International Journal of Genetics
ISSN: 0975-2862 & E-ISSN: 0975-9158, Volume 4, Issue 2, 2012, pp.-85-91.
Available online at http://www.bioinfo.in/contents.php?id=25

FSVML AND GA-FSVML WRAPPER APPROACHES FOR GENE SELECTION AND CLASSIFICATION USING EXPRESSIONS OF VERY FEW GENES

REVATHY N.[1] AND BALASUBRAMANIAN R.[2]

[1]Department of Computer Applications, Karpagam College of Engineering, Coimbatore- 32, TN, India.
[2]PPG Institute of Technology, Coimbatore- 35, TN, India.
*Corresponding Author: Email- revs_kn@rediffmail.com

Received: March 15, 2012; Accepted: April 12, 2012

Abstract- Recently, Gene expression profiling by microarray technique has been effectively utilized for classification and diagnostic guessing of cancer nodules in the field of medical sciences. But the techniques used for cancer classification is still in its lower level. There are various drawbacks in the existing classification techniques such as low testing accuracy, high training time, unreliability, etc. Moreover, microarray data consists of a high degree of noise. Gene ranking techniques such as T-Score, ANOVA, etc are later proposed to overcome those problems. But those approaches will sometimes wrongly predict the rank when large database is used. To overcome these issues, this paper mainly focuses on the development of an effective feature selection and classification technique for microarray gene expression cancer diagnosis for provide significant accuracy, reliability and less error rate. In this paper, Wrapper feature selection approach called the GA-FSVML approach is used for the effective feature selection of genes. In FSVML, the RBF kernel function in SVM is trained using modified Levenberg Marquadt algorithm. This approach proposes a Fast SVM Learning (FSVML) technique for the classification tasks. The experiment is performed on lymphoma data set and the result shows the better accuracy of the proposed FSVML with GA-FSVML classification approach when compared to the standard existing approaches.
Keywords- Feature subset Selection, GA-FSVML, Support Vector Machine, Modified Levenberg Marquardt Learning, FSVML

Citation: Revathy N. and Balasubramanian R. (2012) FSVML and GA-FSVML Wrapper Approaches for Gene Selection and Classification Using Expressions of Very few Genes. International Journal of Genetics, ISSN: 0975-2862 & E-ISSN: 0975-9158, Volume 4, Issue 2, pp.-85 -91.

Introduction

Microarray data analysis has been extensively employed in several analysis over a extensive range of biological domains which consists of cancer classification by class detection and prediction, recognition of the unknown effects of a particular therapy, cancer diagnosis, etc [1,2]. In this research, efficient neural network techniques are used with effective learning algorithms for providing significant cancer classification. The main goal of microarrays is hybridization between two DNA strands, the property of balancing nucleic acid series to chiefly pair with each other by figuring out hydrogen bonds between complementary nucleotide base pairs.Several machine learning techniques have been developed for the examination of microarray data [3, 4]. The grouping of gene microarray 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 microarray 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 [5]. 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.

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 in-

stance, 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 [6]. 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-FSVML 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- Fast Support Vector Machine Learning (GA-FSVML) based "wrapper" approach for feature subset selection was applied to the gene data set. Then the feature selected genes are given to the classifier.

Recent investigations and explorations have described several new characteristic features and the practical applicability of Support Vector Machines (SVMs) in knowledge discovery and data mining. SVMs were widely used to discover informative patterns [7].

Support Vector Machine (SVM) is one of the most extensively used machine learning approaches depending on the statistical learning theory, which uses the structural risk minimization inductive principle with the goal to obtain a good generalization from narrow dataset. A significant feature of the SVM is that this transformation need not be implemented to find out the separating hyperplane in the possibly very-high dimensional feature space, a kernel representation can be used for the purpose of determining the separating hyperplane, where the solution evaluated at the support vectors is written as a weighted sum of the values of certain kernel function.

Thus, in order to obtain significant overall results, this paper uses GA-FSVML feature selection approach for feature selection of the gene and the classifier used in this paper is Fast Support Vector Machine Learning approach [8, 9].

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.

