Measurement 140 (2019) 120–125

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Feasible analysis of gene expression –a computational based classification for breast cancer

V. Nandagopal [a], S. Geeitha [b], K. Vinoth Kumar [c], J. Anbarasi [d,*]

[a] Department of Electrical and Electronics Engineering, Sri Krishna College of Engineering and Technology, Coimbatore, Tamil Nadu, India
[b] Department of Information Technology, Mahendra Engineering College for Women, Tiruchengode, India
[c] Department of Electrical and Electronics Engineering, Karunya Institute of Technology and Sciences, Coimbatore,Tamil Nadu,India
[d] RK Foundation Center for Research, Vellore, TN, India

ARTICLE INFO

Article history:
Received 28 November 2018
Received in revised form 14 January 2019
Accepted 5 March 2019
Available online 4 April 2019

Keywords:
Breast cancer
Gene expression
LASSO Logistic Regression (LLR)
Maximum Likelihood Estimation (MLE)
Expectation Maximization (EM)
Fuzzy

ABSTRACT

Computational based classification of gene expression data for analyzing the genetic pattern provides better clinical prediction for breast cancer. Breast cancer is one of the leading cancers among women all over India. This type of cancer leads to malignant tumor developed in the breast. Recent methodologies have undergone many classification methods to analyze the characteristics of gene expression. This paper focuses on computational method such as fuzzy based logistic regression to predict the expression of the gene data. In order to bring accuracy and to solve the inefficiency in feature selection of gene expression data, LASSO Logistic Regression (LLR), a novel methodology is implemented. For computational tractable, Maximum Likelihood Estimation (MLE) is implied with regression model. Diagnosis and prognosis of breast cancer becomes a great challenge in the medical era. This research work explores the mining technology based algorithm to classify the cancer data using fuzzy methodology by evaluating few samples of gene expression data of breast cancer as a training set and the resultant test data are validated to predict the cancer at the earlier stage. For feasible analysis, Expectation Maximization (EM) algorithm is deployed for unknown or missing parameters of gene data expression.

1. Introduction

Breast cancer [1] is considered to be one of the common female cancers among the human population that represents nearly one fourth of all cancers in India. Though the age estimated cancer rate is lower than other countries, the mortality and morbidity has been significantly increased as predicted in global studies. It is most commonly seen in rural India. Some cases of female breast carcinoma in situ were diagnosed based on the age specific incidence rates from 2003 to 2012 [2].

In the past decade, microarray technology was used that consists of few samples of gene data that includes different types of genes but was not precise enough to predict [3]. Later many models and methods came into era as machine language where specific genes were selected from the pool of different genes using some gene selection techniques [4–6]. Some computational models came into existence to solve the problem of feature selection of gene data. Regression analysis is one of the statistical processes mainly used to estimate the relevance among the variables in a cancer data [7,8]. The correlation coefficient in the independent variables which means the predictors is applied to the consequent result of dependent variables [9].

Generally Regression models do not work well with over fitting in the case of huge covariates. In that case, LASSO Logistic Regression (LLR) model is considered to be good. It is known for its penalized regression model that not only estimates the coefficients of regression model but also it maximizes the log-likelihood method by implementing some constraints. The main feature of LASSO model is that the estimation of the regression coefficients should be sparse which leads to 0 values exactly. And another important feature of LASSO is that it eliminates unwanted covariates [10].

To overcome the challenges faced by microarray technologies, supervised algorithms like Bayesian networks, support vector machine (SVM), Decision tree algorithm, artificial neural network, Regression models, etc, were combined in one or the other form for feature selection of gene pattern [11]. Today machine learning is being explored a lot in the breast cancer classification that provides high accuracy and good diagnostic ability. Among many classification techniques Naïve Bayes (NB) and KNearest neighbor (KNN) techniques play major role in the breast cancer prediction [12]. A novel technique namely multimodal Deep Neural Network

* Corresponding author.
E-mail address: eeeanbarasi@gmail.com (J. Anbarasi).

