May - June 2020
ISSN: 0193-4120 Page No. 4560-4565

Small Round Blue Cell Tumor Classification using Pipeline Genetic Algorithm

Nimrita Koul[1], Sunilkumar S Manvi[2]

[1,2]School of Computing and Information Technology
REVA University, Bangalore, India
[1]nimritakoul@reva.edu.in

Article Info
Volume 83
Page Number: 4560-4565
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract

Computational classification of cancerous tumors is an important research problem in machine learning. A number of approaches have been proposed by researchers to achieve accurate differentiation of samples as cancerous or non- cancerous or to differentiate different stages of a cancer. This process of computational classification has also been successfully carried out by using gene expression data as input. In this paper, we have proposed an evolutionary technique based on genetic algorithms for classification of small round blue cell tumors. This tumor occurs in four subtypes, our method has been able to differentiate these four types with 100% accuracy. The method has been compared with existing methods and has been shown to perform very well with respect to classification accuracy, recall, precision and support.

Keywords: *Gene Expression Data, Small Round Blur Cell Tumor, Cancer Classification, Genetic Algorithm, Evolutionary Feature Selection*

1. Introduction

Abundant availability of high throughput gene expression data[1] from DNA micro array experiments and next generation sequencing technologies has enabled computing researchers to analyze this data through statistical and machine learning methods and derive meaningful insights from it. The focus of this work is on the use of DNA microarray gene expression data. This data contains the expression values of entire set of genes in the genome of an organism being sequences. The number of these genes can be of the order of thousands. The number of genes in human genome is around twenty four thousand. The DNA microarray data is therefore very high in the number of dimensions, 'n', where each gene corresponds to a dimension, while the number of samples, 'p', is very small. This is the typical case of p<<n scenario. Also, before this data can be used for analysis using any machine learning processing pipeline, it must be cleaned, standardized, normalized and the number of dimensions must be reduced to just the relevant dimensions. One very important application which used gene expression datasets is diagnosis, prediction and classification of cancers. The authors in [1] were the first to use gene expression datasets for classification of Leukemia profiles into two subclasses ALL and AML.

The focus of this work is on the use of DNA microarray gene expression data for classification of small round blue cell tumors (SRBCT) into four subclasses by analysis of SRBCT gene expression data set [2] using evolutionary algorithms. In order to perform classification task, the first step is dimensionality reduction through feature selection. Feature selection is used to obtain a subset of most relevant genes that have the highest influence on the class of a sample.

A. Feature Selection

One of the most important steps in machine learning tasks is that of feature selection [1,3]. Researchers have proposed numerous approaches for selecting the subset of most relevant features from the original high dimensional feature set. This not only reduces the curse of dimensionality but also improves the performance of classifier algorithms that are very sensitive to noisy

May - June 2020
ISSN: 0193-4120 Page No. 4560-4565

features, it also makes the computation less expensive. The methods for feature selection have been classified according to the number of features that are evaluated at a time during shortlisting process and according to the principle used for shortlisting of features. If one feature is evaluated at a time, the method is known as univariate feature selection method, if a group of features is evaluated at a time, the method is known as multivariate method. Under univariate feature selection methods, the criteria like signal to noise ratio, threshold number of misclassifications, correlation coefficients, mutual information, information gain, Naïve Bayes global relevance, Euclidean distances [4], median vote relevance, Wilcoxon statistic and t-statistic are used to evaluate each gene and the genes with optimal values of these properties above a threshold are ranked and top ranked genes are retained in the reduced feature subset. Univariate gene selection involves searching a space of '2^n' subsets of genes where 'n' is the dimensionality of original gene set.

