Parallelized Classification of Cancer Sub-types from Gene Expression Profiles Using Recursive Gene Selection

Lokeswari VENKATARAMANA[1], Shomona Gracia JACOB[1]*, Rajavel RAMADOSS[2]

[1] Sri Sivasubramaniya Nadar College of Engineering, Department of CSE, Kalavakkam, Chennai, 603110, India. lokeswaricts@gmail.com, graciarun@gmail.com (**Corresponding author*)

[2] Sri Sivasubramaniya Nadar College of Engineering, Department of ECE, Kalavakkam, Chennai, 603110, India. rajavelr@ssn.edu.in

Abstract: Cancer is a chronic disease that is caused mainly by irregularities in genes. It is important to identify such oncogenes that cause cancer. Biological data like gene expressions, protein sequences, RNA-sequences, pathway analysis, Pan-cancer analysis and structural biomarkers could aid in cancer diagnosis, classification and prognosis. This research focuses on classifying subtypes of cancer using Microarray Gene Expression (MGE) levels. Nature of MGE data is multi-dimensional with very few samples. It is necessary to perform dimensionality reduction to select the relevant genes and remove the redundant ones. The Recursive Feature Selection (RFS) method is proposed as it repeatedly performs the gene selection process until the best gene subset is found. The obtained best subset of genes is further employed for classification using different models and evaluated using 10-fold cross-validation. In order to scale for huge amount of gene expression data, the parallelized classification model was explored on the Spark framework. A comparison was drawn between the non-parallelized classification model on Weka and the parallelized classification model on Spark. The results revealed that the parallelized classification model performs better than non-parallelized classification model in terms of accuracy and execution time. Further, the performance of RFS and parallelized classifier was also compared with previous approaches. The proposed RFS and parallelized classifier outperformed previous methods.

Keywords: Recursive Feature Selection, Gene Selection, Microarray Gene Expression, Parallelized classification, Random Forest.

1. Introduction

All living organisms have cells except Viruses. Humans have trillions of cells and each cell contains a complete copy of genome which is encoded into DNA [1]. A Gene is a section of DNA. Gene denotes how to make a protein. Gene expression is the procedure by which information encoded in a gene is converted into a protein. In cellular organisms, expression of right genes in right order at right time is crucial, particularly during embryonic development and cell differentiation [1]. Gene sequencing involves defining the order of bases and nucleotide units such as Thymine, Guanine, Cytosine and Adenine (T, G, C, A). When there is any change or variation in gene sequences, it results in inflammation of cell, resulting in cancer [9].

A gene is the basic physical and functional unit of heredity. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. There are between 20,000 and 25,000 genes in humans as estimated by The Human Genome Project. Every person has two copies of each gene, one inherited from each parent. Most genes are the same in all people, but a small number of genes are slightly different among people. Cancer is basically a disease of '*genes gone bad*' [3]. Many genes control the way cells grow, divide and die. When there is a fault in this, cell division may go out of control. Different kinds of cancer are caused by different sets of genes. Hence, in order to treat cancer, it is essential to know which of the genes in a cancer cell behave abnormally.

The DNA micro-array is the latest breakthrough in molecular biology, which offers researchers with an approach for monitoring genome-wide expression systematically [9]. Its application in the study of cancer has proved to be successful in revealing the pathological mechanism, with the potential of altering clinical practice through individualized cancer care and ultimately contributing to the battle against cancer [1]. However, the current obstacle and challenge is on how to make use of the tremendous amount of ever-growing micro-array experimental data to specifically explain the cancer mechanism and to better predict the cancer development in the early stage [5-6].

Micro-array data has lot of noisy or irrelevant genes and missing data. This affects the accuracy of predicting a disease. Micro-array gene data is high-dimensional (thousands of genes) but a low-sample dataset. It is highly essential to weed out the irrelevant genes and select only

https://doi.org/10.24846/v27i2y201809

important genes for improving class prediction. Feature selection measures play a prominent role in improving prediction accuracy.

Feature selection uses the subset selection and ranking method to detect the relevant features. The demerits of the ranking approach lie in the fact that the number of attributes to be filtered (the choice of threshold) lies with the user [3]. This may result in a larger number of irrelevant features being considered or certain relevant features being ignored. This could be rectified by the use of subset selection methods that aim at selecting the best subset of features based on the search space, search method and the filtering criterion [7]. Two most widely used feature subset evaluators were Correlation Feature Subset Evaluator (CFS) and Fuzzy Rough Subset Evaluator (FRS) [8].