https://doi.org/10.1016/j.measurement.2019.03.015
0263-2241/© 2019 Elsevier Ltd. All rights reserved.

incorporating the multi dimensional data was implemented for the cancer prediction [13]. The above mentioned techniques proved efficiency in classification of gene expression and also in feature selection. Besides these methods, our proposed work is mainly based on fuzzy based computational methods where the percentage of classification accuracy is increased to some extent that leads to decision making diagnosis at the earliest. Maximum Likelihood Estimation (MLE) is mainly used to acquire more tremendous parameter estimations. In our work, it is mainly used to estimate parameter metrics from the sample gene data to analyze the probability of resulting the supervised data is greater. Expectation Maximization(EM) is used to find the maximum likelihood estimations for the parameters when any data seems to be incomplete or possess some missing data or some hidden variables. It generates the accurate hypothesis for the parameters which distributed in multiple models of gene data.

The proposed system also focuses to solve the efficiency problem by implementing LASSO logistic model.

2. Related works

In the machine learning techniques, classification plays a major role and essential entity in feature selection and prediction of gene expression data. Many researchers have undergone experiments by applying data mining techniques on different data sets of breast cancer to analyze the gene expression and to predict the disease at the earliest stage.

Okun [14] suggested filter feature selection method that reduces the cause of over fitting effect in the gene data classification. Different feature selection methods were used such as Backward Elimination Hilbert-Schmidt Independence Criterion (BAHSIC), Extreme Value Distribution based gene selection (EVD) and Singular Value Decomposition Entropy gene selection (SVDEntropy) [15–17].

Cheng and Lu[18] implemented C5 algorithm associated with bagging and generated more cancer data for training model from the cancer data set and the method attempted to predict the breast cancer survivability.

Pradesh [19] proposed three classifiers namely SVM, IBK and BF and compared and finally proved that the performance of sequential minimal optimization (SMO) algorithm showed a higher value of accuracy.

Jose Maria Celaya Palilla et.al. [20] described a robust machine learning model called LASSO used to identify the subset of features from the cancer data and these features played a vital role in distinguishing between benign and malignant tumor of breast cancer.

Sara Tarek et.al. [21] proposed an efficient ensemble classifier that gradually increases the performance of the classification and improves the confidence of the output. The main idea behind the usage of this classifier was that the output of the resultant variables were very less dependent on a single training set and also its classification performance is outstanding. In our research, thus we implemented LASSO model associated with computational model of MLE so that feature gene selection for the breast cancer

