Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

GA-SVM WRAPPER APPROACH FOR GENE RANKING AND CLASSIFICATION USING EXPRESSIONS OF VERY FEW GENES

N.REVATHY[1], **Dr.R.BALASUBRAMANIAN**[2]

[1]Assistant Professor, Department of Computer Applications, Karpagam College of Engineering, Coimbatore 32, India

[2]Dean of Academic Affairs, PPG Institute of Technology, Coimbatore -35, India

E-mail: [1]revsforu@yahoo.com , [2]reachdeanbs@yahoo.com

ABSTRACT

Recently in the classification and diagnosis of cancer nodules, Gene expression profiling by micro array techniques are playing a vital role. Various researchers have proposed a number of machine learning and data mining approaches for identifying cancerous nodule using gene expression data. But, these existing techniques have certain limitations that do not handle the particular needs of gene micro array examination. Initially, micro array data is featured by a high-dimensional feature space repeatedly exceeding the sample space dimensionality by a factor of 100 or higher. Moreover, micro array data consists of a high degree of noise. Most of the conventional approaches do not adequately handle with the limitations like dimensionality and noise. Gene ranking techniques are later proposed to overcome those problems. Some of the widely used Gene ranking techniques are T-Score, ANOVA, etc. But those approaches will sometimes wrongly predict the rank when large database is used. To overcome these issues, this paper proposes an efficient feature selection technique. Wrapper feature selection approach called the GA-SVM approach is used for the effective feature selection of genes. Then, the selected features are given as input to the classifier. The classifier used in the proposed technique is Support Vector Machine (SVM). The experiment is performed on lymphoma data set and the result shows the better accuracy of classification when compared to the standard SVM with T-Score method.

Keywords: *Feature subset Selection, GA-SVM, Support Vector Machine*

1. INTRODUCTION

The diagnosis of complex genetic diseases like cancer has conventionally been done based on the non-molecular characteristics like kind of tumor tissue, pathological characteristics and clinical phase. DNA micro array method has concerned great attention in both the scientific and in industrial areas. Numerous examinations have been presented on the usage of micro array gene expression examination for molecular categorization of cancer. Several machine learning techniques have been developed for the examination of micro array data [5, 16]. The grouping of gene micro array method and machine learning technique assures new approaches into mechanisms of living schemes. An application field where these methods are likely to create key contributions is the identification of cancers depends on clinical phase and biological activities. Such classifications have a huge contribution on diagnosis and treatment.

Various recent investigations in the field of micro array have discussed the application of feature selection approaches to high-dimensional datasets. These feature selection approaches can be used to choose smaller subsets of interesting genes, supporting the analysis of statistical models while keeping the highest possible degree of the accuracy of models developed on the full dataset [19]. Occasionally, by facilitating statistical learning approaches to concentrate only on highly predictive genes while eliminating redundant variables and irrelevant noise, feature selection techniques can even enhance the accuracy of statistical models.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Feature selection techniques are often partitioned into two types namely filter and wrapper techniques. Filter techniques generally rank each gene individually by certain quality criterion (for instance, the p-value of t-test comparing two populations of interest with regard to the expression levels of the gene in the populations), and then select the subset of genes with the n highest quality criteria. Wrapper techniques employ a search algorithm to compute subsets of the variables as a group, rather than individually [20]. A thorough search through all subsets is clearly not possible—there could be around $2^{25,000}$ variable subsets to consider. Therefore, these search approaches utilizes heuristics to direct their search towards promising candidates.

This approach uses the GA-SVM based Wrapper approach. A feature subset selection is a process that can automatically selects a relevant subset of features and ignores the rest, thus resulting in **a** more comprehensive model. In particular, **a** Genetic Algorithm-Support Vector Machine (GA-SVM) based "wrapper" approach for feature subset selection was applied to the gene data set. Then the feature selected genes are given to the classifier.

Generally, a classifier for this purpose must deal with the following problems:

- The classifier must offer an easy-to interpret measure of assurance for its judgments. Thus, the final diagnosis rests with the medical specialist who evaluates if the confidence of the classifier is highly sufficient.

- The classifier must consider asymmetrical wrong classification costs for false positive and false negative classifications.

