International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Symmetry Based Feature Selection with Multi layer Perceptron for the prediction of Chronic Disease

Sandeepkumar Hegde, Rajalaxmi Hegde

Abstract: Huge amount of Healthcare data are produced every day from the various health care sectors. The accumulated data can be effectively analyzed to identify people's risk from chronic diseases. The process of predicting the presence or absence of the disease and also to diagnosing the various disease using the historical medical data is known as Health Care Analytics. Health care analytics will improve patient care and also the harness practice of medical practitioner. The feature selection is considered as a core aspect of the machine learning which hugely contribute towards the performance of the machine learning model. In this paper symmetry based feature subset selection is proposed to select the optimal features from the Health care data which contribute towards the prediction outcome. The Multilayer perceptron algorithm(MLP) used as a classifier which will predict the outcome by using the features which are selected from the Symmetry-based feature subset selection technique. The chronic disease dataset Diabetes, Cancer, Breast Cancer, and Heart Disease data set accumulated from UCI repository is used to conduct the experiment. The experimental results demonstrate that the proposed hybrid combination of feature selection technique and the multilayer perceptron outperforms in accuracy compare to the existing approaches.

***Index Terms*: Chronic Disease, Feature selection, Healthcare, Machine Learning.**

I. INTRODUCTION

In this technological era, chronic disease is considered as one of the threats to the entire globe. Because around 95% of the nation spending about 3 trillion dollars as health care expenditure in order to cure this disease[1]. According to the new report from the National Center for Chronic Disease, by 2025 67% of the population in India will be a victim of these diseases. The chronic disease is death causing diseases which persist for a longer duration. The disease such as Cancer, Diabetes, Kidney disease, Asthma, Heart Disease, Mental disorder is considered as chronic diseases[2]. Already India is having the largest population of diabetes disease of 35 million in the entire world. One of the remedies could be early predicting the risk of chronic disease so that it can be cured at the earliest.

Feature selection is considered a core part of the machine learning algorithm. The performance of the entire machine learning model depends on the accuracy of the feature selection technique. The filter method, wrapper method, and embedded method are the different types of feature selection technique[3]. The filter-based feature selection Technique is independent of the machine learning model. The wrapper-based feature selection algorithm selects the features based on machine learning classifier. The embedded feature selection uses the strategy of both the filter and wrapper method in selecting the features. These techniques must remove the noise, outliers and irreverent features which will not contribute towards prediction outcome. In this paper novel symmetry based feature subset selection technique in combination with Multilayer Perceptron is proposed in order to early predict the risk of having chronic diseases. The chronic disease data set such as Diabetes data set, Cancer data set, Breast Cancer Data set and Heart Disease Data set is used to conduct the experiment. The Proposed symmetry based feature selection uses a filter approach in selecting the features. The algorithm works based on the concept of significance and redundancy. The chronic disease datasets are given as input to the proposed feature selection algorithm. Each and every feature will undergo the process of significance and redundancy analysis. Only if the significance of the particular features crosses the intended threshold value and if it is not redundant, the particular features are retained. The selected subset of the features is given as input to the multilayer perceptron algorithm. The Multi-Layer Perceptron(MLP) is considered as a family of Artificial Neural Network[4]. The MLP is fed forward neural network technique with three layers in it, input layer, hidden layer, and Output layer. The MLP uses activation functions in order to achieve nonlinearity. It is a Supervised Classification technique, uses backpropagation mechanism in order to train the model. The features selected from the symmetry based feature selection technique given as input to the input layer of MLP. The different weights are assigned to the features based on the feature importance by MLP. The processing of the features will be done in the hidden layers. The output layer of MLP will predict the classification result. The problem of overfitting and underfitting is nullified by using K fold cross validation mechanism[5].

Revised Manuscript Received on 30 July 2019.
* Correspondence Author

Sandeepkumar Hegde*, Department of CSE, NMAM Institute of Technology, Nitte, India.
Rajalaxmi Hegde, Department of CSE, NMAM Institute of Technology, Nitte, India.

The result obtained through the proposed approach is validated using the various measures such as Precision, Recall, F measure, ROC area, and confusion matrix. The experimental results indicated that the proposed approach outperforms in accuracy compared to all the existing approaches.

The paper is organized as follows. Section II focuses on literature work which is already carried out in the same area. Section III explores the proposed methodology. Section IV focuses on experimental results followed by a conclusion.

II. LITERATURE REVIEW

The section explores the various literary work which has been already carried out in the same area.

