Int. J. Advance Soft Compu. Appl, Vol. 13, No. 1, March 2021
Print ISSN: 2710-1274, Online ISSN: 2074-8523

SVM and Naïve Bayes Stacking Approach for Improving Gene Expression Data Classification Using Logistic Regression.

Abdallah Bashir Musa, Mohanad Mohammed, Fuad Abedalrazeq Mussallum, and Murtada Khalafallah Elbashir

Department of Basic Science, Deanship of Preparatory Year and Supporting Studies, Imam Abdulrahman Bin Faisal University,
P.O. Box1982, Dammam 34212, SAUDI ARABIA
abhamad@iau.edu.sa
School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal,
Pietermaritzburg, Private Bag X01, Scottsville 3209, South Africa.
mohanadadam32@gmail.com
Department of Basic Science, Deanship of Preparatory Year and Supporting Studies, Imam Abdulrahman Bin Faisal University,
P.O. Box1982, Dammam, 34212, SAUDI ARABIA
famussallum@iau.edu.sa
College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia
mkelfaki@ju.edu.sa

Abstract

Logistic regression is the foremost statistical classification technique which has many uses in numerous disciplines including machine learning, bioinformatics, and medical research. However, logistic regression classification accuracy is hindered by large data sets. When the number of features exceeds the number of instances, e.g. in the classification of gene expression data, improving logistic regression accuracy has been an important challenge that draws the researchers' attention. Ensemble learning techniques are designed to create a meta-classifier by combining several classifiers that are built on the same data to enhance the machine learning algorithm performance. In this paper, stacking approach is used to improve the accuracy of logistic regression for the classification of gene expression data. The stacking approach is a method in which one meta-classifier learns the output of the combined base classifiers.

For this purpose, support vector machines with linear and radial basis function, and naïve Bayes are used as base classifiers while logistic regression is used as a meta-classifier.

The dimension reduction technique is used for raising the degree of classification accuracy of logistic regression.

Principle component analysis (PCA) is used for reducing the dimension of the data before applying the stacking approach method. Several machine

learning metrics are used for assessing the method: accuracy, sensitivity, specificity, the area under the curve (AUC), kappa and ROC analysis.
The study has demonstrated that applying stacking approach with logistic regression results in improving its accuracy and make it applicable to classify the gene expression data.

Keywords: - *Logistic regression (LR), support vector machines (SVM), naïve Bayes, stacking approach, Principle component analysis (PCA), gene expression data.*

1 Introduction

The Logistic Regression (LR) [1-3] is the most famous statistical technique for performing classification tasks, that has been extensively utilized in many disciplines, involving machine learning [4], and medical studies [5-7]. The benefit of using the logistic regression is that it generates a predicted probability vector for the class label, in addition the LR model can easy be interpreted and well understood. LR is mainly used for binary classification. However, it can be used for multi-class classification and it is known as the multinomial logistic regression.
The rapid increase of the technology of microarrays has generated an enormous gene expression data set. Typically, the gene expression data sets include huge columns of gens with small rows of instances, besides, it also includes a high level of noises.
From a statistical point of view, applying of logistic regression requires the number of the instances of the data set to be relatively larger than the number of the explanatory variables; this restricts the use of logistic regression when the number of features is less the number of instances. Consequently, building a logistic regression model on the gene expression data remains a serious issue and is thought as a great challenge attracting the interest of many researchers. Even though gene expression data has a lot of features, the mainstream of variability in the data can be explained by only few of these features, so that these features are often extracted to create a reliable classifier. Therefore, for the analysis of gene expression data, first we would like to pick or extract only these few important genes from the main data rather than classifying the entire data set of the expression data. Feature selection techniques are aimed to pick the most important features with a reduced dimension from the main entered data, and then the classification method would be applied on these preselected features [8]. Previous studies have proved that, the performance of the classification methods relies on the method that is used for gene selection; accordingly, the classification method's task should be associated with the gene selection method [9]. Gene selection techniques play an important role in performing the classification task. Commonly, the major objective of the feature selection methods is to reduce the complications related to computation and time by generating a few of significant features that include the utmost information within the entire data, which will be used as an input to a classifier for obtaining a high accuracy. The great benefit of features selection methods that it selects few uncorrelated / independent features (genes), this results in solving the issues of over fitting and co-linearity, thus the logistic regression would be applicable to classify the gene expression data. Moreover, applying logistic regression on those relevant selected features will result in a significant improvement of its accuracy. PCA is a well-known features selection method. It calculates the eigenvectors of the covariance matrix of the original input's features. PCA explains the variability of a set

of features in terms of a reduced set of uncorrelated linear space of such features with maximum variance, known as principal components (PCs) [10]. PCA is used to perform the task of dimensionality reduction for the gene expression training data set. After applying the reduction dimension technique, the stacking approach will be applied with the new reduced selected data features. The ensembles have been found to be a good technique that leads to improving the accuracy of the classifier [11, 12]. There are numerous approaches of ensembles that are proposed in machine learning literature. Most of the classifier ensembles studies are focused on generating ensembles using a single learning algorithm [13], such as support vector machines, logistic regression, or neural network.