The CFS hypothesis [10] suggested that the most predictive features needed to be highly correlated to the target class and least relevant to other predictor attributes. The Fuzzy Rough Subset (FRS) evaluator plays a prominent role in case of datasets that contain continuous-valued data [8]. FRS method returns the minimal subset of features based on the Best First Search method [15]. However, the size of the subset is dependent upon the search space and the search method. Moreover, the Fuzzy approach consumes heavy computation time and memory when employed for large datasets with several attributes and a large count of instances [7]. Feature Selection by ranking methods select features by ranking the significance of the attributes based on a specific criterion [15]. The five most widely used feature selection methods by ranking evaluators are Information Gain, Gain Ratio, Symmetric Uncertainty, Chi-Square Significance and ReliefF.

Biological data is prone to grow in multiples and the existing parallelized algorithms could scale for the ever growing data but at the cost of prediction accuracy [6]. Hence it is imperative to propose novel parallelized computational methods with a parallel programming framework that can effectively handle the following issues: 1. Movement of data between nodes in a cluster, 2. Unbalanced workload between nodes, 3. Dependency between results produced by nodes in a cluster, 4. Latency in accessing files from secondary storage devices. 5. Communication and synchronization overhead. This led to a survey on parallelization of data mining algorithms [6]. The well-known parallel programming frameworks are Apache Hadoop & Spark [13, 17]. Hadoop is a parallel programming framework used to process very large datasets. Hadoop consists of a Map Reduce engine that takes care of parallel computing and a Hadoop Distributed File System (HDFS) for distributed storage [6]. The drawback of Hadoop Map Reduce is latency in accessing files or results of iterative tasks from HDFS [6]. Spark is an Apache framework which runs on top of YARN (Yet Another Resource Negotiator) and it has the advantage of in-memory computation with the Resilient Distributed Dataset (RDD). Spark provides Machine Learning library (Spark MLlib) which helps to handle execution of machine learning algorithms on huge amount of data [12-13]. SparkMLlib provides support for various machine learning algorithms to be computed in parallel by using many cores in a processor. In contrast, Weka is an open source data mining software suite [11], that applies serialized computation on data. The focus of this research is to propose a Recursive Feature Selection method for obtaining the best subset of genes and exploit SaprkMLlib for parallelized classification algorithms that can classify subtypes of cancer. The results need to be compared with the performance of the non-parallelized computational methods and previous research outcomes. The following section surveys the work carried out in this field of research.

2. Literature Review

There are more than 100 cancer types based on the organ of occurrence in the body wherein fact cancer first developed or by the type of tissue cell in which they originate (histological type) [1]. According to tissue cell criterion, cancers may be categorized into six major categories: Carcinoma, Sarcoma, Myeloma, Leukemia, Lymphoma and Mixed Types [9]. Based on the primary site of origin, cancers may be of specific types like breast cancer, lung cancer, prostate cancer, liver cancer, renal cell carcinoma (kidney cancer), oral cancer, brain cancer, etc [9]. In cancer medical diagnosis, classification of the tumor types is of supreme importance. An accurate prediction of several

tumor types offers better treatment and toxicity minimization on patients [9].

Effective ensemble classifiers increase not only the performance of the classification, but also the reliability of the results. The motivations beyond using ensemble classifiers are that the results are less dependent on peculiarities of a single training set and that the ensemble system outperforms the performance of the best base classifier [14]. Their proposed method provided three advantages namely enhancing result accuracy, applying the ensemble technique to more cancer types, and mitigating the effect of over-fitting. The proposed system utilized three different feature selection algorithms: (i). Backward Elimination Hilbert-Schmidt Independence Criterion "BAHSIC". (ii). Extreme Value Distribution based gene selection "EVD". (iii). Singular Value Decomposition Entropy gene selection "SVDEntropy". The proposed ensemble system consisted of 5 base classifiers; all base classifiers implemented 3-NN algorithm. Each classifier made use of its own feature selection parameters in order to ensure the diversity of the ensemble. The first three classifiers utilized BAHSIC feature selection algorithm with different number of genes to select. The number of genes selected from genes' pool was a user-defined parameter and it was set to 50, 5 and 25, respectively. The fourth base classifier utilized EVD gene selection algorithm with automated algorithm for defining the number of genes to select; number of genes selected by the algorithm was 49 for Colon dataset, 224 for Leukemia dataset and 5127 for the Breast cancer dataset; the last base classifier utilized SVD entropy gene selection algorithm; the number of genes returned by the algorithm was 240 for Colon dataset, 187 for Leukemia dataset and 1236 for the Breast cancer dataset.

Further, the idea of ensembling was adapted for feature selection. V.Bolón-Canedo [4] proposed an ensemble of filters for classification, aimed at achieving good classification performance together with a reduction in the input dimensionality. The authors tried to overcome the problem of selecting an appropriate method for each problem, as the former was fully dependent on the characteristics of the datasets. Ensemble filters was applied on seven gene expression datasets for classifying cancer subtypes. The classifiers used were C4.5 Decision tree, Naïve Bayes, Instant-based classifier (IB1) and Support Vector Machine (SVM). [4], [5].

