Scientific Journal of Impact Factor (SJIF): 4.14

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 6, June -2016

Classification and Diagnostic Prediction of Cancer Using Support Vector Machine

Sangeeta Sharma[1], Kodanda Dhar Sa [2]

[1,2]*Electronics And Telecommunication Engineering, Indira Gandhi Institute of Technology,Sarang*

Abstract — *The aim of this analysis was to develop a method for classification of cancers to specific diagnostic types based on their gene expression signature by applying Support Vector Machine (SVM).We trained the SVM by utilizing the small, round blue-cell tumors (SRBCTs) as the model. These cancers belong to four distinct diagnostic categories and usually present diagnostic dilemmas in medical study. As their name implies, these cancers are difficult to distinguish by light microscopy, and currently no single test can accurately distinguish these type of cancers. The SVM properly classified the whole samples and identified the genes most relevant to the classification. To test the ability of the trained SVM models to identify SRBCTs, we examined additional blinded samples that were not previously used for the training purpose, and correctly classified them in all cases. This study demonstrates the potential applications of these methods for tumor diagnosis and the identification of candidate targets for therapy. This paper presents architecture of Support Vector Machine classifiers arranged in a binary tree structure for solving multi-class classification problems with increased efficiency.*

Keywords- *Multi-class classification, Principal component Analysis, binary tree architecture, Support Vector Machine, cancer classification and diagnostic prediction of cancer.*

I. INTRODUCTION

Support Vector Machines (SVM) is one of the most encouraging and intelligent techniques for data analysis.SVM gain a huge importance in clinical applications because of its ability to analyze high dimensional gene expression data typically of the tens of thousands of range. Aside from having a strong adaptability, global optimization, and a good generalization performance, the SVMs are suitable for classification of both small and large set of data. The Support Vector Machine successfully adapted in the field of Bioinformatics in order to solve the problems related to the diagnosis of the cancer. This new approach of multiclass SVM classification promises to give better therapeutic measurements to cancer patients by diagnosing the cancer types with improved accuracy.

The small, round blue cell tumors (SRBCTs) of childhood, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS). As their name suggests, these cancers are difficult to categorize by light microscopy, and currently no single test can accurately classify these cancers. Gene-expression profiling using cDNA microarrays permits a simultaneous study of multiple markers, and has been used to classify the cancers into subgroups.

