Diagnosis for Early Stage of Breast Cancer using Outlier Detection Algorithm Combined with Classification Technique

M. Priya, M. Karthikeyan

Abstract: ***Breast cancer is the most dangerous cancers that lead to women in death. Particularly in the developed countries it takes second leading place that increase the chance of death in women. It can be not easily diagnosed by the lab. It has difficult to identifying at the beginning stage. This cancer begins from breast and disseminate to other body parts. It has cured easily if it is identified at beginning stage. The correct classification of benign cancer can prevent from superfluous treatment for patients. This paper focused on diagnosis early stage of the breast cancer based on data mining algorithms. The automatic diagnosis process plays on important role in data mining. The proposed method has a process of three stages. First, data objects are grouped into clusters using k-means clustering algorithm. Size of the dataset has to shrink gently the computation time also reduced. The second stage, the outlier detection (OD) algorithm has used to detect the outliers from the cancer dataset. Finally, diagnose the cancer is either benign or malignant using decision tree classification algorithm. The breast cancer dataset has been used to test the efficiency of the proposed method. The experiments were conducted in breast cancer dataset before and after removal of outliers. Comparison results prove that the proposed method as serves as the better one with high accuracy. This breast cancer research will help with a medical practitioner to diagnose the breast cancer and so that it helps to recover the patients.***

Keywords: ***Accuracy, Breast Cancer, Classification Algorithm, Clustering Algorithm, Data Mining, Outlier Detection.***

I. INTRODUCTION

Breast Cancer is the dangerous which affects women compared to all types of cancers. It has a malignant tumor which is formed by cause of the cells abnormal growth in the breast. The aim of the predictions has to assign the patients into the two classes either a “benign” or a “malignant”. In data mining applications, Prediction of a disease is the most challenging and interesting task. Many tools are used for huge volumes of health data are collected and ready for available in the medical research. Data mining has a unique approach, used to discover new and hidden patterns of knowledge from the large databases. It has the process of finding new patterns from huge data sets that related to the techniques from statistical, machine learning and also in big data [1]. Detection and diagnosis of cancer at early stage has essential to rescue the human life. In 2012, the survey report of the data in 1.7 million women has affected by cancer. Nowadays the medical practitioners are unable to identify at early stage that impacts the death rate [2]. Many different aspects are used for diagnosing breast Cancer as such as doctor’s opinion, a mammography and the study of pathology. Early stage of diagnosis has requirement for an accurate and the reliable method which can be used by practitioner to make a feature of benign breast cancer from the malignant without taking of surgical biopsy [3]. Several researchers have focused in different diagnosis methods for the breast cancer. Breast cancer has a most common disease that the women affect at rural and urban areas in India. Many machine learning techniques are used to diagnosis either the cancer cell is benign or malignant. Benign means non-cancerous or non-life menacing for humans, but malignant has a cancerous that leads the patient in relation to death. Even though improving the technology however 20% of women died every year in the world.

Outlier detection is important and necessary in data mining which is used to preprocessing the dataset that is to improve the accuracy of diagnosis result. In large databases some of the Data objects are not perfect. These data objects are known as outliers which is different from normal data. Hence, these data objects should be removed from normal process in data mining. Accordingly, the outlier detection is also called anomaly detection is necessary to identify the outliers and it is to improve the data objects quality and to obtain correct result of data mining. According to Hawkins, An outlier is “a data objects that deflects so much from other data objects as to provoke suspicions that it was generated by a different mechanism” [4]. Therefore Outliers are probably created due to reading measurement or execution error, etc., and nevertheless it has been considered generally as noise.

In this research paper, the outliers are detected using the outlier detection (OD) algorithm for breast cancer data set. After outlier detection, the preprocessed dataset is classified whether it is benign or malignant cancer using decision tree classification algorithm. Classification includes that data objects are classified to pre-defined class or group. For diagnosing of breast cancer dataset, the decision tree classification is used to check whether output class is benign or malignant with more accuracy when compared to the existing research.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author

M. Priya*, Assistant Professor, Department of Computer Science, PSPT MGR Government Arts & Science College, Sirkali – Puthur (Tamilnadu) India. E-mail: mpriyaau@gmail.com

M. Karthikeyan, Assistant Professor, Division of Computer and Information Science, Faculty of Science, Annamalai University, Annamalai Nagar (Tamilnadu) India. E-mail: karthiaucse@gmail.com

The rest of the paper is organized as follows. Section 2 presented in study of a related work to the breast cancer diagnosis detection. Section 3 described the proposed method of this research such as the outlier detection and classification algorithm. Section 4 represents analysis of experimental results and Finally Section 5 elaborated conclusion and future scope.

