Improvised mRMR feature selection for predicting Breast Cancer

Senthil S, Deepa B G

Abstract: ***The focus of the proposed method is to provide a solution to the problem of predicting the presence of breast cancer for the data in the UCI Repository. The strong ideology of the proposed method is to predict the presence of cancerous information based on the details of parameters from UCI Repository. The feature selection of proposed method tunes certain parameters to select only few features which are most essential and relevant and far away from the redundant information. The output of feature selection algorithm is given to the SVM classifier with various parameters to train and test in the ratio of 90:10, where 90% of information is considered as Training data with proposed method and the rest 10% of data is considered as a Test the data. The proposed method has included the improvement in mRMR feature selection by tuning the parameters of features with respect to feature set. Thus, the proposed statistical approach has yielded a good result of 98.3% accuracy during the testing phase against the training phase over the UCI Wisconsin data repository.***

Index Terms**: *Breast Cancer, Feature Extraction, Improvised mRMR, Classification.***

I. INTRODUCTION

Prevention is better than cure. Identification of cancer at early stages is very important in surviving the lives of the human beings. Thus, we have a new approach of analyzing the data of UCI repository; as such, the analysis can be done at the early stages of the patient's data by incorporating the proposed statistical method. Treatment for cancer can be provided once it is detected. Thus, there are various mechanisms to determine the presence of cancer based on the details of the patient's data. The mechanisms include Digital Mammography and Ultrasound, where the digital mammography includes various methodologies to predict the presence of cancer tissues in human body. The different stages of the prediction are Feature Processing, Feature Extraction, Feature Selection, and Feature Classification. Similarly, the Ultrasound mechanism includes passing the ultrasound rays on human bodies, where the presence of Breast. Cancer is suspected, further processing is carried out to predict the presence of breast cancer or not. There are several research methods proposed by many authors to address the problems of identifying and classifying the information of a data provided in UCI repository. However, there is a need for a method, which improves the accuracy of classifying the data items into various classes like benign or malignant. Thus, we have a new methodology to address the problem of identifying the data's of UCI Repository into various classes of identification.

The organization of entire research paper can be visualized in different sections like section 2 describes the related work, while section 3 presents the proposed Improvised mRMR, section 4 discusses the results and analyses the comparison existing methods with respect to the proposed method. Finally, section 5 concludes the research paper with few contributions.

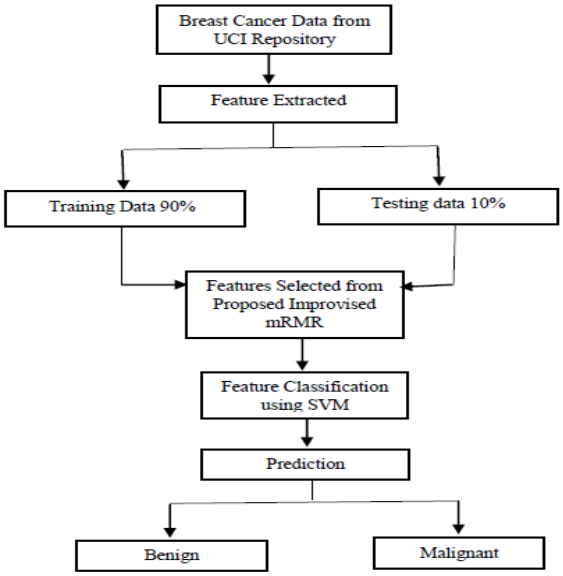

Fig.1. Architecture of proposed Improvised feature selection mRMR method.