Multivariate gene selection methods use combinatorial search over all possible subsets of original features. This search does not consider one feature at a time rather relevance of groups of multiple genes is considered in each search cycle. The search techniques generally applied in multivariate feature selection include – simple forward search, floating search methods, genetic algorithms, and iterative backward search. A prominent example of backward search is recursive feature elimination (RFE) which is often used with SVM[5] classifier as ranking procedure for the genes. Top scoring pair is a method of multivariate feature selection that considers genes in pairs for evaluation. In both classes of methods, the input gene set is split into a training set and a validation set. The classifier is trained on the training set and a gene or a set of multiple genes from training set is evaluated by observing the performance of classifier algorithms on this gene (univariate) or group of genes (multivariate) [10]. The gene subsets what give best performance with the classifier are returned. Since this problem involves two optimization problems viz. selection of gene subsets that maximize classification accuracy and at the same time to minimize the size of selected subset of genes, this problem is often treated as a multi-objective optimization problems. Another way of classifying feature selection methods is that of filter, wrapper and embedded methods. The filter methods which cover most univariate methods of feature selection use the concepts of information theory such as information gain, entropy, RelieF, Gini Index, Chi-square [10], to evaluate features and the selected ones are those with highest evaluation rank. The selected features are then used for the task of classification. Wrapper methods use a fitness function such as mean square error or classification accuracy to rank the features. In wrapper algorithms, an optimization technique is used to optimize the fitness function e.g. to minimize the MSE or to maximize the classification accuracy. These optimization algorithms prevent the need to follow an exhaustive search of feature space hence reduce the time required for feature selection. Some examples of optimization algorithms that have been used by researchers are genetic algorithms, cuckoo search algorithms, whale optimization. Embedded methods have the estimator algorithm built into the model. The wrapper approaches are expensive computationally and prone to overfitting.

B. Genetic Algorithms

Genetic algorithms [7,8,9] are stochastic algorithms which are inspired by the phenomenon of natural selection in biological evolution. The chromosomes of parents which carry the genes, undergo crossover to create new chromosomes with a varying gene sequence, there may happen random mutation which further changes the gene structure of child chromosome in next generation. The cross over, according the theory of evolution, encourages survival of the best feature bearing genes. Borrowing from this concept, genetic algorithms have been formulated as a group of computer science optimization techniques for selection of best solutions. In case of feature selection for cancer classification, the problem to solve using genetic algorithms is the selection of gene subset with maximum classification accuracy. Initial set of features are the population, and each feature is an individual. In each generation, individuals are selected ranked on their fitness value computed by an estimator, the fittest features are combined to form next generation. The next generation may also undergo mutations. Figure 1. Shows the cycle of selection, cross-over, mutation that keeps repeating in each generation till a termination criteria is met or the maximum number of generations is reached.

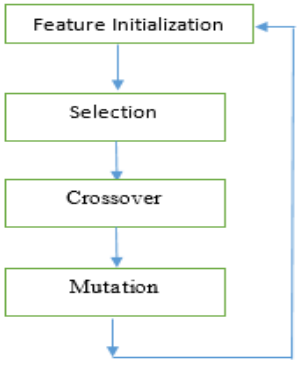

Figure 1: The Initialization, selection, crossover, mutation cyce in genetic algorithms

Rest of this paper is structured ass follows – Section II presents a brief survey of the papers which also use genetic algorithms in various classification tasks, Section III presents the proposed method for classification of SRBCT tumors, Section IV presents the results and discussion, Section V presents the conclusion of the paper.

May - June 2020
ISSN: 0193-4120 Page No. 4560-4565

2. Literature survey

In this section we will present a brief survey of the related works which have used genetic algorithms or other evolutionary algorithms for clasification task applied to gene expression data or related problems.

In [1], authors for the first time demonstrated the use of information theoritic approach viz. signal to noise ratio for accurate classification of Leukemia into ALL and AML subtypes. They were the poineers who launched a series of computational research experiments into cancer classification from gene expression datasests.

In [2], authors used the artificial neural networks as classifier and feature selectors for SRBCT subclasses. The dataset used in our work is from this reference.

