implemented, before Gaussian functions are constructed for each cluster. In addition probabilistic SVM outputs are used for extending binary classification to multiclass classification, in a pair wise manner. In this paper a multiclass classification has been applied to cancer data, represented by SRBCT data set.

Keywords: Support vector machines, Gaussian functions, Ttest, k-Nearest Neighbor, Microarray data, clustering, CKSVM, K-means, Gaussian function and Euclidean Distance.

I. INTRODUCTION

Micro arrays [1] also known as DNA chips, measure the expression level of a large number of genes, under a number of different experimental conditions, where conditions may refer to different points in time, different organs or different tissues or even different individuals. This gene expression data is usually arranged in a data matrix, where each gene corresponds to a row and each condition to one column of the matrix, such that each array element represents the expression level of a gene under a specific condition and is represented by a real number, which is the logarithmic value of the relative abundance of m-RNA of the gene under specific condition and reflects the characteristics of the tissue at molecular level. With the advent of such characteristics of

micro array data, they are proved to be worthy to help in classifying and predicting different types of cancers [2], unlike their morphological counterparts. For example gene expression data set, such as SRBCT [3] has been used to obtain good results in the classification of *lymphoma*, *leukemia*, *liver cancer* and etc. If such a pattern recognition problem is to be treated with supervised machine learning approaches like SVM [4] one will need to deal with the shortage of training samples and high dimensional input features, as gene expression data set is of high dimension and contains relatively small number of samples. The standard SVM requires solving a quadratic programming optimization problem to find a subset of training data points, called support vectors, in order to define the separating hyper plane. But according to complexity of the hyper plane and size of the training data set, the number of support vectors needed for construction of the optimal hyper plane increases [5]. Related to these situations two main problems may occur. First of these results from outliers or undermining genes in training set, represented by gene expression data set and the second one is computational cost problem. Both the first and second problem, can be resolved, by constructing a new training algorithm, which is to be trained through k-NN [6] method, Gaussian function and Euclidean distance measure, instead of quadratic programming which consumes very long time for training of standard SVM. The training algorithm will learn the probability of being support vector for each training data point, which is presented by a normalized Gaussian function, depended on k-NN method and Euclidean distance measure. Further the T-test has been used for gene selection to obtain good classification accuracy by picking out the genes, that benefit the classification most [7].

Since k-NN, Euclidean distance measure, being supporting methods, are all local learning methods, the new training algorithm is built on clusters. Thus local controlling property is being added to the training algorithm. In this study the k-NN method is preferred, as it is useful when training data are inadequate and also gives comparable results to Fuzzy [8] systems. With the advent of such efficiency of k-NN method, probabilistic SVM outputs [9] are used for extending binary classification to

multiclass classification in this study. LS-SVM (least square support vector machine) [10] proposed in literature are used to avoid computational cost problem. However, sparseness capacity of SVM is better than LS-SVM. Since LS-SVM is of equality constraints, instead of inequality constraints and solves linear equations instead of quadratic programming optimization problem, computational cost of SVM is reduced in LS-SVM. Experiments have shown that CKSVM [11] has got an advantage over LS-SVM in terms of efficiency, computational cost and classification performance. Remaining of this paper is organized as follows: Feature reduction, otherwise known as Gene selection is described in second section. The third section presents SVM, LS-SVM and our training algorithm CK-SVM. The fourth section presents experimental results for SRBCT data set. A conclusion is placed in the fifth section of this paper.

II. FEATURE REDUCTION

Among the large number of genes, only a small part may benefit the correct classification of cancers. The rest of the genes have little impact on the classification. Hence, to obtain good classification accuracy, we need to pick out the genes that benefit the classification most. In this regard a statistical method proposed by Welch, known as T-test has been used-test operates in two steps. Where in first step a score based on T-test, TS is calculated for each gene and in second step the gene with largest TS value is placed at the first place, of the ranking list followed by the gene with second largest TS value and so on. Finally, only some top genes in the ranking list are used for classification and the possibility of noisy or undermining genes will be reduced [7].

III. CLUSTERING K-NN SUPPORT VECTOR MACHINE

In this section, SVM and LS-SVM are described briefly. In addition, developed algorithm (CK-SVM) is explained.

