ISSN(O): 2456–026X

Medical Data Classification Using Correlation Based Feature Selection and Multilayer Perceptron Model

Amit Yerpude[1*], Sanjeev Sharma[2]
[1]Assistant Professor, Dept. of Computer Science & Engg., Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India. Email: amit.yerpude@gmail.com
[2]Associate Professor, Dept. of Computer Science & Engg., Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India. Email: sanjeev.sharma1868@gmail.com
*Corresponding Author

Abstract: The classification accuracy in medical diagnosis is critically very important. It is highly reliant on the quality of data and the features use for this purpose. Identifying a good set of features from the set of collected feature is very challenging nowadays because of the available size of data. This challenge can be addressed by applying feature selection method. We create a model for classification based on correlation based feature selection technique and multi layer perceptron model. We select nine different medical dataset and test the performance of the proposed model on it. The experimental results show that the classification accuracy is improved by upto 28%, and the execution time reduces upto 30% to 97% subjected to the different datasets.

Keywords: Classification, Feature selection, Medical dataset, Multilayer perceptron.

I. Introduction

The constraint which need to be in consideration most when we classifying the medical dataset, is identifying a good representative set of features that improve the performance of a classification model [1] [2]. A good feature subset is the set of features that show maximum correlation within the class, yet least correlation with each other. The factors that affect the success of machine learning for a given task are quality of the example dataset, and set of features. Theoretically, having more features should result in more discriminating power. However, practical experience shown that this is not always the case.

In the recent years, many researchers are focusing on feature selection techniques [2] because of the continuously increasing sizes of databases. The problem with these large data is to filter out the data which is useful for extracting knowledge from these databases. Classification (or prediction) is an indispensable part of data mining, machine learning or pattern recognition. A good classifier is able to predict the classes of the unknown patterns and thus produces good classification accuracy. Higher is the accuracy, better will be the prediction for classification models.

This paper is organized in seven sections. In section 2 and section 3, we discussed about different feature selection methods and a brief on correlation based feature selection approach. In section 4, the working of multilayer perceptron model in discussed with its neuron architecture. Section 5 enlighten the structure of proposed methodology, which followed by section 6 which shows the different experimental results. In the last section we conclude the outcomes of our experiments and future enhancements.

II. Feature Selection Methods

Feature Subset Selection is an indispensable pre-processing task in Data Mining. The feature selection methods [3] [4] [5] are typically divided in three classes: Filter method, Wrapper method and Embedded method. Filter type methods [6] [7] [8] [9] select features regardless of the model, based only on correlation with the class to foretell. Filter methods restrain the variables which are of least interest. Remaining variables will be use for the classification or a regression model used to predict data or classify. These methods are particularly effective in terms of computation time and robustness over fitting. Wrapper methods [6] [7] evaluate subsets of variables, unlike filter approaches, to detect the possible relations between variable. Embedded methods are the methods which try to combine the advantages of Filter method and Wrapper method. A learning algorithm takes benefit of its own procedure of selecting variables and performs feature selection and classification concurrently.

Feature subset selection [1] is approach of identifying and removing as much irrelevant and redundant information as possible, which reduces the dimensionality of the data and allow learning algorithms to run faster and more efficiently. In some cases, classification accuracy is improved; whereas in some cases, the result is a more compact and easy to interprete the representation over target. In the filter based feature selection approach [2], the goodness of any feature is evaluated using numerical or essential properties of the dataset. Based on these

properties, a feature is adjudged as the most appropriate feature and is preferred for machine learning model.

III. Correlation Based Feature Selection

Correlation based Feature Selection (CFS) [1] [10] is an algorithm that joints the idea of an appropriate correlation measure and a heuristic search strategy. By evaluating the significance of a subset of features, taking into consideration the individual predictive ability of each feature along with the degree of redundancy between them, subsets of features that are highly correlated with the target class while having low inter-correlation are preferred. Best first, explores the space of attribute subsets by greedy hill climbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Best first may start with the empty set of features and search forward, or start with the full set of features and search backward, or start at any point and search in both directions.

The Fig. 1 shows the basic functioning of Correlation based feature selection model.