II. RELATED WORKS

Priya and Karthikeyan have proposed method that is used to identify data objects as outlier and outlier clusters in a dataset. This algorithm is based on mutual nearest neighbor graph clustering. Using this algorithm the outlier value factor can be found in the database and the outliers and outlier clusters are detected efficiently [5]. Priya and Karthikeyan focus the diagnosis of diabetes using classification techniques. In this work is to diagnostic the chance of diabetes in patients with correct classification of Diabetes. Three classification algorithms can be used to detect diabetes at an early stage [6].

Chen et al. has proposed SVM classifier based on a rough set and this method has achieved a high accuracy of classification in training and test partition [7]. Priya et al. has discussed on outlier detection in health care datasets based on to comparing different clustering algorithms [8]. Priya and Karthikeyan have focused to analyze the performance of outlier detection algorithm using feature bagging technique for health care application. Ensemble classifiers can be improving the overall performance and stability of data mining techniques effectively [9]. Nisihi et al. proposed a new method fuzzy logic combined with knowledge based system for breast cancer data set. Using Principle Component Analysis for data reduction and fuzzy rules CART is used for generation of knowledge based system. They have introduced a reasoning method based on fuzzy rule for classification [10].

Azar et al. has discussed classification algorithms with WBC data set. They have obtained that the PNN algorithm shows better result than MLP [11]. Fatima et al. discussed the Data Mining techniques has to gain its strength because of maintaining the capability of a huge amount of data has combined from many different sources and the background information are integrated [12].

III. MATERIALS AND METHODS

The k-means clustering algorithm has introduced for grouping the dataset, because to shrink the dataset size for reduce the computation time. Then apply outlier Detection algorithm for detect the outliers. Finally the decision tree classifier used to classifying the data into either benign or malignant.

A. Clustering Algorithm

Clustering is a process of grouping the data objects that is getting from input and an output of number of output clusters that are obtained. Hence the data objects in a similar group are one clusters but the dissimilar data objects outside the clusters. So the outside data objects are called outliers. An outlier is defined as a value which does not belongs to any other cluster. Therefore, Clustering algorithm is also used for detection of outliers. In this work, the k-means clustering algorithm is used to partitioning the dataset. The k-means algorithm steps are outlined as follows,

Input: Let D as a Data set, n as a number of observations in dataset D and K as a cluster number.

Output: Clusters K
Step 1: Select k points at randomly that is the initial centroid of cluster.
Step 2: The distance is calculate for each point in data set D with the every cluster center and that is assigned to the nearest cluster.
Step 3: Recalculate the mean for each cluster.
Step 4: Repeated the step 2 and 3 until the cluster centroids has not change.

B. Outlier Detection Algorithm

The outlier detection algorithm is used for preprocessing the original breast cancer data set due to the data object in most of large databases are not perfect because of noise or outliers. The proposed outlier detection algorithm becomes better quality of data and to obtain accurate results. First data objects are grouped into clusters using k-means algorithm, then find the centroid of each cluster. The points Pruned by which the distance from the centroid is less than that the centroid of the clusters respectively. After pruning that for each unpruned data objects in every cluster, we have use LOF (Local Outlier Factor) that tells how far a data objects are deviating from its neighbors. The LOF value is high that point indicates as an outlier and most probably that the data point is deviating faraway from its neighbors. Outlier detection algorithm is described as below,

Input: Clustered data set.
Output: Outlier data Objects
Step 1: Identify the outlier objects, which are not in the clusters and find the centroid of each cluster.
Step 2: Pruning process for each cluster: The points Pruned by which the distance from the centroid is less than that the centroid of the clusters.
Step 3: Detection of outlier: find LOF for all the data points which are unpruned to left in the clusters. If the LOF is larger than the threshold then it will considered as outlier.

