ISSN No. 0976-5697

Volume 2, No. 5, Sept-Oct 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# Cancer Classification in Microarray Data Using Gene Expression with SVM OAA and SVM OAO

Dr. R. Mallika
Director, Department of Computer Applications,
Tamilnadu College of Engineering,
Coimbatore, India.
mallikapanneer@hotmail.com

J. Sumitha*
Research Scholar,
Sri Ramakrishna College of Arts and Science for Women,
Coimbatore, India.
sumivenkat2006@gmail.com

***Abstract:*** Data mining is defined as finding hidden information in a database. Classification is a data mining (machine learning) technique used to predict group membership for data instances. In this research, lymphoma and leukemia cancer detection based on the Support Vector Machine- One against One (SVM-OAO) and the Support Vector Machine – One against All (SVM-OAA) is estimated. For estimating the cancer classifications accurately, information gain ratio is highly effective ranking scheme, and gene subset selection is done using Canberra distance metrics. SVM-OAO and SVM-OAA are used as good classifiers. Many other gene importance ranking schemes and classifiers may also be used in this approach. Classification involves four steps. In the first step, the top genes are selected using a feature importance ranking scheme. In the second step, the gene subset is generated using distance metrics. In the third step, the classification capability of all genes within the subset is classified by a good classifier. In the fourth step, the top genes selected using ranking schemes without subset selection) is classified by a same classifier. A K-nearest neighbor algorithm is applied to fill those missing values. This research suggests a unified criterion for gene ranking and subset selection of genes. In the Micro array technology to find specific cancer related genes that can be used to diagnose and predict cancer stage.

***Keywords:*** Classification, Feature Selection, leukemia and lymphoma, SVM-OAO, SVM-OAA.

## I. INTRODUCTION

This research aims at finding the gene subsets that can ensure highly accurate classification of cancers from micro array data by using supervised machine learning algorithms. SVM-OAO and SVM-OAA classifier are used for cancer classifications [1]. The significance of finding the gene subsets is: 1) It greatly reduces the computational burden and "noise" arising from irrelevant genes. 2) It simplifies gene expression tests to include only a very small number of genes rather than thousands of genes, which can bring down the cost for cancer testing significantly. 3) It calls for further investigation into the possible biological relationship between these small numbers of genes and cancer development and treatment.

Feature selection (also known as subset selection) is a process commonly used in machine learning, wherein a subset of the features available from the data is selected for application of a learning algorithm [2]. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. There have been rows of approaches presented for the development of the cancer diagnosis systems [3]. Pattern recognition can be done by the "One Against One" and "One Against All" strategies in Support Vector Machine. But in order make the recognition or the classification that are implementing a Support Vector Machine model in order to overcome the classification problem effectively in this paper [6].

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A gene ranking approach is presented for the detection of the information gain ratio from a set of data sets. The subset selection of this approach is done by the help of the Canberra distance function metric. The dataset used in this approach for the evaluation process are Lymphoma and Leukemia data set [2] [3]. The evaluation and comparison of the proposed approach is done with the One Against All strategy approach [7]. The experimental result shows that the proposed approach outperforms the existing approaches in terms of the accuracy metrics.

## II. METHODOLOGY

### *A. Existing Systems:*

There have been several pattern recognition approaches for the diagnosis of cancer. In an existing approach, they have used feature ranking algorithms to reduce the feature space of the DARPA data set from 41 features to the six most important features. They used three ranking algorithms based on Support Vector Machines (SVMs), Multivariate Adaptive Regression Splines (MARSs), and Linear Genetic Programs (LGPs) to assign a weight to each feature. Experimental results showed that the classifier's accuracy degraded by less than 1 percent when the classifier was fed with the reduced set of features [3]. Sequential backward search was used to identify the important set of features: starting with the set of all features, one feature was removed at a time until the accuracy of the classifier was below a certain threshold [2]. Even though these approaches provide an agreed result, the accuracy level taken for the computation of the problem solution is only compromising.

