Proc. IEEE Conference on Emerging Devices and Smart Systems *(ICEDSS 2018)*
2-3 March 2018, Mahendra Engineering College, Tamilnadu, India.

A Survey on Breast Cancer Prediction Using Data Mining Techniques

[1]Dona Sara Jacob, [2]Rakhi Viswan, [3]V Manju, [4] L PadmaSuresh, [5]Shine Raj
[1,2,3]PG Scholar, Department of Computer Science and Engineering
[4]Principal, Baselios Mathews II college of Engineering, APJ Abdul Kalam Technological University
[5]Asst.Professor, Department of Computer Science and Engineering
Baselios Mathews II College of Engineering, University of Kerala
Lake View, Muthupilakadu, Sasthamcotta
Kollam, Kerala, India

Abstract - In today's world Breast cancer is one of the main problems faced by women. Identifying cancer is the first stage and is always challenging. Detection and nursing of the breast cancer have become an urgent. Breast cancer, is generally seen tumor in Indian women. Early treatments of breast cancer have become an extremely crucial work to do, not only helps to cure cancer but also helps in curative of its incidence. Today, there are different kinds of methods and data mining techniques and diverse process like knowledge discovery are developed for anticipating breast cancer. As per the survey, we perform a comparison of diverse classification and clustering algorithms. Varied classification algorithms and the clustering algorithm are used. The outcome indicates that the classification algorithms are superior predictors than the clustering algorithms.

Keywords –WEKA; SVM; C5.0; Breast Cancer; PAM; EM; Hyper plane.

I. INTRODUCTION

Now-a-days breast cancer is common among women. It should be predicted at its earlier stages such that proper treatment cures it. For this there is an urgent need for anticipating breast cancer by various data mining methods. Thus, the outcome of data mining algorithms enables us to provide better treatment.

Managing breast cancer among elder women is still challenging because of many reasons, those regarding treatment tolerance, disfigurement in data obtained from elder women, many undergo unsuitable treatment that do not fit the patient's health status. Psychological problems like depression, anxiety, poor self image and use of unhealthy mimicking approaches affect quality of life of patient. Even though proper treatments are available to cure breast cancer, pain and suffering relating available treatment modalities is notable. Chronic, persistent pain acts as an additional stressor for a person already distressed from many psychological, social and medical stressors.

Earlier challenges faced were the prediction of cancer at its earlier stages. As a solution to those challenges differing classifications as well as clustering algorithms were used. These algorithms clearly predict the presence of cancerous cell even in elder women. Now the existing challenge is performance; it took more time in its prediction.

Breast cancer begins with the abnormal growth of some breast cells. These cells divide more rapidly and continue to accumulate than healthy cells do, forming a lump or mass. These cells may grow through your breast to your lymph nodes or to other parts of your body. Breast cancer varies on the basis of age groups, it is less common at a young age (i.e., in their thirties), younger women lean to have more aggressive breast cancer than older women.

II. METHODOLOGY AND ALGORITHMS

A. C5.0

C5.0 is an extended version of c4.5 and ID3 algorithms. It makes use of various splitting algorithms which include entropy based information gain. The sample is split into subsamples until the low-level splits are obtained; splitting is based on various attributes. The low-level samples are examined, and those that do not contribute significantly to the value are removed. This algorithm is quite robust even when the data are missing or contains a large number of inputs. It is simpler than any other models that it is easier to understand. It offers powerful methods to increase accuracy of classification.

B. K-Nearest Neighbor

K-nearest neighbors also called instance- based learning, is a non-parametric learning in which it memorizes the observations to classify the unseen test data. This model is called a lazy model as it does not learn anything during the training phase instead learns only during the testing phase. It compares the test observations with the nearest training observations. It is not much efficient as it takes more time for its computation.

978-1-5386-3479-0/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: ZheJiang Sci-Tech University. Downloaded on May 21,2020 at 08:50:23 UTC from IEEE Xplore. Restrictions apply.

C. NaiveBayes

One of the most effective statistical and probabilistic classification algorithms, Naive Bayes. It predicts class membership probabilities based on whose values we determine to which class a sample belongs. It is carried out in the assumption that the impact of an attribute value on a class does not dependent of other attribute values.

