https://dx.doi.org/10.22161/ijaers.5.5.11 *ISSN: 2349-6495(P) | 2456-1908(O)*

Improved Classification of Breast Cancer Data using Hybrid Techniques

R. Senkamalavalli[1], Dr. T. Bhuvaneswari[2]

[1]Research scholar, CSE Department, SCSVMV University, Enathur, Kanchipuram, India
[2]Asst. Professor, Queen Mary's College, Chennai, India

Abstract*— *Breast cancer is the second leading cancer for women in developed countries including India. Many new cancer detection and treatment approaches were developed. The most effective way to reduce breast cancer deaths is detect it earlier. The frequent occurrence of breast cancer and its serious consequences have attracted worldwide attention in recent years. Problems such as low rate of accuracy and poor self-adaptability still exist in traditional diagnosis. In order to solve these problems, an Ada Boost-SVM classification algorithm, Combined with k-means is proposed in this research for the early diagnosis of breast cancer. The effectiveness of the proposed methods are examined by calculating its accuracy, confusion matrix which give important clues to the physicians for early diagnosis of breast cancer.

Keywords*— *Kmeans, Support Vector Machine, Adaboost, Breast Cancer.

I. INTRODUCTION

In this paper we intend to present a system for diagnosis of breast cancer disease using data mining techniques. The symptoms of breast cancer include mass, changes in shape and dimension of breast. Various diagnostic tests and procedures are available for detecting the presence of breast cancer. Classification of breast cancer data is useful to identify the behavior of the tumor.Tumors can either be malignant or benign. Differentiating a malignant tumor from a benign one is a very Big task due to the structural similarities between the two. Support Vector Machine (SVM) is a classification algorithm used in various applications to classify data. But for big data and imbalanced datasets, it is not suitable to apply SVM, since it leads to computational problems and missing value scenarios. Hence it is highly important to make SVM suitable for the present scenario by modifying the algorithm to adapt to the expectations. In this method, both the training and the prediction of SVM classifiers are done using the cluster centers obtained from the k-means clustering. Misclassifications are treated equally for the entire cluster center. To enhance the accuracy of the classification, we have implemented ADABoost classifier algorithm. ADABoost helps in handling the misclassification of cluster centers using the data points in each cluster as a weight. This approach of ADABoost classifier with SVM can be implemented on imbalanced datasets as well[1] .The main objective of this research is to classify the breast cancer data with high efficient algorithms to obtain the results in a better manner.

II. LITERATURE REVIEW

M. Sewak et.al.,(2007) Support vectors with RBF, polynomial and linear kernel functions were trained using a fraction of the WDBC dataset as a training set. . The classification was then carried out using the majority opinion of the ensemble. This SVM ensemble process yielded 100 percent benign tumor prediction accuracy. [2].

B Zheng et al., (2013) proposed a model that is a hybrid of K-Means and Support Vector Machine. The model is implemented on breast cancer dataset to diagnose cancer based on the extracted features of tumor. Kmeans is used for finding the hidden pattern of tumor and SVM for classification of features. There are two types of tumors: malignant (are cancerous) and benign (can't be cancerous, can be removed). The classifier separates these two types of tumors. The k-means is used for clustering the patterns of the similar tumor based on the features of malignant and benign tumors. The membership function is used to measure the similarity of the data point and the tumor. The results show that the K-means and SVM hybrid model reduces the time required for prediction with higher rate of accuracy[3].

R.K.Kavitha et al.,(2014) Data mining techniques are used to obtain useful information from the large amounts of data which can help the physician for decision making regarding the prognosis. This paper studies the performance comparison of Adaboost algorithm which classifies data as linear combination and CART (Classification and regression trees) which classifies data by constructing decision tree in predicting the survivability of breast cancer patients.[5].

https://dx.doi.org/10.22161/ijaers.5.5.11 *ISSN: 2349-6495(P) | 2456-1908(O)*

Thongkam Jaree et.al.,(2008) ,proposed a method by combining the AdaBoost and random forests algorithms for constructing a breast cancer survivability prediction model. They used random forests as a weak learner of AdaBoost for selecting the high weight instances during the boosting process to improve accuracy, stability and to reduce overfitting problems. The hybrid method performance is evaluated using basic performance measurements (e.g., accuracy, sensitivity, and specificity), Receiver Operating Characteristic (ROC) curve and Area Under the receiver operating characteristic Curve (AUC). Experimental results indicate that the proposed method outperforms a single classifier and other combined classifiers for the breast cancer survivability prediction[7].