In [6] Neuro-fuzzy based interface model is used to predict the diabetes disease. The Pima diabetes data set is accumulated from the National Institute of Diabetes. The feature selection is implemented using Genetic Algorithm. The proposed approach achieved a better result compared to the RNN approach. In[7] the C4.5 based algorithm is used to predict the diabetes disease. The approach achieved an accuracy of 72%. The proposed system did not apply any feature selection algorithm to remove the redundant features. In[8] adaptive SVM based algorithm is used to predict the presence of breast cancer disease. The algorithm achieved an accuracy of 76 %. But the proposed approach had an overfitting issue. In[9]LS-SVM based algorithm used to predict the presence of Knee joint disease. The feature selection is performed using a genetic algorithm and the apriori algorithm. The proposed approach achieved an accuracy of 94%.In[10]relief based feature selection in combination with linear SVM based machine algorithm used to predict the heart disease. The drawback with this approach was dataset contained only 270 sets of instances with 13 attributes in it. In[11]weka data mining tool is combined with clustering methods to predict the heart disease. In[12]RFE in combination with SVM is used to predict the prostate cancer disease. The proposed approach achieved higher accuracy compared to existing approaches. In[13]Genetic based feature selection is proposed in combination with Naïve Bayes classifier to predict the Heart Disease. The proposed method achieved an accuracy of 75%.In[14]ILFS based feature selection is proposed to select the subset of the features from the Heart disease data set based on feature rank and weight. The feature selection is applied in combination with SVM. The data set had a total of 699 instances in it with 13 features. The ILFS based feature selection algorithm selected optimal features for the classification task. The proposed technique achieved good accuracy compared to the existing technique. In[15]sub optimum based feature selection algorithm is proposed in combination with particle swarm optimization to select the optimal features from the Breast cancer data set. The feature selection is performed in combination with SVM, ANN and Bayes Network. The experimental results indicated that the proposed technique achieved better accuracy compared to traditional techniques. In[16]Autoencoder based technique is proposed for the classification of Anti-cancer drug response. The autoencoder in combination with Boruta based algorithm used to select the subset of the features. The classification task was performed by using the random forest algorithm. The proposed technique achieved better efficiency compared to the existing approaches.

The main motivation behind the implementation of the paper is to propose an accurate machine learning model towards the prediction of chronic disease. In India, about 67% of the population is the victim of these diseases. The early prediction of the disease with at most accuracy is very much essential. The existing work proposed with these objectives is not proven to be accurate. The effective feature selection algorithm will always make a huge impact on the performance of the machine learning model. Hence in this paper novel, symmetry-based optimized feature subset selection algorithm is proposed in combination with the Multilayer perceptron algorithm which performed well in prediction accuracy compared to an existing technique.

III. PROPOSED METHODOLOGY

The architecture of the proposed system is shown in Fig 1 below. In the proposed work hybrid approach of Symmetry-based feature selection in combination with Multi-Layer Perceptron(MLP) algorithm is utilized to predict the risk of chronic disease.