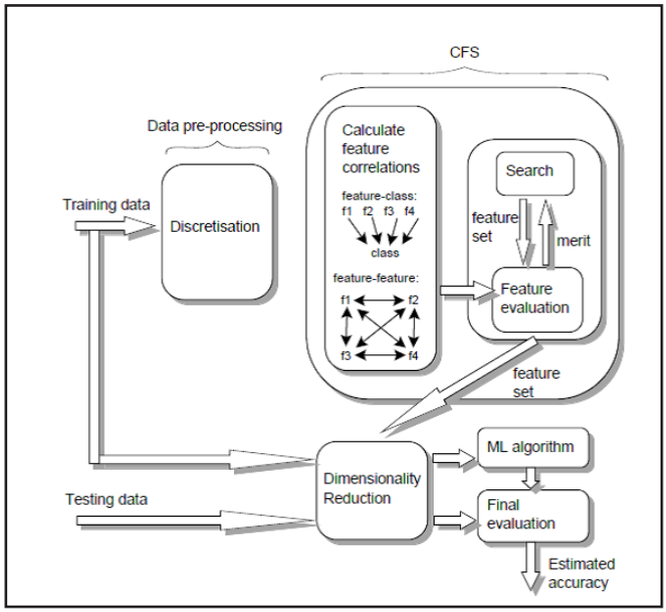

Fig. 1: Model for Correlation Based Feature Selection

IV. Multilayer Perceptron

Neural networks [11] [12] evolved from the working principle of biological nervous system, such as the brain on a human. Neural networks are imitation of biological neuron, as a set of nodes and inter-connections between them. The connections have synoptic weights coupled with them, signifying the "strength" of those connections. A multilayer perceptron [13] is a feed forward neural network model that maps sets of input data onto a set of suitable target output. Back propagation is one of the most widely used learning algorithm for multilayer perceptron in neural networks.

A Classifier that uses back propagation [14] [15] to classify instances, can be built by hand, created by an algorithm or with the combination of these two approaches. The network can also be monitored and customized during training time. The nodes in this network are all sigmoid (except for when the class is numeric in which case the output nodes become un-thresholded linear units). Fig. 2 explains the basic model of a multilayer perceptron model.

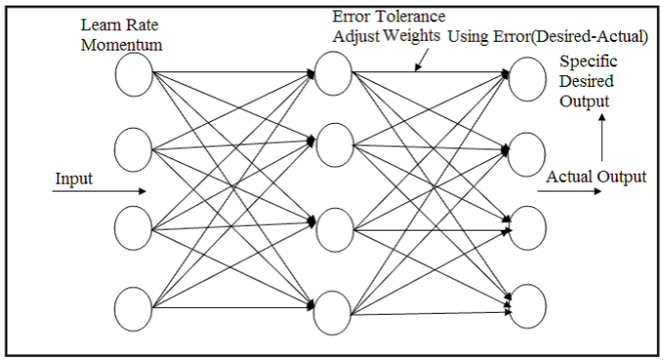

Fig. 2: Basic Model for Multilayer Perceptron

V. Proposed Model for Classification

The working of proposed model is shown in Fig. 3. First we feed the dataset and apply the filter to reduce the number of features, for that we use Correlation Based Feature Selection approach. Then we prepare updated dataset with theses selected features. This updated dataset is use to train the multilayer perceptron network, after training the trained net is use for further classification. We are using 10 cross fold technique for bifurcation of training and testing dataset.

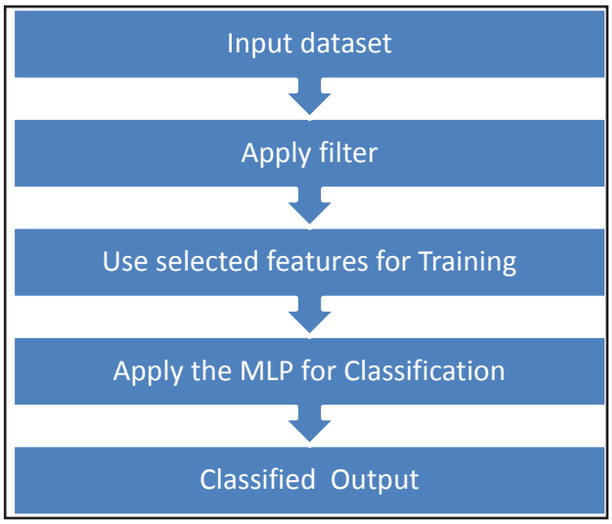

Fig. 3: Flow Diagram of Proposed Model

VI. Experimental Results

To evaluate the performance of the proposed model we use nine different medical datasets, collected from UCI repository

with 2-class and multi class datasets having different number of features [16]. Evaluation is done on the parameters of, number of features reduced, classification accuracy, TP rate, FP rate, precision, recall, F-measure, area under ROC curve and execution time [11].