To achieve this, the micro array gene supplied to the classifier should be consistent. This can be achieved by feature selection of gene accordingly. This paper uses GA-SVM feature selection approach for feature selection of the gene and the classifier used in this paper is Support Vector Machine [6, 11].

2. RELATED WORKS

There are different techniques proposed by different authors for the prediction of cancer regions. Every technique has its own advantages and disadvantages. Some of the existing techniques are presented in this section.

Rui *et al.,* [1] proposed a multiclass cancer classification using semi supervised ellipsoid ARTMAP and particle swarm optimization with gene expression data [7, 9]. It is critical for cancer prediction and treatment to perfectly categorize the site of origin of a cancer. With huge progress of DNA micro array techniques, creating gene expression profiles [8] for various cancer kinds has previously turn out to be a capable way for cancer classification [10]. In addition to research on binary classification like normal versus tumor samples that focuses on various issues from a mixture of disciplines, the discrimination of multiple tumor kinds is also essential. In the meantime, the choosing of genes that are appropriate to definite cancer kinds not only enhances the performance of the classifiers, but also offers molecular insights for treatment. Here, the author utilizes the semi supervised ellipsoid ARTMAP (ssEAM) for multiclass cancer discrimination and particle swarm optimization for informative gene selection. ssEAM is a neural network technique [14] embedded in adaptive resonance theory and applicable for classification purpose. ssEAM characterizes fast, stable, and finite learning and generates hyperellipsoidal clusters, containing complex nonlinear decision boundaries. PSO is an evolutionary algorithm-based method for global optimization. A discrete binary version of PSO is used to represent whether genes are selected or not. The effectiveness of ssEAM/PSO for multiclass cancer diagnosis is illustrated with the help of testing it on three publicly existing multiple-class cancer data sets.

Huilin *et al.,* [2] presents the optimized kernel machines for cancer classification using gene expression data. This technique enhances the performances of the classifiers in classifying gene expression data [15]. Intending to enhance the class separability of the data, the author uses a highly

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

flexible kernel function model, the data-dependent kernel, as the objective kernel to be optimized.

Xiyi *et al.*, [3] given a cancer classification technique by sparse representation using micro array gene expression data. The author presents a novel technique is for diagnosis of cancer with the help of gene expression data by casting the classification difficulty as finding sparse representations of test samples in accordance with the training samples. The sparse representation is effective computed by lscr1-regularized least square.

Runxuan *et al.*, [4] proposed a multi category classification using an extreme learning machine for micro array gene expression cancer diagnosis. The author used the newly created Extreme Learning Machine (ELM) for directing multi category classification in the cancer diagnosis field. ELM neglects drawbacks such as local minima, improper learning rate and over fitting usually faced by iterative learning techniques and completes the training quickly. The author estimates the multi category classification performance of extreme learning machine on three benchmark micro array data sets for cancer diagnosis, namely, the GCM data set, the Lung data set, and the Lymphoma data set.

3. METHODOLOGY

There are two phases included in the proposed technique. In the first phase, every gene in the training data are selected with the help a feature selection technique called GA-SVM Wrapper feature selection approach. In the second phase, the classification ability of every simple combination among the selected genes is tested with the help of a classifier called Support Vector Machine.

Phase 1: *GA-SVM based Wrapper Feature Selection Approach*

Feature subset selection is an optimization problem, which deals with searching the space of possible features to recognize one that is optimum or near-optimal with respect to certain performance measures (e.g., accuracy, learning time, etc.) Wrapper and filter feature selection approaches are available in literature.

This approach uses the randomized wrapper feature selection approach. In particular, genetic algorithm paradigm is selected for randomization and SVM as a base learner in wrapper approach. Alternatively, a population of feature subsets is developed via the process of genetic algorithm and a feature subset is computed via training and testing a SVM with the data set. Genetic Algorithms (GAs) are stochastic search approaches based on the method of natural selection and genetics, and are usually very effective for quick global search of large search spaces in complicated optimization issues. Earlier works have reported the feasibility of GA for wrapper approach to feature subset selection [17]. SVM also suits as a base learner well because of its fast training ability. SVM novelty detector was observed to produce equivalent performance with that of neural network; however, the learning time is much faster than that of neural network. An initial population (genes) is made up of diversified binary strings denoting the features selected. These genes undergo crossover and mutation, assessed by the SVM base learner. Only those genes that are selected based on the particular multi-criteria fitness are put back into the population and the process is repeated for a fixed number of generations. The best solutions are obtained at the end of the complete iterations.