Two vital problems in mammogram analysis for breast cancer in MR-images are described by Behnamghader et al., in [10]. The first is category is between normal and abnormal cases and then, classification between benign and malignant in cancerous cases. This proposed approach obtains textural and statistical illustrative features that are applied to a learning engine depending on the utilization of SVM learning framework to categorize them. The experimental observations provide significant accuracy in both classification problems, that shows the suitable interaction of the features and choosing powerful classifier i.e. SVM leads us to a brilliant outcome.

Jose et al., [11] suggested a Genetic Embedded Approach for gene selection and classification of microarray data classification which focuses on the selection of subsets of relevant genes to attain good classification performance. The author describes a genetic embedded technique that carry out choosing task for a SVM classifier. The key aspect of the proposed technique is that, it focuses the highly specialized crossover and mutation operators that consider the gene ranking information provided by the SVM classifier. The effectiveness of this approach is assessed using three well-known benchmark datasets from the literature, showing highly competitive results.

Rui et al., [12] proposed a multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data [13]. It is critical for cancer prediction and treatment to perfectly categorize the site of origin of a cancer. With huge progress of DNA microarray techniques, creating gene expression profiles for various cancer kinds has previously turn out to be a capable way for cancer classification [14]. 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 semisupervised ellipsoid ARTMAP (ssEAM) for multiclass cancer discrimination and particle swarm optimization for informative gene selection. ssEAM is a neural network technique [15] 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., [16] 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. Intending to enhance the class separability of the data, the author uses a highly flexible kernel function model, the data-dependent kernel, as the objective kernel to be optimized.

Methodology

The important focus of this paper is to use the Fast Support Vector Learning algorithm in which the Radial Basis Function (Gaussian) kernel is trained using the modified Levenberg-Marquardt approach.

Fast SVM Learning (FSVML)

A novel Modified Levenberg-Marquardt like second-order algorithm for tuning the Parzen window σ in a Radial Basis Function (Gaussian) kernel is proposed in this paper. In this scenario, each property has its own sigma parameter connected with it. The values of the optimized σ are then utilized as a gauge for variable selection. Kernel Partial Least Squares (K-PLS) model is applied to a variety of benchmark data sets to assess the efficiency of the second-order sigma tuning process for an RBF kernel. The sigma-tuned RBF kernel model performs better than K-PLS and SVM models with a single sigma value.

Sigma Tuning Algorithm

Metric Q^2 is selected as an error metric, represented as $E(\sigma)$

, which depends on the vector σ , Leave-One-Out (LOO) K-PLS is used to attain an initial Q_0^2 , value based on an initial starting guess for the sigma-vector represented as σ_0 , A second-order gradient descent technique is utilized to reduce the objective function $E(\sigma)$, find the optimal choice for σ. The search process initiates from the initial point $E(\sigma_0) = Q_0^2$, The value of σ , is updated depending on the minimization of the leave-one-out (or alternatively, leave several out) tuning (or validation) error, rather than directly diminishing the training error (Figure 1).

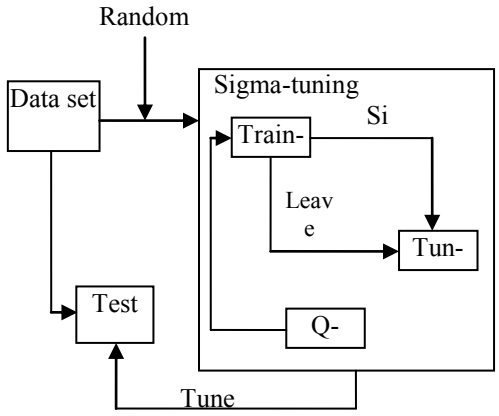

Fig. 1- Process flow for Sigma Tuning

According to Newton's rule for identifying a minimum in a multi-dimensional space, the relation between $E(\sigma)$ and σ at the minimum can be written as:

$$\sigma = \sigma_0 - H^{-1}\nabla E(\sigma_0) \quad (1)$$

where H is the Hessian matrix. $\nabla E(\sigma_0)$ is a vertical vector, as expressed by:

$$\nabla E(\sigma_0) = \nabla E(\sigma)\big|_{\sigma=\sigma_0} = \begin{pmatrix} \frac{\partial E}{\partial \sigma_1}\big|_{\sigma=\sigma_0} \\ \vdots \\ \frac{\partial E}{\partial \sigma_m}\big|_{\sigma=\sigma_0} \end{pmatrix} \quad (2)$$