Bagging learning ensembles, or bootstrap aggregating, proposed by Breiman (1996) and Boosting methods introduced by Freund & Schapire (1996) are the most popular single learning algorithm ensembles methods that have been used in machine learning and statistical researches [14, 15]. The main idea of these methods is generating different classifiers by manipulating the training set. Then the generated classifiers are typically combined by voting or weighted voting.

Recently a new approach is proposed by applying different learning algorithms to a single dataset. Then the predictions of the different classifiers are combined and used by a meta-level-classifier to generate a final prediction. This technique is called “stacking” [16].

Stacking is concerned with combining multiple classifiers generated by using different learning algorithms on a single dataset. In the first phase, a set of base-level classifiers is generated. In the second phase, a meta-level classifier that combines the outputs of the base-level classifiers is learned. To generate a training set for learning the meta-level classifier, a leave-one-out or a cross validation procedure is applied. Applying stacking results has shown to improve the classification performance compared to voting [17]. After applying the reduction dimension technique, the stacking approach will be applied with the new reduced features selection data Combining features selection and stacking approach techniques is potentially expected to improve the classification accuracy of logistic regression. A comprehensive comparison has been done; several machine learning metrics have been used in this study.

2 The classification methods

2.1. Logistic regression

Logistic Regression (LR) [1-3] is a famous statistical classification technique for modeling binary data. Let $x \in R^n$ denote a vector of independent or feature variables and let $y \in \{-1, +1\}$ denote the corresponding binary class label. The logistic model could be defined as:

$$\text{pr } (y/x) = \frac{1}{1+\exp\left(\text{-y}(\beta^T x+\alpha)\right)} = \frac{\exp\left(y(\beta^T x+\alpha)\right)}{1+\exp\left(y(\beta^T x+\alpha)\right)} \quad (1)$$

as Pr(y/x) represent the conditional probability of y associated with the features $x \in R^n$. The logistic model has parameters $\alpha \in R$ and $\beta \in R^n$ which are denoted the intercept term and the weight vector term respectively.$\beta^T x + \alpha = 0$describes a hyperplane in the feature space, on which $pr(y/x) = 0.5$, if the conditional probability $pr(y/x)$larger than 0.5, $\beta^T x + \alpha$

would has the same sign as y, and if the conditional probability $p(y/x)$less than 0.5 , $\beta^T x + \alpha$ would has the other sign of y. Suppose we are given a set of m training data set$\{x_i, y_i\}_{i=1}^{m}$, where $x_i \in R^n$. denote the i-th sample and $y_i \in \{-1, +1\}$denote the associated class label. These training samples data are supposed to be independent. According to the logistic model, the vector of the conditional probabilities corresponding of these samples could be explained as: -

$$\text{pr}\ (\alpha, \beta)_i = \text{p}\ (y_i/x_i) = \frac{\exp y_i(\beta^T x_i + \alpha_i)}{1 + \exp y_i(\beta^T x_i + \alpha_i)} i = 1, \ldots\ldots, m \qquad (2)$$

The likelihood function associated with the samples is$\prod_{i=1}^{m} \text{pr}\ (\alpha, \beta)_i$, and the log likelihood function is:

$$\sum_{i=1}^{m} log\ \text{p r}\ (\alpha, \beta)_i = -\sum_{i=1}^{m} f(\beta^T a_i + \alpha y_i) \qquad (3)$$

Where $a_i = x_i y_i \in R^n$ and f is the logistic loss function which is:

$$f(z) = log(1 + exp(-z) \qquad (4)$$

By using (4), (3) we get the following equation.

$$\sum_{i=1}^{m} log\ \text{p r}\ (\alpha, \beta)_i = -\sum_{i=1}^{m} log(1 + exp - (\beta^T a_i + \alpha y_i))) \qquad (5)$$

The logistic loss is the negative of the log likelihood function. When dividing the logistic loss by number of the training samples we get the logistic loss average,

$$l_{avg}(\alpha, \beta)_i = \frac{1}{m} \sum_{i=1}^{m} log(1 + exp - (\beta^T a_i + \alpha y_i))) \qquad (6)$$

The maximum likelihood estimation method used to determine the parameters β and α from the training data set, through solving the convex optimization problem.

$$\text{minimize} l_{avg}(\alpha, \beta)_i \qquad (7)$$

This optimization problem is known as logistic regression problem (LRP). LRP is a smooth convex optimization problem which could be solved using several techniques such as gradient descent, steepest descent, Newton, quasi-Newton, and conjugate-gradients (CG) methods. Newton method will be used in this paper. After obtaining the maximum likelihood values of α and β, which are the solutions of (7), finally the probability of the two possible outcomes will be predicted entering by entering a new features vector x $\in R^n$, to the associated logistic regression model; the logistic regression classifier is formed as:

$$\varphi(x) = sgn(\beta^T x + \alpha) \qquad (8)$$

Where

$$sgn(z) = \begin{cases} +1 z > 0 \\ -1 z \leq 0 \end{cases}$$

Which picks the more likely outcome, given x, according to the logistic model.