D. Support Vector Machine

SVM model is a machine learning technique which is based on the statistical learning theory. It provides a better classification which generates a more complex boundary between classes. For separating classes it introduces margins on either side of the hyper plane. By maximizing margins, we obtain the largest distances between the hyper plane and samples. In linearly separable classes, SVM finds optimal hyper plane by dividing these classes. Once the hyper plane is found out, we find the support vectors which lie on its margin.

E. K-means

The k- means algorithm is one among the algorithms of partitioning methods. It is the best squared error-based clustering algorithm. , to solve a clustering problem with k clusters the method proceeds as follows: Begin with one cluster and cluster center corresponds to the centroid of dataset. Objects are assigned to clusters based on the similarity of the objects with those in the clusters.

F. Expectation Maximization

An efficient method for estimation from incomplete data is the EM algorithm. In any incomplete dataset, there is indirect evidence about the likely values of the unobserved values. This when combined with some assumptions, generates a predictive probability distribution for missing values. These are averaged in the statistical analysis. This is a common method for matching models to incomplete data.

G. Partition AroundMedoid

Partition around medoid, algorithm is used to find the sequence of objects called medoids that are centrally located in clusters. The algorithm aimed at minimizing the average dissimilarity of objects to their closest selected object. Likewise we can minimize the dissimilarity between the object and closest selected object. The algorithm works in two phases.

In the first phase an initial set of k objects are BUILD.

The second phase exchanges selected object with unselected objects to improve the quality of clustering.

H. Fuzzy C-Mean

This method allows a piece of data belong to one or more clusters .This is commonly used in pattern recognition .Each data point is assigned a membership on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center.

III. LITERATURE SURVEY

Chintan Shah et al [1], three different data mining classification methods are used for the prediction of breast cancer. It considers different parameters for prediction of cancer. But for superior prediction, focus is on accuracy and lowest computing time. Studies filtered all algorithms based on lowest computing time and accuracy and it came up with the conclusion that Naïve Bayes is a superior algorithm compared to decision tree and k-nearest neighbor, because it takes lowest time i.e. 0.02 seconds and at the same time is providing highest accuracy.

Uma Ojha et al [2], WPBC dataset is used for finding an efficient predictor algorithm to predict the recurring or non-recurring nature of disease. This helps Oncologists to differentiate a good prognosis (non- recurrent) from a bad one (recurrent) and can treat the patients more effectively. Eight popular data mining methods have been used, four from clustering algorithms (Kmeans ,EM, PAM and Fuzzy c-means) and four from classification algorithms (SVM, C5.0, KNN and Naive Bayes).The results of these algorithms are clearly outlined in this paper with necessary results. The classification algorithms, C5.0 and SVM have shown 81% accuracy in classifying there occurrence of the disease. This is found to be best among all. On the other hand, EM was found to be the most promising clustering algorithm with the accuracy of 68%. The research shows that the classification algorithms are better predictor than clustering algorithms. The impact factors of various parameters responsible for predicting the occurrence/non-occurrence of the disease can be verified clinically. Further, the identified critical parameters should be verified by applying on larger medical dataset to predict the recurrence of the disease in future.

RunjieShenYuanyuan Yang Fengfeng Shao et al [3], intend to build a model which searches the relationship between breast cancer and its significance. To improve the diagnostic model, a feature selection method INTERACT is used which selects the most important features. Models with and without feature selection are built, among which the model with feature selection provided better accuracy than the other model.

Ahmed Iqbal Pritomet al [4] , Used attribute selection method for improving any classification algorithms. Attribute with its least contribution to dataset often misguides the classification which results in poor prediction. They found SVM giving better performance. After feature selection method, decision tree and Naive Bayes shows better performance under the ROC curve.

S.Padmapriya et al [5], presented a survey of classification simulations which can be used for breast cancer detection using WEKA tool. A discussion on a variety of classification techniques that already exist in real world and the performance accuracy is listed from that. By using that we can decide which

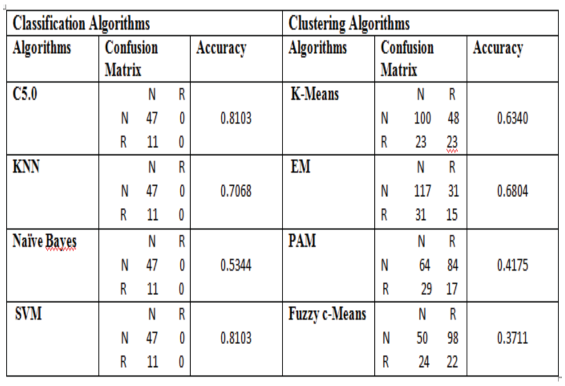
978-1-5386-3479-0/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: ZheJiang Sci-Tech University. Downloaded on May 21,2020 at 08:50:23 UTC from IEEE Xplore. Restrictions apply.

algorithm is best for the WEKA tool for breast cancer detection. It compares different algorithms and found SVM is better having high accuracy and expectation maximization with the least accuracy.

