Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 6 July 2021: 5443 – 5453

Analysis Of Breast Cancer Data Using Machine Learning Techniques

Dr.E.Venkatesan[1], Dr.M.Bennet Rajesh[2], M.Rajendiran[3]

ABSTRACT

Now a days breast cancer in the world is one of the issuessince this disease occurs second place in the developed country.In medical field providing proper awareness, but the ratio of outreach the people understand the cancer awareness is less. The world health organizationreleased an article that breast cancer is one of the challenging problems and most of women affected by breast cancer sometimes one percentage of possibility for men also. The main aim of the study to analyze breast cancer data based on its characteristics and identify the effectiveness of clustering and classification instructions for analyzing breast cancer data. Breast cancer related images, different characteristics, including numerical data and attributes are used in this research work.Patients intake routine, age, lifestyles, occupation, details of diseases that cause problems are taken and then carry out for machine learning based classification and clustering algorithms.And eventually clustering algorithms like k-Means and Expectation-Maximization (EM), and classification algorithms such as J48, Classification and Regression Trees (CART) and Support Vector Machine (SVM) are trained   as well as tested.The overall performance of clustering and classifications are based on sensitivity, specificity, accuracy, error rate and process time.

Keywords:Breast cancer,Clustering, Classification Expectation Maximization and CART.

I. INTRODUCTION

Data mining techniques are used to gain new, concealed and helpful information from data. The numerous mining functions are association rules, classification, prediction and clustering. The process of identifying new forms in heavy data sets to find useful forms of statistics and artificial intelligence and also database control. The superiority of Information Technology has headed up to thetremendous data gathering in many domains, it covers business, communication, technology and disease detection in recent years.Useful facts and intelligence are usually concealed in the from all such cells.Processing vast information and recuperate valid information from it is a tough task. DM is a marvelous tool to grasp this task. The word DM, also familiar as Knowledge Discovery in Databases (KDD) mention to the relevant extrication of tacit, earlier unidentified and useful information from data in a huge warehouse [1,16].

Breast cancer occurs both men and women, but only 1% possible for men. Breast carcinoma is the uttermost common disease in women, in which abnormal rise in the breast mass.A contemporary review in united kingdom proven that breast cancer is not just a trouble for young women, but also for problem of aged woman who have past the time of sixty or seventy.Early

[1]Guest Lecturer, PG Department of Computer Science, RV Government Artscollege, Chengalpattu, India
[2]Assistant professor, Department of Computer Science,Kamarajar Govt.Arts College,Surandai,India
[3]Assistant professor, Department of Computer Science, Goverment Arts College, Chidambaram, India
E-Mail:[1]venkatelumalai12@yahoo.co.in, [2]benraj@gmail.com, [3]rajendranmaha@gmail.com

detection and prevention of breast malignant neoplasm tumor with appropriate treatment can save human life [2,3,4,5,16]. Cancer is a wider spread disease in India. Statistically, the proportion number of cancer patients in India is increased. People, lifestyle modification, food habits and usage of tobacco are the main reason to create cancer tumor.

The tumor is an extraordinary cell development, as it may be benign or malignant. Benign tumors are harmless while malignant tumors can cause cancer and spread in any area in the body [7,8,9,10,11,12,13].Patients can be treated early, using data mining technology to understand the nature of the tumor. Due to various types of cancer, including breast cancer and more problems arise in detecting that,but in data mining technology easy to finding cancer cell.Malignant neoplastic tumors can occur in one or both breasts, it can be life threatening.The inside of the female breast contain predominantly of fatty and fibrous connective tissues. [14,15].

The main objective of this research is to accurately diagnose and discover the algorithm performance of breast cancer datasets using clustering and classification technologies in data mining. This research report has been coordinated with these little initiation, part one the construction of the study is planned as pursue, Part two analyzes the techniques and methods and Section three expansion results clustering and categorization. Lastly, section four concludes the research work through the potency of its methodologies [19].