2.2 Support vector machine

Support vector machine (SVM) [27-29] may be a somewhat new-found classification technique. It is attracted tons of attention within the previous couple of years. The idea of SVM is as follows: input vectors x are mapped to an extremely high dimension feature space z through nonlinear mapping $\emptyset(x), z = \emptyset(x)$. In this space, an optimal separating hyperplane is built. For a given training dataset with n samples $(x_1, y_1), (x_2, y_2), (x_3, y_3) \ldots\ldots\ldots, (x_n, y_n)$,where x_i is a feature vector in a d-dimensional feature space Rd and $y_i \in \{1, -1\}$ is the corresponding class label. The task is to obtain a classifier with a decision function $\mathrm{f}\,(\mathrm{x}, \mathrm{w}, \mathrm{b}) = w^T \emptyset(x) + b$, SVM retains an optimal hyperplane with the maximal margin that separates the data points into two classes.

2.3 Naïve Bayes classifier

Naïve Bayes classifier [30] is familiar popular algorithm in statistic and machine learning that have been found to perform perfectly [31]. Naïve Bayes has been commonly applied in statistics and machine learning research areas. For a given X (nxp) training data set where n is the number of training data and p the number of features, these training data need to be classified according to $C\ where\ c \in \{1, -1\}$. Naïve Bayes classifier is dependent on the so-called Bayes Theorem.

3 Material and Methods

3.1. Datasets

Six datasets from different types of cancers are been used in this paper, downloaded from gene expression omnibus (GEO) public platform (https://www.ncbi.nlm.nih.gov/geo/), with accession numbers GSE98708, GSE133385, GSE140684, GSE96669, GSE71799], and GSE115313. The GSE98708 consist of 47312 probes taken from 102 samples with primary tumor and 83 PDX [18]. Sample type were used as class type for the classification process. The GSE133385 dataset has 28940 probes from 111 patients with moderate-to-severe atopic dermatitis (AD) which is the most commonly inflammatory skin disease [19]. The GSE140684 dataset has 54143 probes taken from 152 adult samples with moderate-to-severe atopic dermatitis [20]. The GSE96669 dataset has 48107 probes from 132 patients with junctional cancers of gastric and oesophageal origin [21]. The GSE71799 has 54675 probes taken from 134 samples for identifying of molecular signatures of cystic fibrosis disease status with plasma-based functional genomics [22]. Finally, the GSE115313 dataset consists of 53617 probes from 84 patients with colon cancer [23]. Table 1 summaries the gene expression data information used in this study.

Table 1: shows the explanatory analysis of the gene expression data.

Accession no	*Platform*	*No of samples*	*No of genes*	*After filtration*	*After t-test*	*PCs*
GSE115313	*GPL16686*	*84 (42/42)*	*53617*	*15686*	*5050*	*52*
GSE71799	*GPL570*	*134 (31/103)*	*54675*	*18864*	*2968*	*82*
GSE96669	*GPL10558*	*132 (67/65)*	*48107*	*18706*	*4716*	*71*
GSE140684	*GPL570*	*152 (76/76)*	*54143*	*2459*	*266*	*55*
GSE133385	*GPL570*	*111 (81/30)*	*28940*	*6589*	*156*	*40*
GSE98708	*GPL10558*	*102 (83/19)*	*47312*	*10945*	*2156*	*31*

3.2. Preprocessing steps

Features reduction using principal component analysis.

Principal component analysis [24,25] is that the most traditional standard linear technique for dimensionality reduction. Even though PCA could be a traditional linear technique, the new nonlinear techniques do not vanquish the traditional PCA on real life tasks [26]. It uses an orthogonal transformation to convert a collection of observation vectors of probably correlated features into a set of linearly uncorrelated features known as PCs.

This transformation is defined in such that the first principal component has the biggest possible variance. Since applying of logistic regression needed features to be uncorrelated, PCA would be the most effective possibility because it yields reduced uncorrelated new features.

4 Stacking ensemble approach

The main aim of the ensemble learning is to improve the accuracy of the base models to perform specific classification task [32]. In general, every classifier vote, these votes are combined to predict the final class label. Consequently, the classifier obtained through the ensemble learning techniques is outperform the single classifier. There are various types of ensemble learning approaches, such as simple average, weighted average, majority voting, weighted voting, boosting, and bagging ensemble stacking. Some of these methods work out based on combining models typically from the same type while others combining models typically from different types of classifiers, Recent work suggests models which use stacked ensemble classifiers perform especially well on training and independent testing data [33-3٦]. The stacking ensemble is used in this paper.

Stacking Ensemble is an ensemble method that workout by combining different types of classifiers. The final optimal classification performance is obtained by combining the models that been built by each classifier. The stacking ensemble is consisted of two phases, the first one consisting of the base classifiers while the second phase is consisting of a meta-classifier whichreceives the prediction of the base classifiers as an input to perform the final classification. The stacking approach is applied for the reduced data after using PCA. The framework of the study is shown below in Fig1.