II. RELATED WORK

Supervised feature selection approaches are those data which are labeled. Traditional supervised methods such as Fisher Score [1] [2] [3] [4] rank features individually according to the criterion, which can not consider the correlation among different features. Linear discriminant analysis (LDA for short) [5] [6] [7] [8] was proposed to elevate features by maximizing the ratio between the class scatter and within class scatter. Unfortunately, LDA suffers from the small sample size problem because it needs to calculate the inverse matrix of within class scatter, which is singular when the number of training samples is smaller than the dimensionality of the data [9] [10] [11] [12]. To avoid this problem, maximum margin criterion(MMC for short) based algorithm is proposed in [13] [14] [15] [16],

which uses a linear combination of traces between class scatter and within class scatter in the objective function and introduces a constraint of orthogonal weight matrix. However, all supervised methods have the common limitation of the requirement of sufficient labeled data, which is very expensive to obtain in practice. The performances of such supervised methods, however, usually drop dramatically when the labeled training data are Scarce Semi-supervised feature selections, by contrast, exploit not only labeled but also unlabeled training data. As a result, semi-supervised methods are able to select features by utilizing unlabeled data when there is limited number of labelled data. Among others, graph Laplacian based semi-supervised methods assumes that most data examples lie on a low dimensional manifold, such as semi-supervised Discriminant Analysis (SDA) [17] [18] [19] [20] [21] [22]. In graph Laplacian based methods, graph Laplacian matrix is introduced to harness the unlabeled samples. However, they are usually less efficient on handling large-scale data because of the time-consuming computation of the graph [23] [24] [25] [26]. Therefore, it is necessary and important to study unsupervised feature selection2.1. The paper should have the following structure

III. PROPOSED METHOD

The proposed method consists of various stages of data analysis, such as feature extraction, feature selection and feature classification. We have identified a methodology of statistical machine learning algorithm to analyze and understand certain relevant properties of attributes, which are most essential and uniquely distinguishes from other set of features. The feature extraction involves loading of data into attributes of a proposed system, as such it is most suitable and useful for further processing stages.

3.1. Feature Extraction

Either the feature extraction plays a very important role in machine learning algorithm, as the features extracted from images or data needs to be validated based on the features extracted from the data attributes. Further, the feature extraction involves representing each column of data items from UCI repository into different attributes like Mean radius, texture mean, perimeter mean, smoothness mean, compactness mean and various other features are represented by attributes. These attributes information is extracted into the individual variables of attributes to measure the accuracy of the statistical approach.

3.2. mRMR Feature Selection

The proposed method has improvised the feature selection algorithm mRMR by tuning certain set of features selected from a set of features obtained from features extracted. The feature selection approach plays a vital role in selecting the important features, which are not mutually exclusive, correlation, and similarity scores. The feature selection algorithm is mainly dependent on these three features like similarity scores, correlation, and mutually exclusive nature of the attributes.

The mutually exclusive property of the proposed method consists of tuning of parameters of variables like shown in eq. (1).

$$S(P,c) = \frac{1}{\sqrt{P}} \sum_{P_i \in 1}^{c} I(P_i, c) \qquad (1)$$

Sample S (P, c) represents the features P that belongs to class c and the class c indicates whether the selected features are same or different. P_i Indicates the features considered for validation of features, whether they exhibit the features of mutually exclusion.

$$P(i,j) = \sum_{i \neq j} P(i,j) \qquad (2)$$

It is clear from the above eq. (2) that mutually exclusive property is exhibited. The redundancies of all features are calculated as per eq. (3).

$$M(i,j) = \frac{1}{\sqrt[w]{P}} \sum_{w \in i,j=1}^{m,n} P(i_m, j_n) \qquad (3)$$

The existing feature selection algorithm does not explains how the weights are to be used along with the summated features, but the proposed improvised feature selection mRMR algorithm presents how the weights along with summated results will give boosting to the selection of relevant data from features extracted as per (4).

$$\frac{1}{\sqrt[w]{P}} \qquad (4)$$

Eq.(4), indicates the weights calculated from the summated features, when both indexes i and j are similar in nature, the weights are used along with the summated features extracted from the data. This weight makes a contribution of filtering redundant features from a set of features of a data and provides relevant features of data items.