### *B. Proposed System:*

In our proposed system, a new approach of handling the pattern recognition is proposed. Although it is having many existing approaches in the case of diagnosing the cancer infected gene, it is presented a Support Vector Machine

(SVM) based model for this problem. A Support Vector Machine (SVM) is a new learning. SVM has several interesting properties for pattern recognition. The most popular decomposing strategy is probably the "one against all", which consists of building one SVM per class, trained to distinguish the samples in a single class from the samples in all remaining classes. Another popular strategy is the "one against one", which builds one SVM for each pair of classes. Here in this paper it undergoes the pattern recognition approach by means of certain methods. Initially the feature selection is done by the gene ranking concept on the specified data set. The information gain ratio measure is used for the gene ranking approach. Later the subset selection of the ranked data will be takes place. In this approach the Canberra distance function is used for the subset selection process. Next the classification using the SVM based on the "One Against All" strategy is takes place for the data sets taken. The two datasets used here are Lymphoma and leukemia [3]. Next another classification approach of SVM based on the "One Against One" strategy is presented. Finally using the dataset specified the results are evaluated and compared the SVMs both strategies. The experimental results show that the SVM based on the "One Against All" performs better than the "One Against One" approach in terms of the accuracy [7].

The following steps are used for both lymphoma and leukemia dataset to find accuracy for each method.

*Steps for Subset Selection*

*Step1:* Preprocess the data using KNN impute.

*Step2:* Rank all the genes (M*N) using Information Gain Ratio.

$$\text{IGR (Ex, f)} = \frac{Gain(Ex,f)}{SplitInfo(Ex,f)} \text{ where}$$

$$Gain(Ex,f) = Entropy(Ex) - \sum_{v \in Values(f)} \frac{|Ex,v|}{|Ex|} * Entropy(Ex,v),$$

$$Ex,v = \{x \in Ex / value(x,f) = v.$$

*Step3:* For top selected gene pair, find the Canberra distance.

$$d^{CAD}(i,j) = \sum_{k=0}^{n-1} \frac{|y_{i,k} - y_{j,k}|}{|y_{i,k}| + |y_{j,k}|}$$

*Step4:* For a given threshold value, select the gene subsets.

*Step5:* Use SVM-OAO and SVM-OAA classification.

*Steps for Without Subset Selection*

*Step1:* Preprocess the data using KNN impute.

*Step2:* Rank all the genes (M*N) using Information Gain Ratio.

*Step3:* Use SVM-OAO and SVM-OAA classification.

## III. RESULTS AND DISCUSSION

In this paper the results are evaluating and comparing the proposed approaches with the existing systems. In this case, the proposed approach with existing is compared in terms of accuracy. And the experimental results shows that the approach outperforms better than the existing, i.e., the SVM approach using "One Against All" strategy works better than previous approach and "One Against One" strategy in terms of accuracy. Computational time for SVM-One Against All is slightly increased with few milli seconds when compared to SVM-One Against One But the accuracy is much better for SVM-One Against All when compared to the SVM-One Against One.

Table 1: Accuracy results for Leukemia Dataset using SVM-OAO and SVM-OAA

| TOP GENES | SVM - OAA | | SVM - OAO | |
|---|---|---|---|---|
| | WITH SUBSET | WITHOUT SUBSET | WITH SUBSET | WITHOUT SUBSET |
| 1000 | 96% | 75% | 95% | 71% |
| 2000 | 97% | 74% | 91% | 72% |
| 3000 | 95% | 76% | 93% | 71% |
| 4000 | 96% | 75% | 95% | 74% |
| 5000 | 99% | 76% | 91% | 75% |

In the table 1, Leukemia dataset have with subset selection and without subset selection accuracy with both SVM OAA and SVM OAO. The results shows support vector machine –One Against All algorithm is good in accuracy. It also shows that the accuracy of with subset selection is good than the accuracy of without subset selection