The Distributed Feature Selection for the process of distributing the feature selection process was proposed by V.Bolón-Canedo [6]. It distributed the data by features, i.e. according to a vertical distribution, and then performed a merging procedure that updated the feature subset according to improvements in the classification accuracy. The effectiveness of their proposed method was tested on microarray data, that was characterized by a high number of gene expressions but with a small sample size. The results related to eight microarray datasets revealed that the execution time was considerably shortened whereas the performance was maintained or even improved compared to the standard algorithms applied to the non-partitioned datasets.

Research by Das et al. [7], presented an algorithm which, based on a horizontal partition (by samples), performed feature selection in an asynchronous fashion with a low communication overhead by which each peer could specify its own privacy constraints. A vertical partition of the data (by features) to generate the diverse components of an ensemble [16] was also present in the literature. However in those cases, feature selection was not applied to the different partitions of the data and therefore the model might have been constructed even on irrelevant features. More recently, Banerjee and Chakravarty [3] proposed a distributed feature selection method evolved from a method called Virtual Dimension Reduction, enabling the partition of data both vertically and horizontally. Zhao et al. [18] presented a distributed parallel feature selection algorithm based on maximum variance preservation. The algorithm could read data in a distributed form and performed parallel feature selection in both symmetric multiprocessing modes via multithreading and massively parallel processing. The proposed method included performing a number of fast filters over several partitions of the data and combining features into a single subset of features. Thus, the dataset D was separated into several small disjoint subsets D_i. The filter was applied to each of them, generating

a corresponding selection S_i. After all the small datasets D_i had been used (which could be done in parallel, as all of them were independent of each other), the combination method built the final selection S as the result of the filtering process. The partitioned dataset consisted of dividing the original dataset into several disjoint subsets of approximately the same size. The dataset was split vertically. Two different methods were used for partitioning the data: (i) performing a random partition and (ii) ranking the original features before generating the subsets. The second option was introduced in order to try to improve the performance obtained by the first one. By having an ordered ranking, features with similar relevance to the class would be in the same subset that facilitated the task of the subset filter that was to be applied later. Those two techniques for partitioning the data would generate two different approaches for the distributed method: *Distributed Filter (DF)* with the random partition and *Distributed Ranking Filter (DRF*) associated to the ranking partition. The performance of the distributed filter was tested over eight DNA microarray datasets. The feature selection algorithms used were Correlation-based Feature Selection (CFS), Consistency-based Filter [8], INTERACT algorithm, Information Gain filter and ReliefF. The classifiers used to classify cancer subtypes were C4.5 Decision Tree, Naïve Bayes, K-NN and SVM.

Distributed Feature Selection and ensemble classifiers motivated the authors of this research work to propose a new method of feature selection which partitions the data along feature (gene)-wise and performs gene selection recursively until only one subset of best genes was obtained. The obtained genes were used to construct the parallelized classification model for classifying subtypes of cancer. These computational methods are applied on Microarray Gene Expression (MGE) data. Further the proposed method was inspired by the Rank-Weight Feature Selection method proposed by Ramani R. G et al. [15] which made the most of the filtering capacity of more than one feature selection algorithm to identify an optimal set of predictive genes and generated higher prediction accuracy over five major cancer types from MGE data. The filtered features (genes) were weighted based on the number of feature relevance algorithms reporting them to be significant. The results of the proposed Recursive Feature Selection (RFS) and parallelized classification model were compared with the results of the RWFS method [15] non-parallel classification algorithms to depict the dominance of the proposed method.

3. Materials and Methods

Microarray gene expression (MGE) dataset for five different cancer types have been collected from AI Orange labs, Ljubljana [2].

Datasets include: 1. Brain Tumor, 2. Gastric Cancer, 3. Glioblastoma, 4. Lung Cancer and 5. Childhood Leukemia. The platform from which dataset was collected is Affymetrix Human Genome Array. All gene expression datasets consisted of continuous values. The description of cancer gene expression data is tabulated in Table 1.

Methods

The following methods were explored for classifying cancer using MGE data.

3.1 Correlation Feature Subset Selection (CFS) Attribute Evaluator Method

The removal of irrelevant and redundant information often improves the performance of machine learning algorithms [10]. Feature subset selection is the process of identifying and removing as much of the irrelevant and redundant information as possible.