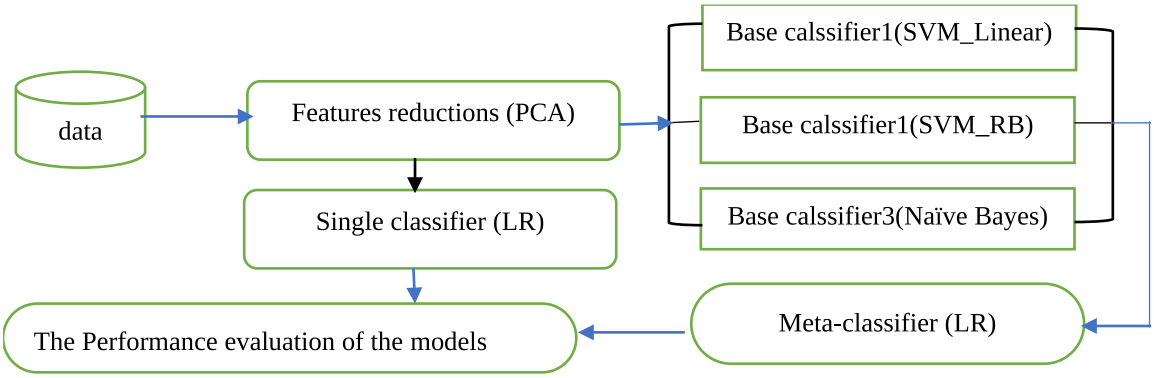
data
Features reductions (PCA)
Single classifier (LR)
Base calssifier1(SVM_Linear)
Base calssifier1(SVM_RB)
Base calssifier3(Naïve Bayes)
Meta-classifier (LR)
The Performance evaluation of the models

Fig 1. The framework of the study.

5 Methods evaluation

In this study, we considered different types of performance measures to evaluate our methods. These measures are accuracy, kappa statistic, sensitivity, specificity, balanced error rate (BER), and area under the ROC curve (AUC). Ten-folds cross-validation with three repetitions used to calculate these metrics. In ten-folds cross-validation the datasets were divided into ten parts, then each nine folds combined to be used for training the models, and the remaining fold used as a test and validate the models. This process iterated ten times and the average of the ten iterations is calculated. Furthermore, for more evaluation, we plotted the receiver operating (ROC) curve which will give a different point of view that might help in assessing the models.

6 Results and Discussions

6.1. Results

The results were obtained using caret package in R statistical software version (3.6.3) for applying the classifiers to classify the gene expression data. We have used logistics regression as a top layer in the stack, the base layer classifiers were support vector machines with linear and radial basis function, and naïve Bayes. We compared the performance of the single and stacked logistics regression. Table 2 below shows the performance in terms of accuracy, kappa, sensitivity, specificity, balanced error rate, and AUC metrics. Overall, we noticed that the logistics regression classification performance has been enhanced using the stacking ensemble approach as shown in Table 2.

Table 2. A comparison of single and stacking based logistic regression.

Datasets	Metrics					
	Accuracy	Kappa	Sensitivity	Specificity	BER	AUC
			Single logistic regression			
GSE98708	*0.97 (0.94, 0.99)*	*0.90 (0.84, 0.97)*	*0.97 (0.94, 0.99)*	*0.96 (0.87, 1.00)*	*0.06*	*0.97*
GSE133385	*0.72 (0.67, 0.77)*	*0.33 (0.22, 0.44)*	*0.77 (0.71, 0.82)*	*0.58 (0.47, 0.68)*	*0.34*	*0.67*
GSE140684	*0.95 (0.93, 0.97)*	*0.90 (0.86, 0.94)*	*0.96 (0.93, 0.98)*	*0.94 (0.90, 0.97)*	*0.05*	*0.95*
GSE96669	*0.89 (0.86, 0.92)*	*0.79 (0.73, 0.85)*	*0.88 (0.82, 0.92)*	*0.91 (0.86, 0.95)*	*0.11*	*0.89*
GSE17799	*0.96 (0.94, 0.98)*	*0.90 (0.85, 0.95)*	*0.98 (0.92, 1.00)*	*0.96 (0.93, 0.98)*	*0.07*	*0.97*
GSE115313	*0.62 (0.55, 0.68)*	*0.23 (0.12, 0.34)*	*0.85 (0.77, 0.91)*	*0.38 (0.30, 0.47)*	*0.35*	*0.62*

Stacking using logistic regression						
GSE98708	*0.98 (0.97, 0.99)*	*0.94 (0.91, 0.97)*	*0.99 (0.98, 1.00)*	*0.94 (0.90, 0.97)*	*0.03*	*0.97*
GSE133385	*0.84 (0.82, 0.86)*	*0.58 (0.53, 0.64)*	*0.91 (0.89, 0.93)*	*0.65 (0.59, 0.71)*	*0.20*	*0.78*
GSE140684	*0.96 (0.95, 0.97)*	*0.93 (0.91, 0.95)*	*0.96 (0.94, 0.97)*	*0.97 (0.95, 0.98)*	*0.04*	*0.96*
GSE96669	*0.89 (0.87, 0.91)*	*0.79 (0.75, 0.82)*	*0.89 (0.87, 0.92)*	*0.89 (0.86, 0.92)*	*0.11*	*0.89*
GSE17799	*0.99 (0.98, 0.99)*	*0.96 (0.95, 0.98)*	*0.96 (0.93, 0.98)*	*1.00 (0.99, 1.00)*	*0.01*	*0.98*
GSE115313	*0.82 (0.79, 0.85)*	*0.64 (0.59, 0.70)*	*0.76 (0.72, 0.80)*	*0.88 (0.84, 0.91)*	*0.17*	*0.82*

The results above were supported and validated by plotted the ROC curves for all the single and stacked logistics regression in each dataset (see Fig 2, 3, 4, 5, 6 and 7) below. The ROC provides more insight into the performance of the methods used.

