Enhanced SVM based Ensemble Algorithm to Improve the Classification for High Dimensional Data

Kavitha S.
Dept. Computer Science
Karpagam University

M. Hemalatha, PhD
Dept. Computer Science
Karpagam University

ABSTRACT
Microarrays are novel biotechnological technology that is being used widely in cancer research. By allowing the monitoring of expression levels in cells for thousands of genes simultaneously, microarray experiments may lead to a more complete understanding of cell's function. This is due to the fact that the physiology of an organism is generally associated with changes in gene expression patterns, thus leading to a finer and more reliable classification.

Microarray data is an arrangement of points in rows and columns. Out of the various techniques of data mining, classification and clustering are two processes that have great potential in microarray data analysis. This research work focuses on using machine learning classification algorithms for predicting the presence or absence of cancer. A classification model for microarray data analysis consists of three major steps, namely, preprocessing, gene selection and identification or prediction of genetic defect. The preprocessing step consists of cleaning algorithms like normalization, missing value handling routines which enhance the quality of the gene microarray data and help to improve the subsequent steps. Gene selection is a process where a set of informative genes is selected from the gene expression data in a form of microarray dataset.This process helps improve the performance of the classifier. The third step, classification, is a process to classify microarray data into several predefined classes that have its own characteristics.

Keywords
SVM, Microarraydata, classification, Gene Selection

1. INTRODUCTION
As microarray technology is capable of producing massive amounts of genetic data, the need for new or enhanced techniques that can mine and discover biologically meaningful knowledge in large data sets is the current need of molecular biologists.

Analysis of microarray data is performed to bring forward techniques to identifying new cancer classes and assigning tumors to known classes. It is hybridization-based technique that allows simultaneous analysis of thousands of samples on a solid substrate. Microarray data analysis offers numerous advantages like (i) simultaneous analysis of thousands of genes (ii) Discovery of gene functions (iii) Identification of drug targets (iv) Understanding of diseases and (v) Clinical studies and field trials.

As microarray technology is capable of producing massive amounts of genetic data, the need for new or enhanced techniques that can mine and discover biologically meaningful knowledge in large data sets is the current need of molecular biologists. Data mining is defined as the science of extracting useful information from large data sets or databases (Sharma, 2010).

Methods to improve the performance of the classifier for microarray data analysis are a continuous research problem attracting several researchers, academicians and industrialists (De Paz *et al.*, 2011; Chen, 2012). Several of these proposals improve gene selection step of the classification model as a means to improve the classification efficiency. Alternatively, this research improves each of three steps of the classification model as a means to improve the overall performance of the classification model. While considering methods to improve existing algorithms, methods that combine the advantages of various techniques have gained more attention. The solutions proposed in this research work are based on this fact and are more compatible with microarray data in their identification of malignant and benign tumour.

2. MOTIVATION
Cancer is a serious and potentially life-threatening illness worldwide and relates to a collection of related diseases, all of which are abnormal growth of tissue that invade and destroy healthy tissues, including organs. Previously, cancer classification has always been morphological and clinical based. These conventional cancer classification methods are reported to have several limitations in their diagnostic ability (Nguyen *et al.*, 2015). It has been suggested that specifications of therapies according to tumor types differentiated by pathogenetic patterns may maximize the efficacy of the patients (Revathi and Sumathi, 2014). Also, the existing tumor classes have been found to be heterogeneous and comprises of diseases that are molecularly distinct and follow different clinical courses.

As all microarray datasets are high dimensional datasets, there is a need to develop new approaches and associated learning classifiers that can address the complications of the complex microarray datasets. Machine learning classifiers that can scale effectively with the increase in the dimensions of the feature space are highly desirable in this respect and have motivated this research to choose this as the research topic.

In order to gain a better insight into the problem of cancer classification, systematic approaches based on global gene expression analysis can be used. The expression levels of genes are known to contain the keys to address fundamental problems relating to the prevention and cure of diseases, biological evolution mechanisms and drug discovery. The recent advent of microarray technology has allowed the simultaneous monitoring of thousands of genes, which motivated the research work to focus on the development of cancer classification model using gene expression data.