The following equation specifies the merit of a feature subset S that consisted of 'k' features.

$$Merit_{S_k} = \frac{k\overline{r}_{cf}}{\sqrt{k + k(k-1)\overline{r}_{ff}}} \quad (1)$$

where $\overline{r}_{cf}$ was the average value of all feature to class correlations, and $\overline{r}_{ff}$ was the average value of all feature to feature correlations. The CFS criterion was defined as follows.

$$CFS = \underset{S_k}{MAX}\left[\frac{r_{cf1} + r_{cf2} + \ldots + r_{cfk}}{\sqrt{k + 2(r_{f1f2} + \ldots + r_{fifj} + \ldots + r_{fkf1})}}\right] \quad (2)$$

where r_{cfi} and r_{fifj} variables are referred to as correlations. The attributes that portrayed a high

Table 1. Microarray Gene Expression Dataset Description

Gene Dataset	No. of Genes	Total Samples	Target Class	Class wise samples	Cancer Sub-types
Brain Tumor	7129	40	5	10 10 10 4 6	1. Medulloblastoma 2. Malignant glioma 3. Rhabdoid tumor 4. Normal cerebellum 5. Primitive Neuroectodermal tumor
Gastric Cancer	4522	30	3	8 5 17	1. Normal gastric tissue 2. Diffuse gastric tumor 3. Intestinal gastric tumor
Glioblastoma	12625	50	4	14 7 14 15	1. Classic Glioblastoma 2. Classic Oligodendroglioma (CO) 3. Nonclassic Glioblastoma (NG) 4. Nonclassic Oligodendroglioma (NO)
Lung Cancer	10541	34	3	17 8 9	1. Squamous cell carcinoma 2. Adenocarcinoma 3. Normal lung tissue
Childhood Leukemia (Acute Lymphoblastic Leukemia)	8280	60	4	13 21 16 10	1. Mercaptopurine alone (MP) 2. High-dose methotrexate (HDMTX) 3. Mercaptopurine and low-dose methotrexate (LDMTX_MP) 4. Mercaptopurine and high-dose methotrexate (HDMTX_MP)

correlation to the target class and least relevance to each other were chosen as the best subset of attributes. The attributes filtered by the CFS subset evaluator method were given as input to classifier algorithms.

3.2 Predictor Models

The attributes selected by the CFS were evaluated by constructing two predictor models namely Decision Tree classifier and Random Forest. The choice of classification model was based on the results of previous research [14]. Decision Tree can be built in parallel easily and it naturally handles both categorical and numeric features [17]. Decision Tree uses measures such as Information Gain, Gain Ratio or Gini Index to select the first best attribute to form the root of the tree. Further instances were split accordingly and second best attribute was found. This process repeats until all instances belong to the same class label. The construction of multiple decision trees forms a Random forest by splitting dataset into sub-samples and using each sub-sample for a decision tree model. The ensemble of decision trees was used as a base classifier to construct Random Forest. The number of trees constructed to form a random forest was 100. Finally the constructed model was validated by feeding the test sample to the constructed predictor model. A 10-fold cross-validation was used to evaluate the parallelized random forest model on Spark framework. The server used for computations was RackServer R220. The ensuing section details the proposed Recursive Feature Selection Method.

4. Recursive Feature Selection

A Recursive Feature Selection (RFS) method was proposed to iteratively select important genes (features) from high-dimensional MGE data. Figure 1 depicts the overall process of gene selection based on RFS method. Initially MGE data is extracted from the Artificial Intelligence Orange labs, Ljubljana [2] for five cancer types namely Brain cancer, Glioblastoma, Gastric cancer, Lung cancer and Childhood Leukemia. The raw dataset has to be pre-processed for eliminating missing values and converting discrete class labels into numeric class labels. The pre-processed dataset was split vertically exploiting the *Distributed Filter (DF)* with the random partition as mentioned in Distributed Feature Selection [6]. Each partition consists of 4000 genes