Sri bala et.al.,(2016) Machine learning provides better prediction methodologies for diseases in health care management. Ensemble learning is nothing but group of classifiers which in reality yielding better results rather than the existing results. To produce the better results we use collection of classifiers called ensembles. They have implemented ensemble methods to improve the better prediction for breast cancer to classify the breast tissue as in the form of carcinoma and fibroadinoma .Along with existing classifiers like J48Naive Bayes, random forest and SMO. We implemented ensemble classifiers like Adaboosting, bagging and stacking or blending methods with them, in reality it is showing better accuracies[10].

III. PROPOSED SYSTEM

The proposed method is designed with SVM and k-means clustering called as the KM-SVM. KM-SVM is a fast algorithm to increase the processing speed of training and the prediction of SVM classifiers using the cluster centers received from the k-means clustering. The misclassifications are treated equally in each cluster center. To enhance the accuracy of the proposed method, we introduce the ADABoost classifier algorithm which handles the misclassification cluster centers by assigning penalties. The SVM method along with ADABoost can be applied on imbalanced datasets as well.

The extracted correlation features are placed in ascending order for the given data and also given in the form of SVM classifier. The misclassified data obtained from the first level of classification is samples using N sample method and then sent to the classifier again for an accurate classification

The preprocessing done with k-means algorithm by finding the cluster centers. The Benign and Malignant tumors are again checked with svm classifier in order to overcome the misclassification. Boosting is done at the end so that all the output weak learners are clubbed to form a strong learner. Boosting concentrates more on the misclassified examples or to the examples that have higher prediction errors.

A THE BASIC K-MEANS ALGORITHM

The k-means algorithm is a simple process where K initial centroid is selected. The value of K is specified by the user as the number of clusters required. Now points are selected close to the centroid and these points are the clusters of the centroid. The centroid in each cluster will be updated with the points assigned to the cluster. This process continues until no point changes in the cluster and the centroid remains the same.

B SVM

SVM- Support Vector Machines was first proposed by Vladimir Vapnik. It's a new learning method proposed for binary classification. The main objective of this algorithm is to find a hyper plane which separates the D-Dimensional data into two perfect classes. Later, SVM was introduced for kernel induced feature space that considers higher dimensional space where the data can be classified. So it's a challenge to classify data which is possible to be present in two classes of data.