In the proposed GA-SVM wrapper technique, a Gaussian kernel is utilized for the induction approach, i.e. SVM**,** and the parameters were tuned via some heuristic technique. GA was employed with the following settings. The chromosome is a binary string where each bit represents whether the equivalent feature is available (1) or absent *(0).* The population size was usually set at 30**,** but when the population diversity resulted in an unacceptable performance, it was modified up to 50. The crossover rate of 0.6 and the mutation rate of 0.01-0.02 were adopted with equivalent methods being two-point crossover and uniform mutation, respectively. Selection offers the powerful force in the evolutionary process, and the selection pressure is vital. At the primary stage of evolution, a low selection pressure is chosen for a wide exploration

Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of the search space. At the end of evolution, where the population is near convergence, a high selection pressure is used to exploit the most promising regions of the search space [18]. As for the sampling space, a regular one was selected which has the size of the particular population and is made up of all the offspring and only segments of parents. The sampling mechanism follows the probabilistic roulette wheel selection. To discriminate among the similar strong individuals in the last 10%-20% generations, a linear scaling technique was applied to handle the selection probability.

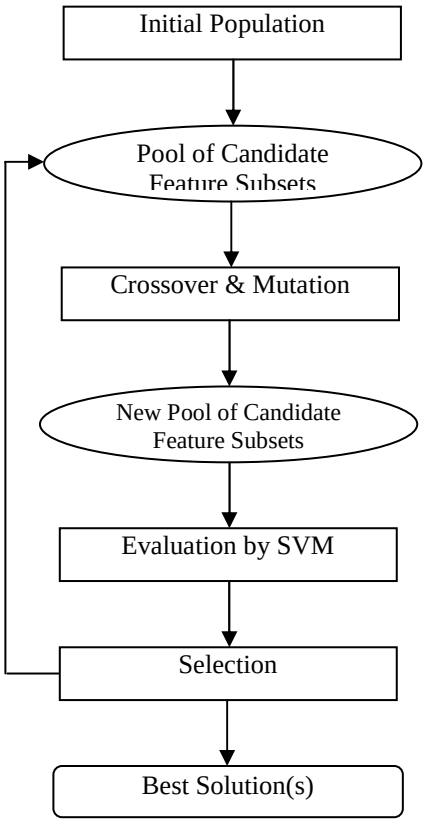

Figure 1: GA-SVM Wrapper Feature Subset Selection

The fitness function integrated three different criteria, i.e. the accuracy of the novelty detector, the learning time used, and the dimension reduction ratio. One definitions of the fitness function emphasized more on the accuracy:

$$Fitness(x) = \frac{1}{DimRat(x)} + \frac{1}{100 \times LrnT(x)} + 10 \times ACC(x)$$

Where $Fitness(x)$ denotes the fitness of the feature subset represented by x, Acc(x) represents the test accuracy of the SVM novelty detector using the feature subset represented by x and $LrnT(x)$ is the time taken to train the SVM.

Though, the test accuracy is the only vital criterion, dimension reduction ratio and training time is also included into the fitness function in that when the model show comparable results, the model with least training time which is vital in practical application, and the feature subset with the smaller dimension which is less susceptible to introduce irrelevant or redundant features, are more preferred. In fact, proper tradeoff values among the multiple objectives have to be based on the knowledge of the problem domain or the experimental results.

Phase 2: *Classification using Support Vector Machines*

Support Vector Machines (SVMs) [12, 13] is a kind of classifier that is a set of associated supervised learning techniques especially for classification. SVM will create a separating hyper plane in the space, one that increases the boundary between the two data sets. To establish the boundary, two parallel hyper planes are created, one on every side of the separating hyper plane between the two data sets. For SVM, a data point is represented as a p dimensional vector, and it is required to distinguish whether it can split such points with a p − 1-dimensional hyper plane. This is called a linear classifier.