Table 2: Accuracy results for Lymphoma

| TOP GENES | SVM – OAA | | SVM – OAO | |
|---|---|---|---|---|
| | WITH SUBSET | WITHOUT SUBSET | WITH SUBSET | WITHOUT SUBSET |
| 1000 | 97% | 74% | 93% | 73% |
| 2000 | 97% | 75% | 94% | 74% |
| 3000 | 99% | 73% | 92% | 71% |
| 4000 | 95% | 74% | 92% | 72% |
| 5000 | 96% | 75% | 93% | 72% |

Dataset using SVM-OAO and SVM-OAA

In the table 2, Lymphoma dataset have with subset selection and without subset selection accuracy with both SVM OAA and SVM OAO. The results shows support vector machine –One Against All algorithm is good in accuracy. It also shows that the accuracy of with subset selection is good than the accuracy of without subset selection.

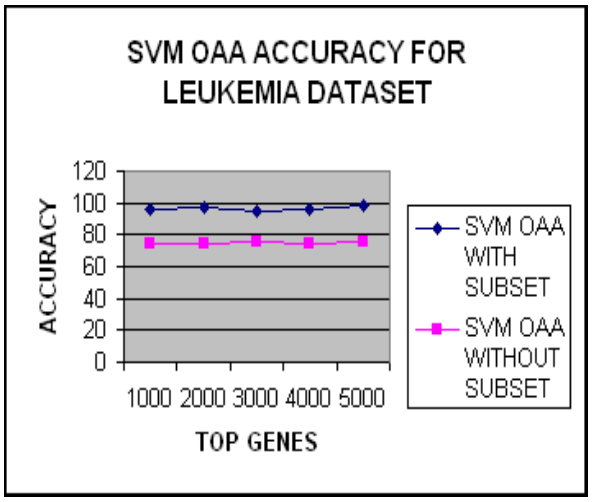


Figure 1: Accurcy graph plotted with no. of Top Genes selected for Leukemia Dataset using SVM-OAA

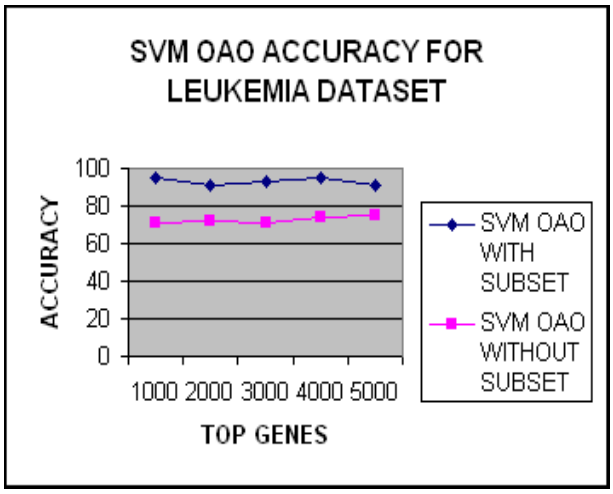


Figure 2: Accuracy graph plotted with no. of Top Genes selected for Leukemia Dataset using SVM- OAO

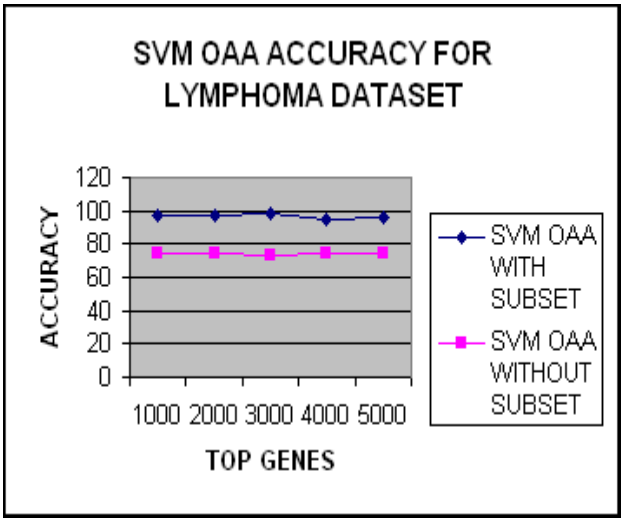


Figure 3: Accuracy graph plotted with no. of Top Genes selected for Lymphoma Dataset using SVMOAA