Jahanvi Joshi et al [6] , paper presented a survey of classification simulations which can be used for breast cancer detection using WEKA tool. A variety of classification techniques that already exist in real world are discussed. By using that we can decide which algorithm is best for the WEKA tool for breast cancer detection.

TABLE I. COMPARISON TABLE



Classification Algorithms			Clustering Algorithms		
Algorithms	Confusion Matrix	Accuracy	Algorithms	Confusion Matrix	Accuracy
C5.0	N R N 47 0 R 11 0	0.8103	K-Means	N R N 100 48 R 23 23	0.6340
KNN	N R N 47 0 R 11 0	0.7068	EM	N R N 117 31 R 31 15	0.6804
Naïve Bayes	N R N 47 0 R 11 0	0.5344	PAM	N R N 64 84 R 29 17	0.4175
SVM	N R N 47 0 R 11 0	0.8103	Fuzzy c-Means	N R N 50 98 R 24 22	0.3711

Accuracy= (TP+TN)/(TP+TN+FP+FN)

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

Algorithms	**Advantages**	**Disadvantages**
C5.0	Does not require domain knowledge. Support huge dimensional data	Categorical output. One output attribute
K-NN	Implementation easy for parallel implementation work with local info.	Requires large storage area. May be slow in classifying tuples
Naïve Bayes	High accuracy speed and for large data set .	Conditional independence in class for rules
SVM	Maximize the margin between two classes in the feature space characterized by a kernel function.	Difficult to incorporate background knowledge .It is sensitive to outliers
k-means	Ease of implementation and high speed performance. Measurable and efficient in large data collection.	Selection of optimal number of clusters is difficult. Selection of the initial centroids is random
Expectation Maximization	Ease of implementation and high speed performance. Measurable and efficient in large data collection	Selection of optimal number of clusters is difficult. Selection of the initial centroids is random
PAM	PAM is more robust in the presence of noise and outliers.	PAM works efficiently for small data set but does not scale well for large data set.
Fuzzy c-means	Allows a data point to be in multiple clusters. It is unsupervised and always.	Long computational time. Sensitivity to noise. Need to determine membership cutoff value converges

CONCLUSION

From the above comparisons, it is concluded that the classification algorithms works better than the clustering algorithms in predicting breast cancer. Andin the classification algorithms the SVM and C5.0 came up with better performance. The best algorithm for predicting breast cancer is purely based on the accuracy of the algorithm.

REFERENCE

[1] [1] Chintan Shah; Anjali G. Jivani “Comparison of data mining classification algorithms for breast cancer prediction”,pp.1- 4,2013.

[2] [2] Uma Ojha; Savita Goel “A study on prediction of breast cancer recurrence using data mining techniques” 7th International Conference on Cloud Computing, Data Science& Engineering - Confluence,pp.527- 530,2017

[3] [3] RunjieShenYuanyuan Yang Fengfeng Shao “Intelligent Breast Cancer Prediction Model Using Data Mining Techniques” , vol.1,pp.384-387,2014

[4] [4]Ahmed Iqbal Pritom; Md.Ahadur Rahman Munshi; Shahed Anzarus Sabab; Shihabuzzaman Shihab.“Predicting breast cancer recurrence using effective classification and feature selection technique”,pp.310-340,2016.

[5] [5] S. Padmapriya, M. Devika, V. Meena, S.B. Dheebikaa& R. Vinodhini, “ Survey on Breast Cancer Detection Using Weka Tool”, imperial journal of inter disciplinary research, vol-2, issue-4, 2016

[6] [6] Jahanvi Joshi, RinalDoshi, Jigar Patel, Ph.D,”Diagonosis of Breast Cancer using Clustering Data Mining Approach” International Journal of Computer Applications,vol-101,no. 10,2014.

Authorized licensed use limited to: ZheJiang Sci-Tech University. Downloaded on May 21,2020 at 08:50:23 UTC from IEEE Xplore. Restrictions apply.