In [3], authors, have used Bhatacharya distance as a measure of feature fitness to perform classification. This is a filter methods with univariate analysis.

In [5], authors have presented a floating search method for feature selection.These methods use backtracking to remove wrongly selected features. The performance was shown to be better that comparative methods. However, the computational time was higher.

In [6], authors have proposed a filtering technique based on various ranking methods for selection of strong genes that are indicative of cancers. They have compared it with combinatorial search methods and found it to perform better.

In [7], authors have used a genetic algorithm for feature selection from electric data.

In [8], authors have created a hybrid of neural network and genetic algorithm fro feature selection from microarray gene expression datasets.

In [9], authors have proposed a robust hybrid algorithms between support vector machine and genetic algorithm for feature selection from genomic data.

In [10], the authors have presented an exhaustive survey of the univariate and multivariate methods for feature selection from gene expression datasets.

In [11], authors have presented techniques for multivariate feature selection to construct an optimal feature subset for classification problems.

In [12] and [13], authors have presented a survey of evolutionary algorithms that can be used for feature selection for classification tasks. They have presented comparison on basis of classification accuracy and computational times.

In [14], authors have used genetic algorithm for feature selection using artificial neural network pattern classifiers.

In [15], authors have used a mixiumum relevance minimum redundancy appraoch for feature selection from biological datasets. This was used with SVM classifier as evluator.

3. Proposed Approach

In this section we present the proposed method and experimental setup of feature selection and classification on SRBCT dataset.

The entire feature set of original data set with 2803 features are treated as first generation of the genes. These genes are evaluated for fitness using the logistic regression estimator, the top ranking genes are selected and allowed to cross over. Local mutations are performed on the crossed over genes in the second generation. And this generation is also evaluated with estimator score. This process continues till a minimal subset of genes with maximum score is obtained. This subset is returned.

The implementation was carried out in Python 3.7 on Windows 10 machine. The input dataset contains 2308 genes and 62 samples. There are four classes of tumors in the dataset namely – Neuroblastoma, Ewing's family of tumors, non-Hodgkin lymphoma and rhabdomyosarcoma. Dataset is already normalized and standardized with removal of zero values and missing gene values. The train test ratio was 80:20 and the 5 fold cross validation was used to calculate the cross validation mean square error values for each feature. The one over rest method for converting multiclass classification was used with Logistic regression estimator to the genetic algorithm. Number of generations was 40, 50 and 60. The experiment was repeated 20 time and average values of all performance parameters are reported in results.

The proposed algorithm for feature selection and classification is as follows –

1. Generate the first generation of features as initial population from SRBCT dataset.
2. Split the population into training and testing set.
3. Evaluate the fitness of each feature in training set using a Logistic regression estimator the cross-validation mean square error has been taken as the fitness function.
4. The features with highest fitness, i.e. lowest value of MSE are selected for next generation.
5. Perform cross over among the selected genes
6. Allow random mutations
7. Evaluate fitness function, if it is optimal, return the current set of genes and stop
8. Else continue from step 2
9. Use the returned features for comparison with other classification algorithms

Figure 2 shows the proposed approach for feature selection for classification of SRBCT cancers..