2. MATERIALS AND METHODS

Most researchers in medical data analysis, uses two main data mining techniques like clustering and classification.Clustering and classification algorithms have suggested for many scholars in the field of business and other department of fraud detection and scientific discovery, because this algorithm provides accurate results. This research work refers to the segregation based clustering algorithm, that is k-Means and Expectation-Maximization (EM) algorithm then using classification algorithms J48, Classification and Regression Trees (CART) and Support Vector Machine (SVM).

This operation time in apiece method detects and examine are compared to overlapping for best result.Nowadays, data mining techniques are very useful in the medical field for diagnosing diseases and many researchers coming up with new techniques for this every year.The main purpose of this research method is to consciously perform comparable performance in cluster and classification algorithms based on sensitivity, profiling, accuracy and error rate.In this study have three phases, phase one preprocessing the breast cancer data set, phase two inputs the data into the algorithms to measure the parameters like sensitivity, specificity, accuracy and rate and phase three finding performance.

2.1. Details of Data Set

In this study used three types of breast cancer data sets such as normal, benign and malignant and also used fourteen attributes. The breast cancer dataset was collected from a private Diagnostic Centre and hospital in Tamil Nadu, totally 200 breast cancer data set wasusedfor analysis, they are 100 benign and 100 malignant. Fourteen attributes are shown in the below table.

Table 1. Details of the breast cancerData Set

S.No	Variables	Details

1	Age	Age in year
2	Sex	Male and Female
4	Blood cell counts	WBC,RBC,hemoglobin and platelets
5	Blood Pressure	mm Hg on admission to the hospital
8	Blood chemistries	Various organs are healthy and functioning properly during treatment
9	Lifestyles	Living environment
10	Occupation	Working environment
11	Patients food habits	Vegetable and non vegetable
12	Monitoring disease risk	Treatment process
13	Positive	Malignant and Benign
14	Negative	Normals

2.1 Classification Algorithms

Classification in Data Mining maps the most important work order into predefined goals. The aim to create a classification depend on test cases with attributes to interpret things to assign a set of substance.Then the section will compute a group of attributes based on the values and create a decision tree in which each non-leaf node represents a trial or conclusion.So the scanning for the classification process from the sow node and travels till it reaches a leaf node. A result will be made on reaching the end node [2].

2.2 J48 method

J48 works to develop the Quinlans C4.5 algorithm an organized decision tree. Each feature should be divided into smaller pieces based at one end. J48 it receives the default information, which is actually an effect of the data by selecting an attribute. Attribute to decision-making most standardized information uses the resource. Small subgroups are provided the algorithm. Events that stop when the rift is a subgroup be held by the identical class.J48 creates a result node using one of the significance seen in the class. J48 decision tree can handle specification properties, lose or missed attribute value data and attribute costs. The clarity can be increased by shear [2].

2.2.1 The method step

Step 1: If the incident is in identical group, the leaf is labeled the same class.

Step 2: The potential communication to every attribute is calculated and the gain of the details is carryout from the tests on the attribute.

Step 3: The special impute is elected based on the present selection parameters.

2.3Classification and Regression Tree (CART)

In 1984, Leo Breiman, Jerome Friedman, Richard Olsen and Charles Stone are together evolves a general method for generating Classiication and Regression Tree (CART) from simple data to statistical models.CART is strong because it has data that is completely unfinished and has estimated and input features.There will be rules that human beings can read about the tree that the CART has created. A set of sample questions about data features this process will guide to data reduction and carry on till few stop measure is reached.CART efficient of running everything like numerical and type parameters.The Gini code standards how a given parameters divide testing models as the target class. This is where the binary splitting properties occurs.This is the uttermost broadly applied analytical procedures. This gives an authority of unequal binary conclusion.

2.3.1 The Method step

Step1: The initial is how possible to separate the attributes.
Step2: The second, determining what stopping order must be.
Step3: Finally, how nodes are allocated to classes.

2.4 Support Vector Machines

Support Vector Machine (SVM) idea made basically founded decision planes are describing the end limits.The idea of judgment is one that divided between a set of objects that have characteristics of truth of class members. Uses a collection of data, including the regular SVM and predicts it two possible classes comprises the input. Example of SVM representations in examples as points in spaced spaces, examples of separated types are divided by a clear spacing as wide as possible.[17].