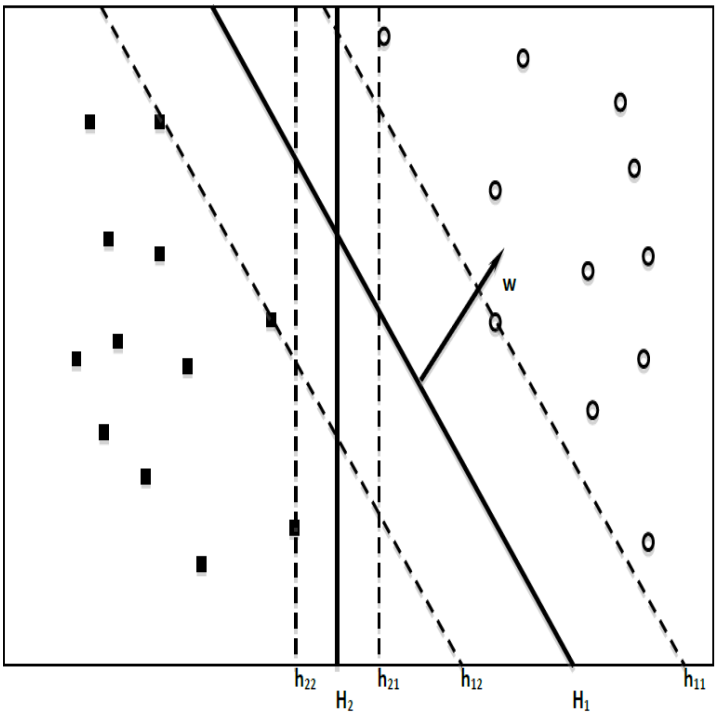
II. SUPPORT VECTOR MACHINE

Support vector machine is a machine learning method that is widely used for data analyzing and pattern recognizing. The algorithm was invented by Vladimir Vapnik ,Boser, Guyon in 1992 and the current standard incarnation was proposed by Corinna Cortes and Vladimir Vapnik. Support vector machine (SVM) is gaining popularity for its ability to classify noisy and high dimensional data. SVM is a statistical learning algorithm that classifies the samples using a subset of training samples called support vectors. They belong to a family of both generalized linear and non-linear classifiers. The idea behind SVM classifier is that it creates a feature space using the attributes in the training data. It then tries to identify a decision boundary or a hyper-plane that separates the feature space into two halves where each half contains only the training data points belonging to a category this is shown in Fig.1.

In Fig.1 the circular data points belong to one class and square points belong to another class. SVM tries to find a hyper-plane (H1 or H2) that separates the two categories. As shown in figure there may be many hyper-planes that can separate the data. Based on “maximum margin hyper-plane” concept SVM chooses the best decision boundary that separates the data.

Each hyper-plane (Hi) is associated with a pair of supporting hyper-planes (hi1and hi2) that are parallel to the decision boundary (Hi) and pass through the nearest data point. The distance between these supporting planes is called as margin. In the figure, even though both the hyper-planes (H1 and H2) divide the data points, H1 has a bigger margin and

tends to perform better for the classification of unknown samples than H2. Hence, bigger the margin is, the less the generalization error for the classification of unknown samples is. Hence, H1 is preferred over H2.



"Figure.1 Decision boundary and margin of SVM classifier"

For a linear SVM the equation for the decision boundary is

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \quad (1)$$

where, **w** and **x** are vectors and the direction of **w** is perpendicular to the linear decision boundary. Vector **w** is determined using the training dataset. For any set of data points (**xi**) that lie above the decision boundary the equation is

$$\mathbf{w} \cdot \mathbf{x_i} + b = k, \text{ where } k > 0, \quad (2)$$

and for the data points (**xj**) which lie below the decision boundary the equation is

$$\mathbf{w} \cdot \mathbf{x_j} + b = k`, \text{ where } k` < 0. \quad (3)$$

By rescaling the values of **w** and b the equations of the two supporting hyper planes (h11and h12) can be defined as

$$h_{11} : \mathbf{w} \cdot \mathbf{x} + b = 1 \quad (4)$$
$$h_{12} : \mathbf{w} \cdot \mathbf{x} + b = -1 \quad (5)$$

The distance between the two hyper planes (margin "d") is obtained by

$$\mathbf{w} \cdot (\mathbf{x_1} - \mathbf{x_2}) = 2 \quad (6)$$
$$d = 2/\|\mathbf{w}\| \quad (7)$$

The objective of SVM classifier is to maximize the value of d. This objective is equivalent to minimizing the value of $\|\mathbf{w}\|2/2$. The values of **w** and b are obtained by solving this quadratic optimization problem under the constraints

$$\mathbf{w} \cdot \mathbf{x_i} + b > 1 \quad \text{if } y_i = 1 \quad (8)$$
$$\mathbf{w} \cdot \mathbf{x_i} + b < -1 \quad \text{if } y_i = -1 \quad (9)$$

where y_i is the class variable for $\mathbf{x_i}$.

Imposing these restrictions will make SVM to place the training instances with $y_i = 1$ above the hyper plane h_{11} and the training instances with $y_i = -1$ below the hyper plane h_{12}.

To improve the SVM models in order to distinguish cancers in each of the four SRBCTs classes, we used gene-expression data from cDNA microarrays consist of 6567 genes. The 63 training samples included both tumor biopsy material (13EWS and 10 RMS) and cell lines (10 EWS, 10 RMS, 12 NB and 8 Burkitt lymphomas (BL; a subset of NHL)).Then the whole data set was first quality filtered and reduced the number of genes to 2308. Principal component analysis (PCA) further reduced the dimensionality, and we found that using the 10 PCA components per sample as inputs and four cancer category as outputs (EWS, RMS, NB or BL) produced improved SVM models.