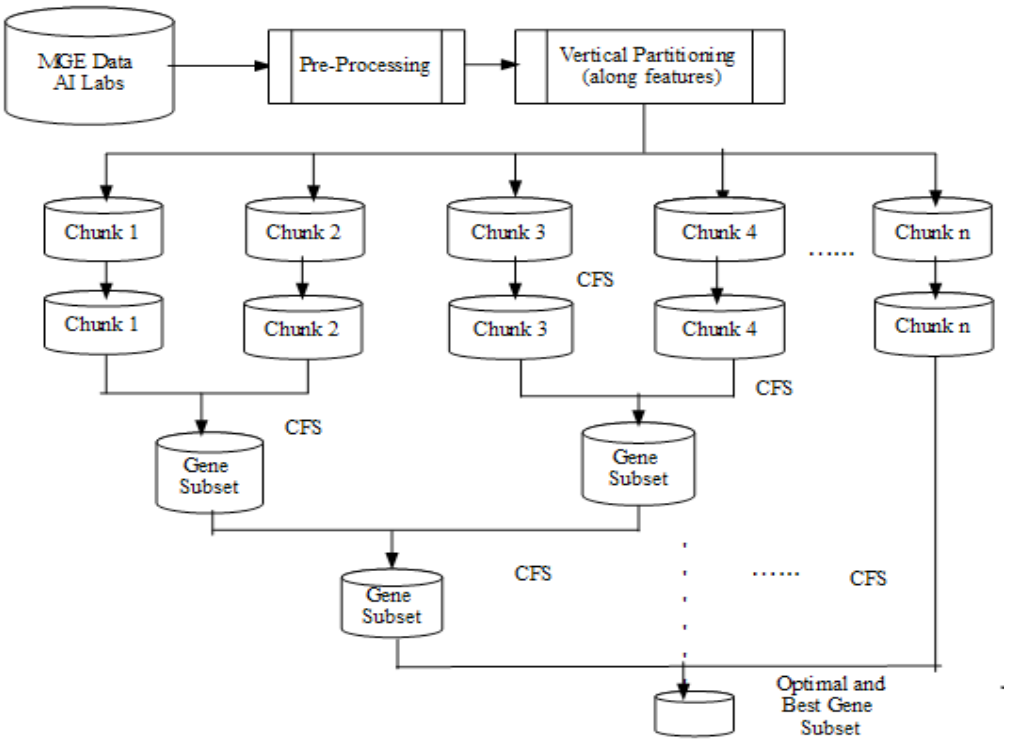


Figure 1. Recursive Selection of Features (Genes)

which was purely an experiment-based decision. Correlation Feature Subset Selection (CFS) was applied on each chunk of data to obtain the best feature subset. This results in finding intra-chunk relevant features. The relevance among chunks has to be found by applying CFS across chunks of data. This process has to be repeated until only one chunk of optimal feature set is obtained, hence the name Recursive Feature Selection (RFS). Space Complexity depends on the number of features. The space complexity is *O(N),* where *N* is the number of genes (features). Let N be the number of chunks formed due to vertical partitioning. Considering the time complexity, the first time, the algorithm runs *N* times, second time it executes *N/2* times, the third time it runs *N/4* times and so on until it reaches 1. So, the time complexity of the algorithm in *O(N).*

5. Ensemble of Classifiers for Improved Classification of Cancer Sub-types

The optimal features obtained from RFS have to be evaluated using parallelized classification model. The ensemble of classifiers was explored with the purpose of improving classification of cancer subtypes from MGE data. The ensemble method utilized for classification was Random Forest. Apache Spark, a Parallel Programming framework was employed to parallelize the task of model construction and classification. Initially parallelized decision tree was constructed using SparkMLlib. The number of trees considered for building forest was 100.

The majority voting of decision tree was found, to obtain the final classification. Figure 2 portrays the steps involved in classifying sub-types of cancer from MGE data. The results of RFS were given to training phase where the classifier model was built in parallel. Parallelized Random Forest was constructed using Spark framework and non-parallelized Random Forest was constructed using Weka, a data mining software. The evaluation method used was 10-fold cross-validation. As the vertical split chosen was *Distributed Filter (DF)* with the random partition, the prediction result varied in every iteration. Hence, the 10-fold cross-validation result was executed five times and the average results of five iterations was obtained. The performance of the classification model was thus evaluated. The final results obtained for the RFS-based parallelized and non-parallelized classifiers were compared with previous results (Rank-Weight feature Selection method).

Table 2. Comparison of the Performance of Proposed Computational Methods with previously reported methods

MGE Data	Total number of Instances	Total number of Genes	Previously reported results		Proposed RFS and Parallelized Classifier	
			No. of Genes selected by RWFS [15]	Accuracy in % (10-Fold CV)	No. of Genes selected by RFS	Accuracy in % (10-Fold CV)
Brain 5C	7129	40	3	77.5	57	93
Gastric3C	4522	30	4	93.3	25	92
Glio4C	12625	50	5	90	66	96
Lung3C	10541	34	3	94.1	32	98
Child4C	8280	60	6	65	17	64