3. RESEARCH PROBLEM
Given a microarray dataset, $G_{N\times(M+c)}$ having 'N' tissue samples of 'M' genes and each sample belonging to any one of 'c' classes, the research problem is to construct an integrated gene selection and classification model of the form

$O = f(P \rightarrow F \rightarrow C(G))$, where f(.) is a model that applies a sequential set of operations, P, F and C, to obtain the output O that classifies the input data as malignant or benign tumour. Here P indicates the preprocessing operation that transforms the incomplete dataset into a complete version, F is the operation that selects an optimal subset of genes, C is the classification model that performs binary classification and $\rightarrow$ denotes the sequential application of operations. The main focus of this research work is to perform microarray data classification for identifying benign and malignant tumours. The classification is enhanced through the improved missing value handling algorithm and integrated gene selection and classification model.

Thus, this paper proposes an optimized microarray classification model build using improved missing value handling algorithm, enhanced gene selection algorithm and enhanced ensemble SVM classifier.

4. . METHODOLOGY

The main focus of this research work is to perform microarray data classification for identifying benign and malignant tumours. The classification is enhanced through the improved missing value handling algorithm and integrated gene selection and classification model. The contributions of this research work are listed below.

Missing Value Handling : Two major issues with the frequently used KNNImpute (K-Nearest Neighbour Imputation) algorithm for handling missing values in microarray data is it's high time complexity and performance degradation with high rate of missing values. This issue is solved by incorporating a cluster-based BPCA (Bayesian Principal Component Analysis) method as pre-handler and combining its result with a sequential KNN imputation method.

The two methods used for gene selection are filter-based and wrapper-based. this proposed work presents a sequential algorithm that combines groups genes according to their information correlation coefficient and then on each group applies pre-pruning multiple-filter-based approach. Then the final optimal gene subset is generated and classified using an integrated Genetic Algorithm (GA)-based SVM classifier.

- Classification using conventional Ensemble SVM classifier has three major issues when applied to microarray data classification. The issues are involved with the training speed, choice of kernel and its parameters and usage of correct gene subset. These issues are solved by using a pruning method, Extreme Learning Method and ensemble classifier respectively.

4.1 Missing Value Handling

Missing values are defined as the situation when an attribute or feature in a dataset has no associated data value. Generally, missing values mining are handled using the three manners, namely, (i) discard the missing attribute value (ii) maximum likelihood approaches and imputation of missing attributes value. This research work proposes an imputation based method, which replaces a new value wherever the data is missing in the database, by first establishing known relationships among the complete values of the dataset, to assist in missing data estimation. Later, by using this established relationship, a classification task is performed with complete observation and incomplete cases with imputed values. Among the various methods available, the usage of KNN-Imputation (KNNI) method is more predominant (Garcıa-Laencina *et al.*, 2009). However, as mentioned earlier, the algorithm has two major issues, namely, high search time and performance degradation when the dataset has high rate of missingness.

To solve these problems, this work proposes an enhanced Sequential K-Nearest Neighbour (SKNN) Imputation algorithm replaced by optimal and appropriate values.

4.2 Gene Selection

Given a set of genes G and a target variable T, the task of gene selection is to find minimum set G′ that achieves maximum classification performance of T. Gene selection in microarray data classification provided multifolded advantages and some of them are listed below.

- Improve performance of classification algorithm by using useful features
- Enforces scalability feature
- Better understanding of the domain

The first step groups similar genes using a grouping technique to obtain p groups. Pre-pruning is performed on each group using multiple filter-based algorithm to obtain 'q' reduced gene sets. As Genetic algorithms (GA) are best known for its ability to efficiently search large space with little a priori knowledge, the final step uses GA with SVM for designing the wrapper based feature selection method. The multiple filter-based algorithm is designed using two filters, namely, Fisher Criterion Score (F-Score) Method and Normalized Conditional Mutual Information (NCMIGS).

