Classification of Micro-array Data in Apache Spark Framework

1. Wafaa S. Albaldawi, 2 Rafah M. Almuttairi
stud.wafaa.shakir@uobabylon.edu.iq
Rafah@uobabylon.edu.iq
1) Ministry of Higher Education and Scientific Research, Department of Computer Science, College of Science, University of Kerbala, Karbala, Iraq
1,2) Ministry of Higher Education and Scientific Research, Department of Software, College of Information Technology, University of Babylon, Hillah, Iraq

Abstract

Apache Spark is an emerging huge information analytics technology. Machine learning (ML) frameworks engineered on Spark are more ascordable compared with traditional ML frameworks. We tend to build SVMwithSGD (SVM with Stochastic Gradient Descent) and LinearRegressionWithSGD models by using Spark Python API (PySpark) to classify normal and tumor microarray samples. Microarray measures expression levels of thousands of genes in a very tissue or cell kind. Feature extraction and cross-validation are used to make sure effectiveness. The SVMwithSGD and LinearRegressionWithSGD models achieve associate degrees accuracies quite eightieths.

This paper presents a study of feature selection methods effect, using a filter approach, on the accuracy and time consumed of supervised classification of cancer. A comparative evaluation among different selection methods: Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Locally Linear Embedding (LLE) is carried out with SVMwithSGD or LogisticRegressionWithSGD classifier, using the datasets of prostate, cancer, lung and Huntington's Disease samples. The classification results using SVMwithSGD and LogisticRegressionWithSGD (LGWithSGD) classifiers show that the SVMwithSGD classifier can present the highest accuracy and much time when compared with LGWithSGD. The results show that when we have classified with SVMWithSGD, PCA and SVMWithSGD is the best combination for analyzing the Borovecki, Gordon, and Chowdary datasets. While ICA and SVMWithSGD in the Singh and Chin datasets. Moreover, the results illustrate that when we have classified with LGWithSGD, PCA and LGWithSGD is the best combination for analyzing the Borovecki and Gordon datasets. While ICA and LGWithSGD in the Chowdary and Singh datasets. LLE and LGWithSGD is the best for analyzing Chin dataset.

KeyWords : PySpark, Spark, Machine Learning, Support Vector Machine, Feature Extraction, Feature Scaling, Dimension Reduction, Principal Component Analysis, Cross-Validation, independent component analysis, Stochastic Gradient Descent, Locally Linear Embedding.
Developments in biotechnology have allowed molecular biology to measure the information contained in the genes, by microarray (DNA microarrays), with the hope to provide diagnostic tools (recognition healthy tissue / cancer tissue or distinction between different types of cancer) of prostate, lung, breast cancer or Huntington's disease.

Available data contains a high number of features (genes), and a limited number of samples. The data processing needs to reduce the number of genes to provide a subset of relevant features and build a classifier predicting tumor or normal class of the sample.

In this paper, we will focus on the study of the effect of feature selection methods: PCA, ICA and LLE on the rate of supervised classification, using SVMWithSGD and LogisticRegressionWithSGD classifiers.

To do this, we will work on datasets of prostate, lung and breast cancer and Huntington's disease samples. Afterwards, we extract and select relevant features for the supervised classification, using three selection methods: PCA, ICA and LLE. Finally we will test the results of the selection operation on cancers classification using SVMWithSGD and LogisticRegressionWithSGD classifiers.

Background

Microarray dataset is an example of a high throughput data characterized with more genes than gene expression levels. With the large amount of data, a new problem surfaces about extraction useful information from them. Therefore, a number of feature selection methods and classification techniques have been illustrated in literature. Data mining and machine learning techniques have been applied in many computer applications for some time. It would be natural to use some of these techniques to assist in drawing inference from the volume of information gathered through microarray experiments.

Wahid et al 1 have proposed a correlation based overlapping score (COS) feature selection technique. Then, they have compared it with state-of-the-art of gene selection techniques by using microarray datasets. The boosting, random forest and k-nearest neighbor (kNN) classifiers have been evaluated in their experiment. They have found out that COS has the minimum misclassification errors obtained via evaluated algorithms and more steady than other gene selection techniques.