2.5 Clustering Algorithm

The clustering substance depends on parameter the correlation among the set of objects using the for away activity.Therefore the outcome of clustering is a pair of clusters, where the material inside a one cluster resembles each other, rather than the cluster resistance. [21]. Cluster analysis, medical image section, information, analysis and image processing, including many widely used in applications.Clustering is as well as called data segmentation applications because clustering divided as groups of huge data according to their harmony, Clustering is particularly the uttermost general unpracticed data mining mode to un identified framework installed in a dataset. Clustering is the process to convert a crew of conceptual things into classes of alike substance. A cluster of data substance can be considered as a group. [20].

2.5.1 Expectation-Maximization algorithm

EM algorithm handles data mining very efficiently by taking every clustering aspects like data substance. This clearly enables features such as multiplicity, average, and sub-clustering double-clause statistics. It explains more accurately up to the sub-cluster of data element and is smaller sensitive to the practice of summarizing data. The created clusters are very near to the primary ones.General EM method is a general action plan for increase the probability. That's an expansion probability evaluate can be gained.EM method of unfinished data analysis thus very useful technique, one divergent example of which is cluster examine if the class index is a remark as lost values.Its fundamental concept is in connection with the given defective data trouble, an entire-data trouble for which the topmost probability measure is computationally changed.

$$\wedge\,(p,\lambda) = -\sum_{i-1p_i}^{n} log_2\, p_i - \lambda(\sum_{i=1}^{n} p_i - 1) \qquad (1)$$

Where, p is an open set of attributes subject to constraints[18].

2. 6 The k-Means method

The k-Means method, is used to clear up k-Means clustering problem. Initial step in this method is to choose the sum of clusters. It is mandatory that the sum of clusters should match the data. Wrong choice in the number of clusters does not go through the entire action. In general,the efficiency of clustering can be identified compared with the performance of another clustering, and the performance of the k-means clustering should be compared with the performance of other clustering, so that k-means performance can be identified.Then the center of the clusters should be started. Each data point must be the closest cluster and each clustering position is a pair of average to all data points inclusion to the cluster.This action should be frequent until merge.If there are $n$data points $xi, i = 1 \ldots n$ that have to be partitioned into k clusters. The goal is to assign a cluster to each data point. K-Means is a clustering method that aims to find the positions $\mu i, i = 1 \ldots . k$ of the clusters that minimize the distance from the data points to the cluster.

Implicit objective function in k-Means measures sum of the distances of the observations from their cluster centroids, called Within-Cluster-Sum-of-Squares (WCSS). This is computed as

$$k = \sum_{j=1}^{k} \quad \sum_{i=1}^{n} || \mathrm{x}_i{}^{(j)-c_{(j)}} ||^2 \qquad (2)$$

Where $||x_i{}^{(j)} - c_{(j)}||^2$is a chosen distance measure between a data point$x_{(i)}{}^{(j)}$ and the cluster center $c_{j,}$ is an indicator of the distance of the n data points from their respective cluster centers. The algorithm is composed of the following steps:

Step 1: Keep k dot in the gap described by the substances that are being clustered. These points describe start crew centroids.

Step 2: Allocate every substance to the group that has the close centroid.

Step 3: While all substance has an allocated, reevaluate the attitude of k centroids.

Step 4: Continue steps 2 and 3 until the centroids no longer act. This process a division of the objects into groups from which the metric to be decreased can be calculated. The k-means is an easy clustering method that has been enlarged to more difficulty domains [6,18].

3. EXPERIMENTAL RESULTS CLUSTERING AND CLASSIFICATION

The experimental results describes the accuracy results of classification and clustering algorithms which are shown in Table 2 to Table 8 and figure 1 to figure 8 describes the performance measures of various algorithms.Here clustering techniques like K-means provided 99% than EM 90.5% and classification algorithm like SVM provided better results 90.4523% than other classifiers like J48 and CART.