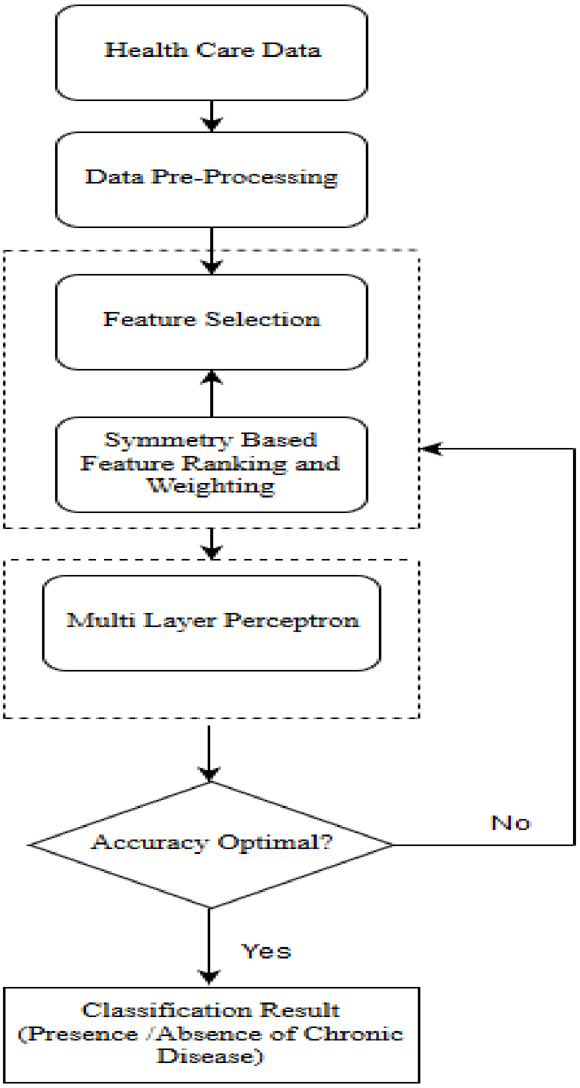

Health Care Data
Data Pre-Processing
Feature Selection
Symmetry Based Feature Ranking and Weighting
Multi Layer Perceptron
Accuracy Optimal?
No
Yes
Classification Result (Presence /Absence of Chronic Disease)

Fig. 1.Proposed Architecture Diagram

The feature selection is considered as a core aspect of machine learning. The performance of the machine learning model depends on the fitness of the feature selection algorithm. The feature selection algorithm must extract only the optimized features which impact on the prediction result. The Health care dataset such as Heart Disease, Breast Cancer, Lung Cancer, Diabetes dataset is given as input to the proposed system. The dataset will undergo the process of Data preprocessing which will remove the noise, outlier if any from the given data. The preprocessed data is given as input to the proposed Symmetry Based Feature Ranking and Weighting based Feature selection system.

The proposed symmetry based feature selection algorithm works based on the concept of information gain, entropy, and merit of feature subset.

The entropy of feature subset can be calculated using equation 1 below.

$$Entropy(f) = \sum_{i=1}^{c} -p_{fi}\log_2 p_{fi} \qquad (1)$$

The information gain between the given features f1 and f2 can be represented as

$$InfoGain(f_1|f_2) = entropy(f_1) - entropy(f_1|f_2) \qquad (2)$$

The merit of each feature subset can be calculated using the formula 3 below.

$$M_{fs} = \frac{n\,m_{cf}}{\sqrt{n+n(n-1)m_{ff}}} \qquad (3)$$

M_{fs} is the merit of feature subset
n is the number of features.
m_{cf} is mean of the class feature correlation
m_{ff} is mean of the feature-feature correlation

Hence symmetry between the two features can be obtained using the equation (4).

$$Symmetry(f_1, f_2) = 2 * \frac{InfoGain(f_1|f_2)}{entropy(f_1)+entropy(f_2)^1} \qquad (4)$$

The symmetry based feature selection uses the concept of mutual information in order to derive the optimal features from the given set of features. The mutual information represents the relationship between each of the features and effect of these in the target class. Hence the goodness of each feature is identified by measuring the correlation between each feature and target class concept. The gain of the each feature is computed using the concept of Information gain. The computation of the information gain involve the entropy calculation by subtracting the entropy of class label for full data set with conditional entropy of the features. The entropy calculation will give the frequency count of each class label.

The proposed symmetry based feature selection algorithm is shown in the Fig 2 . Initially, the entropy of each feature is computed using equation 1. The purity of each feature is obtained by measuring the entropy value using equation 2. The merit of each subset of the feature is measured and symmetrized relationship between the set of features are obtained using the equation 4.The feature selection system will select the optimal features from the given dataset which will impact the prediction output.

Algorithm: Symmetry Based Feature Selection

```
Let S denotes the symmetric matrix
E represents the entropy of features
F represents the Feature
Calculate entropy using
    $Entropy(f) = \sum_{i=1}^{c} -p_{fi}\log_2 p_{fi}$
Calculate the merit of the feature using
    $M_{fs} = \frac{n\,m_{cf}}{\sqrt{n+n(n-1)m_{ff}}}$
Calculate information gain IG(DC|DA)
DC denotes the class label
DA denotes the attribute values
for each $DC_i \in$ DC do
    Compute P(DC[i])
        Fe = F+P(DC[i])*$\log_2$(P(Dc[i]))
        F←Fe
end for
for each attribute ai ∈ DA
    compute P(a[j])
    count = F+P(a[j])*$\log_2$(P(a[j]))
    F←count
end for
for each DCi do
    for each aj do
        find P(DC[i]|a[j])
        FS = F+P(DC[i]|a[j]) * $\log_2$ P(DC[i]|a[j])
        F←FS
    end for
end for
compute the symmetric matrix using
$Symmetry = 2 * \frac{InfoGain(f_1|f_2)}{entropy(f_1) + entropy(f_2)^1}$
```

Fig. 2.Proposed Symmetry Based Feature Selection Algorithm

The selected features are given as input to the Multilayer perceptron based classifier algorithm as shown in Fig 3 below. In the proposed system multilayer perceptron is constructed using 3 layers namely input layer, hidden layer, and output layer.