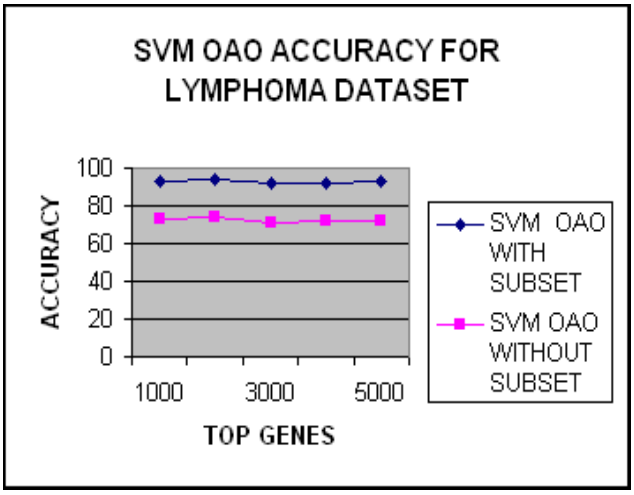


Figure 4: Accuracy graph plotted with no. of Top Genes selected for Lymphoma Dataset using SVM- OAO

## IV. CONCLUSION

In this paper, a novel approach is presented to select the best features for the recognition of the cancer. The Support Vector Machine classifiers such as 'One against One' and 'One against All' are proposed in this paper. The performance evaluation of the proposed system was evaluated and compared with existing approaches. The experimental result shows that this approach of using one against All strategy of SVM works well than One Against One strategy and other existing approaches in terms of the accuracy. Computational cost of time of SVM-One against All is slightly increased with few milli seconds. But the accuracy of SVM-One Against All works much better when compared to the SVM-One Against One.

## V. REFERENCES

[1] Ying Lu, Jiwei Han., "Cancer classification using gene expression data", 2003.

[2] Lipo Wang, Feng Chu, Wei Xie., "Accurate cancer classification using expressions of very few genes. Bioinformatics", Vol 4, 2007.

[3] M. S. Mohamad, S. Omatu, S. Deris, M. Yoshioka., "A Three-Stage method to select informative genes from gene expression data in classifying cancer classes. Intelligent Systems, Modelling and Simulation "2010.

[4] Chen, Y., Dougherty, E.R., & Bittner, M.L. "Ratio-based decisions and the quantative analysis of cDNA micro array images". *Biomedical Optics* **2**,364-374(1997).

[5] Alizadeh, "Distinct types of Diffuse Large B-Cell Lymphoma Identified By Gene Expression Profiling." Nature 403:503-511(2000).

[6] JuiHis, Fu ChihHsiung, Huang SingLing Lee, "A Multi-class SVM classification system based on methods of self – learning and Error filtering".

[7] Jonathan Milgram, Mohamed Cheriet, Robert Sabourin, "One against One" or "One Against All": Which One is Better for Handwriting Recognition with SVMs "year 2006.

[8] Ship, "Diffuse Large B-Cell Lymphoma outcome prediction by gene expression profiling and supervised machine learning." Nat. Med. 8, 68-74 (2002).

[9] Hamosh A., Scott," Online Mendelian inheritance in Man (OMIM): A knowledge base of human genes and genetic disorders. Hodgkin's lymphomas: clinical features of the major histologic subtypes" year 1998; vol 16, pp: 2780-2795.

[10] Rosenwald and Alizadeh, "The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. Cancer Cell", Vol 3,185-197, Feb 2003.