Amidst the preference of classification methods, the performance of SVM is excellent than the other methods best for the particular data and also assures that quality of the classification methods.Among these, two clustering algorithms such as k-Means and Expectation-Maximization (EM), the K- Means arebetter algorithms based on the parameters as accuracy analysis clustering. Overall, this Research work carried out by classification algorithm performance compared to clustering techniques, k-Means algorithms best for breast cancer data.

Table 2. Statistics Measures of J48 Algorithm

CCI	88.9447 %
ICC	11.0553 %
KS	0.7786
MBE	0.1516
RMSE	0.3132
RAE	30.3406 %
RRSE	62.6496 %

The table 2 and figure 1shows different statistic measures and shows that correctly classified event is more than the misclassified event.

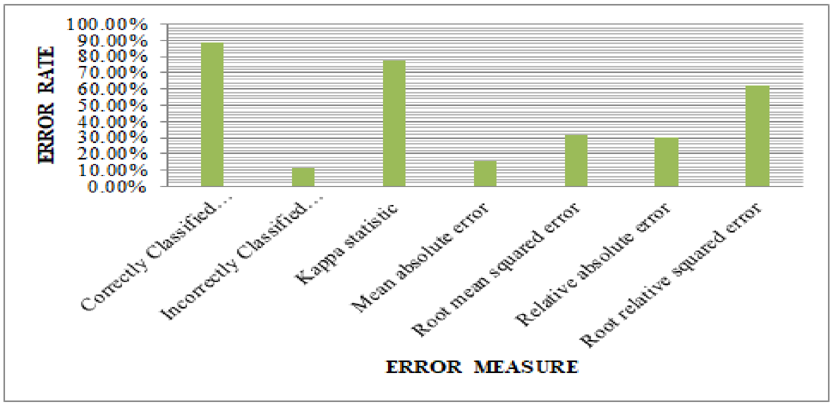

Figure 1. Statistical Measures of J48 Algorith

Table 3. Error Measures of J48 Algorithm

Class	**True Positive**	**False Positive**	**Precision**	**Recall**	**F-Measure**	**MCC**	**ROC Area**	**PRC Area**
Malignant tumor	0.902	0.124	0.885	0.902	0.893	0.779	0.912	0.865
Benign tumor	0.876	0.098	0.895	0.876	0.885	0.779	0.912	0.920
Weighted Average	0.889	0.111	0.890	0.889	0.889	0.779	0.912	0.892

Table 3 and figure 2shows a variation error measure available, butcompares both error measures in weighted average false positive rateislesser than true positive rate. Specifically, in benign and malignant tumor the true positive rate is high.