C. Decision Tree Classifier

Decision Trees has a tree structure, where every node defined as a test on an attribute and the branch node represents a test outcome. It has a predictive technique for analyzing the output class variable from a dataset on different given input attributes. In this connection, the tree has constructed from the training dataset. It has been pruned heuristically and to prevent overfitting the data, that tended to present a classifying error on the testing data. It takes the place of the post-pruning method that removes the branch nodes from a tree. The pruning algorithm has been calculated the error rate for every branch node in the tree if the non-leaf node has been pruned. If the error rate is minimal to the pruning node then the branch node has kept, otherwise is has pruned. The basic steps for the Decision Trees algorithm as follows,

Input: The data set after detecting outliers.
Output: The output class variable either benign or malignant
Step1: Check the data objects belongs to the root node of the tree that denotes a leaf then the leaf is labeling with the same node.
Step 2: Calculate the gain information for each attribute, if result from a specified by a test on every attribute.
Step 3: Based on selection criterion the best attribute has selected for branching, and then it will be labeled with an appropriate class.

D. Proposed Method

This is the combined techniques such as clustering, outlier detection and classification for diagnosing early stage of breast cancer. The k-means clustering algorithm has used to grouping the dataset, because to shrink the dataset size. Then apply Outlier Detection algorithm for detect the outliers. After detection of outlier, then the data objects has been classified as either benign or malignant class using the decision tree classifier. The proposed method is described as below.

Input: The Breast Cancer data set.

Output: Class variable either Benign or Malignant cancer with better accuracy
Step 1: Using k-means clustering algorithm for Pre-processing the data
Step 2: Applying the Outlier Detection algorithm for removing the outliers.
Step 3: Next, the outlier detected dataset has been uploaded in WEKA tool for analysis.
Step 4: The classification algorithm has applied in two data sets.
Step 5: The data object has classified as either benign or malignant cancer.

IV. RESULTS AND DISCUSSIONS

A. Environment

Several tools are developed for Data analysis. WEKA tool is used in this work, for diagnosing early stage of breast cancer and testing the performance. University of Waikato in New Zealand has designed the Waikato Environment for Knowledge Analysis (WEKA) tool. This tool consists of a collection of data mining tasks especially for preprocessing, clustering, classification, feature selection and visualization.

B. Dataset Description

Breast Cancer dataset has been used for recording the measurements of breast cancer cases. The data set contains the Dimensions of 214 data sample medical records (objects). Each record contains 10 attributes which are considered as risk factors for the occurrence of cancer. There are two classes labeled as 0 and 1, to diagnosis of cancerous and non-cancerous. The data set attributes except the class variable represent bio-physical features of the cyst biopsy and they are represented in integer values. The Detailed statistics of BC dataset has shown in Table- I. It has 214 data objects, 10 attributes and 2 classes, Non-cancerous (Benign) or Cancerous (Malignant) which of 123 records in Non-cancerous and 91 records in cancerous. Apply k-means algorithm for group the dataset then outlier detection algorithm has used to detect the outliers. In the total of 214 objects in this dataset, of which 20 objects are outlier objects. The 20 outliers are removed from the data set, and then remaining 194 data objects are formed a new data set called BC without outliers. It contains 194 data objects, of which 115 records in Non-cancerous and 79 records in cancerous. Table- II shown the detailed statistics of BC dataset without outliers.

Table- I. Detailed statistics for BC dataset

Attribute	Value	Mean	Std. Dev.
Clump Thickness	1 – 10	4.589	3.032
Uniformity of Cell Size	1 – 10	3.355	3.062
Uniformity of Shape	1 – 10	3.472	3.036
Marginal Adhesion	1 – 10	2.897	2.895
Single Epithelial Cell Size	1 – 10	3.701	2.606
Bare Nuclei	1 – 10	3.85	3.702
Bland Chromatin	1 – 10	3.776	2.1
Normal Nucleoli	1 – 10	3.238	3.248
Mitoses	1 – 10	1.888	2.062
Class	2 - 4	Non-cancerous (benign) 123	Cancerous (malignant) 91

Table- II Detailed Statistics for BC dataset without Outliers.