The MLP with one hidden layer is a mapping between the input vectoriv and output vectorov shown in the equation(5) below.

$$f: R^{iv} \to R^{ov} \qquad (5)$$

Hence the matrix representation of the same can be denoted as

$$f(x) = af_2\left(bw^{(2)} + w^{(2)}\left(af_1\left(bw^{(1)} + w^{(1)}x\right)\right)\right) \qquad (6)$$

In the equation (6) above af_1 af_2 and are the activation function, $bw^{(1)}$ and$bw^{(2)}$ are the bias vectors, $w^{(1)}$ and $w^{(2)}$ are weight matrices. In the proposed approach the sigmoid is used as an activation function in order to achieve the non-linearity.

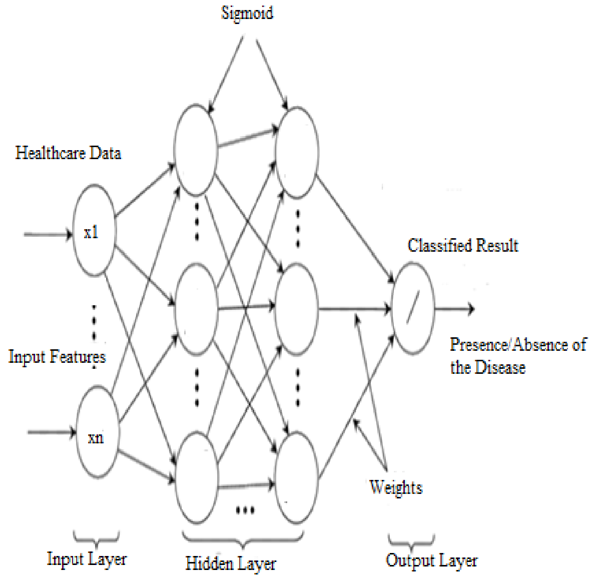


Fig. 3.Multi-Layered Perceptron Architecture

In the proposed system multilayer perceptron is constructed using 3 layers namely input layer, hidden layer, and output layer.

The MLP with one hidden layer is a mapping between the input vectoriv and output vectorov shown in the equation(7) below.

$$f: R^{iv} \rightarrow R^{ov} \qquad (7)$$

Hence the matrix representation of the same can be denoted as

$$f(x) = af_2\left(bw^{(2)} + w^{(2)}\left(af_1\left(bw^{(1)} + w^{(1)}x\right)\right)\right) \qquad (8)$$

In equation (8) above af_1 and af_2 are the activation function, $bw^{(1)}$ and$bw^{(2)}$ are the bias vectors, $w^{(1)}$ and $w^{(2)}$ are weight matrices. In the proposed approach the sigmoid is used as an activation function in order to achieve the non-linearity.

The set of features $f_1,f_2….f_n$ are given as input to the perception. The various weights $w_1,w_2..w_n$ are assigned and adjusted in order to train the system using the different earning rate using the equation (9) below.

$$w = w + LR * (expected - predicted) * x \qquad (9)$$

Where LR is nothing but the learning rate set by the MLP.

The learning will occur in MLP by adjusting weights after each processed data based on the error between expected output and actual output.

$$e_i n = a_i(n) - ex_i(n) \qquad (10)$$

The error in the output node i at the nth data point can be derived by computing the difference between a_i the actual outcome andex_i the expected outcome.

The weights in each of the neuron nodes are adjusted which minimize the error rate of entire output using equation 11 below.

$$e(n) = \frac{1}{2}\sum_i e_i^{\,2}(n) \qquad (11)$$

As discussed in the section above the nonlinear activation is performed in MLP by using the sigmoid function.

$$Sigmoid\ \sigma(f) = \frac{1}{1+e^{-f}} \qquad (12)$$

The sigmoid function is shown in the Fig 4 which maps the machine learning models in the range between 0 and 1.The sigmoid function most suitable for Binary Classification problem.