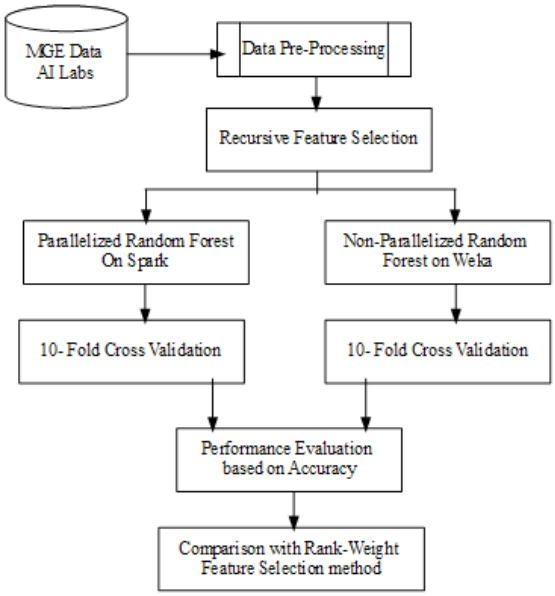


Figure 2. Parallelized Computational Methods for Classification

6. Results and Discussion

The results of this research work will be discussed in two sections. (i). Improved performance of the proposed parallelized Random Forest with the RFS method and its comparison to previous work and (ii). The performance of RFS-based parallelized classification will be compared with RFS-based non-parallelized classification.

6.1 Comparison of proposed method with previous research

The previously proposed Rank-Weight Feature Selection method [15] identifies the important and optimal features by ranking each feature obtained from six different feature selection algorithms. When this technique was applied to five different cancer datasets, it resulted in filtering out a very small numbers of genes as significant ones. These genes were evaluated using ten classification algorithms. On the other hand, the Recursive Feature Selection unearthed the important genes by repeatedly applying CFS across chunks of data. The selected genes were evaluated using parallelized Random Forest classifier and the sub-types of cancer were predicted with high accuracy. Table 2 depicts the number of genes (features) selected by RWFS and RFS. Total number of genes in Table 2 is the feature vector length. The selected features were considered as significant, since it improved classification accuracy. The selected features were validated by constructing the classifier model and evaluating them using the 10-fold cross-validation. The MGE data namely Brain5C represents brain cancer with five diagnostic classes, Gastric3C with three diagnostic classes, Glio4C with four diagnostic classes, Lung3C with three diagnostic classes and Child4C with four diagnostic classes. The classification accuracy was tabulated. It was obvious from the table that RFS-based parallelized Random Forest algorithm drastically improved the classification accuracy for Brain cancer from 77.5% to 93%. The classification accuracy increased from 90% to 96% and 94.1% to 98% for Glioblastoma and lung cancer respectively. However, considering the Gastric cancer and Childhood Leukemia data, the obtained accuracy was more or less equal to the previous results. The limitation of the proposed method is that it selects many number of genes (features) compared to previous method (RWFS) for yielding improved classification accuracy.

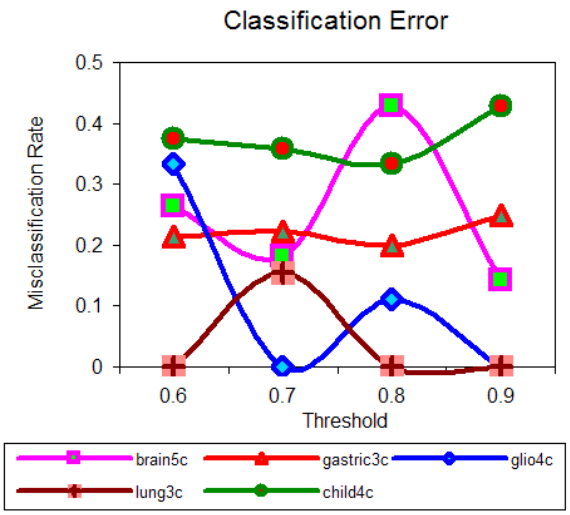
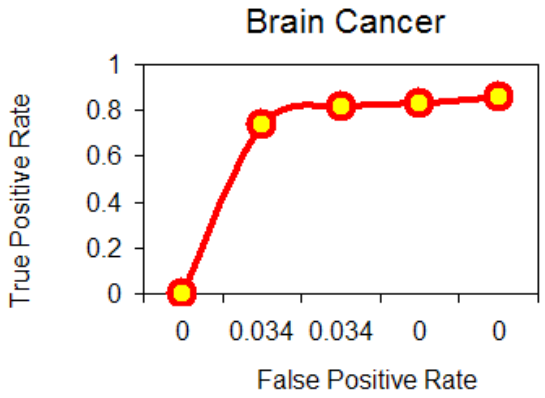
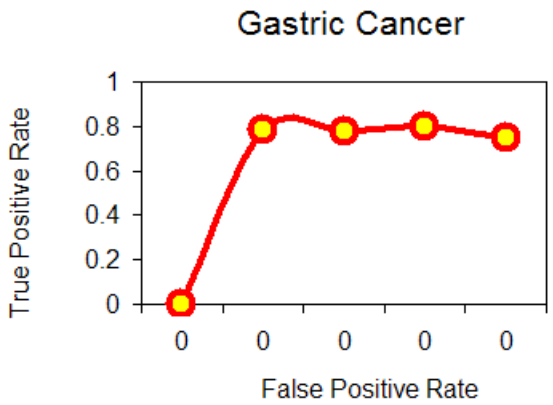