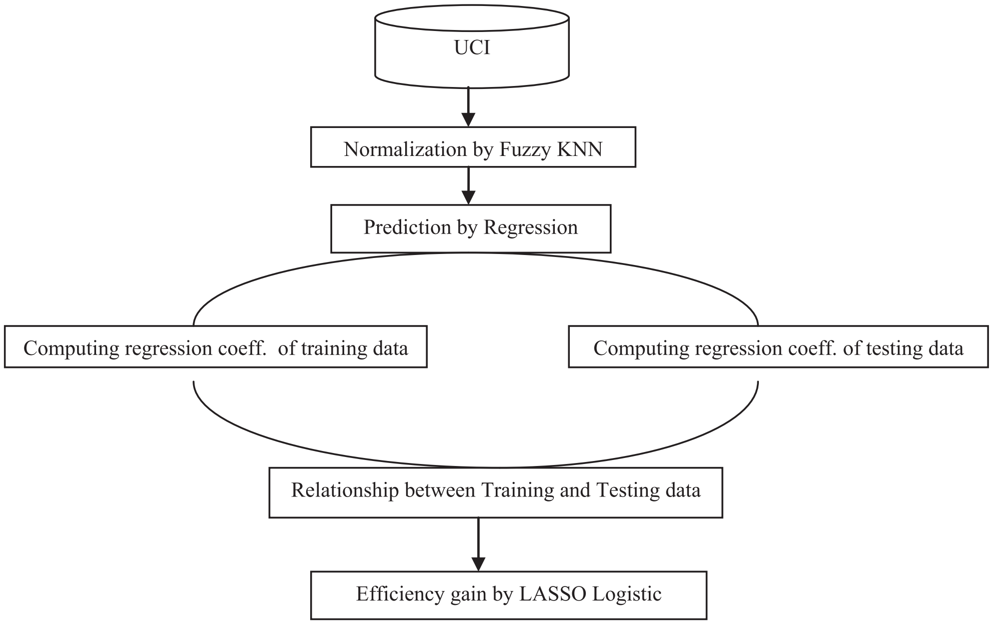

UCI
Normalization by Fuzzy KNN
Prediction by Regression
Computing regression coeff. of training data
Computing regression coeff. of testing data
Relationship between Training and Testing data
Efficiency gain by LASSO Logistic

Fig. 1. Schematic block diagram of Proposed Architecture.

Table 1
Sample Breast Cancer Dataset with two types of Classification.

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
48	23.5	70	2.707	0.467409	8.8071	9.7024	7.99585	417.114	#1
83	20.69049	92	3.115	0.706897	8.8438	5.429285	4.06405	468.786	#1
82	23.12467	91	4.498	1.009651	17.9393	22.43204	9.27715	554.697	#1
68	21.36752	77	3.226	0.612725	9.8827	7.16956	12.766	928.22	#1
86	21.11111	92	3.549	0.805386	6.6994	4.81924	10.57635	773.92	#1
45	21.30395	102	13.852	3.485163	7.6476	21.056625	23.03408	552.444	#2
45	20.83	74	4.56	0.832352	7.7529	8.237405	28.0323	382.955	#2
49	20.95661	94	12.305	2.853119	11.2406	8.412175	23.1177	573.63	#2
34	24.24242	92	21.699	4.924226	16.7353	21.823745	12.06534	481.949	#2
42	21.35991	93	2.999	0.687971	19.0826	8.462915	17.37615	321.919	#2

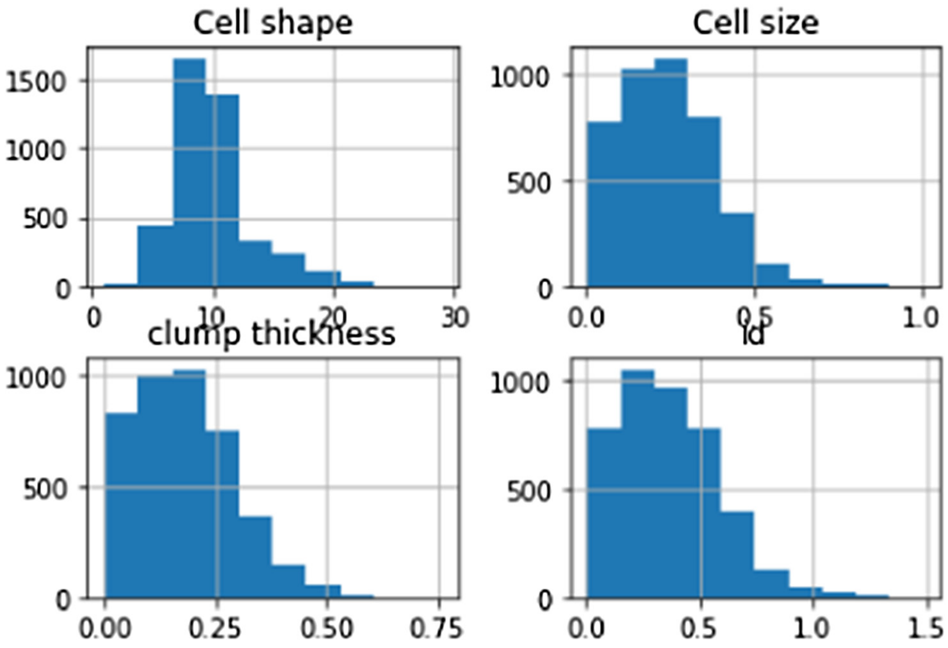
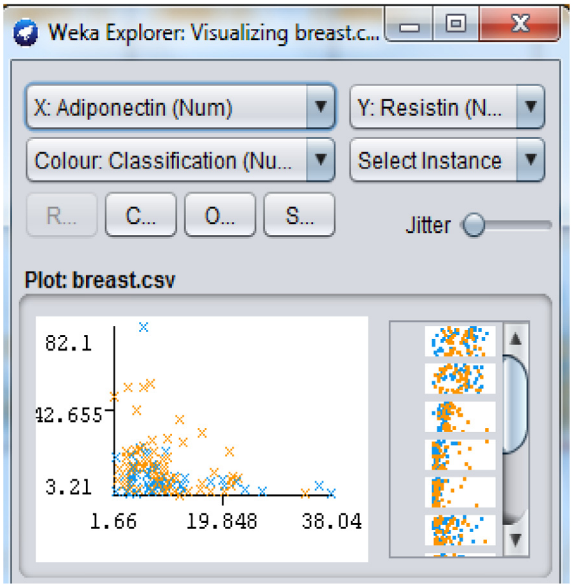