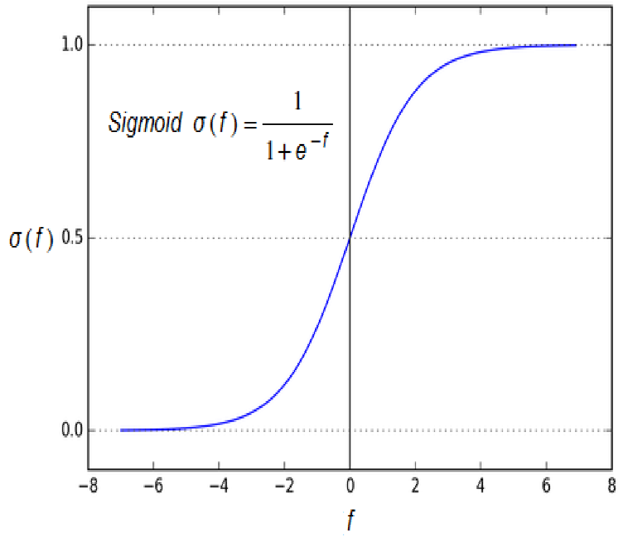


Fig. 4.Sigmoid Function

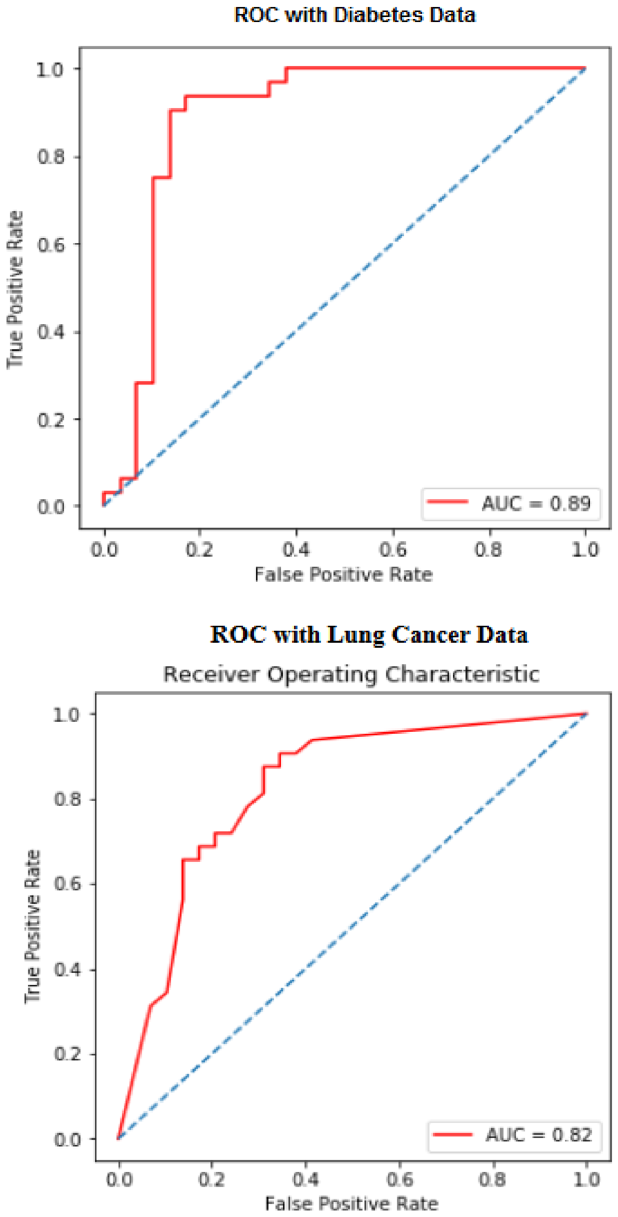
The sigmoid function is considered as differential function because slope of the sigmoid lies between two points. It is non linear in nature which will give analog like activation. It is considered as smoothing gradient. It can be used with machine learning model which can predict the probability. The sigmoid is commonly used activation function in MLP which will take the input in terms of real value and squashes it to 0 and 1. Since most of the Disease prediction problems belong to Binary classification i.e Presence or absence of the disease, in the proposed system sigmoid is chosen as the activation function for MLP.

If the accuracy obtained through the proposed system is not optimal, a subset of the features are selected from the remaining set and classification task is performed using the MLP system. The task is repeated recursively until optimal accuracy is obtained.

IV. RESULTS AND DISCUSSION

The experiment is conducted using the Disease data set Lung Cancer, Heart Disease, Breast Cancer, Diabetes Data sets taken from the UCI repository. The Datasets are passed as input to the Proposed system. The Data sets will be preprocessed to remove the noise and outliers. The optimal features from the Datasets are filtered using the proposed symmetry based feature selection technique. The predictions are carried out by using the MLP based supervised classifier technique.