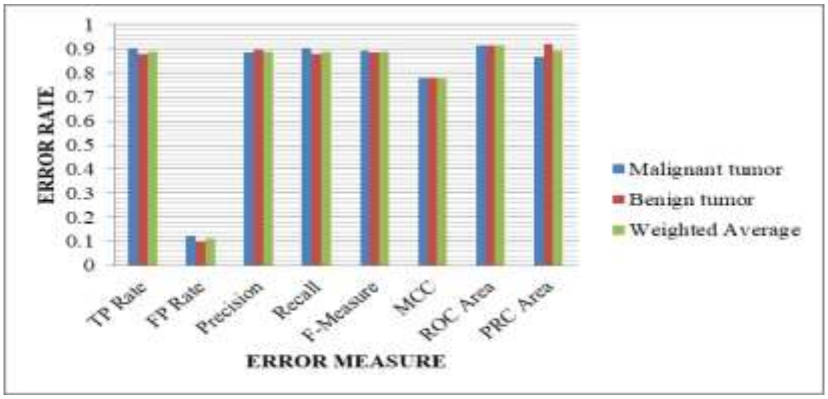

Figure 2. Error Measures of J48 Algorithm

Table 4. Statistical Measures of CART Algorithm

CCI	75.3769 %
ICC	24.6231 %
KS	0.5016
MBE	0.2807
RMSE	0.3983
RAE	56.1671 %
RRSE	79.6607 %

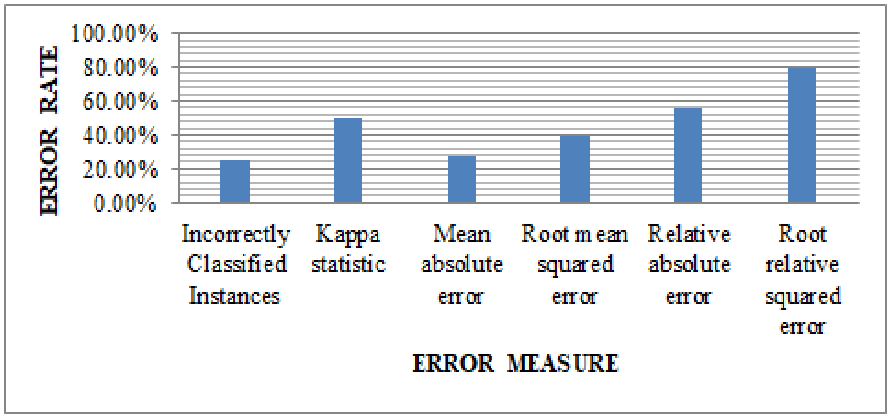

Figure 3. Statistical Measures of CART Algorithm

Table 4 and figure 3 shows the calculation of truly classified instances and falsely classified instances, in both measures correctly classified instances is higher.

Table 5. Error Measures of CART Algorithm

Class	True Positive	False Positive	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Malignant tumor	0.980	0.485	0.680	0.980	0.803	0.564	0.838	0.844
Benign tumor	0.515	0.020	0.962	0.515	0.671	0.564	0.838	0.853
Weighted Average	0.754	0.258	0.817	0.754	0.739	0.564	0.838	0.849

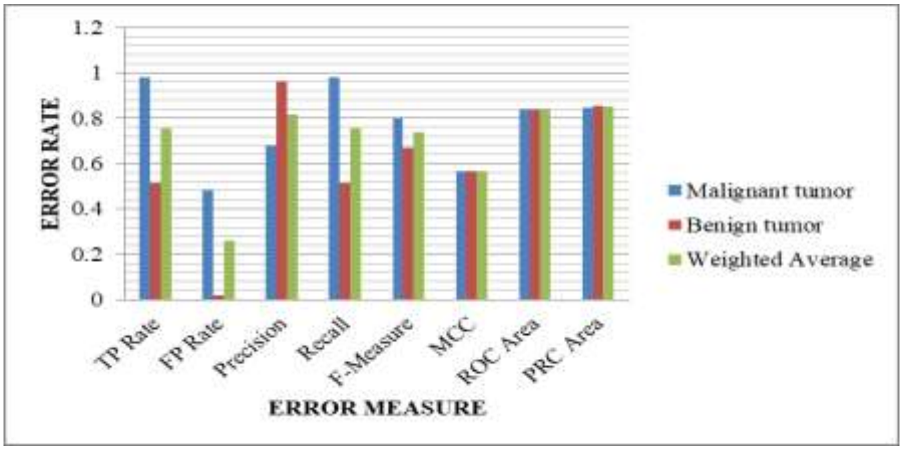

Figure 4. Error Measures of CART Algorithm

Table 6. StatisticsMeasures of SVM Algorithm

CCI	90.4523 %
ICC	9.5477 %
KS	0.8093
MBE	0.0955
RMSE	0.309

RAE	19.104 %
RRSE	61.8065 %

The table 6 and figure 4shows in SVM classification algorithm, correctly classified instances is better than compare to table 2 j48 and table 4 CART classificationalgorithmstatistic measures. So the SVM methodis best for classification.