Fig. 2. Gene Expression Level of Cancer Data.


Weka Explorer: Visualizing breast.c...

X: Adiponectin (Num)

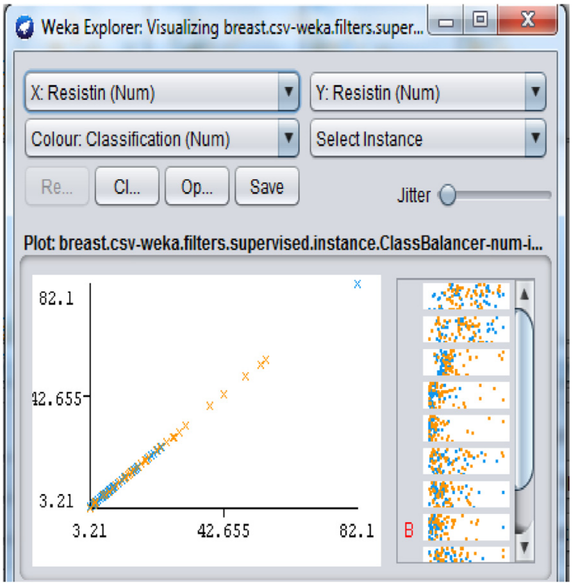
Y: Resistin (N...

Colour: Classification (Nu...

Select Instance

Plot: breast.csv

Fig. 3. Classification Analysis of Adiponectin and Resistin.


Weka Explorer: Visualizing breast.csv-weka.filters.super...

X: Resistin (Num)

Y: Resistin (Num)

Colour: Classification (Num)

Select Instance

Plot: breast.csv-weka.filters.supervised.instance.ClassBalancer-num-i...

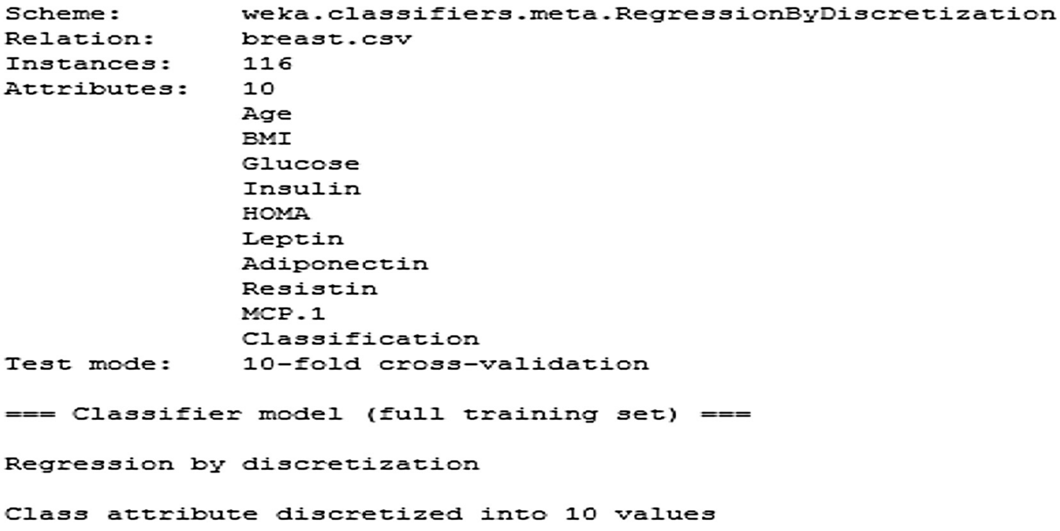
Fig. 4. Representation of Resistin Adiponectin and Resistin Level of Breast Cancer Patient.