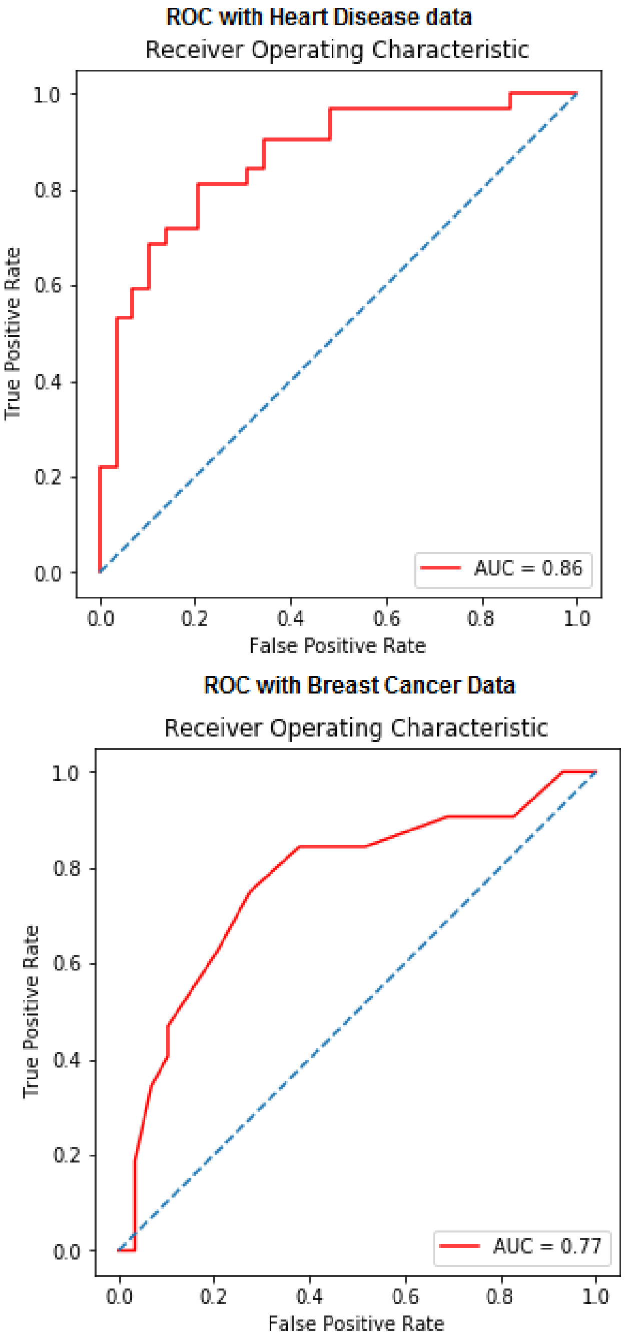
In order to avoid the problem of Data overfitting, underfitting or biased model, the data sets will undergo the process of K fold cross-validation. The Value of K is chosen as 1.The accuracy obtained with the proposed system is validated using the various measures such as Root mean square, True positive rate(TP rate), True negative rate(TN rate), Precision, Recall, F-measure and ROC area.


ROC with Diabetes Data

ROC with Lung Cancer Data

Fig. 5.ROC Curves with Diabetes and Lung Cancer Data

The Receiver Operating characteristics(ROC) is considered as one of the essential performance measure for the machine learning Classifier. The ROC can be visualized using the Area Under Curve(AUC).The ROC can be calculated based on the number of True Positive Cases and False Positive Cases obtained with the Machine learning classifier. The True positive cases are plotted in the Y axis and False Positive Cases are plotted in the X axis. The Best ROC rate for any classifier is 1 and poor ROC rate is 0.The ROC illustrates the capability of machine learning classifiers to separate between the different classes.


ROC with Heart Disease data

ROC with Breast Cancer Data

Fig. 6.ROC Curves with Heart and Breast Cancer Data

The ROC curve obtained with the proposed approach is as shown in the Fig 5 and Fig 6 above. The ROC rate 0.89 and 0.82 is achieved with Diabetes and Lung cancer data where as with the Heart Disease and Breast Cancer , ROC rate of 0.86 and 0.77 is recorded. The performance of the proposed approach is also measured using the other parameter such as Precision , Recall, F measure, True Positive rate and false positive rate.