May - June 2020
ISSN: 0193-4120 Page No. 4560-4565

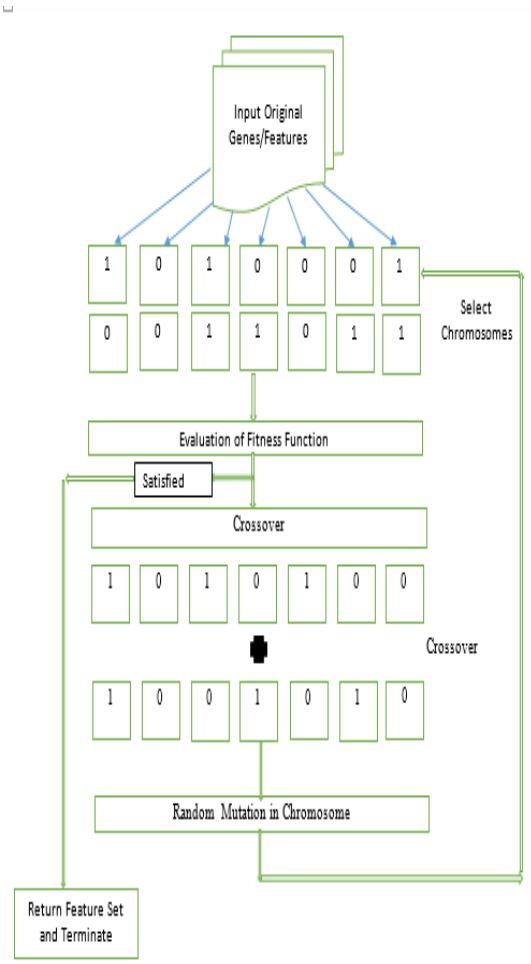

Figure 2: Schematic Representation of Proposed Approach

4. Results and Discussion

In this section, we will cover the significant results and interesting findings of the experiment as well as analyze them. We have used a train test ratio of 80:20 with five fold cross validation for computation of validation loss of each gene. Since the dataset contains multiple classes, we have used the one versus rest strategy to transform the input problem of multiclass classification to binary classification. The "ovr" method for multiclass applied to Logistic regression estimator to the genetic algorithm. Number of generations was 40, 50 and 60. The experiment was repeated 20 time and average values of all performance parameters are reported in results. Figure 3 shows the comparison of classification accuracy obtained with five feature sets. The original full feature set with dimensionality of 2308 had a classification accuracy of 100%, the feature set obtained with application of proposed algorithm also showed an accuracy of 100% with a cardinality of 50. Which is a 40 fold dimensionality reduction. Figure 4 shows the training time comparison for four classification methods, the time is reported in seconds. As shown in figure 4, the proposed method has highest training time of 23 seconds against a training time of 0.09 seconds for SVM. Figure 5 has three insets. Top of figure 5 shows comparison of learning curves on SRBCT dataset using Naïve Bayes and proposed method. Middle part of figure 5 shows the model scalability comparison of Naïve Bayes with proposed method and bottom inset of figure 5 shows the comparison of performance of the two classifiers. As shown in figure 5, the proposed method performs better in terms of learning curve, scalability and performance, however, the training time for the proposed algorithm is comparatively higher.

The proposed algorithm has shown a classification accuracy of 100% on the used dataset with just 30 genes out of an original of 2308 genes which is a dimensionality reduction of 98.8%. The performance is at par with the standard methods in the field.