As support vector machines are linear classifier that has the capability of finding the optimal hyper plane that increases the separation among patterns, this characteristic creates support vector machines as a potential means for gene expression data examination purposes. The 5 fold cross validation (CV) is performed for support vector machine in the training data set to adjust their constraints. First, the entire data set is split into training (F1) and testing (F2) data by random. The genes are ranked with the help of samples of F1. The combination (FC1) is produced with the help of 2 genes from 20. Then FC1 is arbitrarily split into 5

Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

folds (fc1, fc2, fc3, fc4 and fc5). Among these folds one fold is chosen for testing. The other 4 folds are used as a classifier for SVM. This combination is produces continuously and stops only when the better accuracy is achieved. At last with the fitted SVM, the prediction can be carried out.

4. EXPERIMENTAL RESULTS

The experimentation on the proposed method is carried on lymphoma data set. In the lymphoma data set, there are 42 samples obtained from Diffuse Large B-cell Lymphoma (DLBCL), nine samples from Follicular Lymphoma (FL), and 11 samples from Chronic Lymphocytic Leukemia (CLL). The whole dataset contains the expression data of 4026 genes. Some data may be lost in the dataset because of some error. For filling those lost values k-nearest neighbor technique is used.

Initially, the 62 samples are split randomly into 2 groups: 31 samples for testing, 31 samples for training. Based on the enrichment scores in the training set, the whole sets of 4026 genes are ranked. Then, 200 genes with highest rank are chosen. Finally, the genes are passed to the SVM classifier for classification.

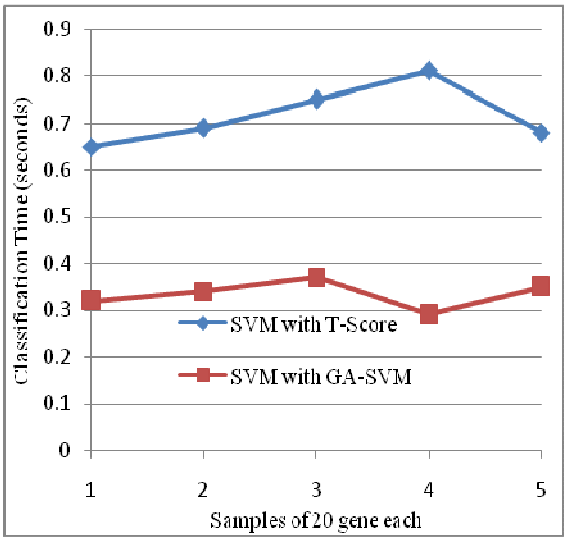

Figure 2: Classification Time for Different Gene Samples

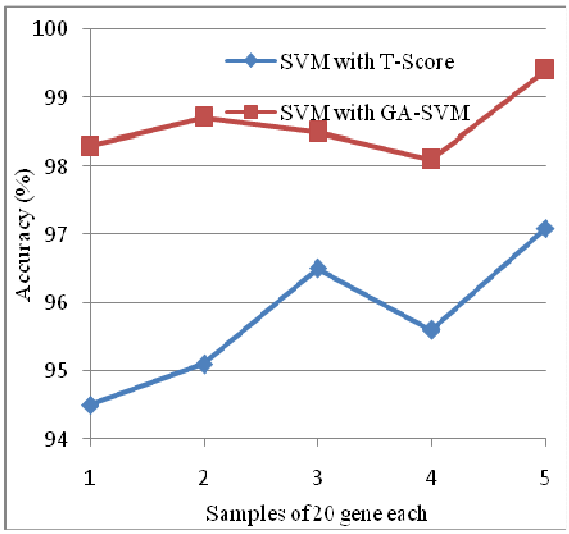

Figure 3: Accuracy of Classification for Different Gene Samples

Figure 2 shows the resulted classification time for different gene samples. It can be observed that the proposed SVM with GA-SVM technique takes lesser time for classification when compared to the SVM technique with T-Score. Figure 3 shows the obtained accuracy for classification. It is clear from the figure that the proposed SVM with GA-SVM technique resulted in better accuracy for all the samples used for classification.

5. CONCLUSION

Cancer research has become one of the vital areas of research in the field of medical sciences. Earlier cancer categorization techniques focused on morphological and clinical analysis. These previous cancer classification techniques had several drawbacks in their diagnostic capability. To overcome those drawbacks in cancer classification, efficient technique in accordance with the global gene expression examination have been evolved. The expression level of genes holds the solutions to overcome basic drawbacks related to the prevention and treatment of cancer. The micro array gene data must be preprocessed for classification with significant accuracy using the classifier. The feature selection technique is used to support that task. This paper uses an efficient wrapper feature selection algorithm called GA-SVM. The selected features from the GA-SVM approach are given as input to the SVM classifier. Then the classifier is trained

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

with that data. Finally, the classification of gene for identifying the cancer is performed. The experiment is performed with the help of lymphoma data set. The experimental result shows that the proposed SVM with GA-SVM technique results in better accuracy and consumes less time for classification when compared to the SVM with T score technique.