is feasible and with the implementation of regression technique high classification accuracy with low error rate is possible.

Geeitha et.al. [22] implemented a novel model called Enhanced Bat Optimization (EBO)-SVM for gene selection that handles large dimensional dataset associated with the Hilbert-Schmidt Independence Criterion (HSIC) algorithm for selecting the relevant genes.

3. Proposed methodology

In the proposed paper, classification technique is mainly used. The classification algorithms play an extensively vital role in health care analysis, especially in analyzing gene expression and it can be very useful in predicting and discovering the genetic characteristics of breast cancer disease. In this methodology, the classification model is built on predefined data sets of labeled classes and attributes. Since classification model in the data mining technique are much capable of classifying a huge quantity of data and much specialized in predicting the categorical class labels to classify the cancer data based on the training set. Many supervised methods [23] such as C4.5, Multi layer perceptron (MLP), Naïve Bayes and J48 were implemented for predicting the cancer gene, classifying based on feature selection and to estimate the degree of certainty to some extent. In order to provide better accuracy in feature selection LASSO Logistic regression model is deployed and to solve the optimization problem, regularization technique is implemented. For obtaining better results in predicting the cancer gene, computational method namely MLE is implemented along with the regression model.



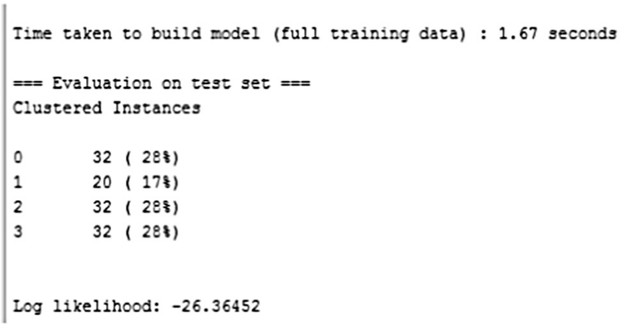
```
Scheme:       weka.classifiers.meta.RegressionByDiscretization
Relation:     breast.csv
Instances:    116
Attributes:   10
              Age
              BMI
              Glucose
              Insulin
              HOMA
              Leptin
              Adiponectin
              Resistin
              MCP.1
              Classification
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Regression by discretization

Class attribute discretized into 10 values
```

Fig. 5. Classification using Regression Model.



```
Time taken to build model (full training data) : 1.67 seconds

=== Evaluation on test set ===
Clustered Instances

0       32 ( 28%)
1       20 ( 17%)
2       32 ( 28%)
3       32 ( 28%)


Log likelihood: -26.36452
```

Fig. 6. Evaluating Test set using MLE.

The proposed methodology undergoes three phases (Fig. 1) such as Normalization using fuzzy based KNN. After preprocessing, feature selection and prediction by Linear Regression (LR) and third phase detects the victim gene of breast cancer by LASSO logistic regression algorithm. Then MLE computational method is derived for the resultant regression method. Expectation Maximization algorithm is implied for feasible solution. In this paper, the breast cancer dataset is gathered from https://archive.ics.uci.edu/ml/machine-learning-databases/00451/.