The feature M (i, j) represents the redundant features represented by the weights of features $P(i_m, j_n)$. The weights w are calculated by applying the condition $w \in i,j = 1$.

$$f_p = \max_{p} \left[\frac{1}{\sqrt{P}} \sum_{P_i \in 1}^{c} I(P_i, c) - \frac{1}{\sqrt[w]{P}} \sum_{w \in i,j=1}^{m,n} P(i_m, j_n) \right] \qquad (5)$$

Where, f_p indicates the features selected from a relevant data represented by individual attributes. Thus, we have defined a variable f_p to find the most relevant correlated features of an attribute that exhibits the maximum correlation with minimum redundancy. The optimization of features selected from the features set is done with the help of a similarity measure exhibited by the features f_p.

$$F_s = \max_{c=1,0} [I(p,c) + f_p] \qquad (6)$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Where F_s is the optimized features selected from the features set selected from a feature extraction stage. Thus, the optimization of features is done by making the union of the features set $F_p$with the original class information.
Thereby, we shall see the optimized results of the proposed method. The results of the features are selected with more optimization. The feature selection is made by tuning the parameters of the attributes like gaining the weights of the equation, while calculating the features from a raw feature. Thus, we have been able to select the most relevant features and from a set of features extracted.

3.1.1 Improvised mRMR Feature Selection Algorithm

Input: The set of raw feature are given as input to the system.
Description: The purpose of the proposed Feature Selection Algorithm is to optimize the features selected from a raw data
Output: The optimized features.

Algorithm

Begin
Step.1 Select features from raw data by $S(P, c)$
Step. 1.1. Obtain the class information c
Step. 2 [Assign weights]
Step.2.1 Extract redundant features $M(i,j)$
Step.2.2 Assign weights w with M as per (3)
Step.2.3 Select features with maximum relevancy f_p
Step.2.4 Selected features f_p> number of samples S, go to step 2
Step.3 [Optimization]
Step.3.1 Determine F_s.
End

3.3. Features Classification

Classification is one of the prominent and important algorithms in machine learning, as it is useful in classifying the features of a data selected from feature selection algorithm. There are various feature classification algorithm such as naïve bayes, SVM and nearest neighbor. This entire algorithm has a certain set of advantages. SVM is the supervised machine learning classification algorithm.
The features selected from features selection algorithm is used as input to feature classifier SVM, which classifies the selected features into two different classes like Benign and Malignant. Benign indicates the early stages of the cancer and malignant indicates the later stages of the cancer. Thus, the proposed method uses 2 classes of classification. The classification of features are done by tuning certain percentage of features in the ratio of 90:10, where 90% of the features are considered for training and the rest 10% is considered for testing. The SVM Classifier has been tuned with certain set of parameters in the ratio of Training versus Testing in the ratio of 90:10, which has summarily yielded an accuracy of 98.3% over a UCI repository.

IV. RESULTS AND DISCUSSION

The proposed method has yielded an accuracy of 98.3%, as it has data items of different attributes of UCI repository; the SVM Classifier has performed the task of classifying the attributes of UCI repository. The UCI repository contains the information of attributes and its class labels. Thus, we have been able to predict the class of information accurately with up to 98.3%.The SVM Classifier has been tuned with certain set of parameters in the ratio of Training versus Testing in the ratio of 90:10, which has summarily yielded an accuracy of 98.3% over a UCI repository.

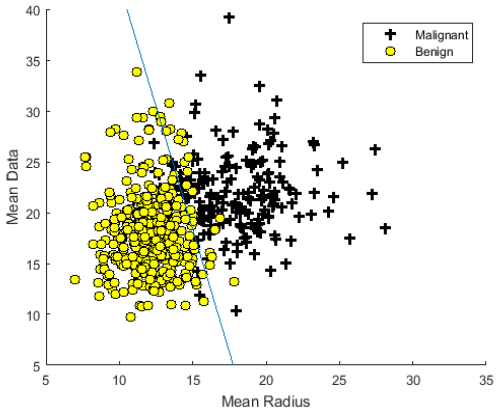

Fig 2. Result of SVM Classifier result of Classification of SVM has yielded an accuracy of determining the Benign and Malignant breast cancer tissues data obtained from the proposed method of statistical method.The UCI repository consists of attributes like Mean Radius, which has been considered along x-axis and the other parameter mean data attribute is represented along y-axis, and has been indicated along X-Y plane to show the accuracy of the proposed method. Further, the proposed method statistical approach has been used to exhibit the strength of the feature selection approach. The feature selection algorithm has provided relevant data, while minimizing the irrelevant data items from a list of attributes. Further, the black coloured samples indicate the features of malignant separated from the features of benign indicated in yellow colour. The above Fig.1. Represents the samples separated based on certain parameters of the proposed method; as such it clearly separates the two classes of features.