III. METHOD

SUPPORT VECTOR MACHINE IN BINARY TREE ARCHITECTURE (SVM-BTA):

Here, we applied Multiclass SVM to decipher gene expression signatures of SRBCTs and used them for diagnostic classification. This approach adopts the multiple SVMs arranged in a binary tree structure technique. A SVM in each node of the tree is trained using two of the classes then all the samples in the node are assigned to two different sub nodes derived from the previously selected classes by similarity. This step goes on repeating itself on every node until each node contains only one class samples. The main problem that should be considered sincerely here is training time because one has to test all the samples at every node to find out which classes the samples should be assigned to which sub node while building the tree. The proposed classifier architecture SVM-BTA (Support Vector Machines with Binary Tree Architecture), takes advantage of both efficient computation of the tree architecture and high classification precision of SVMs.Using this architecture, N–1 SVMs are needed to be trained for an N-class problem but only $[\log_2 N]$ SVMs are essential to classify a sample.

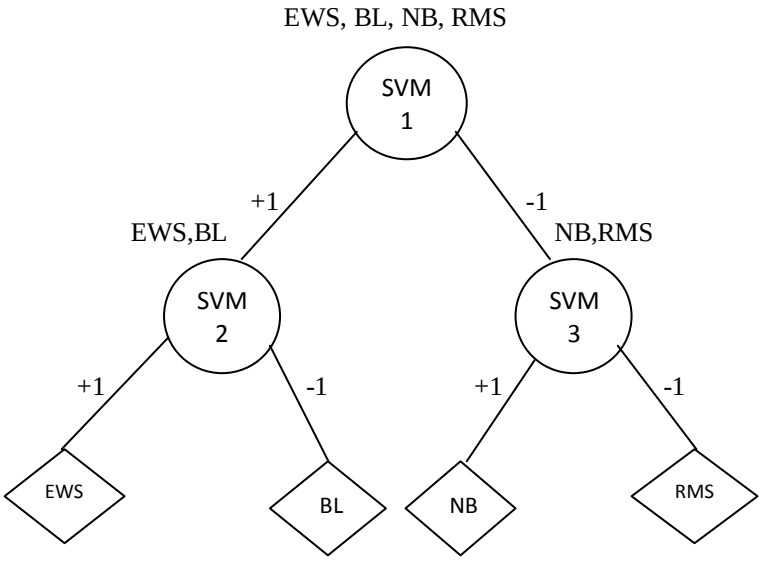


Fig.2. Schematic illustration of the analysis process (SVM-BTA)

Fig 2, is the schematic illustration of the SVM-BTA process. Here, N=4 number of classes are there that's three SVMs are required in order to train and classify the cancer types. At each node of the binary tree, a decision is being made about the assignment of the input pattern into one of the two possible groups characterized by transferring the pattern either to the left or to the right sub-tree. Each of these clusters may have multiple classes. This is repeated downward the tree until the sample reaches to a leaf node that represents the class it has been referred to.

To maintain the consistency between the way SVM calculates the decision hyperplane and the clustering model, the clustering model utilizes distance measures at the kernel space, not at the input space. Because of this, all training samples are modified with the same kernel function that is to be used for the training phase.

The SVM-BTA method that consists of two major steps: (1) calculating the clustering of the classes , and (2) combining a SVM at each node of the taxonomy obtained by (1).Next, out of the 83 experiments, 20 test experiments were set aside and the rest 63 samples are used for the training experiment and the process of clustering is carried out.

In the second step, each SVM is combined to a node and trained with the elements of the two groups of the corresponding node. Fig 2 illustrates the clustering of 4 classes, the SVM classifier in the root is trained by considering

samples from the classes {EWS,BL} as positives elements and samples from the classes {*NB,RMS*} as negative elements. The SVM classifier in the left child of the root is then trained by considering samples from the class {*EWS}* as positives and samples from the class *{BL}* as negative element. Similarly, the SVM classifier in the right child of the root is then trained by considering samples from the class {*NB}* as positives and samples from the class *{RMS}* as negative element. This theory is repeated for each SVM associated to a node in the taxonomy. This will result in training only *N−1* SVMs for solving a *N*-class problem. Next the 20 test experiments were taken for testing purpose and classified using SVM-BTA.