The performance measures for the evaluation of the classification results, is based on TP/TN, is the number of True Positives/Negatives instances and FP/FN is the number of False Positives/Negatives instances.

Precision is a proportion of predicted positives which are actual positive:

$$Precision = \frac{TP}{TP + FP}$$

Recall is a proportion of actual positives which are predicted positive:

$$Recall = \frac{TP}{TP + FN}$$

Precision and recall measures are utilized to find the best method, but it is not easy to make decision. Thus, F-measure was used to get a single measure to evaluate results.

The F-measure is the harmonic mean of precision and recall:

$$F - measure = \frac{2TP}{2TP + FN + FP}$$

In a ROC curve [17] the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups.

TABLE I: DIFFERENT DATASETS USED

Actual Dataset			
Dataset	Number-of-Instances	Number-of-Features	o/p classes
Brest-can	286	10	2
Diabetes	768	9	2
Hepatitis	155	20	2
Hypothyroid	3772	30	4
Lung-cancer	32	57	2
Lymph	148	19	4
Primary-tumor	339	18	22
Dermatology	366	35	6
Heart-statlog	270	14	2

TABLE II: NUMBER OF FEATURES AFTER APPLYING CORRELATION BASED FEATURE SELECTION METHOD

Number-of-Features-Reduced		
Dataset	Actual	Processed
Brest-can	10	6
Diabetes	9	5
Hepatitis	20	11
Hypothyroid	30	6
Lung-cancer	57	9
Lymph	19	11
Primary-tumor	18	13
Dermatology	35	20
Heart-statlog	14	8

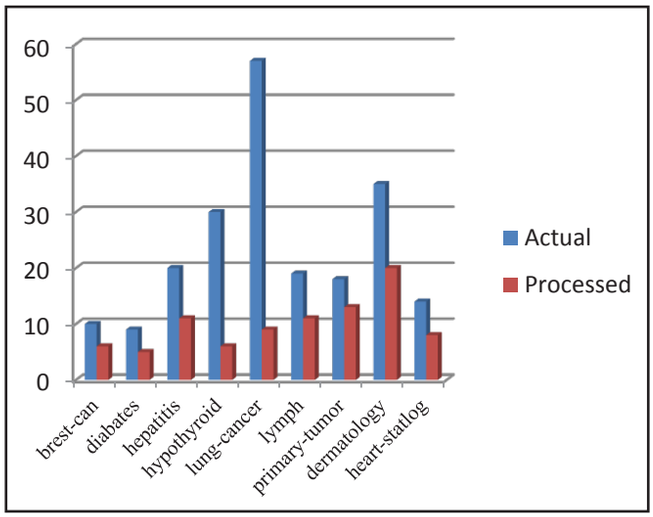

Fig. 4: Number of Features after Applying Correlation Based Feature Selection Method

TABLE III: COMPARISON OF CLASSIFICATION ACCURACY BEFORE AND AFTER PROCESSING

Classification-Accuracy			
Dataset	Actual	Processed	Improvement (%)
Brest-can	64.6853	71.678	10.81
Diabetes	75.3906	75.5208	0.17
Hepatitis	80	84.5161	5.65
Hypothyroid	94.1676	96.1029	2.06
Lung-cancer	65.625	84.375	28.57
Lymph	84.4595	81.7568	-3.20
Primary-tumor	38.3481	40.413	5.38
Dermatology	96.1749	96.4481	0.28
Heart-statlog	78.1481	82.2222	5.21