A. Support Vector machine

SVM is a kernel based classification method. Kernel functions are used for mapping from N dimensional input space to higher dimensional feature space. Thus, SVM tries to find optimal separating hyper plane which has maximum margin linearly in the feature space via quadratic programming represented by following formula.

$$D(x) = (w \cdot x) + w_0 \quad (1)$$

Each training data point in input space is considered as an N dimensional vector X and its label is +1 or -1. Training set with 'n' samples for SVM is represented by $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Here x and y. And each separating hyper plane has to provide following inequalities for both classes.

$$(w \cdot x_i) + w_0 \geq (1 - \xi_i), \text{ if } (y_i = +1) \quad (2)$$

$$(w \cdot x_i) + w_0 \leq (-1 - \xi_i), \text{ if } (y_i = -1) \quad (3)$$

Or

$$y_i \{ (w \cdot x_i) + w_0 \} \geq 1 \quad (4)$$

Here ξ_i are slack variables. To find the optimal hyper plane, one has to minimize the following formulae with respect to (4).

$$C \sum_{i=1}^n \xi_i + \frac{1}{2} \|w\|^2 \quad (5)$$

Here C is a regularization parameter between complexity and classification accuracy. By transforming (5) in to an optimization problem with Lagrange multipliers in dual form, following problem can be obtained.

$$w(x) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,k=1}^n \alpha_i y_i \alpha_k y_k k(x_i, x_k) \quad (6)$$

Here $\sum_{i=1}^n \alpha_i y_i = 0$ is to be maximized, subjected to $\sum_{i=1}^n \alpha_i y_i = 0$. According to, the Lagrange multipliers, decision function can be built as follows:

$$f(x) = \text{sign} \left[\sum_{i=1}^{sv} \alpha_i y_i (x, x_i) + b \right] \quad (7)$$

In this study RBF kernel function is used, and it can be defined as follows:

$$k(x, x_i) = \exp(-\gamma \|x - x_i\|^2) \quad (8)$$

Here γ (gamma) is a constant specified by the user. With above description of support vector machine, now we will describe, LS-SVM, which is an improvement to SVM in the next section.

B. Least Square Support Vector Machine

LS-SVM tries to minimize the primal cost function subject to equality constraints, instead of inequality ones. Therefore, LS-SVM solves a set of linear equations instead of computational cost quadratic programming problem [11].

C. Proposed Model for Cancer Classification Using CK-SVM

In the proposed model, for cancer classification, the training data set so obtained from gene selection via T-test, is mapped into Higher Dimensional feature space via RBF kernel function, being used in the CK-SVM algorithm. The algorithm uses K-means [12] [13] clustering scheme that is to be carried out according to given number of clusters for both classes (class1, class2), in the feature space. Finally, the point which has a smaller standard deviation value than another is selected as a reference point and clustering scheme is completed. To reduce the computational cost problem, the proposed algorithm will be trained through k-NN method, using Gaussian func-

tion and Euclidean distance measure, instead of quadratic

Data	Kernel parameter	Tolerance	Threshold values	Number of clusters
SRBCT	.001 ,0.1, 0.2	0.09	1.1, 1.1, 1.6, 1.9	4

programming [11]. The outline of the proposed model is pictorially given in Fig. 1.



Figure 1. Steps involving classification, using CK-SVM

IV. EXPERIMENTAL RESULTS AND TESTING

A. Experimental Results

The CK-SVM mentioned in third section of this paper will be applied to the SRBCT data set. The entire SRBCT data set includes the expression data of 2308 genes. Properties of SRBCT data set after being processed through T-test are summarized in table¹. In this study, RBF kernel function will be selected as a mapping function. RBF

Data Set	Number of training data	Number of testing data	Features	output
SRBCT	63	25	30	4

kernel parameter, threshold values of elimination from data set and tolerance rate for being support vector are determined on trial way to achieve the best performance. Values of these parameters according to structure of the training algorithm are presented in table².

Table¹ SRBCT data set specification

Table² Parameters for training CK-SVM

With above experimental information, about SRBCT data set, the degree of testing accuracy for CK-SVM with a RBF kernel function is shown graphically in Fig 2.

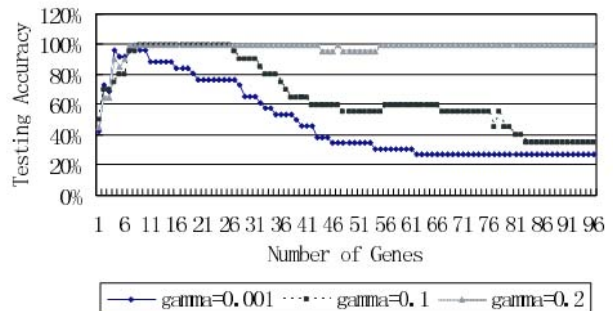


Figure 2. Testing result for SRBCT data set using a RBF kernel

Since used data set (SRBCT), has four output levels, as evident from table¹, the classification method described in the third section of this paper must be extended to multiclass classification case. If there are more than two classes in the data set, binary SVM are not sufficient to solve the whole problem. To solve multiclass classification problem, we should divide the whole problem into a number of binary classification problems.