Figure.3 Misclassification Rate for MGE data

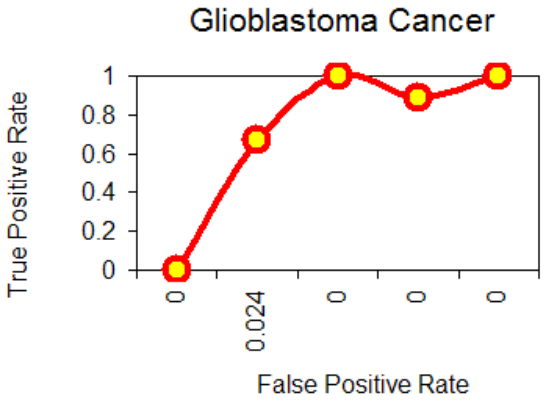
(a)



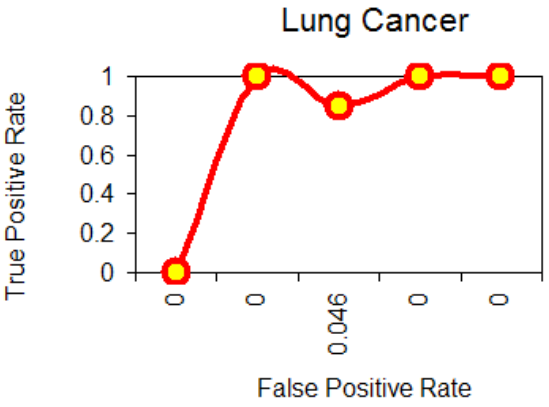
(b)



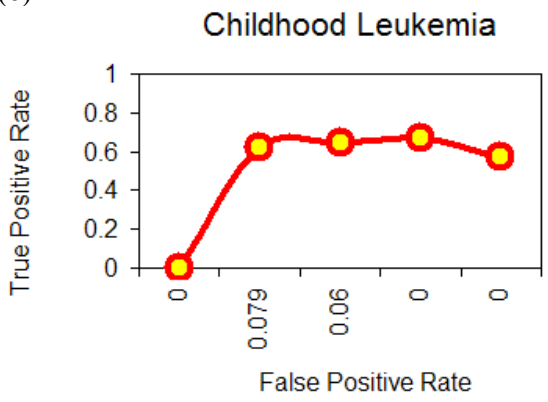
(c)



(d)



(e)



a: Brain cancer, b: Gastric Cancer, c: Glioblastoma, d: Lung Cancer, e: Childhood Leukemia

Figure 4. ROC curves for MGE data

Figure 3 shows the misclassification rate of the Random Forest classifier for each type of cancer. Receiver Operating Characteristics (ROC) is the plot between False Positive Rate (FPR) and True Positive Rate (TPR) for different threshold values [14]. Figure 4(a) to 4(e) represent the ROC curves for Random Forest classifier with the threshold values 0.6, 0.7, 0.8 and 0.9 which was applied on all five MGE datasets.

6.2 RFS-based parallelized Vs non-parallelized classification.

In order to scale for huge amount of data, parallelized machine learning algorithms were discovered. The challenge of utilizing these parallelized machine learning algorithms is that they focus on reducing execution time at the cost of accuracy. Hence, this research aimed at selecting the significant genes prior to parallel classification. The accuracy of RFS-based parallelized Random Forest was compared with that of RFS-based non-parallel Random Forest. Figure 5 illustrates the classification accuracy of

non-parallel Random Forest classifier compared to parallel Random Forest classifier. Parallelized classification results outperform those of the non-parallelized classifier for all four cancer types except for Childhood Leukemia. This is attributed to the fact that the distinguishing characteristics between the different subtypes are very minimal. This could be further studied and investigated in the future work.

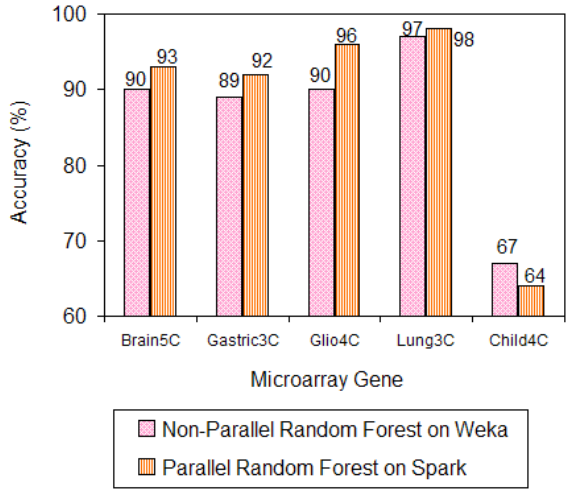


Figure 5. Performance of RFS method with non-parallel and parallelized classification method.