Attribute	Value	Mean	Std. Dev.
Clump Thickness	1 – 10	4.428	2.999
Uniformity of Cell Size	1 – 10	3.17	2.983
Uniformity of Shape	1 – 10	3.299	2.947
Marginal Adhesion	1 – 10	2.722	2.763
Single Epithelial Cell Size	1 – 10	3.521	2.444
Bare Nuclei	1 – 10	3.608	3.585
Bland Chromatin	1 – 10	3.639	2.045
Normal Nucleoli	1 – 10	3.139	3.21
Mitoses	1 – 10	1.887	2.103
Class	2 - 4	Non-cancerous (Benign) 115	Cancerous (Malignant) 79

C. Performance Evaluation

In the stability of the proposed method, the performance is measured and evaluated on datasets. Using the evaluation measures like precision, recall, accuracy, F-Measure, and kappa statistic, for classifying the results. Then the measures are described as in Table- III.

Table- III: Evaluation Measures

Measures	Definitions	Formula
Accuracy (A)	Determines the correctly classifying objects accuracy	$A = \frac{TP + TN}{\text{Total no. of samples}}$
Precision (P)	Determine exactness of the Classifier	$\text{Precision}(P) = \frac{TP}{TP+FP}$
Recall (R)	To measure the completeness of the classifiers or sensitivity	$\text{Recall}(R) = \frac{TP}{TP+FN}$
F-Measure	F-Measure can be harmonic mean of precision and recall	$F_{Measure} = 2 * \left[\frac{(Precision * Recall)}{Precision + Recall}\right]$

TP - True Positive, FP - False Positive, FN - False Negative, TN - True Negative

Kappa Statistic: It is another evaluation measure which is difference between an observed agreement and expected agreement. The value ranges from 0 to 1, and it produce a value 1 for perfect agreement. Equation 1 is used to compute kappa statistic value.

$$K = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \qquad (1)$$

where, p_o is the observed related agreement, and p_e is probability chance of agreement.

D. Result Analysis

In the experimental analysis, 10-fold cross validation has used. The data set has randomly partitioned in to 10 equal sizes. Out of all the partition, one has used in the testing of data and remaining nine has training of data. This technique is repeated k times, until each partition has taken as testing of data at least once. Results are acquired from the repetitions and they are averaged or merged to given an individual estimation. This method helps to increasing the accuracy. The experimental results are listed in Table- IV.

Table- IV: Performance measures

Breast Cancer dataset	Kappa statistic	Precision	Recall	F-Measure	Accuracy %
With outlier	0.962	0.982	0.981	0.981	98.13
Without outlier	0.9787	0.990	0.990	0.990	99.01

From the Table- IV, the proposed method yields a classification accuracy of 98.13% for BC data set and 99.01% for BC dataset without outliers. Hence combination of the proposed model has been used as an optimistic and consistent method for helping medical practitioner for making process of diagnosis for the breast cancer as either cancerous or non-cancerous. Fig. 1 shows the graphical representation of performance measures for the dataset.

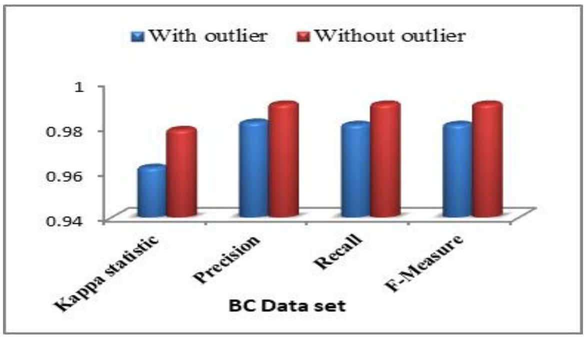

Fig. 1. (a) Performance measures for the datasets

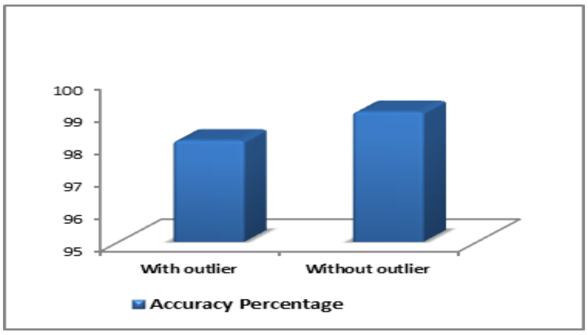

Fig. 2. (b) Accuracy Percentage for the datasets

V. CONCLUSION

The diagnosis for early stage of breast cancer is necessary in challenges to the medical field. Incorporated data mining algorithms are obtained better accuracy. Thereby this paper proposed a combined method of outlier detection algorithm with classification model. Throughout this paper, three techniques such as clustering, outlier detection and classification algorithms have been studied and the performance has evaluated using various measures. This method obtains an accuracy of 98.13% for BC dataset with outlier and 99.01% of accuracy for BC dataset without outlier based on the 10-fold cross validation technique. As the result determines combined Meta-algorithm has achieved high accuracy compared with other models. So, the overall conclusion observed that the combined algorithms have found to be more efficient for the diagnosis of breast cancer when compared to that of other methods. Furthermore, the perfect classification has been done by incorporating the high-risk level and optimal data mining technique to disseminate for the patients treatment from the common people.