https://dx.doi.org/10.22161/ijaers.5.5.11 *ISSN: 2349-6495(P) | 2456-1908(O)*

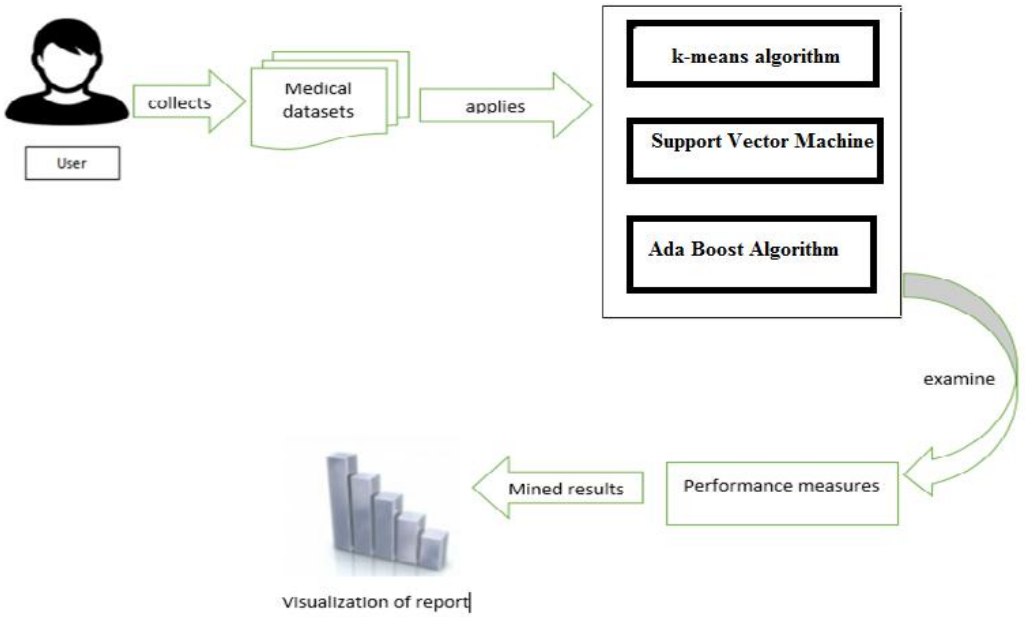

Fig.1: Proposed Architecture

C ADABOOST Classifier

Boosting is the concept of converting a weak learner to a strong learner. It is the process of combining all weak learners to form a single strong rule. Each time when the base learning algorithm is applied it generates weak prediction rules through an iteration process. After conducting several iterations, the boosting algorithm combines all weak rules to form a single strong prediction rule. Below are the steps used for choosing the right distribution:

Step 1: The base learner is applied to distribute and assign equal weight to each observation.

Step 2: If any prediction error is observed then a higher attention is paid for observations having error. Now, the next base learning algorithm is applied.

Step 3: Step 2 is repeated until higher accuracy is achieved by the base learning algorithm.

At the end, all the output weak learners are clubbed to form a strong learner. Boosting concentrates more on the misclassified examples or to the examples that have higher prediction errors.

IV. RESULT AND DISCUSSION

The basic phenomenon used to classify the Wisconsin diagnosis breast cancer data using matlab and compare the accuracy obtained using this technique with other techniques. The below table shows the accuracy comparison. The kmeans, correlation svm and adaboost combined technique used in this research yields higher accuracy when compared to other techniques.

V. CONCLUSION

The proposed novel algorithm was experimented on the Breast cancer database. The simulation results proved that the approach achieved a very high accuracy rate than the existing methods used in literature. We also demonstrated a certain level of accuracy in the classifier, and for finding accurate results there must be sufficient preprocessing of data done. Missing data, data imbalance and other peculiar cases are to be considered in order to derive an accurate result. Finally we also demonstrated that we can attain accuracy in diagnosing breast cancer disease using the K-means classifier,adaboost and Support Vector Machines. It is being applied to classify images into two sectors as with tumor and without tumor. New cases will be analyzed in the future studies.

https://dx.doi.org/10.22161/ijaers.5.5.11 *ISSN: 2349-6495(P) | 2456-1908(O)*

Table.4.1: Comparison of the methods and accuracy of previous studies and this study.(WDBC data sets)

Author (year)	Method	Accuracy (%)
J.Abonyi et al., (J.Abonyi and Szeifert, 2003)	supervised fuzzy clustering	95.57
I.Gadaras et al., (I. Gadaras, 2009)	Fuzzy rule classification	96.08
Chunekar et al., (Chunekar, 2009)	Jordan Elman neural network	98.25
M.Darzi et al., (Mohammad Darzi, 2011)	CBRGenetic	97.37
L.Chuang et al., (Li-Yeh Chuang, 2011)	CatfishBPSO	98.17
Mert et al., (A. Mert and Akan, 2014)	LOO, PNN	97.01
Zheng et al., (B. Zheng and Lam, 2014)	K-SVM	97.83
Subrata Kumar Mandal(2015)	Logistic Regression 97.90	97.9
Priyanka Jain et al.,(2016)	Svm	80
This study	Kmeans+modified svm+adaboost	98.85