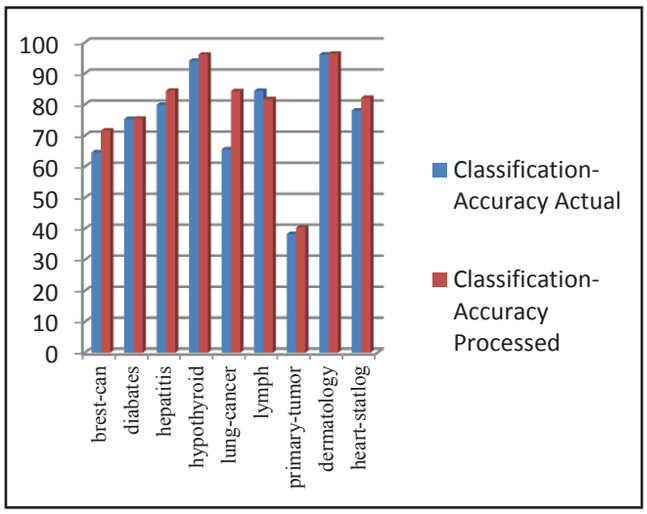

Fig. 5: Comparison of Classification Accuracy Before and After Processing

TABLE IV: COMPARISON OF TP RATE BEFORE AND AFTER PROCESSING

TP-Rate		
Dataset	Actual	Processed
Brest-can	64.70%	71.70%
Diabetes	75.40%	75.50%
Hepatitis	80.00%	84.50%
Hypothyroid	94.20%	96.10%
Lung-cancer	65.60%	84.40%
Lymph	84.50%	81.80%
Primary-tumor	38.30%	40.40%
Dermatology	96.20%	96.40%
Heart-statlog	78.10%	82.20%

TABLE V: COMPARISON OF FP RATE BEFORE AND AFTER PROCESSING

FP-Rate		
Dataset	Actual	Processed
Brest-can	48.90%	45.90%
Diabetes	31.40%	31.10%
Hepatitis	37.60%	34.10%
Hypothyroid	42.90%	26.10%
Lung-cancer	54.00%	26.40%
Lymph	15.70%	17.30%
Primary-tumor	6.50%	6.10%
Dermatology	0.70%	0.70%
Heart-statlog	21.80%	19.40%

TABLE VI: COMPARISON OF PRECISION BEFORE AND AFTER PROCESSING

Precision		
Dataset	Actual	Processed
Brest-can	64.80%	70.10%
Diabetes	75.00%	75.20%
Hepatitis	80.70%	84.20%
Hypothyroid	93.50%	95.90%
Lung-cancer	64.50%	84.00%
Lymph	83.70%	81.10%
Primary-tumor	36.80%	39.60%
Dermatology	96.20%	96.50%
Heart-statlog	78.40%	82.40%

TABLE VII: COMPARISON OF RECALL BEFORE AND AFTER PROCESSING

Recall		
Dataset	Actual	Processed
Brest-can	64.70%	71.70%
Diabetes	75.40%	75.50%
Hepatitis	80.00%	84.50%
Hypothyroid	94.20%	96.10%
Lung-cancer	65.60%	84.40%
Lymph	84.50%	81.80%
Primary-tumor	38.30%	40.40%
Dermatology	96.20%	96.40%
Heart-statlog	78.10%	82.20%

TABLE VIII: COMPARISON OF F-MEASURE BEFORE AND AFTER PROCESSING

F-Measure		
Dataset	Actual	Processed
Brest-can	64.70%	70.50%
Diabetes	75.10%	75.30%
Hepatitis	80.30%	84.30%
Hypothyroid	93.80%	96.00%
Lung-cancer	65.00%	84.10%
Lymph	83.40%	80.80%
Primary-tumor	37.30%	39.90%
Dermatology	96.20%	96.40%
Heart-statlog	78.20%	82.10%

TABLE IX: COMPARISON OF AREA UNDER ROC CURVE BEFORE AND AFTER PROCESSING

ROC-Area		
Dataset	Actual	Processed
Brest-can	0.623	0.619
Diabetes	0.793	0.809
Hepatitis	0.823	0.863
Hypothyroid	0.893	0.97
Lung-cancer	0.676	0.918
Lymph	0.92	0.915
Primary-tumor	0.784	0.781
Dermatology	0.997	0.998
Heart-statlog	0.839	0.842

TABLE X: COMPARISON OF EXECUTION TIME BEFORE AND AFTER PROCESSING AND IMPROVEMENT IN TIME

Execution-Time(in-seconds)			
Dataset	Actual	Processed	Improvement (%)
Brest-can	17.34	4.38	74.74
Diabetes	2.03	1.29	36.45
Hepatitis	1.22	0.55	54.92
Hypothyroid	78.27	10.19	86.98
Lung-cancer	8.38	0.23	97.26
Lymph	3.98	1.7	57.29
Primary-tumor	13.21	11.31	14.38
Dermatology	81.98	30.63	62.64
Heart-statlog	1.63	0.56	65.64