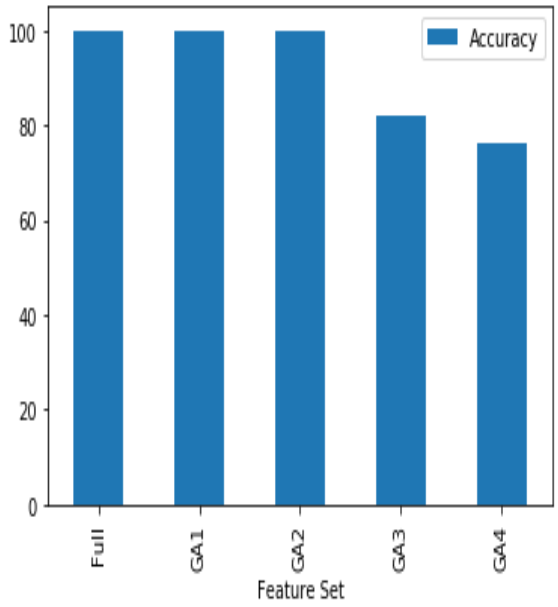

Figure 3: Classification Accuracy with 4 sets of features

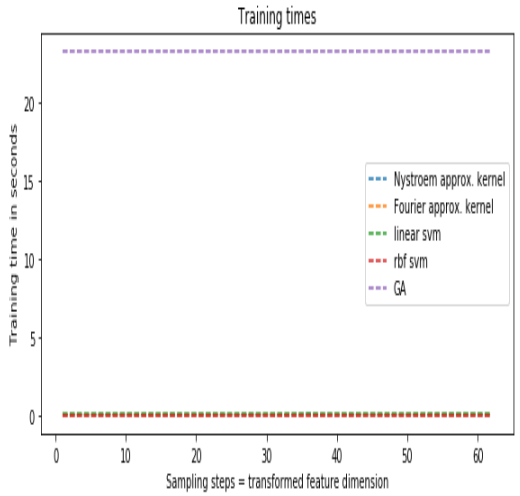

Figure 4: Comparison of Training Times with 5 different classifiers

May - June 2020
ISSN: 0193-4120 Page No. 4560-4565

5. Conclusion

In this paper, we proposed an algorithm using genetic algorithm and logistic regression for classification of cancer gene expression data for SRBCT cancer. From 2308 original features, we obtained a reduced feature set of size 50 with 100% classification accuracy by applying the proposed method. We presented the comparison with SVM, Naïve Bayes to show that the proposed method performs better. In future, we wish to apply measures to bring down the training times and apply the method on other datasets.

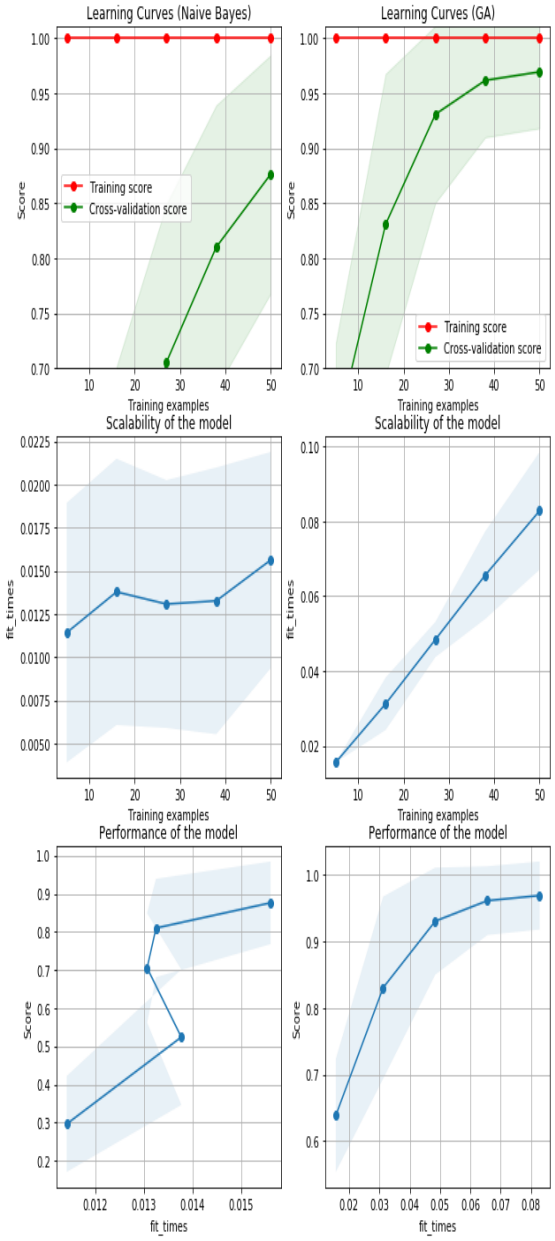

Figure 5: Comparison of Learning Curves, Scalability and Performance with Naïve Bayes classifier and proposed classifier.

Acknowledgment

Authors would like to thank Department of Science and Technology (DST), Government of India, for financially supporting this research work under the scheme DST-ICPS 2019.