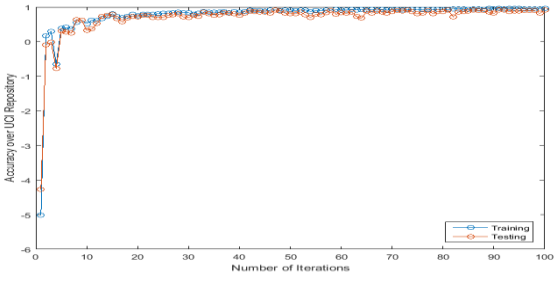

Fig 3. Result of Improvised mRMR feature selection algorithm.

The result of training versus Testing is indicated in Figure.3. It is clear from the above graphical representation that the testing of 10% of attributes and the 90% of training attributes has yielded accuracy from the above pink line and the blue line. The X-axis represents the number of iterations the proposed algorithm runs for the purpose of yielding the optimized results. Pink colored line indicates the percentage of data items or attributes considered for testing and the blue colored lines indicate the percentage of data attributes considered for training of data items using the improvised mRMR feature selection algorithm. The table.1 presents the results of existing methods with respect to the proposed method in numerical values.

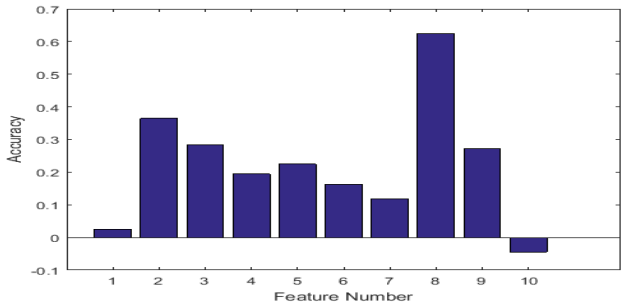

Fig 4. Ratio of Training versus Testing and its results of accuracy are shown above

Table 1. Comparison of proposed method with few existing methods

Methods	Results of Accuracy
Multiple classifier System	67.40%
Colour analysis using K-means clustering technique	71.30%
Thermal infrared image analysis	76.80%
Wavelet based thermogram analysis	84.90%
Improvised mRMR feature selection	**98.30%**

The table-1 shows the comparison of the classification of proposed improvised mRMR feature selection algorithm with respect to other approaches. Further, the graph shown in Figure.4 indicates that 90 % of attributes are considered for training of UCI data and the rest 10 % is considered for testing of data.The below figure.5 states that the percentage of training data considered is 90% of attributes in runtime with different possible set of attributes, the rest of the data attributes 10% of the data are considered for testing the data based on the trained data attributes to test the efficacy of the proposed method.

The results of the classification of proposed improvised mRMR feature selection algorithm with respect to other approaches has shown significant improvement than other existing methodologies.

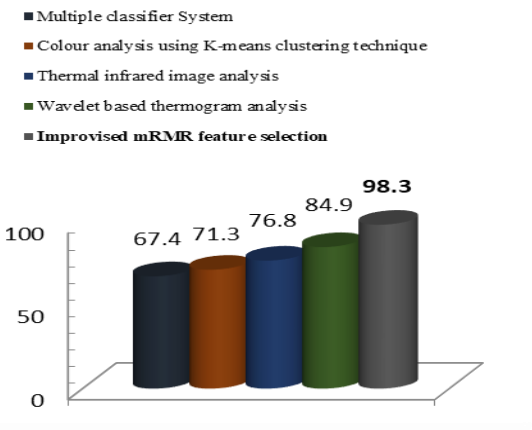

Fig 5. Comparison of proposed method with few existing methods.