REFERENCES

1. Ubeyli, E. D., "Implementing automated diagnostic systems for breast cancer detection", *Expert Systems with Applications*, vol. 33, pp. 1054–1062, 2007.
2. www.breastcancerindia.net, Statistics of Breast Cancer in India.
3. Akay, M. F, "Support vector machines combined with feature selection for breast diagnosis", *Expert Systems with Applications*, vol. 36, pp. 3240–3247, 2009.
4. Hawkins. D.M, *Identification of Outliers*, Chapman and Hall, London, 1980.
5. Priya. M and M. Karthikeyan, "An Efficient Cluster Based Outlier Detection Algorithm", *Journal of Engineering and Applied Sciences*, vol. 14, no. 23, PP. 8699-8704, 2019.

6. Priya. M and M. Karthikeyan, “Data Mining Technique for Diabetes Diagnosis using Classification Algorithms”, *International Journal of Recent Technology and Engineering*, vol. 8, no. 4, PP. 9044- 9049, 2019.
7. Chen, H.-L., Yang, B., Liu, J., & Liu, D.-Y., “A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis”, *Expert Systems with Applications*, vol. 38, pp. 9014–9022, 2011.
8. M. Priya and M. Karthikeyan, “A Comparative Study Of Clustering Algorithms For Outlier Identification”, *International Journal for Research in Engineering Application & Management*, vol. 04, no. 07, pp. 391-396, 2018.
9. Priya. M and M. Karthikeyan, “Performance Evaluation of Ensemble Method Based Outlier Detection Algorithm”, *International Journal of Research in Advent Technology*, vol. 7, no. 3, pp. 1376 – 1380, 2019.
10. Nayak, T., Dash, T., Rao, D. C., & Sahu, P. K., “Evolutionary neural networks versus adaptive resonance theory net for breast cancer diagnosis”, *In Proceedings of the International Conference on Informatics and Analytics*, Article no. 97, ACM, August, 2016.
11. Azar, A. T., & El-Said, S. A., “Performance analysis of support vector machines classifiers in breast cancer mammography recognition”, *Neural Computing and Applications*, vol. 24, no. 5, pp. 1163–1177, 2014.
12. Fatima, M., Pasha, M., “Survey of Machine Learning Algorithms for Disease Diagnostic”, *Journal of Intelligent Learning Systems and Applications*, vol. 09, pp. 1–16, 2017.

AUTHORS PROFILE

M. Priya, is currently working in as an Assistant Professor in the Department of Computer Science at PSPT MGR Government Arts & Science College, Sirkali – Puthur -609108. She has received her Bachelor’s and Master Degree in Computer Science from Bharathidhasan University, Tiruchirappalli and M.phil., from Annamalai University. She is presently pursuing Ph.D in Computer Science in Annamalai University and area in the research of Data Mining. Her research interest includes Data Mining and Big Data analytics. She is published more than 5 research articles published in UGC and Scopus indexed journal. She has been participated more than 20 National and International seminars / conference / workshops and also presented more than 16 papers for various National and International level conferences.

Dr. M. Karthikeyan, working as Assistant professor in Division of Computer and Information Science, Annamalai University, India. He completed his M.Sc. [Computer Science] from Bharathiar University and M.Phil [Computer Science] and Ph.D from Annamalai University in 2005 and 2014 respectively. He is having 19 years of teaching experience. His area of interest is Data Mining, Digital Image Processing, and Artificial Neural Networks. He has published more than 25 research papers in various reputed journals and conferences. And also more than 25 papers presented in various international and national level seminars and conferences. Under his guidance more than 10 students completed the M.Phil. Degree and is presently 8 scholars pursuing Ph.D. under his guidance.