After rearranging, the equation can be reorganized as

$$H\Delta\sigma = -\nabla E(\sigma_0) \quad (3)$$

where $\Delta\sigma = \sigma - \sigma_0$. In order to efficiently proceed towards a converged solution, a Levenberg-Marquardt approach will be utilized.

Because of the approximate evaluation of the Hessian, a heuristic coefficient α will be introduced in the iterative updating procedure for the elements of σ leading to:

$$\sigma = \alpha\Delta\sigma + \sigma_0 \quad (4)$$

The value of α is set to 0.5 which turns out to be a robust choice based on hundreds of experiments with this algorithm on different datasets. Due to the drawbacks of Levenberg-Marquardt algorithm, this paper uses modified Levenberg Marquardt Algortihm.

Modified Levenberg Marquardt Algortihm

A Modified Levenberg-Marquardt algorithm is used for training the neural network [19]. Considering performance index is $F(w) = e^T e$ using the Newton method the equation obtained is as follows:

$$W_{K+1} = W_K - A_K^{-1}.g_K \quad (5)$$

$$A_k = \nabla^2 F(w)\big|_{w=w_k} \quad (6)$$

$$g_k = \nabla F(w)\big|_{w=w_k} \quad (7)$$

$$[\nabla F(w)]_j = \frac{\partial F(w)}{\partial w_j} = 2\sum_{i=1}^{N} e_i(w).\frac{\partial e_i(w)}{\partial w_j} \quad (8)$$

he gradient can write as:

$$\nabla F(x) = 2J^T e(w) \quad (9)$$

Where

$$J(w) = \begin{bmatrix} \frac{\partial e_{11}}{\partial w_1} \frac{\partial e_{11}}{\partial w_2} \cdots \frac{\partial e_{11}}{\partial w_N} \\ \frac{\partial e_{21}}{\partial w_1} \frac{\partial e_{21}}{\partial w_2} \cdots \frac{\partial e_{21}}{\partial w_N} \\ \vdots \\ \frac{\partial e_{KP}}{\partial w_1} \frac{\partial e_{KP}}{\partial w_2} \cdots \frac{\partial e_{KP}}{\partial w_N} \end{bmatrix} \quad (10)$$

$J(w)$ is called the Jacobian matrix.

Then, the Hessian matrix is to be found. The k, j elements of the Hessian matrix yields as:

$$[\nabla^2 F(w)]_{k,j} = \frac{\partial^2 F(w)}{\partial w_k \partial w_j} = 2\sum_{i=1}^{N}\left\{\frac{\partial e_i(w)}{\partial w_k}\frac{\partial e_i(w)}{\partial w_j} + e\right. \quad (11)$$

The Hessian matrix can then be expressed as follows:

$$\nabla^2 F(w) = 2J^T(W).J(W) + S(W) \quad (12)$$

$$S(w) = \sum_{i=1}^{N} e_i(w).\nabla^2 e_i(w) \quad (13)$$

If $S(w)$ is small assumed, the Hessian matrix can be approximated as:

$$\nabla^2 F(w) \cong 2J^T(w)J(w) \quad (14)$$

Using equations (4) and (12), the Gauss-Newton method is obtained as follows:

$$W_{k+1} = W_k - [2J^T(w_k)\cdot J(w_k)]^{-1}2J^T(w_k)e(w_k)$$
$$\cong W_k - [J^T(w_k)\cdot J(w_k)]^{-1}J^T(w_k)e(w_k) \quad (15)$$

The advantage of Gauss-Newton is that it does not require calculation of second derivatives. There is a problem the Gauss-