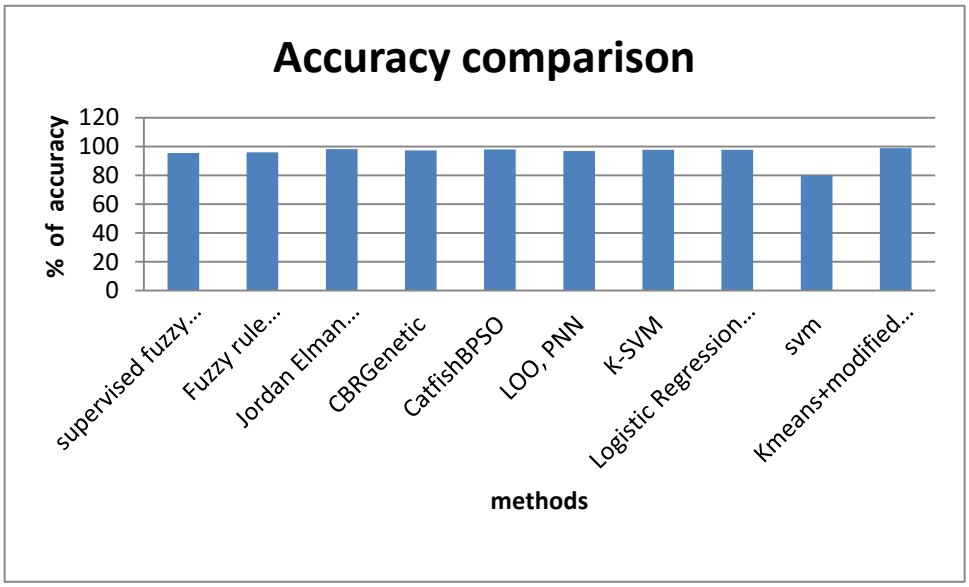

Fig.4.2: Accuracy comparison graph of Wisconsin breast cancer dataset

REFERENCES

[1] Suna.Y,(2007) 'Cost-sensitive boosting for classification of imbalanced data', sci2s.ugr.e

[2] Sewak.M , et.al (2007),' International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)', Iowa City, IA, pp. 32-37.

[3] Zheng. B, et.al (2014). ' Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms', Expert Systems with Applications, 41(4), 1476-1482.

[4] Guanjin Wang ,et.al(2016), 'Tackling missing data in community health studies using additive LS-SVM classifier', IEEE Journal of Biomedical and Health Informatics , Issue: 99.

[5] Kavitha R.K., et.al(2014), 'Breast Cancer Survivability Predictor Using Adaboost and CART Algorithm', International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 1,pp351-353.

[6] GhadaSaad , et.al(2016), 'ANN and Adaboost application for automatic detection of microcalcifications in breast cancer', The Egyptian Journal of Radiology and Nuclear Medicine, Volume 47, Issue 4, PP 1803-1814.

https://dx.doi.org/10.22161/ijaers.5.5.11 *ISSN: 2349-6495(P) | 2456-1908(O)*

[7] ThongkamJaree, et.al(2008). 'Boost Algorithm with Random Forests for Predicting Breast Cancer Survivability', Proceedings of the International Joint Conference on Neural Networks, PP 3062 - 3069.

[8] Delen.D, et.al(2005) , 'Predicting breast cancer survivability: a comparison of three data mining methods', journal -Artificial Intelligence in Medicine, vol. 34, pp. 113-127.

[9] Jaedong Lee , et.al(2014) , 'K-means clustering based SVM ensemble methods for imbalanced data problem', 2014 Joint 7th International Conference on Advanced Intelligent Systems (ISIS) .

[10] Sreebala et.al(2016), 'Efficient Ensemble Classifiers for Prediction of Breast Cancer', IJARCSSE 6.

[11] Soltani Sarvestani A , et.al(2010) ,'Predicting breast cancer survivability using data mining techniques', Software Technology and Engineering (ICSTE), 2nd International Conference , Volume: 2.

[12] Vimal Kumar Dubey et.al(2016), ' Hybrid Classification Model of Correlation-based Feature Selection and Support Vector Machine', IEEE International Conference on Current Trends in Advanced Computing (ICCTAC) .

[13] Sumaiya Farzana G et.al(2016)' Domain independent model for data prediction and visualization', International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 5, Issue 4.

[14] Bedy Purnama et.al(2015), 'A classification of polycystic Ovary Syndrome based on follicle detection of ultrasound images', 3rd International Conference on Information and Communication Technology (ICoICT).

[15] Hsin-Ta Li et.al(2016), 'Lower-limb motion classification for hemi paretic patients through IMU and EMG signal processing', 2016 International Conference on Biomedical Engineering (BME-HUST) .