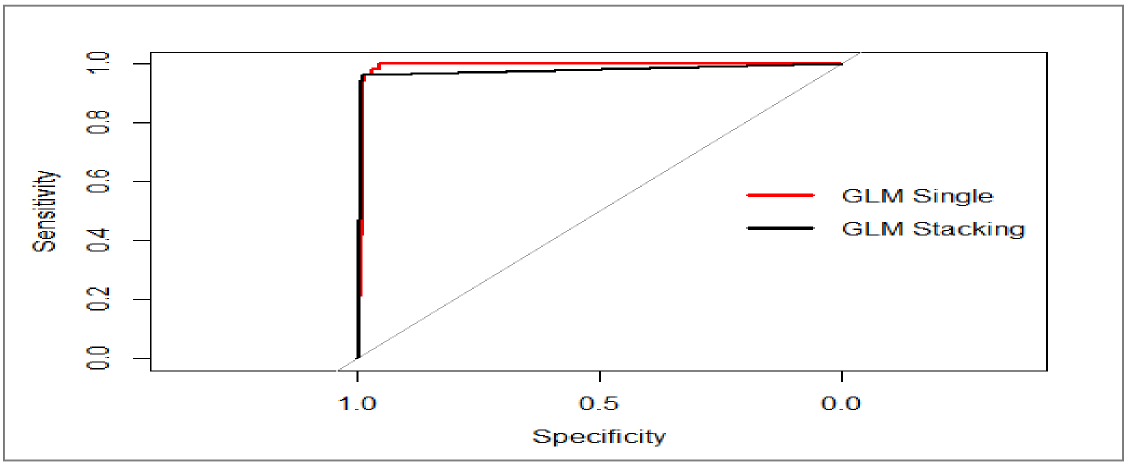

Fig 2: ROC curve for GSE98708 dataset.

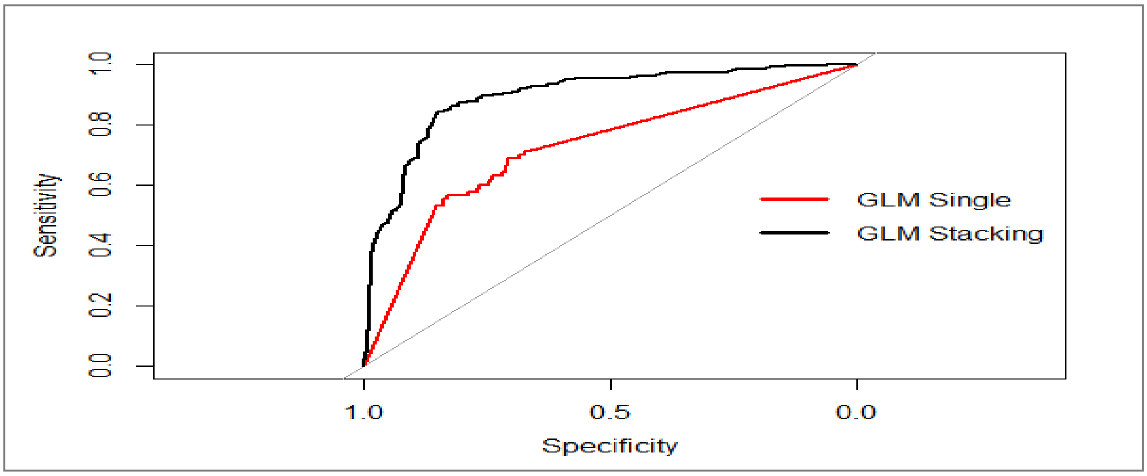

Fig 3: ROC curve for GSE133385 dataset.

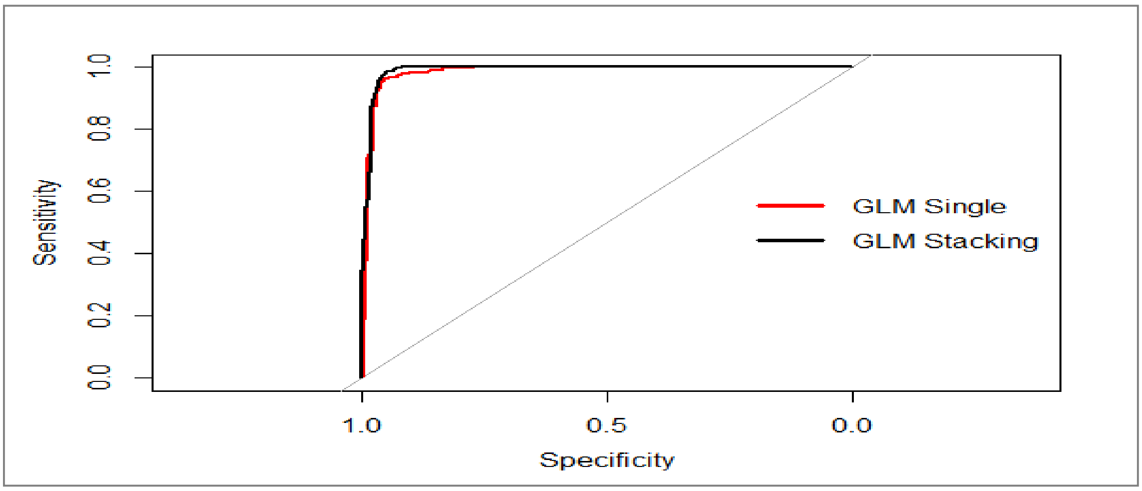

Fig 4: ROC curve for GSE140684 dataset.