Newton method is the matrix $H = J^T J$ may not be invertible. This can be overcome by using the following modification.
Hessian matrix can be written as:

$$G = H + \mu I \tag{16}$$

Suppose that the eigen values and eigenvectors of H are $\{\lambda_1, \lambda_2, \ldots\ldots, \lambda_n\}$ and $\{z_1, z_2, \ldots\ldots, z_n\}$ Then:

$$\begin{aligned} Gz_i &= [H + \mu I]z_i \\ &= Hz_i + \mu z_i \\ &= \lambda_i z_i + \mu z_i \\ &= (\lambda_i + \mu)z_i \end{aligned} \tag{17}$$

Therefore the eigenvectors of G are the same as the eigenvectors of H and the eigen values of G are $(\lambda_i + \mu)$. The matrix G is positive definite by increasing μ until $(\lambda_i + \mu) > 0$ for all i therefore the matrix will be invertible.
This leads to Levenberg-Marquardt algorithm:

$$w_{k+1} = w_k - [J^T(w_k)J(w_k) + \mu I]^{-1} J^T(w_k)e(w_k) \tag{18}$$

$$\Delta w_k = [J^T(w_k)J(w_k) + \mu I]^{-1} J^T(w_k)e(w_k) \tag{19}$$

As known, learning parameter, μ is illustrator of steps of actual output movement to desired output. In the standard LM method, μ is a constant number. This work modifies LM method using μ as:

$$\mu = 0.01 e^T e \tag{20}$$

Where e is a $k \times 1$ matrix therefore $e^T e$ is a 1×1 therefore $[J^T J + \mu I]$ is invertible.

Therefore, if actual output is far than desired output or similarly, errors are large so, it converges to desired output with large steps. Likewise, when measurement of error is small then, actual output approaches to desired output with soft steps. Thus, error oscillation reduces greatly. Modified Levenberg-Marquardt Algorithm for Learning is used for learning of the RBF kernel function in SVM which provides significant performance in cancer classification. This technique reduces the amount of time taken in learning procedure.

Microarray Gene Selection and Classification Approach

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-FSVML Wrapper feature selection approach in which the RBF kernel function in SVM is trained using Modified Levenberg Marquardt algorithm. In the second phase, the classification ability of every simple combination among the selected genes is tested with the help of a classifier called FSVML. Thus, both the selection and the classification technique provide better results.

Phase 1- GA-FSVML 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. Figure 2 shows the flowchart of GA-FSVML wrapper feature subset selection.
This approach uses the randomized wrapper feature selection approach. In particular, genetic algorithm paradigm is selected for randomization and FSVML 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 FSVML 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]. FSVML also suits as a base learner well because of its fast training ability. FSVML 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 FSVML 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.

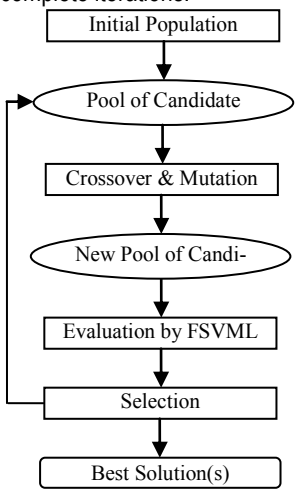

Fig. 2- GA-FSVML Wrapper Feature Subset Selection

In the proposed GA-FSVML 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 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. 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.

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:

$$Fitness(x) = \frac{1}{DimRat(x)} + \frac{1}{100 \times LrnT(x)} + 10 \times ACC(x) \quad (21)$$

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

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 Fast Support Vector Machines Learning (FSVML) Approach

SVM is a significant machine learning technique which is based on artificial intelligence. SVM is an efficient approach for training classifiers depending on various functions such as polynomial functions, radial basis functions, neural networks etc. In Support Vector Machine (SVM), the classifier is generated using a hyper linear separating plane.

To establish the boundary, two parallel hyperplanes are created, one on every side of the separating hyperplane 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 hyperplane. This is called a linear classifier.

To build a SVM classifier, a kernel function and its parameters need to be chosen. In this work, the following kernel function has been applied to build SVM classifiers:

Radial basis function $K(x,z) = \exp\left\{-\frac{\|x-z\|^2}{2\sigma^2}\right\}, \sigma$

is the width of the function.