Aydadenta and Adiwijaya2 have discussed some researches about SVM, ANN, Naive Bayes, kNN, C4.5, and simulation of Random Forest algorithm with Relief technique to reduce the dimension. Their result shows that the accuracy of Random Forest algorithm higher than other evaluated classification algorithms based on microarray data.2

Sardana et al have proposed ClusterQGA hybrid approach. In this approach, they have used clustering for gene selection. Then, they have applied Quantum Genetic Algorithm to find out minimum set of relevant and non-redundant genes. Additionally another wellness work is proposed to decrease number of gene without yielding the accuracy. The adequacy of the proposed approach in contrast with existing strategies as far as order precision and number of features has been tentatively settled for both binary and multi-class freely accessible malignancy microarray datasets. The proposed approach decrease the calculation time of Quantum Genetic Algorithm for the used datasets.3
Vanitha et al. have used Mutual information (MI) between the genes and the class label for selecting the informative genes on two cancer microarray datasets. Then, they have used SVM classification algorithm. Moreover, Leave-one-Out Cross Validation (LOOCV) is utilized for model evaluation. The results have shown that dimension of the selected features has been reduced and improved classification accuracy has been obtained when compared to other methods.

In 2019, Shukla has combined multilayer and $F$-score methods for filtering the redundant and noisy genes in 10 microarray of numerous tumor datasets. Moreover, author has applied adaptive genetic algorithm based on multi population strategy with support vector machine and naïve Bayes (NB) classifiers for selecting the extremely reasonable genes on the previous step datasets. The complete outcomes disclose that first step remarkably effects on the effectiveness of the adaptive genetic algorithm with multi population strategy and upgrade the capacity of the proposed approach as far as combination rate and arrangement quality. The tests results show the prevalence of the proposed technique when contrasted with other standard coverings in regards to classification accuracy and ideal number of genes.

In addition, Fajila has used two steps for feature selection in Leukemia classification. First step, removing redundant and irrelevant gene by using multi-objective evolutionary algorithm. The next step, multi objective evolutionary based gene subset selection highly influences the perfect classification. Author has compared the performance of the proposed approach against the standard genetic algorithm and evolutionary algorithm by using three Leukemia microarray datasets. Then, achieving the perfect Leukemia classification model with few significant genes.

In 2020, Abdelnabi et al. have used information gain (IG) to select the most relevant features to the disease from the input patterns to avoid over fitting. Then, they have applied the grey wolf optimization (GWO) algorithm to reduce the selected genes. The SVM cancer type classification was the final step of their work. The Breast and Colon datasets were used to apply the proposed method. They have used classification accuracy to evaluate the model. They have found out that the proposed methodology is able to enhance the steady of the classification accuracy and the feature selection.

Zhong et al. have investigated a distance-based for microarray gene selection for two-group classification problem. They have implemented Bhattacharyya distance to measure the dissimilarity in gene expression levels between groups. Then, they have used the selected marker genes to build the SVM model. Simulation studies and two real data analysis have been used for illustration of the performance of the gene selection and classification methods.

The proposed algorithm was implemented and evaluated on central nervous system, binary leukemia, 3 classes leukemia, 4 classes leukemia, lymphoma, mixed lineage leukemia, and small round blue cell tumor microarray cancer datasets using the naive Bayes, radial basis function network, instance-based classifier, decision-based table, and decision tree classifiers. An average increase in the prediction accuracy of 5.1% was observed on all datasets averaged over all classifiers. The average reduction in training time was 2.86 seconds. The performance of the proposed method was also compared with information gain, gain ratio, and symmetric uncertainty based feature selection filters. The results were impressive when all classifiers were used on all the datasets.

Passi et al. presented and compared the performance of the various methodologies of genes subset selection methods dependent on Wrapper and Markov Blanket models for the five microarray cancer datasets. They have depended on Memetic algorithms (MAs) for feature selection method and Minimum Redundant Maximum Relevant (MRMR) for feature subset selection method hybridized by genetic search optimization techniques. They compared the performance of Markov blanket model with most basic classification algorithms for those set of features. The results showed
that the performance measures of classification algorithms based on Markov Blanket model mostly offer better accuracy rates than other types of traditional classification algorithms for the cancer Microarray datasets.