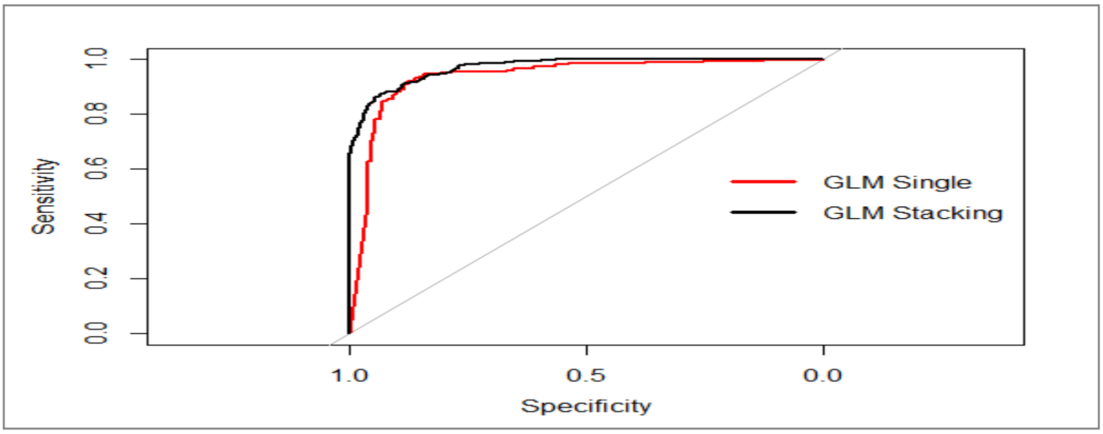

Fig 5: ROC curve for GSE96669 dataset.

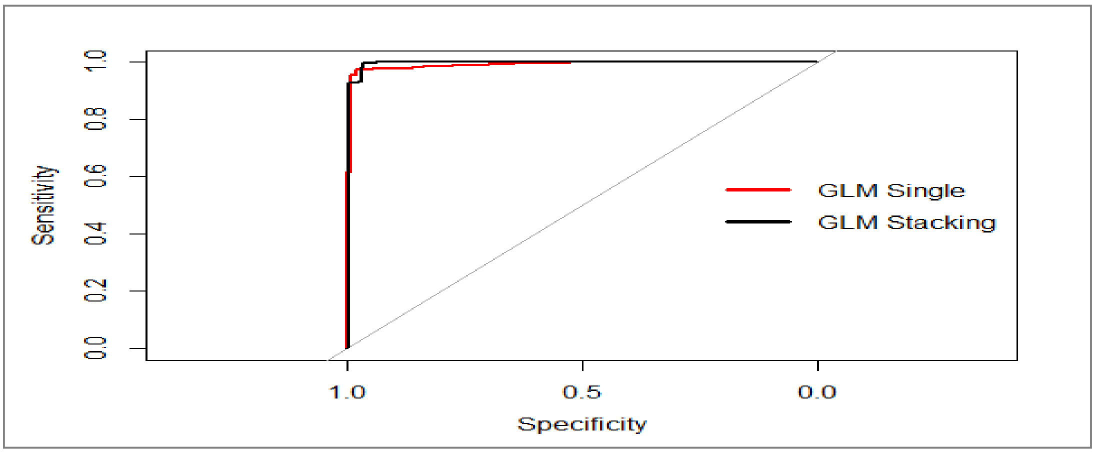

Fig 6: ROC curve for GSE17799 dataset.

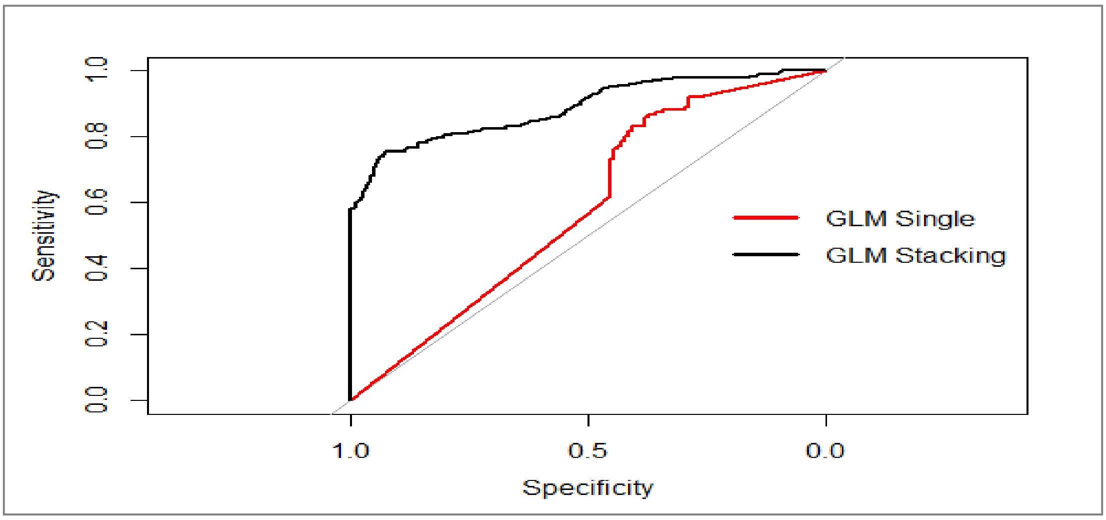

Fig 7: ROC curve for GSE115313 dataset

6.2. Discussion

Microarray techniques produce massive amount of gene expression data. The use of gene expression in a classification problem has been faced many challenges because this kind of data has a different structure from other commonly used data. Microarray techniques often produce data with small samples size and each sample has a large number of genes (variables). However, in the case of gene expression data where $p > n$ most of the statistical methods fail. Therefore, to alleviate this problem dimensionality-reduction step has been performed using principle component analysis method. This procedure ensures that we get only a small number of variables that contain most of the information, variation, and bring out strong pattern in the dataset.