A kernel in Support Vector Machine is looked upon as a similarity measure to recode the input data. The kernel is used along with a map function. Identifying the best options for the kernel function and parameters is a challenging task, when applied to real dataset.

Usually, the recommended kernel function [18] for nonlinear problems is the Gaussian radial basis function, because it resembles the sigmoid kernel for certain parameters and it requires less parameters than a polynomial kernel. The kernel function parameter γ and the parameter C, which controls the complexity of the decision function versus the training error minimization, can be determined by running a 2 dimensional grid search, which means that the values for pairs of parameters (C, γ) are generated in a predefined interval with a fixed step. The performance of each combination is computed and used to determine the best pair of parameters.

Proposed Learning Algorithm for RBF Kernel

A novel Modified Levenberg-Marquardt like second-order algorithm for tuning the Parzen window σ in a Radial Basis Function (Gaussian) kernel is proposed in this paper. In this scenario, each property has its own sigma parameter connected with it. The values of the optimized σ are then utilized as a gauge for variable selection. Kernel Partial Least Squares (K-PLS) model is applied to a variety of benchmark data sets to assess the efficiency of the second-order sigma tuning process for an RBF kernel. The sigma-tuned RBF kernel model performs better than K-PLS and SVM models with a single sigma value.

Modified Levenberg-Marquardt Algorithm for Learning is used for learning of the RBF kernel function in SVM which provides significant performance in cancer classification which is described in section 3.1. This technique reduces the amount of time taken in learning procedure.

FSVML with GA-FSVML Algorithm Description

Step 1: The raw gene input data is given to the GA-FSVML feature selection technique.

Step 2: Features are selected based on the GA-FSVML based Wrapper Feature Selection Approach.

Step 3: Features selected genes are given as input to the SVM classifier.

Radial basis function $K(x,z) = \exp\left\{-\frac{\|x-z\|^2}{2\sigma^2}\right\}, \sigma$ is the width of the function.

Step 4: Modified Levenberg Marquat learning is used for training the RBF kernel in SVM.

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 test-

ing (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 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 FSVML. This combination is produces continuously and stops only when the better accuracy is achieved. At last with the fitted FSVML, the prediction can be carried out.

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) [20], 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 FSVML classifier for classification.

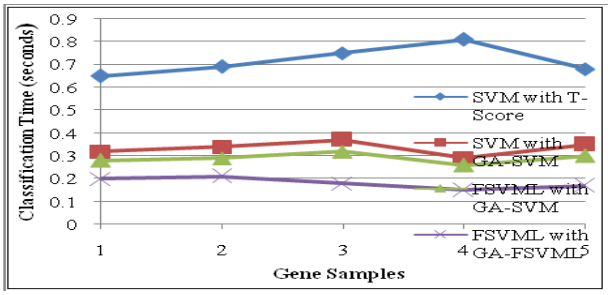

Fig. 2- Classification Time for Different Gene Samples

Figure 2 shows the resulted classification time for different gene samples. It can be observed that the proposed FSVML with GA-FSVML technique takes lesser time for classification when compared with other approaches like SVM with T-Score, SVM with GA -SVM and FSVML with GA-SVM approaches.

The comparison of accuracy of the classification approaches such as SVM with T-Score, SVM with GA-SVM, FSVML with GA-SVM and FSVML with GA-FSVML is shown in figure 3. It is clear from the figure that the proposed FSVML with GA-FSVML technique resulted in better accuracy for all the samples used for classification.