The effectiveness of the selected feature subset was evaluated using numerous classifiers on five biological datasets and five UCI datasets of a varied dimensionality and number of instances. The experimental results emphasize that the proposed method provides additional support to the significant reduction of the features and outperforms the existing methods. For microarray datasets, they found the lowest classification accuracy is 61.24% on SRBCT dataset and highest accuracy is 99.32% on large B-cell lymphoma (LBCL). In UCI datasets, the lowest classification accuracy is 40.04% on the Lymphography using k-nearest neighbor (k-NN) and highest classification accuracy is 99.05% on the Ionosphere using support vector machine (SVM).11

The new approach was based on a support vector machine (SVM) algorithm to measure the classification accuracy for selected genes. Authors evaluated the performance of the proposed mRMR-ABC algorithm by conducting extensive experiments on six binary and multi class gene expression microarray datasets. Furthermore, they compared the proposed mRMR-ABC algorithm with previously known techniques. They implemented two of these techniques for the sake of a fair comparison using the same parameters. These two techniques are mRMR when combined with a genetic algorithm (mRMR-GA) and mRMR when combined with a particle swarm optimization algorithm (mRMR-PSO). The experimental results prove that the proposed mRMR-ABC algorithm achieves accurate classification performance using small number of predictive genes when tested using both datasets and compared to previously suggested methods. This shows that mRMR-ABC is a promising approach for solving gene selection and cancer classification problems.12

An effective feature selection method, called supervised locally linear embedding and Spearman’s rank correlation coefficient (SLLE-SC2), was proposed which was based on the concept of locally linear embedding and correlation coefficient algorithms. Supervised locally linear embedding have taken into account class label information and improves the classification performance. Furthermore, Spearman’s rank correlation coefficient was used to remove the co expression genes. The experiment results obtained on four public tumor microarray datasets illustrate that the method was valid and feasible.13

In the research14, a Principal Component Analysis (PCA) dimension reduction method that includes the calculation of variance proportion for eigenvector selection was used. For the classification method, a Support Vector Machine (SVM) and Levenberg-Marquardt Backpropagation (LMBP) algorithm were selected. Based on the tests performed, the classification method using LMBP was more stable than SVM.

In 2018, Li et al15 introduced a more efficient implementation of linear support vector machines and improved the recursive feature elimination strategy and then combine them together to select genes. Besides, they proposed a simple re-sampling method to preprocess the datasets, which makes the information distribution of different kinds of samples balanced and the classification results more credible. Moreover, the applicability of four common classifiers was also studied in the paper. Extensive experiments were conducted on six most frequently used microarray datasets in this field, and the results show that the proposed methods have not only reduced the time consumption greatly but also obtained comparable classification performance.

Hadoop MapReduce programming model was used. The obtained accuracy of the improved algorithm yields better result when compared with SVM without MapReduce. The paper presented a MapReduce algorithm for classification using Support Vector Machines, aiming to accumulate time and obtain high accuracy on big gene datasets.16
Hadoop is a well-known and distributed file system framework that provides a parallel environment to run the experiment. The capability of Hadoop and statistical power of R have been leveraged to parallelize the available preprocessing algorithm called RMA to efficiently process microarray data. The experiment has been run on cluster containing 5 computers, while each node has 16 cores and 16 GB memory. It compared efficiency and the performance of parallelized RMA using Hadoop with parallelized RMA using affyPara package as well as sequential RMA. The result showed the speed-up rate of the proposed approach outperforms the sequential approach and affyPara approach.

In the paper, various statistical methods based on MapReduce were proposed for selecting relevant features. After feature selection, a MapReduce based K-nearest neighbor (mrKNN) classifier was employed to classify microarray data. These algorithms were successfully implemented in a Hadoop. A comparative analysis was done on these MapReduce based models using microarray datasets of various dimensions. From the obtained results, it was observed that these models consume much less execution time than conventional models in processing big data.

Kumar et al proposed various feature selection techniques based on statistical tests by using Apache Spark framework. Moreover, Artificial Neural Network (ANN) based on Spark framework (sf-ANN) were proposed, which runs on a scalable cluster with multiple computers. The performance of sf-ANN were tested with the help of microarray datasets of various dimensions. A detailed comparative analysis in terms of execution time was presented on sf-ANN classifier based on Spark framework and conventional system (data is stored on a standalone machine) respectively, in order to examine its performance.

The authors proposed to execute parallel Chi-Square gene selection algorithm on Spark, selected genes are evaluated using parallel logistic regression and support vector machine (SVM) for Binary classification on Spark Machine Learning library (Spark MLlib) and compare the accuracy of prediction and classification respectively. The results show that parallel Chi-Square selection followed by parallel logistic regression and SVM provide better accuracy compared to accuracy obtained with complete set of gene expression data.