4.3 Enhancement of SVM Classifier

Classification or prediction is the task of predicting normal or cancerous gene for gene instances. In this research, SVM classifier is enhanced for this purpose. SVM classifier has gained wide acceptance because of their high generalization ability for a wide range of applications in image processing, pattern recognition and data mining domains. SVM requires the use of an interactive process such as quadratic programming to identify supports vectors. When the number of samples in the training set is high, identification of potential SVs is complex. Reduction in number of SVs used during classification has a direct impact on the speed of SVM. Heuristic methods like partitioning training set have been used for this purpose. However, these methods provide no guarantee that this method would not remove potential SVs. Thus, the conventional SVM classifier has the following issues and this research work provides solutions to them as a method of enhancing the operation of SVM.

5. EXPERIMENTAL RESULTS

The experiments were conducted in three stages. experiments were designed to evaluate the performance of the proposed missing value handling algorithm. This stage used two parameters, namely, Normalized Root Mean Square Error (NRMSE) and speed of handling the missing values.

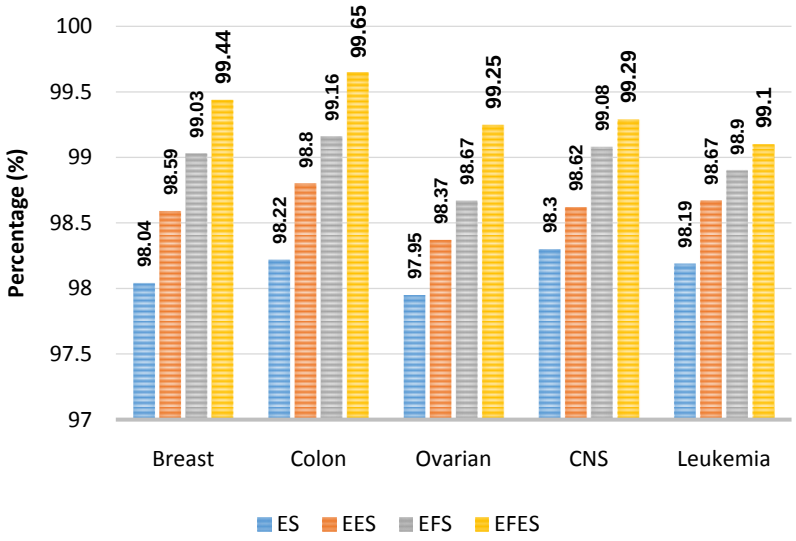

Fig 1: Accuracy (%)

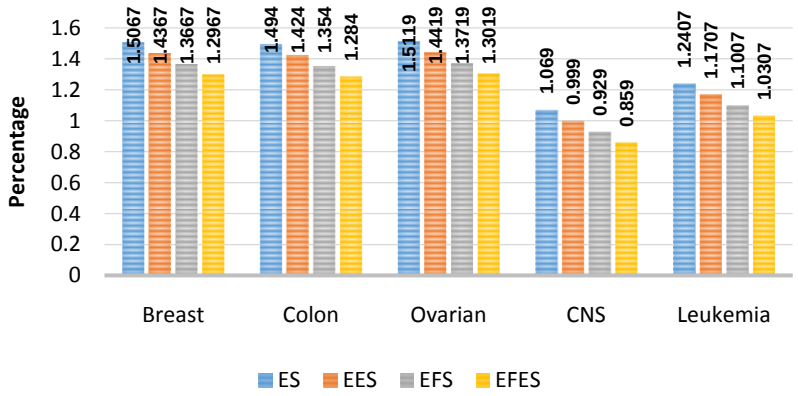

Fig2:Errorrate(%)

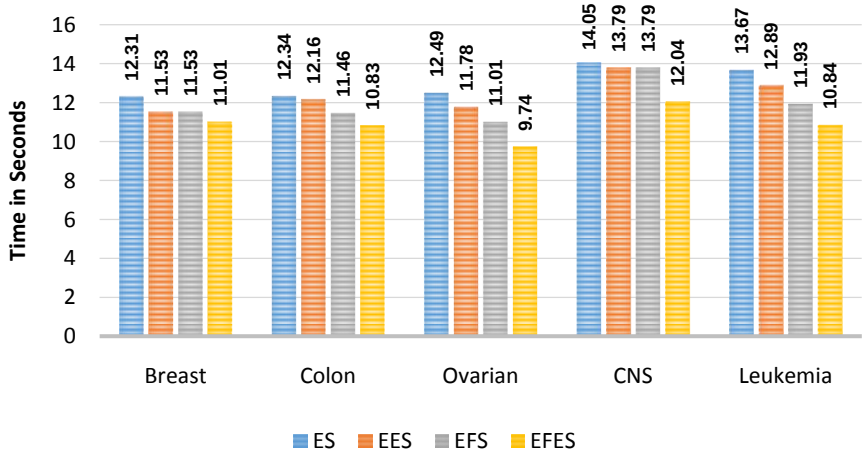

Fig3:Speed(Seconds)