Usually, there are two approaches [14]. One is the “one against all” scheme and the other is the “one against one” scheme. In one -against- all extension approach, one class is separated from other classes at each training phase. Therefore, in the case of multi-class classification with classes, whole training data set will be trained times. This strategy needs very long training time especially for very large data sets. The Second approach, which is one-against-one, as proposed by *Krebel* is pair wise [11]. To classify classes, numbers of SVMs are trained based on this approach. At each training time one class is separated from another class. Therefore, training time of this approach is shorter than one-against-all approach. In our study, this extension method has been selected because of this property.

B. Testing an Unknown datum with CK-SVM

When an unknown datum is taken, firstly it is assigned to the nearest cluster. In testing case, weighted averaging method is implemented on Euclidean distances between unknown datum and support vectors as per following formula for each cluster independently.

$$wa = \frac{1}{n_{sv}} \sum_{i=1}^{n_{sv}} w_i EU(x_i, x) \quad (9)$$

Here w_i reflects the weighted averaging value and n_i reflects the number of support vectors in each cluster. $d(x, x_i)$ represents the Euclidean distance between testing datum and i^{th} support vector, w_i represents the weight of being support vector for the i^{th} support vector. For extension from binary classification to multiclass classification, these weighted distances instead of the output values, w_i and n_i are used. The unknown datum is assigned to the class which has minimal weighted distance.

V. CONCLUSION

In this paper we touched upon both directions from experimental viewpoints to find a good solution to cancer classification problem by using gene expression data. Selecting important genes helps to determine a new input space, in which the samples are more likely to be correctly classified. Previous studies show that the T-test has a better selection criterion in comparison to other known approaches for gene selection. This paper also describes the procedure for constructing cluster based SVM, i.e. CK-SVM. In this regard we have introduced a cluster based simple and fast training algorithm to solve outliers and computational cost problem. In addition, CK-SVM has provided efficiency for fast classification and continuous outputs via weighted distances for multiclass classification. Our training method CK-SVM has provided equal and better classification performance in comparison to other existing SVM based classifier such as LS-SVM in shorter running time for cancer classification problem.

REFERENCES

- [1] M. Schena, D. Shalon, R.W.Davis and P. O.Brown, Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray, *Science* 267 (1995) 467– 470.
- [2] X. Chen, S. T. Cheung, S. So, S. T. Fan and C. Barry, Gene Expression Patterns in Human Liver Cancers, *Molecular Biology of Cell* 13 (2002) 1929– 1939.
- [3] J. Khan, J.S. Wei, M.Ringner, L.H.Saal and M.Ladanyi, Classification and Diagnostic Prediction of Cancers using Gene Expression Profiling and Artificial Neural Networks, *Nature Medicine* 7 (2001) 673– 679.
- [4] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.
- [5] Zhan, Yiqiang, & Shen, Dinggang (2005). Design Efficient Support Vector Machine for Fast Classification. *Pattern Recognition*, 38, 157–161.
- [6] Bontempi, G., Birattari, M., & Bersini, H. (1998). Recursive Lazy Learning for Modeling and Control, in *Machine Learning* : (10th European conference on machine learning) (pp. 292– 303). April 21–23, 1998, Chemnitz, Germany.
- [7] Feng Chu & Lipo Wang, Application of Support Vector Machine to Cancer Classification with Microarray Data. *International Journal of Neural systems*, Vol. 15, No. 6 (2005) 475-484: World Scientific.
- [8] Bontempi, G.Bersini, H & Birattari M. (2001). The Local Paradigm for Modeling and Control: From Neuro-fuzzy to lazy learning. *Fuzzy sets and systems* (Vol. 121). Elsevier.
- [9] Ana M.B Nikolik D & Curfs L. M. G. (2004). Probabilistic SVM Outputs for Pattern Recognition using Analytical Geometry. *Neuro computing*, 62, 293– 303.
- [10] Suykens, J.A.K. & Vandewalle J. (1999). Least Squares Support Vector Machine Classifier. *Neural Processing Letters*, 9(3), 293–300.
- [11] Commack, E., & Arslan, A. (2006). A Support Vector Machine using Lazy Learning Approach for Multi-Class Classification. *Medical Engineering & Technology*, 30(2), 73– 77.
- [12] R. Dubes and A. Jain, *Algorithms for Clustering Data*. Prentice Hall, 1988.
- [13] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 254-323, Sept. 1999.
- [14] S. Knerr, L. Personnaz and G. Dreyfus, *Single Layer Learning Revisited: A Stepwise Procedure for Building and Training Neural Network*, in *Neuro Computing: Algorithms, Architectures and Applications* J. Fogelman (Springer-Verlag, 1990).