IV. RESULT AND DISCUSSION

Table 1 shows the result of the classified test samples. From table 1 we can analyze that all the blinded test samples and properly classified and the diagnosis of the cancer type is being done accurately except the sample label 1, according to histological diagnosis sample label 1 belongs to NB type of cancer. Now, we have calculated the Confusion Matrix as per the output generated by the SVM.

"Table ***1. SVM Diagnostic Prediction"***

SAMPLE LABEL	SVM CLASSIFICATION	SVM DIAGNOSIS	HISTOLOGICAL DIAGNOSIS
1	EWS	EWS	NB
2	EWS	EWS	EWS
3	RMS	RMS	RMS
4	EWS	EWS	EWS
5	BL	BL	BL
6	NB	NB	NB
7	RMS	RMS	RMS
8	EWS	EWS	EWS
9	NB	NB	NB
10	BL	BL	BL
11	NB	NB	NB
12	RMS	RMS	RMS
13	BL	BL	BL
14	EWS	EWS	EWS
15	EWS	EWS	EWS
16	EWS	EWS	EWS
17	RMS	RMS	RMS
18	NB	NB	NB
19	RMS	RMS	RMS
20	NB	NB	NB

CONFUSION MATRIX:

A confusion Matrix is a table that is often used to describe the performance of a classification model on a set of the test data for which the true values are known. Table 2 represents the confusion matrix of the given test samples.

"***Table 2. Confusion Matrix"***

	EWS	BL	NB	RMS
EWS	6	0	0	0
BL	0	3	0	0
NB	1	0	5	0
RMS	0	0	0	5

ACCURACY RATE:

Accuracy Rate gives the value of overall how often the classifier is correct which can be calculated from the confusion matrix.

$$\text{Accuracy Rate} = \frac{6+3+5+5}{20} = 0.95$$

Accuracy (%) = 95%

MISCLASSIFICATION RATE:

Misclassification Rate can be calculated by confusion matrix simply by the equation,

Misclassification Rate= (false positive + false negative) ÷ (Total number of test samples)

So,

$$\text{Misclassification Rate} = \frac{1}{20} = 0.05$$

DISCUSSION:

Cancers are currently identified with the help of the histology and immunohistochemistry based on their morphology and protein expression, respectively. However, poorly differentiated cancers are not easy to diagnose with the help of conventional histopathology. Here we used a method for the diagnostic classification of cancers from their gene-expression signatures and analyze the genes that that are responsible to this classification. We used the SRBCTs of childhood as a model because these cancers occasionally present diagnostic difficulties, i.e, it creates a diagnostic dilemmas in clinical practice. Such as in case of Ewing sarcoma, it is diagnosed by immunohistochemical evidence of MIC2 expression18 and lack of expression of the leukocyte common antigen CD45 (excluding lymphoma), muscle-specific actin or myogenin (excluding RMS). However, dependence on detection of MIC2 alone can lead to incorrect diagnosis as MIC2 expression occurs occasionally in other tumor types also including RMS and NHL.

Here we have approached this problem using SVM based models. Due to the limited amount of training data and the high performance to be achieved, we limited our analysis to linear SVM models. Although we have used the linear methods, our method can easily use nonlinear features of expression data if at all it's required. Although SVM analysis leads to identification of genes that are responsible for a specific type of cancer, strength of this study is that it does not require genes to be exclusively associated with a single cancer type. This also allows the classification based on complex gene-expression patterns.

As the main goal of this analysis was to make the most effective classification of these cancers, we used a precise quality filter to use only the genes which shows good measurement results for all the samples. This may remove certain genes that are highly expressed in some cancers, but not expressed in other cancers, or may not appear to be expressed because of an artifact in a particular cDNA spot. However, we found that this quality filtration produce more vigorous prediction models and led to the identification of these type of cancers.