Table 1: Code Description

Code	Description
ES	**Conventional Ensemble SVM**
EES	**Ensemble ELM-SVM**
EFS	**Ensemble Fast SVM**
EFES	**Ensemble FES**

6. CONCLUSION

Microarray datasets tend to be small in sample size due to the cost associated with the assays. There are many more gene expression measurements (e.g. 54,000 transcripts) available than samples. This huge number makes gene/feature selection a crucial step while designing and building a classifier. In this research work, an integrated algorithm that combines missing value algorithm, gene selection and classification method into a single framework is proposed. Experimental results proved that all the proposed methods are efficient and enhance the process of microarray data classification and can safely be used by molecular biologists during knowledge discovery and analysis.

7. REFERENCES

[1]. Alshamlan, H.M., Badr, G.H. and Alohali, Y. (2013) A study of cancer microarray gene expression profile: Objectives and approaches, Proceedings of the World Congress on Engineering, Vol. II, Pp. 1-6.

[2]. Alshamlan, H.M.., Badr, G. and Alohali, Y. (2015) mRMR-ABC: A

[3]. Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling, BioMed Research International, Vol. 2015, Article ID 604910, Pp. 1-15.

[4]. Bolon-Canedo, V., Sanchez-Marono, N. and Betanzos, A.A. (2010) An ensemble of filters and classifiers for microarray data classification, Pattern Recognition, Vol. 45, Pp. 531-539.

[5]. Boulesteix, A., Porzelius, C. and Daumer. M. (2008) Microarray-based classification and clinical predictors. *Bioinformatics, Vol.* 24, No. 15, Pp. 1698-1706.

[6]. Chen, C.K. (2012) The classification of cancer stage microarray data, Computer Methods and Programs in Biomedicine, Vol. 108, Issue 3, Pp. 1070-1077.

[7]. De Paz, J.F., Bajo, J., Vera, V., Corchado, J.M. (2011) MicroCBR: A case-based reasoning architecture for the classification of microarray data, Applied Soft Computing, Vol. 11, Issue 8, Pp. 4496-4507.

[8]. Garcia-Laencina, P.J., Sancho-Gomex, J.L., Figueiras-Vidal, A.R. and Verleysoen, M. (2009) K-nearest Neighbours and mutual information for simulataneous classification and missing data imputation, Neurocomputing, Elsevier, vol. 72., pp. 1483-1493.

[9]. Hassan, M.M., Dowd, A.A., Ibrahim, F.I., Mohamed, A.H., Kaheel, H.H. and Hassan, M.A.,(2014) In silico analysis of single nucleotide polymorphisms (SNPs) in human HLA-A and HLA-B genes responsible for renal transplantation rejection, European Academic Research, Vol. II, Issue 3, Pp.3627-3646.

[10]. Huang, G.B., Zhu, Q.Y. and Siew, C.K. (2006) Extreme learning machine: Theory and applications, Neurocomputing, Vol. 70, Pp. 489–501.

[11]. Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D. and Bijnens, L. (2012) Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R: Order-Restricted Analysis of Microarray Data, Springer Publishing Company Incorporated.

[12]. Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015). Hierarchical Gene Selection and Genetic Fuzzy System for Cancer Microarray Data Classification. PLoS ONE, 10(3), e0120364.

[13]. Revathi, T. and Sumathi, P. (2014) A Successive Feature Selection Algorithm for Gene Ranking, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 2, Pp. 5744-5748.

[14]. Sharma, L.S. (2010) Commerce education in north-east India, (Singh, R.A. Ed.), The role of information technology in Commerce, Chapter 9, Mittal Publications, Pp. 125-134.