3.1. Preprocessing using fuzzy based KNN

The preprocessing of the gene data set is carried out using fuzzy KNN method with ease to proceed with the feature selection. After normalizing the cancer data, a few gene data are taken for experiment. Totally there are 116 instances with 10 attributes where 10 samples of cancer gene data (Table 1) with two classifications where #1 represents benign condition and #2 represents malignant tumor and #2 represents is divided into vectors V = {v1, v2, ..., vn}ϵ F_n into c (1 < c < n) as fuzzy subsets. The fuzzy members [24] can be shown as U, where U_{ij} can be considered as degree of fuzzy membership of sample vector vi in class i. The matrix U comprises two constraints, such as

$$\sum_{i=1}^{c} \mathrm{U_{ij}} = 1 \tag{1}$$

$$\mathrm{U_{ij}V}[0,1], 0 < \sum_{j=1}^{n} U_{ij} < n \tag{2}$$

where the Eq. (1) ensures the degree of gene samples are obtained across all the classes and its summation is equal to 1, whereas the second equation indicates that degree of fuzzy gene data member for all the classes lie between the intervals of either greater than zero or lesser than one or equally placed.

3.2. Cancer gene prediction using Linear regression

LR is mainly defined for discriminating gene pattern and also for predicting the cancer genes. Though many algorithms provide better prediction, it is well known for its calibration and for finding the dependency between benign and malignant cancer data. During data modeling, a graph is constructed with the training data set where nodes are represented as genes and edges are represented as interrelationship between a pair of gene. In the graphical construction, the set Ei,j exists only when there exists a relationship between node i and node j. Here adjacency matrix is framed $A = [G_{i,j}]_{nxn}$, where n is the number of nodes in the graph. The gene element $G_{i,j}$ can have any value either 0 or1 that depends upon the directed edge from one gene node i to another gene node j respectively.

3.3. Gene detection by LASSO logistic model with MLE

In this model, an assumption diagnosis of malignant breast cancer data was taken as dependent variable, X and it was symbolized as 0 for null benign cancer and 1 for appeared malignant cancer. Thus the probability of cancer can be obtained in the covariates of $y_{i.}$ This is calculated as:

$$\mathrm{Pb(X = 1|yi)} = \frac{exp(\beta_0 + \beta_{1yi1} + \cdots\cdots\cdots + \beta_{kYik}}{1 + exp(\beta_0 + \beta_{1yi1} + \cdots\cdots\cdots + \beta_{kYik}}$$

where yi = (yi1, yi2,. . .. . ..yik) named as covariates of i^{th} observation and β0 is considered as intercept and it falls between 1 and k in βj where j lies between 1 and k and it is the corresponding coefficient of i^{th} covariate.

4. Results and experiments

In this paper, performance evaluation of LASSO regression model and formerly proposed algorithms for feature selection of cancer data for two types of classification are described. The regression algorithm is generated using 10 attributes with 116

Table.2
Performance Comparison Analysis.

Methods	CORCOF	MAE	RMSE	RAE(%)	RRSE(%)	Time
MLP	2.733	5.413	7.175	108.58%	143.07	43 s
Linear Regression	4.295	3.997	4.548	80.16	90.68	85 s
Logistic Regression	3.768	3.358	5.211	67.35	103.91	0.07 s
LASSO logistic Regression	4.442	3.849	5.016	77.21	100.01	0.04 s

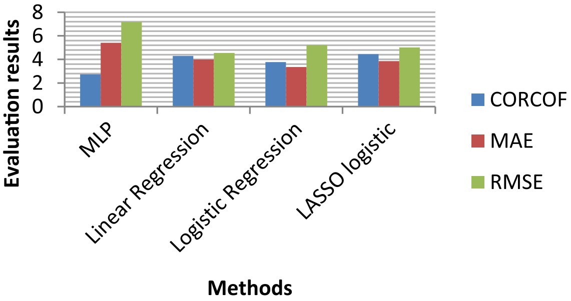


Fig. 7. Performances of Three Measures Vs. Classifiers.