REFERENCES:

[1] Rui Xu, Anagnostopoulos, G.C. and Wunsch, D.C.I.I., "Multiclass Cancer Classification Using Semi supervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.4, No.1, Pp. 65-77, 2007.

[2] Huilin Xiong and Xue-Wen Chen, "Optimized Kernel Machines for Cancer Classification Using Gene Expression Data", Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Pp. 1-7, 2005.

[3] Xiyi Hang, "Cancer Classification by Sparse Representation using Microarray Gene Expression Data", IEEE International Conference on Bioinformatics and Biomeidcine Workshops, Pp. 174-177, 2008.

[4] Runxuan Zhang, Huang, G.B., Sundararajan, N. and Saratchandran, P., "Multicategory Classification Using An Extreme Learning Machine for Microarray Gene Expression Cancer Diagnosis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 4, No.3, Pp. 485 – 495, 2007.

[5] Brown, M., "Knowledge Based Analysis of Micorarray Gene Expression Data by using Support Vector Machines", In Proc. of the National Academy of Sciences, Vol. 97, Pp. 262–267, 2000.

[6] Xiaogang Ruan, Jinlian Wang, Hui Li and Xiaoming Li, "A Method for Cancer Classification Using Ensemble Neural Networks with Gene Expression Profile", The 2nd International Conference on Bioinformatics and Biomedical Engineering, Pp. 342-346, 2008.

[7] Berns, A., "Cancer: Gene Expression in Diagnosis", Nature, Pp. 491–492, 2000.

[8] Alizadeh, A., "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling", Nature, Pp. 503–511, 2000.

[9] Campbell, V., Li, Y. and Tipping, N. "An Efficient Feature Selection Algorithm for Classification of Gene Expression Data", 2001.

[10] Dubitzky, W., Granzow, M. and Berrar, D., "Comparing Symbolic and Subsymbolic Machine Learning Approaches to Classification of Cancer and Gene Indentification", Kluwer Academic, 2002.

[11] Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M. and Haussler, D. "Support Vector Machine Classification and Validation of Cancer Tissue Samples using Microarray Expression Data", Bioinformatics, 2001.

[12] Fujarewicz, K., Kimmel, M. and Rzeszowska-Wolny, J., "Improved Classification of Gene Expression Data using Support Vector Machines", Journal of Medical Informatics and Technologies, Vol. 6, Nov 2001.

[13] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., "Gene Selection for Cancer Classification using Support Vector Machines", Machine Learning, 2000.

[14] Khan, J., Wei, J., Ringner, M. and Saal, L., "Classification and Diagnostic Prediction of Cancers using Gene Expression Profiling and Artificial Neural Networks", Nature Medicine, 2001.

[15] Ramaswamy, S., Tamayo, P. and Rifkin, R., "Multiclass Cancer Diagnosis using Tumor Gene Expression Signatures", PNAS, Pp. 15149–15154, 2001.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[16] Zhang, H., Yu, C., Singer, B. and Xiong, M., "Recursive Partitioning for Tumor Classification with Gene Expression Microarray Data". PNAS, Pp. 6730–6735, 2001.

[17] I. Yang. V. Hanavar, "Feature subset selection wings a genetic algorithm", Feature Extraction, In Comrrucrion, and Subset Selection A Darn Mining Perrpective. Momdo, H. and Liu, H. (Ed.J New York: Kluwer, 1998.

[18] L. Goh, Q. Song, and N. Kasabov. A novel feature selection method to improve classification of gene expression data. In Proceedings of the Second Asia-Pacific Conference on Bioinformatics, pages 161–166, Australian Computer Society, Darlinghurst, Australia, 2004.

[19] Enzhe Yu and Sungzoon Cho, "GA-SVM Wrapper Approach for Feature Subset Selection in Keystroke Dynamics Identity Verification" IEEE, 2003.