Alghunaim and Al-Baity addressed the problem of breast cancer prediction in the big data. They considered two varieties of data, namely, gene expression (GE) and DNA methylation (DM). The objective of the paper is to scale up the machine-learning algorithms that were used for classification by applying each dataset separately and jointly. For this purpose, they chose Apache Spark as a platform. They selected support vector machine (SVM), decision tree, and random forest to create nine models that help in predicting breast cancer. They conducted a comprehensive comparative study using three scenarios with the GE, DM, and GE and DM combined, in order to show which of the three types of data would produce the best result in terms of accuracy and error rate. Moreover, they performed an experimental comparison between Spark and Weka platforms in order to show their behavior when dealing with big data. The experimental results showed that the scaled SVM classifier in the Apache Spark environment exceed the other classifiers, as it achieved the highest accuracy and the lowest error rate with the GE dataset.

Jędrzejowicz et al investigated a Gene Expression Programming (GEP) based ensemble classifier constructed using the stacked generalization concept. The classifier has been implemented with a view to enable parallel processing with the use of Spark and SWIM an open source genetic programming library. The classifier has been validated in computational experiments carried out on datasets. This paper presents the existing technology of microarray gene expression and classify the cancer genes using machine learning algorithms. A logical design was presented using supervised classification and gene selection model. This model can improve the process and method of identifying and classifying cancer disease using gene expression data.
Venkataramana et al.\textsuperscript{23} focused 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 was proposed as it repeatedly performs the gene selection process until the best gene subset has been found. The obtained best subset of genes was further employed for classification using different models and evaluated using 10-fold cross validation. In order to scale for big 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. Furthermore, the performance of RFS and parallelized classifier was also compared with previous approaches. The proposed RFS and parallelized classifier exceed previous methods.

Therefore, we decide to choose apache spark framework to perform classification of microarray data because it can be considered as a big data due to features’ numbers larger than microarray samples to reduce consumed time and enhanced the performance.

Dataset and Methods

Microarray data

The five Microarray datasets in CSV tarballs for MATLAB and Python. The datasets are downloaded from this link https://github.com/kivancguckiran/microarray-data/tree/master/csv. Each data set is listed by the first author on the original paper.

![Sample Size](image)

Fig. 1 the five datasets sizes.

Methods

There are many steps to build the model using PySpark ML as following:

- Read feature file obtained from the selected features step into a Spark RDD dataset.
- Parallelize the ML procedure using the Spark MapReduced framework.
- Import the PySpark ML library.
- Use the SVMWithSGD method, which is an optimization-based linear SVM.
- Use the LogisticRegressionWithSGD method, which is an optimization-based Logistic Regression
SVM with SGD use the stochastic gradient descent (SGD) method to train the classifier. The algorithm is used with feature selection ICA, PCA and LLE. SVM with SGD only work for two class problems.

Two typical cross-validation methods (namely k-fold cross-validation and leave-one-out validation) have been widely used in microarray data classification evaluation. Comparing to the k-fold cross-validation method, the LOOCV method is more applicable due to the small sample size of microarray data.

We have used 4-fold cross-validation to verify the generated SVM, Logistic Regression models. The steps are:

1. Generate the cross-validation datasets (75% training and 25% testing) using the training data.
2. Run SVM with SGD, LogisticRegressionWithSGD models on the cross-validation datasets (use the training part to train the SVM with SGD and LogisticRegressionWithSGD models.

We use the testing part to assess the models accuracies. Then, we use the generated models to make predictions on new data by performing the following steps:

1. Make predictions on the 20% prediction data.
2. Transform the prediction data into features, according to the feature extraction rules.
3. Get feature file that has the first column is sample ID.
4. Make predictions on the feature file using the generated SVM with SGD or LogisticRegressionWithSGD model.
5. Implement the Spark MapReduce framework to parallelize predictions.

It is divided into training and test set, further independent component analysis is performed to reduce the dimensionality of train data. For ICA, the FastICA algorithm software package for python is applied.

Analysis of large Microarray datasets has become a challenging task, as the number of genes available in technical research sets and the number of test samples for an experimental set increases. The computational and analytical challenges including feature selection and classification of the dataset into correct group or class. In this paper, PCA, ICA and LLE based on Spark framework are used to select the pertinent features. After feature selection, the classifier SVM with SGD and Logistic Regression with SGD based on Spark framework are applied to classify the Microarray dataset. A detail comparative analysis in terms of execution time and accuracy is done on these feature selection and classifier methodologies that are based on Spark framework.