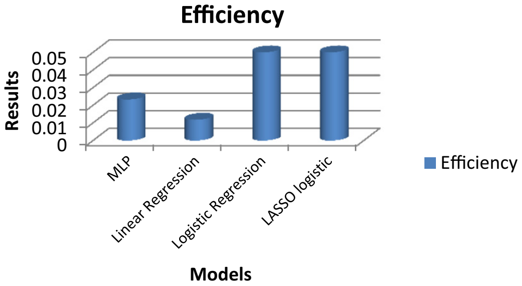


Fig. 8. Efficiency based Time Complexity Vs. Classifiers.

instances among which the gene expression level of cancer data is described in the (Fig. 2).

In this section, classifier algorithm is implemented using WEKA tool to evaluate the predictive model in which the samples of original data set is divided into training and test data and validated with the training data. During preprocessing technique, the data is visualized for accurate classification analysis. (Fig. 3) depicts the classification analysis of two attributes namely Adiponectin and Resistin in which the malignant tumor is identified by decreasing and increasing level of two parameters respectively. Similarly, (Fig. 4). shows the graphical representation of increasing level of Resistin in malignant tumor.

The third phase undergoes LASSO regression model with EM algorithm to measure the evaluate the performance using 10-fold cross validation test (Fig. 5) to evaluate the effectiveness of our novel classifier in terms of Correlation coefficient (CORCOF), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE) . Maximum Likelihood Expectation is deployed on the test set and some of the clustered instances are identified (Fig. 6)

From the Table 2, we can notice that LASSO logistic regression takes 0.04 s for processing the model compared to other models. In the previous proposed model MLP classifier were used. Though trained data set were processed with lower value of incorrect classified instances it lagged in efficiency. By implementing logistic regression efficiency were proved up to 94% but still variation exists between Logistic and LASSO logistic in efficiency by 0.05%. The performance analysis of three models in (Fig. 7) and (Fig. 8) are represented with correlation coefficient and different error rates which proves that LLR is better than other classifiers

5. Conclusion

In this paper, computational methodology namely fuzzy based Logistic regression was introduced for feature selection of cancer gene data. The system also employed a novel model called LASSO Logistic Regression(LLR) derived from the logistic regression for learning the gene pattern in an accurate method and also to identify the victim genes in the cancer data classification. In our proposed algorithm, Logistic regression along with LASSO model was implemented as it is well known for its increased log-likelihood model by implementing some constraints and eliminating some unnecessary covariates. Other models such as Multilayer Perceptron (MLP), Linear Regression were compared with the proposed model that showed the remarkable difference in the classification accuracy with the training data set from the breast cancer data. It was found that the mean classification accuracy on breast cancer dataset with 116 instances were implemented with this novel model was 94.05% of accuracy and gained efficiency with the lower error rate. Though MLE was deployed for mathematical analysis and EM algorithm feasible prediction, there arrived some backlogs that led to ignore some attributes. In the proposed work, cancer prediction is analyzed only with two attributes. In the absence of missing data other attributes are considered for prediction which produces less accuracy results that leads to elimination of that particular data. So this issue can be dealt with the future work where cancer prediction may be carried out with even with the absence of some specific attributes by considering the remaining attributes by implementing machine learning tools.

References

[1] Shreshtha Malvia, Sarangadhara Appalaraju Bagadi, Uma S. Dubey, Sunita Saxena, Epidmiology of breast cancer in Indian Women, Asia –Pac. J. Clin. Oncol. 13 (4) (2017) 289–295, Wiley online Library.

[2] R.L. Siegel, K.D. Miller, Jemal A. Cancer Statistics 00 00 2016. 2016, 1 24 10.3322/caac.21332.

[3] E. Alba, J. García-Niet, L. Jourda, E-G. Talbi Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. In: Evolutionary computation, CEC 2007. pp. 284–9, IEEE congress,2007.

[4] S. Ghora, A. Mukherjee, S. Sengupta, PK. Dutta. Multicategory cancer classification from gene expression data by multiclass NPPC ensemble. In: Systems in medicine and biology (ICSMB), 2010 international conference on. IEEE, pp.41-8, 2010.

[5] S.-B. Guo, M.R. Lyu, T.-M. Lok, Gene selection based on mutual information for the classification of multi-class cancer, Comput Intell Bioinforma (2006) 454–463, Springer.