In this study, we compared classification of microarray gene expression data using single logistic regression and stacked logistic regression after using the principle component analysis to reduce the data dimension. In addition, we presented an improvement of logistic regression method using the stacking concept based on support vector machines (with radial basis function and linear kernels) and Naïve Bayes as base layer classifiers, where the logistic regression is used as a top-layer classifier. The logistic regression receives the prediction from the base layer classifier, which enhances the logistic regression classification performance. Subsequently, we have tested this approach using 6 gene expression datasets. The performance has been measured using repeated ten-fold cross-validation with 3 repetitions, and the comparison is based on six metrics: accuracy, kappa, sensitivity, specificity, BER, and AUC.

From Table II above, we have observed that the logistic regression performance is enhanced by the stacking compared to single logistic regression for 5 data sets out of 6. In the GSE96669 dataset the logistic regression performance remains the same with 0.89 of accuracy. Overall, stacking ensemble approach improves the logistic regression accuracy with 0.062 in average. Furthermore, the results have been proven by the ROC curves, as shown in Fig 2, 3, 4, 5,6 and 7.

Overall, we have implemented an ensemble learning using stacking concept for gene expression data. Logistic regression method was used as top layer and support vector machines (radial and linear kernels) and Naïve Bayes were used as base layer in the stack to help in improving the logistic regression classification performance.

In this study, we observed a significant increase in the logistic regression performance. In addition, stacking consistently performed well in terms of the classification performance measures used.

7 Conclusion

This paper proposes the implementation of PCA and stacking ensemble approach with logistic regression for classification of gene expression. The principal components analysis has been used as a dimension reduction method of the gene data. Logistic regression method has been used as top layer while support vector machines (SVM) with radial and linear kernels as well as Naïve Bayse have been used as base layer in the stacking approach, six gene expression data sets have been used. The study concludes that applying stacking ensemble with logistic regression results in improving the accuracy of classification of gene expression data after applying the principle components analysis (PCA) for data reduction. Logistic regression classifies the gene expression data successfully when combining data reduction techniques.

References

1. Hosmer DW, Lemeshow S (2000) Applied logistic regression. Wiley series in probability and statistics, 2nd edn. Wiley, New York.
2. Menard S (2002) Applied logistic regression analysis, 2nd edn. Sage publications Inc, UK.
3. Ryan TP (2008) Modern regression methods, 2nd edn. Wiley, New York.
4. Rymarczyk, T., Kozłowski, E., Kłosowski, G. and Niderla, K., 2019. Logistic regression for machine learning in process tomography. Sensors, 19(15), p.3400.
5. Wu, H., Yang, S., Huang, Z., He, J. and Wang, X., 2018. Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, pp.100-107.
6. Patel, C.J., Bhattacharya, J. and Butte, A.J., 2010. An environment-wide association study (EWAS) on type 2 diabetes mellitus. PloS one, 5(5), p.e10746.
7. Tirzite, M., Bukovskis, M., Strazda, G., Jurka, N. and Taivans, I., 2018. Detection of lung cancer with electronic nose and logistic regression analysis. Journal of breath research, 13(1), p.016006.
8. Musa, A.B., 2014. A comparison of ℓ1-regularizion, PCA, KPCA and ICA for dimensionality reduction in logistic regression. International Journal of Machine Learning and Cybernetics, 5(6), pp.861-873.
9. Lee, J.W., Lee, J.B., Park, M. and Song, S.H., 2005. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis, 48(4), pp.869-885.
10. Musa, A.B., 2014. Gene expression data classification with kernel independent component analysis. Research Journal of Mathematical and Statistical Sciences ISSN, 2320, p.6047.
11. Opitz ,D., & Maclin, R. (1999). Popular ensemble methods: an empirical study, Journal of Artificial Intelligence Research, 11, 169-198.
12. Pal, M. (2007). Ensemble learning with decision tree for remote sensing classification. Proceedings of World Academy of Science, Engineering and Technology, 26, 735–737.
13. Dietterich T.G. (2000) Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg

14. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
15. Freund, Y. and Schapire, R.E., 1996, July. Experiments with a new boosting algorithm. In icml (Vol. 96, pp. 148-156).
16. Džeroski, S. and Ženko, B., 2004. Is combining classifiers with stacking better than selecting the best one?. Machine learning, 54(3), pp.255-273.
17. Zenko, B. and Dzeroski, S., 2002, August. Stacking with an extended set of meta-level attributes and mlr. In European Conference on Machine Learning (pp. 493-504). Springer, Berlin, Heidelberg.
18. Corso, S., Isella, C., Bellomo, S.E., Apicella, M., Durando, S., Migliore, C., Ughetto, S., D'Errico, L., Menegon, S., Moya-Rull, D. and Cargnelutti, M., 2019. A Comprehensive PDX Gastric Cancer Collection Captures Cancer Cell–Intrinsic Transcriptional MSI Traits. Cancer research, 79(22), pp.5884-5896.
19. Pavel, A.B., Song, T., Kim, H.J., Del Duca, E., Krueger, J.G., Dubin, C., Peng, X., Xu, H., Zhang, N., Estrada, Y.D. and Denis, L., 2019. Oral Janus kinase/SYK inhibition (ASN002) suppresses inflammation and improves epidermal barrier markers in patients with atopic dermatitis. Journal of Allergy and Clinical Immunology, 144(4), pp.1011-1024.
20. Khattri, S., Brunner, P.M., Garcet, S., Finney, R., Cohen, S.R., Oliva, M., Dutt, R., Fuentes-Duculan, J., Zheng, X., Li, X. and Bonifacio, K.M., 2017. Efficacy and safety of ustekinumab treatment in adults with moderate-to-severe atopic dermatitis. Experimental dermatology, 26(1), pp.28-35.
21. Bornschein, J., Wernisch, L., Secrier, M., Miremadi, A., Perner, J., MacRae, S., O'Donovan, M., Newton, R., Menon, S., Bower, L. and Eldridge, M.D., 2019. Transcriptomic profiling reveals three molecular phenotypes of adenocarcinoma at the gastroesophageal junction. International journal of cancer, 145(12), pp.3389-3401.
22. Levy, H., Jia, S., Pan, A., Zhang, X., Kaldunski, M., Nugent, M.L., Reske, M., Feliciano, R.A., Quintero, D., Renda, M.M. and Woods, K.J., 2019. Identification of molecular signatures of cystic fibrosis disease status with plasma-based functional genomics. Physiological genomics, 51(1), pp.27-41.
23. Del Puerto-Nevado, L., Minguez, P., Corton, M., Solanes-Casado, S., Prieto, I., Mas, S., Sanz, A.B., Gonzalez-Alonso, P., Villaverde, C., Portal-Nuñez, S. and Aguilera, O., 2019. Molecular evidence of field cancerization initiated by diabetes in colon cancer patients. Molecular oncology, 13(4), pp.857-872.
24. Jolliffe IT (2002) Principle components analysis, 2nd edn. Springer, Verlag
25. Escabias M, Aguilera AM, Valderrama MJ (2004) principal component estimation of functional logistic regression: discussion of two different approaches. J Nonparametric Stat 16(3–4): 365–384
26. van der Maaten LJP, Postma EO, van den Herik HJ (2008) Dimensionality reduction:a comparative review. Neurocomputing
27. Wang L (ed) (2005) Support vector machines theory and applications. Springer, Berlin
28. Kecman V (2001) Learning and soft computing: support vector machines, neural networks, and fuzzy logic models. MIT, Cambridge
29. Musa, A.B., 2013. Comparative study on classification performance between support vector machine and logistic regression. International Journal of Machine Learning and Cybernetics, 4(1), pp.13-24.
30. Berrar, D., 2018. Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 403.
31. Friedman, N., Geiger, D. and Goldszmidt, M., 1997. Bayesian network classifiers. Machine learning, 29(2), pp.131-163.

32. Friedman, J., Hastie, T. and Tibshirani, R., 2001. The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
33. Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.
34. Mishra, A., Pokhrel, P. and Hoque, M.T., 2019. StackDPPred: a stacking based prediction of DNA-binding protein from sequence. Bioinformatics, 35(3), pp.433-441.
35. Chen, C., Zhang, Q., Yu, B., Yu, Z., Lawrence, P.J., Ma, Q. and Zhang, Y., 2020. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. Computers in Biology and Medicine, 123, p.103899.
36. Xiong, Y., Wang, Q., Yang, J., Zhu, X. and Wei, D.Q., 2018. PredT4SE-stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. Frontiers in Microbiology, 9, p.2571.

Notes on contributors.

Abdallah B. Musa: Received the B.Sc. degree in Computer/Statistics from University of Gezira, Sudan, in 2000, The M.Sc. degree with distinction in Applied Statistics from University of Gezira in 2006. Doctor of Engineering in Management sciences and Engineering from Hebei University, Baoding, China in 2014. Area of research applied Statistics and Machine Learning.

Mohanad Mohamed: Received the B.Sc. degree in Computer/statistics from University of Gezira, Sudan, The M.Sc. degree from South Africa, Currently Ph.D. student in School of Mathematics, Statistics and Computer sciences, university of KwaZulu_Natal, Area of research gene expression data classification and Machine learning.

Fuad A. Mussallum: Received the bachelor’s degree in mathematics from Yarmouk university, Jordon. Master and doctor degree in Measurements and Evaluations from Yarmouk university, Jordon. Assistant professor in Department of Basic sciences, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia. Area of teaching and research are statistics and mathematics.

Murtada K. Elbashir: Received the B.Sc. degree with distinction in Computer/statistics from University of Gezira, Sudan, in 2000, The M.Sc. degree with distinction in computer information systems from the University of the Free State, South Africa in 2003 and the Ph.D. degree in Computer Science and technology from Central South University, Changsha, China in 2013. Research interests include bioinformatics and machine learning