Result

The output file is labeled points’ format, with features being dense vectors. Each row represents a microarray sample. The first column is label (sample class: 0 or 1). The other columns are features (first 10 principal components). We have used SVM with SGD and LinearRegressionWithSGD.
methods to train the classifiers. Fig.2 and Fig.3 show training errors of SVM with SGD and LinearRegression with SGD respectively.

**Fig. 2 SVM with SGD training errors when we build it with ICA, PCA and LLE (ICASVM, PCASVM and LLESVM).**

**Fig. 3 LinearRegression with SGD training errors when we build the model with ICA, PCA and LLE (ICALG, PCALG and LLELG).**

Fig.2 and Fig.3 show that the training errors as following:

- **Chowdary dataset:** training error in ICALG and PCALG more than in ICASVM and PCASVM.
- **Singh and Gordon datasets:** training error in PCALG more than in PCASVM.
• Chin dataset: training error in PCASVM more than in PCALG.

• Borovecki dataset: training errors in SVMwithSGD and LinearRegressionwithSGD equals to 0 with all feature selection methods.

• All datasets: training error in LLESVM and LLELG are same.

Also, using the testing part to assess the model accuracy. The results show the averages of accuracies as shown in the Fig.4, Fig.5 and Fig.6.

Fig.4 Comparison between average of Accuracy of LGWithSGD and SVMWithSGD Classifiers with ICA Method as Feature Selection.

Fig.4 shows that we obtained higher accuracies from SVMWithSGD Classifier with ICA than LGWithSGD with the same feature selection in all five datasets. Therefore, ICA can be a good choice with SVMWithSGD.

Fig.5 Comparison between the average of Accuracy of LGWithSGD and SVMWithSGD Classifiers with PCA Method as Feature Selection.
Fig. 5 shows that we obtained higher accuracies from SVMWithSGD Classifier with PCA than LGWithSGD with the same feature selection in Borovecki, Singh and Chowdary datasets. Therefore, PCA can be a good choice with SVMWithSGD in these datasets because of the averages of accuracies are higher than 70%. While in Chin and Gordon datasets, we obtained higher accuracies from LGWithSGD Classifier with PCA than SVMWithSGD with the same feature selection.

Fig. 6 shows that we obtained higher accuracies from SVMWithSGD Classifier with LLE than LGWithSGD with the same feature selection in Borovecki dataset. While in Chin, Singh, Chowdary, Gordon datasets, we obtained higher average of accuracies from LGWithSGD Classifier with LLE than SVMWithSGD with the same feature selection. Therefore, LGWithSGD Classifier with LLE is a good choice because of the averages of accuracies are higher than 70% in the five datasets.

Fig. 7 shows the comparison between Times Consumed in LGWithSGD and SVMWithSGD Classifiers with ICA Method.
Fig. 8 Comparison between Time Consumed in LGWithSGD and SVMWithSGD Classifiers with PCA Method.

Fig. 8 Comparison between Times Consumed in LGWithSGD and SVMWithSGD Classifiers with LLE Method.

Fig. 6, Fig. 7 and Fig. 8 are shown that SVMWithSGD Classifier consumed much time than LGWithSGD in all feature selection methods for all five datasets. Moreover, Fig. 7 illustrates that SVMWithSGD classifier with ICA which are applied on Chin dataset, have consumed much more than SVMWithSGD classifier with (PCA, LLE) which are applied on the same dataset as well. The most average models accuracies are more than 70. Therefore, we use the generated models to make predictions on new data.

Conclusion

To conclude, when we have compared the SVMWithSGD and LGWithSGD by training errors metric. SVMWithSGD is a better choice than LGWithSGD with all datasets and with all feature selection methods.

By measuring accuracy, PCASVM is the best choice in analysis Chowdary and Borovecki datasets. PCALG is good option in analysis Gordon dataset while we have to avoid it in Chowdary and Singh because the average accuracies less than 70%. For the same reason, avoiding LLESVM in classification of Singh dataset. Therefore, ICASVM is the best choice in analysis Singh dataset. For Chin dataset, LLELG has the highest average accuracy.

With time measuring, LGWithSGD is a better choice than SVMWithSGD with all datasets and in all feature selection methods. In addition, we have to avoid to use ICASVM specially. SVMWithSGD is better than LGWithSGD in Borovecki, Singh, and Chowdary datasets when the measured matrices are training errors and accuracies. While LGWithSGD is best than SVMWithSGD in Gordon and Chin datasets when the measured matrices are accuracy and time.