Table I. Performance Statistics I

Data Set	Precision	Recall	F measure	TP Rate	FP Rate
Diabetes	0.780	0.866	0.821	0.866	0.13
Lung Cancer	0.923	0.857	0.889	0.858	0.14
Breast cancer	0.744	0.826	0.783	0.826	0.17
Heart Disease	0.801	0.833	0.801	0.833	0.16

The table 1 above shows the performance statistics when the health care datasets are passed as input to the proposed system. As shown in the table II below, performance of the proposed system is also validated using the different measure such as kappa statistics, Mean absolute error, ROC area, PRC area and Root mean square.

Table II. Performance Statistics II

Data Set	**Kappa statistics**	**Mean absolute error**	**ROC area**	**PRC area**	**Root mean square error**
Diabetes	0.431	0.313	0.890	0.801	0.414
Lung Cancer	0.323	0.115	0.820	0.930	0.280
Breast cancer	0.452	0.348	0.770	0.748	0.525
Heart Disease	0.5779	0.2281	0.860	0.842	0.4271

The confusion matrix obtained with the Diabetes data shown in table III. It indicates that out of 768 instances total 579 instances are classified correctly and 189 instances are in correctly predicted.

Table III. Confusion Matrix With Diabetes Data

True Label	Negative	True Negative	416	False Positive	84
	Positive	False Negative	105	True Positive	163
		Negative		Positive	
		Predicted Label			

As indicated in the confusion matrix in the table IV, out of 286 instances 194 instances are correctly predicted with the Breast Cancer Data

Table IV. Confusion Matrix With Breast Cancer Data

True Label	Negative	True Negative	125	False Positive	25
	Positive	False Negative	31	True Positive	89
		Negative		Positive	
		Predicted Label			

The confusion matrix of lung cancer data set shown in table V indicates that out of 32 test set instances 26 instances are predicted correctly and only 6 instances are wrong predicted.

Table V Confusion Matrix With Lung Cancer Data

True Label	Negative	True Negative	2	False Positive	2
	Positive	False Negative	4	True Positive	24
		Negative		Positive	
		Predicted Label			

The confusion matrix of heart data is shown in table VI below. The data set had total 270 instances. Total 214 instances are predicted correctly.

Table VI Confusion Matrix With Heart Disease Data

True Label	Negative	True Negative	166	False Positive	35
	Positive	False Negative	57	True Positive	28
		Negative		Positive	
		Predicted Label			

The accuracy obtained with proposed system is compared with different machine learning classifier.

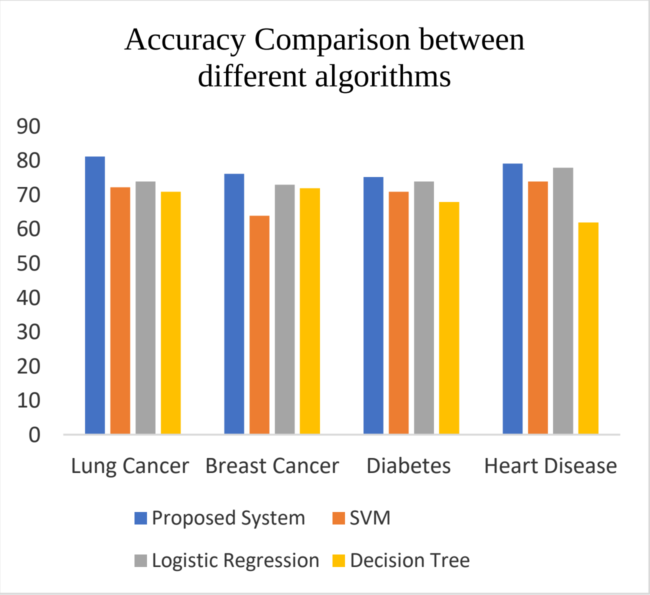


Fig. 7.Accuracy Comparison between Algorithms

Retrieval Number: B2658078219/19©BEIESP
DOI: 10.35940/ijrte.B2658.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The accuracy obtained with the proposed system is compared with different machine learning classifier. The accuracy comparison graph between different techniques is shown in Fig 7 above. The graphs indicate that the proposed system obtained higher accuracy compared with the existing approaches.