However, we expect that this method can be elaborated by the use of more number of the classes and larger sample sets for training. Although we achieved high accuracy and specificity for diagnostic classification, we believe that with larger arrays and more samples it will be possible to improve the accuracy rate of these models for purposes of diagnosis in clinical practice. Here, training is relatively easy as compared to neural network (NN) because here no local minima are required. There is a tradeoff between classifier complexities and here error can be controlled explicitly. The only limitations of the Support Vector Machine are that we only need to choose a "good" kernel function so that we can get more accurate classification.

Future applications of this method will include studies to classify cancers according to their stages and biological behavior in order to predict diagnosis and thereby use the direct therapy.

V. CONCLUSION

Cancer diagnosis is one of the most emerging medical applications of gene expression microarray technology. Multiclass support vector machines (MC-SVMs) are the most effective classifiers in performing accurate cancer diagnosis from gene expression data. The performance of SVMs on a given data is largely dependent upon the methods of feature extraction. The proposed Support Vector Machines in Binary Tree Architecture (SVM-BTA) method was designed to provide superior accuracy by utilizing the decision tree architecture. Clustering of the samples uses the distance measures at the kernel space and is used to convert the multi-class problem into the binary tree, in which the

binary decisions are made by the SVMs. SVM-BTA is becoming more favorable as compared to other methods as the number of classes increases the classified result will be more accurate.

VI. REFERENCES

[1] McManus, A.P., Gusterson, B.A., Pinkerton, C.R. & Shipley, J.M. The molecular pathology of small round-cell tumours—relevance to diagnosis, prognosis, and classification. *J. Pathol.* 178, 116–121 (1996).
[2] Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
[3] Alizadeh, A.A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
[4] Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., et al. (2001).Classification and diagnostic prediction of cancers using gene expressing profiling and artificial neural network. Nature Medicine, 7, 673–679.
[5] Bittner, M. *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536–540 (2000).
[6] Lindsay I Smith, A tutorial on Principal Components Analysis.
[7] Bennett, K. P., Cristianini, N., Shawe-Taylor, J., & Wu, D. (2000). Enlarging the margins in perceptron decision trees. *Machine Learning*, *41*, 295–313.
[8] Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.
[9] Aguilar-Ruiz, J. S., & Divina, F. (2006). Biclustering of expression data with evolutionary computation. IEEE Transactions on Knowledge and Data Engineering, 18(5), 590–602.
[10] V. Vapnik, *The Nature of Statistical Learning Theory*, second ed., Springer, New York, 1999.
[11] J. Weston, C. Watkins, Multi-class support vector machines, *Proceedings of ESANN99*, M. Verleysen, Ed., Brussels, Belgium, 1999.
[12] B. Fei, J. Liu, Binary Tree of SVM: A New Fast Multiclass Training and Classification Algorithm, *IEEE Transaction on neural networks*, Vol. 17, No. 3, May 2006.
[13] S.S.Keerthi, O.Chapelle, D.DeCoste. \Building Support Vector Machines with ReducedClassifier Complexity", Journal of Machine Learning Research, 2006.
[14] Phan, R. Moffitt, J. Dale, J. Petros, A. Young, M. Wang, Improvement of SVM Algorithm for microarray analysis using intelligent parameter selection, Conf. Proc. IEEE Eng. Med. Biol. Soc. 5 (2005) 4838–4841.
[15] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.
[16] K. Crammer, Y. Singer, On the algorithmic implementation of multiclass kernel- based vector machines, Journal of Machine Learning Research 2 (2001) 265– 292.
[17] E. Domany, Cluster analysis of gene expression data, Journal of Statistical Physics 110 (3–6) (2003) 1117–1139.
[18] Zhang L, Zhou W, Su TT, Jiao LC. Decision tree support vector machine. Int JArtif Intell Tools 2007;16(1):1–16.
[19] Wang L, Zhu J, Zou H. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*,2008, **24**(3): 412−419