Table 2. Comparison of proposed method with few methods of classification without feature selection, and with feature selection.

Methods	Results of Accuracy
SVM Without feature selection	95.50%
SVM + mRMR With feature selection	96.90%
SVM + improvised mRMR feature selection	**98.30%**

Classification of features without feature selection has yielded an accuracy of 95.5%, classification of features with mRMR feature selection has given an accuracy of 96.9%, while classification accuracy of SVM with improvised mRMR feature selection has produced an even better accuracy than the SVM with mRMR feature selection.
Thus, we state that any research algorithm with improvised mRMR yields good classification accuracy. Our proposed method on Wisconsin data has yielded an accuracy of 98.3%. The below Figure.6 indicates clearly that the proposed method has yielded good results over a data, which shall be extended to other research works.

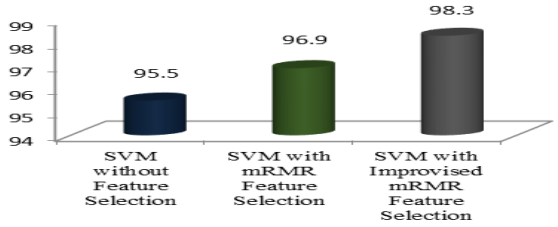

Fig. 6. Comparison of methods with feature selection, with improvised feature selection andwithout feature selection algorithm.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. CONCLUSION

The proposed research contributes to the subject of breast cancer identification by providing the improvised mRMR feature selection method, thereby, we have been able to select on few significant features, which are most essential and relevant is selected from a raw set of features. Further, the proposed method uses the classification by support vector machine, which is linear in nature, with the help of improvised mRMR feature selection algorithm and feature classification (SVM) method the data sets are classified as benign and malignant. The proposed method has produced an accuracy of 98.3%, which shows the significance of the proposed method with respect to classification.