[6] Xue B, Zhang M, Browne W, Yao X. A survey on evolutionary computation approaches to feature selection. 2015

[7] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley & Sons, 2014.

[8] J.F. Hair, Multivariate Data Analysis, Prentice Hall, 2010.

[9] Stefano Morotti, Eleonora Grandi, Logistic regression analysis of populations of electrophysiological models to assess proarrythmic risk, MethodsX 4 (2017) 25–34, Science Direct, Elsevier.

[10] Sun Mi Kim, Yongdai Kim, Kuhwan Jeong, Heeyeong Jeong, Jiyoung Kim, Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography, Ultrasonography 37 (1) (2018) 36–42.

[11] H. Alshamlan, G. Badr, Y. Alohali, A comparative study of cancer classification methods using microarray gene expression profile, in: Proceedings of the first international conference on advanced data and information engineering, Springer, 2014, pp. 389–398, DaEng- 2013.

[12] Meriem Amrane, Saliha Oukid, Ikram Gagaoua, Tolga Ensari, "Breast Cancer Classification using machine learning", ,IEEE Explore, Conference at Istanbul, Turkey, 2018 Electric Electronics, Computer Science, Biomedical Engineerings Meeting(EBBT), DOI:10.1109/EBBT.2018.8391453, 18-19 April 2018.

[13] Dongdong Sun, Minghui Wang, Ao Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data", IEEE/ACM Transactions on Computational Biology and Bioinformatics (Early Access), pp.1-1, 2018.

[14] O. Okun, Feature selection and ensemble methods for bioinformatics: algorithmic classification and implementations, Med. Inform. Sci. Ref. (2011).

[15] Song L, Smola A, Gretton A, Borgwardt KM, Bedo J. Supervised feature selection via dependence estimation. In: Proceedings of the 24th international conference on machine learning. ACM; 2007. p. 823–30. 2007, June.

[16] W. Li, F. Sun, I. Grosse, Extreme value distribution based gene selection criteria for discriminant microarray data analysis using logistic regression, J. Comput. Biol. 11 (2–3) (2004) 215–226.

[17] R. Varshavsky, A. Gottlieb, M. Linial, D. Horn, Novel unsupervised feature filtering of biological data, Bioinformatics 22 (14) (2006) e507–e513.

[18] L. Ya-qin, W. Cheng, and Z. Lu, "Decision tree based predictive models for breast cancer survivability on imbalanced data," pp. 1–4, 2009.

[19] A. Pradesh, Analysis of feature selection with classification : breast cancer datasets, Indian J. Comput. Sci. Eng. 2 (5) (2011) 756–763.

[20] José María Celaya Padilla, Jorge Armando Ortiz Murillo, María Del Rosario Martínez Blanco, Luis Octavio Solís Sánchez, Rodrigo Castañeda Miranda, Idalia Garza Veloz1-, Margarita Martínez Fierro1, José Manuel Ortiz Rodríguez, Breast Cancer tumor classification using LASSO method selection Approach, Proceedings of the ISSSD, 2016.

[21] Sara Tarek, Reda Abd Elwahab, Mahmoud Shoman, Gene expression based cancer classification, Egyptian Informatics Journal, Science direct, Elsevier, Vol.18, pp.151-159, 2016.

[22] S. Geeitha, M. Thangamani, "Incorporating EBO-HSIC with SVM for Gene Selection Associated with Cervical Cancer Classification", Journal of Medical Systems, Springer, Vol.42, No.11, 2018.

[23] F. Akinsola Adeniyi, M.A. Sokunbi, F.M. Okikiola, I.O. Onadokun, Data Mining For Breast Cancer Classification, Int. J. Eng. Comput. Sci. 6 (8) (2017) 22250–22258, ISSN:2319–7242.

[24] M.R. Nikoo, R. Kerachian, M.R. Alizadeh, A fuzzy KNN-based model for significant wave height prediction in large lakes, Oceanologia 60 (2) (2018) 153–168.