References

[1] Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E: Molecular classification of cancer: dass discovery and class prediction by gene expression monitoring. *Science* 1999, 286: 531–537.

[2] J. Khan,etal.,Classificationanddiagnosticprediction ofcancersusingexpression profilingandartificialneuralnetworks,Nat.Med.7(2001)673–679.

[3] Bhattacharyya C, Grate LR, Rizki A, Radisky D, Molina FJ, Jordan MI, Bissell MJ, Mian IS: Simultaneous classification and relevant feature Identification in high-dimensional spaces: application to molecular profiling data. *Signal Processing* 2003, 83(4):729–743.

[4] Cho S, Won H: Machine learning in DNA microarray analysis for cancer classification. *Proceedings of the First Asia-Pacific bioinformatics Conference* 2003.

[5] Pudil P, Novovicova J, Kittler J: Floating search methods in feature selection. *PRL* 1994, 15: 1119–1125.

[6] Silva P, Hashimoto R, Kim S, Barrera J, Brandao L, Suh E, Dougherty E: Feature selection algorithms to find strong genes. *Pattern Recognition Letters* 2005, 26(10):1444–1453

[7] [8].Srinivas,L.M.Patnaik,Adaptiveprobabilitieso fcrossoverandmutationingenetic algorithms,IEEETrans.Syst.ManCybern.24(4)(1 994)656–667.

[8] F. Tan, et al. , A genetic algorithm based method for feature subset selection, Appl. Soft Comput. 11–20, 2007

[9] D. L. Tong, A. C. Schierz, Hybrid genetic algorithm-neural network: feature extraction for un preprocessed microarray data, Artif. Intell. Med. Vol 53, no. 1, pp. 47–56, 2011

[10] L.Li, et al. , A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset, Genomics 85 (1) (2005)16–23.

[11] D. Mishra, B.Sahu, Feature selection for cancer classification: a signal-to-noise ratio approach, Int. J. Sci. Eng. Res. 2 (4) (2011)1–7.

May - June 2020
ISSN: 0193-4120 Page No. 4560-4565

[12] Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
[13] Izenman, A.J., 2013. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, second ed. Springer, New York,USA.
[14] Xue, B., Zhang, M., Browne, W.N., 2015. A comprehensive comparison on evolutionary feature selection approaches to classification. Int. J. Comput. Intell. Appl. 14 (2).
[15] Xue, B., Zhang, M., Browne, W.N., Yao, X., 2016. A survey on evolutionary computation approaches to feature selection. IEEE Trans. Evol. Comput. 20 (4), 606–626.
[16] Yang, J., Honavar, V., 1998. Feature subset selection using a genetic algorithm. In: Feature Extraction, Construction and Selection: A Data Mining Perspective. Springer, New York, USA, pp. 117–136.
[17] Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. 5, 1205–1224.
[18] [18]. Aman Sharma, Rinkle Rani, " C-HMOSHSSA: Gene selection for
[19] cancer classification using multi-objective meta-heuristic and
[20] machine learning methods", Computer Methods and Programs in
[21] Biomedicine 178 (2019) 219–235
[22] Y. Zhang , D. W. Gong, J. Cheng , Multi-objective particle swarm optimization ap- proach for cost-based feature selection in classification, IEEE/ACM Trans. Com- put. Biol. Bioinf. (TCBB) 14 (1) (2017) 64–75.
[23] M.A. Tawhid , K.B. Dsouza , Hybrid binary bat enhanced particle swarm opti- mization algorithm for solving feature selection problems, Appl. Comput. Inf. (2018).
[24] Moslehi, F., Haeri, A. A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection. J Ambient Intell Human Comput 11, 1105–1127 (2020). https://doi.org/10.1007/s12652-019-01364-5
[25] Nádia Junqueira, MartarelliMarcelo, Seido Nagano, "Unsupervised feature selection based on bio-inspired approaches" Swarm and Evolutionary ComputationVolume 52, February 2020, 100618