7. Conclusion

Microarray Gene Expression (MGE) data plays a key role in diagnosing tumors and genetic diseases. Machine learning strategies are mostly used to identify significant genes and classify tumor sub-types. The major limitation is the high-dimensional nature of MGE data that misleads researchers when it comes to prediction. It is imperative to perform feature (gene) selection prior to the classification of MGE data. Moreover the processing of high-dimensional data would be computationally intensive and requires much labor and time. The proposed Recursive Feature Selection (RFS) method partition data vertically and applies feature selection separately on each chunk and also across chunks to obtain the optimal and best subset of genes. Further the best subset of genes was used to construct an ensemble of classifiers. Parallelized Random Forest was constructed using SparkMLlib. The constructed classifier was evaluated using the 10 fold cross-validation. The results reveal the fact that the RFS-based parallelized Random Forest classifier outperforms the RWFS-based non-parallel classifiers reported previously. The future scope is to exploit the hybrid feature selection algorithm in order to further reduce the number of genes required for classification without compromising on classification accuracy. To a great extent, it would be a good initiative to analyze the Next-Generation Sequence (NGS) data which is huge with the help of parallelized computational methods.

Acknowledgements

This research work was carried out as part of project funded by Science and Engineering Research Board (SERB), Department of Science and Technology (DST) funded project under Young Scientist Scheme – Early Start-up Research Grant- titled “Investigation on the effect of Gene and Protein Mutants in the onset of Neuro-Degenerative Brain Disorders (Alzheimer’s and Parkinson’s disease): A Computational Study” with Reference No- SERB – YSS/2015/000737.

REFERENCES

1. Alshamlan, H. M., Badr, G. H. & Alohali, Y. (2013). A study of cancer microarray gene expression profile: objectives and approaches. In *Proceedings of the World Congress on Engineering*, *vol. 2* (pp. 1-6).

2. Artificial Intelligence Orange Labs. Ljubljana (2018). Available online at: <http://www.biolab.si/supp/bi-cancer/projections/> (07 June 2018, date last accessed).

3. Banerjee, M. & Chakravarty, S. (2011). Privacy preserving feature selection for distributed data using virtual dimension. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 2281-2284). ACM.

4. Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. (2014). Data classification using an ensemble of filters, *Neurocomputing*, *135*, 13-20.

5. Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. (2012). An ensemble of filters and classifiers for microarray data classification, *Pattern Recognition*, *45*(1), 531-539.

6. Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. (2015). Distributed feature selection: An application to microarray data classification, *Applied soft computing*, *30*, 136-150.

7. Das, K., Bhaduri, K. & Kargupta, H. (2010). A local asynchronous distributed privacy preserving feature selection algorithm for large peer-to-peer networks, *Knowledge and information systems*, *24*(3), 341-367.

8. Dash, M. & Liu, H. (2003). Consistency-based search in feature selection, *Artificial intelligence*, *151*(1-2), 155-176.

9. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, *286*(5439), 531-537. Elsevier.

10. Hall, M. A. (1999). *Correlation-based feature selection for machine learning,* Doctoral dissertation. The University of Waikato.

11. Hall, M., Frank, E. et al. (2009). The WEKA data mining software: an update, *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

12. Meng, X., Bradley, J. et al. (2016). Mllib: Machine learning in apache spark, *Journal of Machine Learning Research*, *17*(34), 1-7.

13. Parallel Programming Framework Apache Spark (2018). <http://spark.apache.org/> (11 June 2018, date last accessed).

14. Piao, Y., Piao, M. & Ryu, K. H. (2017). Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles, *Computers in biology and medicine*, *80*, 39-44.

15. Ramani, R. G. & Jacob, S. G. (2013). Benchmarking classification models for cancer prediction from gene expression data: A novel approach and new findings, *Studies in Informatics and Control*, *22*(2), 133-142.

16. Rokach, L. (2009). Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography, *Computational Statistics & Data Analysis*, *53*(12), 4046-4072.

17. Ryza, S., Laserson, U. et al. (2015). *Advanced Analytics with Spark: Patterns for Learning from Data at Scale*. O'Reilly Media, Inc.

18. Zhao, Z., Zhang, R., Cox, J., Duling, D. & Sarle, W. (2013). Massively parallel feature selection: an approach based on variance preservation, *Machine learning*, *92*(1), 195-220.