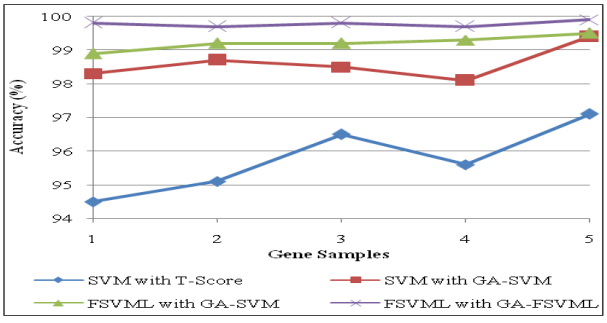

Fig. 3- Accuracy of Classification for Different Gene Samples

Conclusion

The importance of cancer diagnosis and classification in the field of medical sciences has been increasing day by day. Several researches have been done in gene ranking and classifications to develop a novel approach as the existing approaches have lot of drawbacks such as low testing accuracy, high classification time, etc. This paper focuses on developing an efficient gene selection and classification techniques. The microarray 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-FSVML. The selected features from the GA-FSVML approach are given as input to the FSVML classifier in which the modified Levenberg-Marquardt Algorithm is used for learning of the RBF kernel function in SVM. Then the classifier is trained 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 FSVML with GA-FSVML technique results in better accuracy and consumes less time for classification when compared to other existing techniques taken into consideration.

References

[1] Tusher V.G., Tibshirani R. and Chu G. (2001) *Significance Analysis of Microarrays applied to the Ionizing Radiation Response*, 98 (51), 16-21.

[2] Shalon T.D. (1995) *DNA Microarrays: A New Tool for Genetic Analysis, Stanford University. Ph.D. thesis.*

[3] Brown M. (2000) The National Academy of Sciences, 97, 262-267.

[4] Zhang H., Yu C., Singer B. and Xiong M. (2001) *Recursive Partitioning for Tumor Classification with Gene Expression Microarray Data*, 6730-6735.

[5] Goh L., Song Q. and Kasabov N. (2004) *The Second Asia-Pacific Conference on Bioinformatics*, 161-166.

[6] Enzhe Yu and Sungzoon Cho (2003) *GA-SVM Wrapper Approach for Feature Subset Selection in Keystroke Dynamics Identity Verification*.

[7] Mallika R. and Saravanan V. (2010) *World Academy Of Science, Engineering And Technology*, 62.

[8] Xiaogang Ruan, Jinlian Wang, Hui Li and Xiaoming Li. (2008) *The 2nd International Conference on Bioinformatics and Biomedical Engineering*, 342-346.

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

[10] Behnamghader E., Ardekani R.D., Fatemizadeh E. *(2007) 5th International Symposium on Image and Signal Processing and Analysis (ISPA 2007)*, 98 - 101.

[11] Hernandez J.C., Duval B. and Jin-Kao HaoGuyon, Weston J., Barnhill S. and Vapnik V. (2007) *A Genetic Embedded Approach for Gene Selection and Classification of Microarray Data*.

[12] Berns A. (2000) *Nature*, 491-492.

[13] Dubitzky W., Granzow M. and Berrar D. (2002) *Comparing Symbolic and Subsymbolic Machine Learning Approaches to*

Classification of Cancer and Gene Indentification, Kluwer Academic.
[14]Khan J., Wei J., Ringner M. and Saal L. (2001) *Nature Medicine*.
[15]Huilin Xiong and Xue-Wen Chen (2005) *The 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 1-7.
[16]Yang I. and Hanavar V. (1998) *Feature subset selection wings a genetic algorithm.*
[17]Cristianini N. and Shawe -Taylor J. (2000) An Introduction to Support Vector Machines and Other Kernel based Learning Methods.
[18]Suratgar A.A., Tavakoli M.B. and Hoseinabadi A. (2005) *World Academy of Science, Engineering and Technology,* 46-48.

About Authors

Revathy N. had completed B.Sc., Computer Science in the year 2000 and Master of Computer Applications (MCA) in the year 2003 under Bharathiar University. Completed M.phil. Computer Science from Alagappa University in the year 2005. Currently pursuing Ph.d. and the area of research is Neural Networks. Other areas of interest are Mobile Computing, Data Mining and Artificial Intelligence At present working as an Assistant professor in the Department of Computer Applications at Karpagam College of Engineering at Coimbatore-32 and published 3 papers in International Journals, presented 3 papers in International Conferences and 30 papers in National Conferences.

Balasubramanian R. had completed M.Sc. and PhD and his area of interest in Networking, Data Mining. He had published of papers in International and National level journals. At present working as a Dean of Academic Affairs, PPG Institute of Technology, Coimbatore-35, India. Has organized National level conference and chaired technical sessions in conferences/workshops.