The most predictions are correct because the robust ML models have constructed, their predictions are mostly trustable. We treat positive predictions as highly risky instances.
References

1. Wahid A, Khan DM, Iqbal N, Khan SA, Ali A, Khan M, et al. Feature selection and classification for gene expression data using novel correlation based overlapping score method via Chou's 5-steps rule. Chemometrics and Intelligent Laboratory Systems. 2020;199:103958.

2. Aydadenta H, Adiwijaya. On the classification techniques in data mining for microarray data classification. Journal of Physics: Conference Series. 2018;971:012004.

3. Sardana M, Agrawal R, Kaur B. A hybrid of clustering and quantum genetic algorithm for relevant genes selection for cancer microarray data. International Journal of Knowledge-based and Intelligent Engineering Systems. 2016;20(3):161–73.

4. Vanitha CDA, Devaraj D, Venkatesulu M. Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection. Procedia Computer Science. 2015;47:13–21.

5. Shukla AK. Identification of cancerous gene groups from microarray data by employing adaptive genetic and support vector machine technique. Computational Intelligence. 2019;36(1):102–31.

6. Fajila MNF. Gene Subset Selection for Leukemia Classification Using Microarray Data. Current Bioinformatics. 2019;14(4):353–8.

7. Abdelnabi MLR, Jasim MW, El-Bakry HM, Taha MHN, Khalifa NEM. Breast and Colon Cancer Classification from Gene Expression Profiles Using Data Mining Techniques. Symmetry. 2020;12(3):408.

8. Zhong W, Lu X, Wu J. Feature Selection for Cancer Classification Using Microarray Gene Expression Data. Biostatistics and Biometrics Open Access Journal. 2017;1(2).

9. Mazumder DH, Veilumuthu R. An enhanced feature selection filter for classification of microarray cancer data. ETRI Journal. 2019;41(3):358–70.

10. Passi K, Nour A, Jain CK. Markov blanket: Efficient strategy for feature subset selection method for high dimensional microarray cancer datasets. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2017;

11. Shukla AK, Singh P, Vardhan M. A New Hybrid Feature Subset Selection Framework Based on Binary Genetic Algorithm and Information Theory. International Journal of Computational Intelligence and Applications. 2019;18(03):1950020.

12. Alshamlan H, Badr G, Alohaly I. mRMR-ABC: A Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling. BioMed Research International. 2015;2015:1–15.

13. Xu J, Mu H, Wang Y, Huang F. Feature Genes Selection Using Supervised Locally Linear Embedding and Correlation Coefficient for Microarray Classification. Computational and Mathematical Methods in Medicine. 2018;2018:1–11.

14. Adiwijaya, Wisesty UN, Lisnawati E, Aditsania A, Kusumo DS. Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification. Journal of Computer Science. 2018;14(11):1521–30.

15. Li Z, Xie W, Liu T. Efficient feature selection and classification for microarray data. Plos One. 2018;13(8).

16. Jenifer XR, Lawrance R. Classification of microarray data using SVM mapreduce. 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). 2017;

17. Sahlabadi A, Muniyandi RC, Sahlabadi M, Golshanbaqhy H. Framework for Parallel Preprocessing of Microarray Data Using Hadoop. Advances in Bioinformatics. 2018;2018:1–9.
18. Kumar M, Rath NK, Rath SK. Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier. Journal of Biomedical Informatics. 2016;60:395–409.
19. Kumar M, Ray RB, Rath SK. Spark based classification of microarray data using scalable artificial neural network. International Journal of Data Mining and Bioinformatics. 2017;19(4):312.
20. Lokeswari YV, Jacob SG. Prediction of Child Tumors from Microarray Gene Expression Data Through Parallel Gene Selection and Classification on Spark. Advances in Intelligent Systems and Computing Computational Intelligence in Data Mining. 2017;:651–61.
21. Alghunaim S, Al-Baity HH. On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context. IEEE Access. 2019;7:91535–46.
22. Jędrzejowicz J, Jędrzejowicz P, Wierzbowska I. Implementing Gene Expression Programming in the Parallel Environment for Big Datasets’ Classification. Vietnam Journal of Computer Science. 2019;06(02):163–75.
23. Venkataramana L, Jacob SG, Ramadoss R. Parallelized Classification of Cancer Sub-types from Gene Expression Profiles Using Recursive Gene Selection. Studies in Informatics and Control. 2019;27(2):213–22.
24. Bartenhagen C, Klein H, Ruckert C, Jiang X, Dugas M. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. BMC Bioinformatics. 2010;11(1).