V. Conclusion

In this paper symmetry based feature selection technique is proposed in combination with Multilayer perceptron Algorithm to predict the early risk of chronic disease. The experiment is conducted using chronic disease data sets Lung Cancer, Diabetes, Breast cancer, and Heart Disease. The proposed symmetry based feature selection technique filtered the optimal features from the Data set which is given as input to the MLP based classifier. The prediction results showed that the proposed system performed better in accuracy as well as in terms of the other performance measures such as confusion matrix, ROC, Precision and Recall compared to the existing approaches. The limitation of the approach is, the proposed technique has experimented with the smaller data set As future work the same approach can be extended with the larger data set and analysis on the performance can be made.

References

1. Jain, Divya, and Vijendra Singh. "Feature selection and classification systems for chronic disease prediction: A review." *Egyptian Informatics Journal* 19.3 (2018): 179-189.
2. Gharibdousti, Maryam Soltanpour, et al. "Prediction of chronic kidney disease using data mining techniques." *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE), 2017.
3. Reddy, N. Satish Chandra, et al. "Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction." *International Journal of Innovative Computing* 9.1 (2019).
4. Heidari, Ali Asghar, et al. "Ant Lion Optimizer: Theory, Literature Review, and Application in Multi-layer Perceptron Neural Networks." *Nature-Inspired Optimizers*. Springer, Cham, 2020. 23-46.
5. He, Jinbo, and Xitao Fan. "Evaluating the Performance of the K-fold Cross-Validation Approach for Model Selection in Growth Mixture Modeling." *Structural Equation Modeling: A Multidisciplinary Journal* 26.1 (2019): 66-79.
6. Alby, S., and B. L. Shivakumar. "A prediction model for type 2 diabetes using adaptive neuro-fuzzy interface system." *Biomedical Research (0970-938X)* (2018).
7. Saxena, Kanak, and Richa Sharma. "Diabetes mellitus prediction system evaluation using c4. 5 rules and partial tree." *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions)*. IEEE, 2015.
8. Gürbüz, Emre, and Erdal Kılıç. "A new adaptive support vector machine for diagnosis of diseases." *Expert Systems* 31.5 (2014): 389-397.
9. Jac Fredo, Agastinose Ronickom, et al. "Classification of normal and knee joint disorder vibroarthrographic signals using multifractals and support vector machines." *Biomedical Engineering: Applications, Basis and Communications* 29.03 (2017): 1750016.
10. Wu, Yunfeng, et al. "Classification of knee joint vibration signals using bivariate feature distribution estimation and maximal posterior probability decision criterion." *Entropy* 15.4 (2013): 1375-1387.
11. Mirmozaffari, Mirpouya, Alireza Alinezhad, and Azadeh Gilanpour. "Heart disease prediction with data mining clustering algorithms." *Int'l Journal of Computing, Communications & Instrumentation Engg (IJCCIE), ISSN* (2017): 2349-1469.
12. Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." *Machine learning* 46.1-3 (2002): 389-422.
13. Le, Hung Minh, Toan Dinh Tran, and L. A. N. G. Van Tran. "Automatic Heart Disease Prediction Using Feature Selection And Data Mining Technique." *Journal of Computer Science and Cybernetics* 34.1 (2018): 33-48.
14. Roffo, Giorgio, et al. "Infinite latent feature selection: A probabilistic latent graph-based ranking approach." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.
15. Xue, Bing, Mengjie Zhang, and Will N. Browne. "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms." *Applied soft computing* 18 (2014): 261-276.
16. Lu, Xiaolu, et al. "Autoencoder based Feature Selection Method for Classification of Anticancer Drug Response." *Frontiers in Genetics* 10 (2019): 233.

AUTHORS PROFILE

Mr. Sandeepkumar Hegde is working as Assistant Professor in the Department of Computer Science &Engg at NMAM Institute of Technology. He has over 8 years of Teaching experience. He also served as Assistant System Engineer at Tata Consultancy services. He completed his B.E degree in Information Science& Engg and Master Degree in Computer Science& Engg from Visvesvaraya Technological University, Belagavi. He is Currently pursuing Ph.D. from Visvesvaraya Technological University. He has published various research papers in National and International Journals and conferences..

Mrs.Rajalaxmi Hegde is working as Assistant Professor in the Department of Computer Science &Engg at NMAM Institute of Technology. She has over 6 years of Teaching experience. She completed her B.E degree in Information Science& Engg and Master Degree in Computer Science&Engg from Visvesvaraya Technological University, Belagavi. She is Currently pursuing Ph.D. from Visvesvaraya Technological University. She has published various research papers in National and International Journals and conferences. She is member of ISTE.