REFERENCES

1. A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artificial intelligence 97 (1997) 245-271.
2. H. Liu, H. Motoda, Feature selection for knowledge discovery mining, Springer Science & Business Media, 2012. 3. I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of machine learning research 3 (2003) 1157-1182.
3. Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, H. Liu,
4. Advancing feature selection research, ASU feature selection repository (2010) 1-28. 5. P. Langley, Selection of relevant features in machine learning, in: Proceedings of the AAAI Fall symposium on relevance, 1994, pp. 245-271.
5. P. Langley, Elements of machine learning, Morgan Kaufmann, 1996.
6. J.L. Crowley, A.C. Parker, A representation for shape based on peaks and ridges in the difference of low pass transform, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1984) 156-170.
7. Z.L. Sun, D.S. Huang, Y.M. Cheun, Extracting nonlinear features for multispectral images by FCMC and KPCA, Digital Signal Processing 15 (2005) 331-346.
8. Z.L. Sun, D.S. Huang, Y.M. Cheung, J. Liu, G..B. Huang, Using FCMC, FVS, and PCA techniques for feature extraction of multispectral images, IEEE Geoscience and Remote Sensing Letters 2 (2005) 108-112
9. A. Khotanzad, Y.H. Hong, Rotation invariant image recognition using features selected via a systematic method, Pattern Recognition 23 (1990) 1089-1101.
10. N. Vasconcelos, Feature selection by maximum marginal diversity: optimality and implications for visual recognition, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, pp. 762-769.
11. N. Vasconcelos, M. Vasconcelos, Scalable discriminant feature selection for image retrieval and recognition, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2004.
12. J.Y. Choi, Y.M. Ro, K.N. Plataniotis, Boosting color feature selection for color face recognition, IEEE transactions on image processing 20 (2011) 1425-1434.
13. A. Goltsev, V. Gritsenko, Investigation of efficient features for image recognition by neural networks, Neural Networks 28 (2012) 15-23.
14. D.L. Swets, J.J. Weng, Efficient content-based image retrieval using automatic feature selection, in: Proceedings of International Symposium on Computer Vision, 1995.
15. D.L. Swets, J.J. Weng, Using discriminant eigenfeatures for image
16. retrieval, IEEE Transactions on pattern analysis and machine intelligence 18 (1996) 831-836.
17. E. Rashedi, H. Nezamabadi-Pour, S. Saryazdi, A simultaneous feature adaptation and feature selection method for content-based image retrieval systems, Knowledge-Based Systems 39 (2013) 85-94.
18. D.D. Lewis, Y. Yang, T.G. Rose, F. Li, Rcv1: A new benchmark collection for text categorization research, Journal of machine learning research 5 (2004) 361-397.
19. L.P. Jing, H.K. Huang, H.B. Shi, Improved feature selection approach TFIDF in text mining, in: Proceedings of International Conference on Machine Learning and Cybernetics, 2002, pp. 944-946.
20. S. Van Landeghem, T. Abeel, Y. Saeys, Y. Van de Peer, Discriminative and informative features for biomolecular text mining with ensemble feature selection, Bioinformatics 26 (2010) 554-560.
21. R. Khezri, R. Hosseini and M. Mazinan, a fuzzy rule-based expert system for the prognosis of the risk of development of the breast cancer , IJE TRANSACTIONS A: Basics Vol. 27, No. 10, (October 2014) 1557-1564.
22. A. Safari, R. Hosseini, M. Mazinani, A Novel Type-2 Adaptive Neuro Fuzzy Inference System Classifier for Modelling Uncertainty in Prediction of Air Pollution Disaster, IJE TRANSACTIONS B: Applications Vol. 30, No. 11, (November 2017) 1746-1751.
23. R. Khezri, R. Hosseini, M. Mazinani, A Fuzzy Rule-based Expert System for the Prognosis of the Risk of Development of the Breast Cancer, IJE transactions a: basics vol. 27, no. 10, (October 2014) 1557-1564.
24. H. Hamidi*, A. Daraei, Analysis of Pre-processing and Post-processing Methods and Using Data Mining to Diagnose Heart Diseases, IJE TRANSACTIONS A: Basics Vol. 29, No. 7, (July 2016) 921-930.
25. G. Stein, B. Chen, A.S. Wu, K.A. Hua, Decision tree classifier for
26. Network intrusion detection with GA-based feature selection, in: Proceedings of the 43rd ACM Southeast conference, 2005, pp. 136-141.
27. F. Amiri, M.R. Yousefi, C. Lucas, A. Shakery, N. Yazdani, Mutual information-based feature selection for intrusion detection systems, Journal of Network and Computer Applications 34 (2011) 1184-1199

AUTHORS PROFILE

Dr. S. Senthil, Professor and Director, School of Computer Application was awarded with Doctoral Degree by Bharathiar University for his dissertation on Lossless Preprocessing Algorithms for effective text compression. He has completed his B.Sc (Applied Sciences – Computer Technology) from P.S.G College of Technology, MCA from Bharathidasan University, M.Phil in Computer Science from Manonmaniam Sundaranar University and Ph.D in Computer Science from Bharathiar University. He has been qualified in State Eligibility Test conducted by Bharathiar University. At present he is guiding 8 Ph.D scholars in the fields of Data mining and Networks. He has 18 years of experience in teaching. His areas of interest include RDBMS, Data Mining, Data Compression, Computer Networks and Data Structures. He has published 30 research papers in various reputed National and International Journals. He has presented a paper entitled "Lossless Preprocessing Algorithms for better Compression" in an IEEE International Conference at Zhangjiajie, China. He was also the recipient of the best paper awards, at an International Conference on “Wisdom Based Computing” at Thiruvananthapuram and at a National Conference on “Transforming India through Digital Innovations” at Guru Shree Shantivijai Jain College for Women, Chennai.

Mrs. Deepa, Assistant Professor, holds MCA in Computer Applications from VTU and B.Sc. in Computer Science from Kuvempu University. She has 7 years of teaching experience. She has presented papers in National level conferences, published technical papers in International journals. She is pursuing Ph.D in Data Mining (Medical diagnosis).