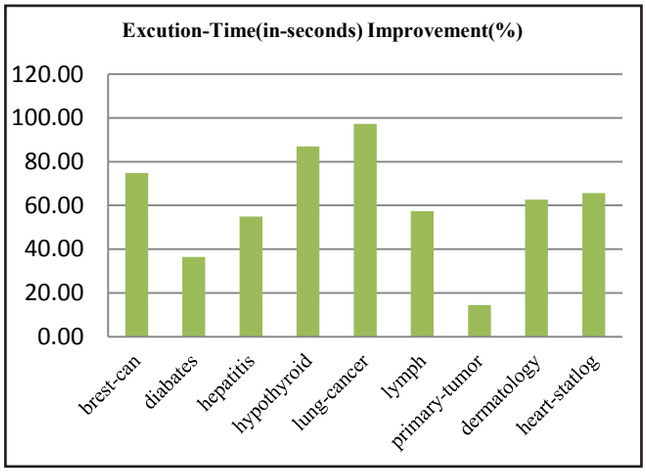

Fig. 6: Improvement in Execution Time with Proposed Model

VII. CONCLUSION AND FUTURE SCOPE

Experimental results are shown in Table II to X. The results of Table II and III show that, the proposed model reduces the numbers of features upto 50% without affecting the classification accuracy in 8 out of 9 cases except the lymph dataset for which the classification accuracy is decreased by 2%. The Tables IV, V, VI, VII, VIII and IX compares the performance of model over tradition model and show the significant improvement in TP-Rate, FP-Rate, Precision, Recall, F-Measure and ROC area. Result of Table X shows that the proposed model is faster and more accurate then the tradition model, in 7 out of 9 cases it is even more than 50%.

The work can be extended by applying other type of filters and classifiers, further the results can be improved by applying hybrid models, and by the use of nature inspired algorithms.

REFERENCES

[1] M. A. Hall, *Correlation-based Feature Selection For Machine Learning*, 1999.

[2] A. K. Saxena, and V. K. Dubey, "Hybrid classification model of correlation-based feature selection and support vector machine," In *IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, Bangalore, India, 10-11 March 2016.

[3] R. Porkodi, "Comparison of filter based feature selection algorithms: An overview," *International Journal of Innovative Research in Technology & Science*, vol. 2, no. 2, pp. 108-113, Oct. 2017.

[4] K. Sutha, and J. J. Tamilselvi, "A review of feature selection algorithms for data mining techniques," *International Journal on Computer Science and Engineering*, vol. 7, no. 6, pp. 63-67, June 2015.

[5] T. Z. Phyu, and N. N. Oo, "Performance comparison of feature selection methods," In *MATEC Web of Conferences,* Published by EDP Sciences, pp. 1-4, 2016.

[6] N. Cueto-Lopez, and R. Alaiz-Rodrıguez, "Assessing feature selection techniques for a colorectal cancer prediction model," In Springer International Publishing, Leon, Spain, pp. 471-481, 6-8 September 2017.

[7] J. Brownlee. (6 October 2014) https://machinelearningmastery.com. [Online]. Available: https://machinelearningmastery.com/an-introduction-to-feature-selection/

[8] P. Yildirim, "Filter based feature selection methods for prediction of risks in hepatitis disease," *International Journal of Machine Learning and Computing*, vol. 5, no. 4, August 2015.

[9] M. A. Ambusaidi, X. He, P. Nanda, and Z. Tan, "Building an intrusion detection system using a filter-based feature selection algorithm," *IEEE Transactions on Computers*, November 2014.

[10] J. Alhajjaj, B. Alnajrani, I. Elaalami, A. Alqahtani, N. Aldhafferi, T. O. Owolabi, and S. O. O. R. Alyami, "Investigating the effect of correlation based feature selection on breast cancer diagnosis using artificial neural network and support vector machines," IEEE, 2017.

[11] P. Yildirim, "Chronic kidney disease prediction on imbalanced data by multilayer perceptron," In *2017 IEEE 41st Annual Computer Software and Applications Conference*, pp. 193-198, 2017.

[12] S. Haykin, *Neural Networks - A Comprehensive Foundation*, 2nd ed. Englewood Cliffs: Prentice-Hall, 1998.

[13] S. K. Biswas, M. C. Urmi, A. Siddique, and M. M. A. Mia, "An algorithm for training Multilayer Perceptron (MLP) for image reconstruction using neural network without overfitting," *International Journal of Scientific & Technology Research*, vol. 4, no. 2, pp. 271-275, February 2015.

[14] P. E. Hart, D. G. Stork, and R. O. Duda, *Pattern Classification*, New York, Wiley, 2001.

[15] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford, Oxford University Press, 1995.

[16] (July, 2008) http://repository.seasr.org. [Online]. Available: http://repository.seasr.org/Datasets/UCI/arff/

[17] https://www.medcalc.org. [Online]. Available: https://www.medcalc.org/manual/roc-curves.php