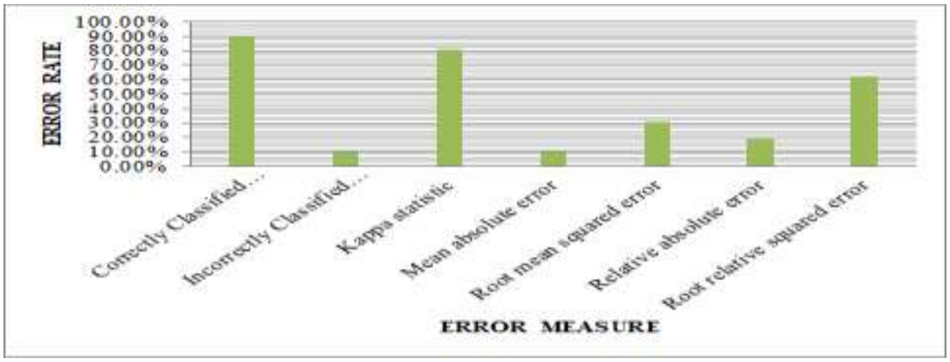

Figure 5: Statistics Measures of SVM Algorithm

Table 6. Error Measures of SVM Algorithm

Class	True Positive	False Positive	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Malignant tumor	0.873	0.062	0.937	0.873	0.904	0.811	0.905	0.883
Benign tumor	0.938	0.127	0.875	0.938	0.905	0.811	0.905	0.851
Weighted Average	0.905	0.094	0.907	0.905	0.904	0.811	0.905	0.867

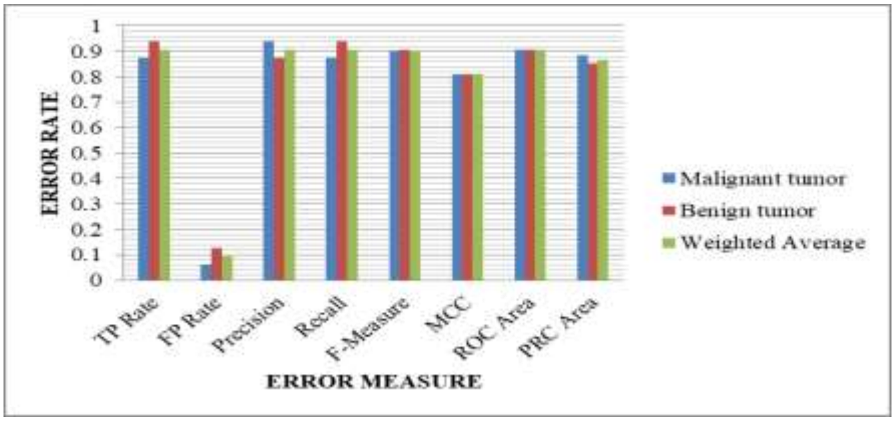

Figure 6. Error Measures of CART Algorithm

Table 8. Result of Clusteringand classification algorithms forbreast cancer Data Set

S.No	Clustering Algorithm	Accuracy %	Classifications algorithm	Accuracy %
			J48	88.9447 %
1	K-Means	99%	CART	75.3769 %
2	EM	90.5%	SVM	90.4523 %

Table 8 shows the results of accuracy for both clustering and classification algorithms.Here k Means provided better results than other clustering techniques and SVM provided better results than other classification techniques.The figure 7shows the accuracy comparison chart of two clustering algorithm's. Herek-Means provided better results than EM algorithm.

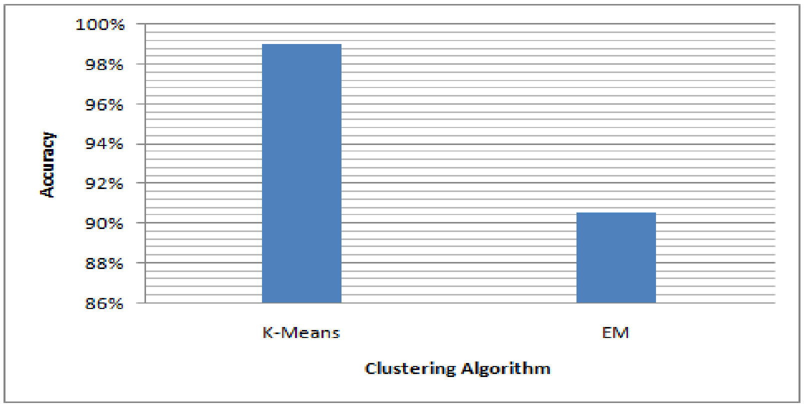

Figure 7. Comparison of breast cancer dataset using Clustering Algorithms

The figure 7 shows the results of clustering techniques, hereK-Means provided better results than EM algorithm..

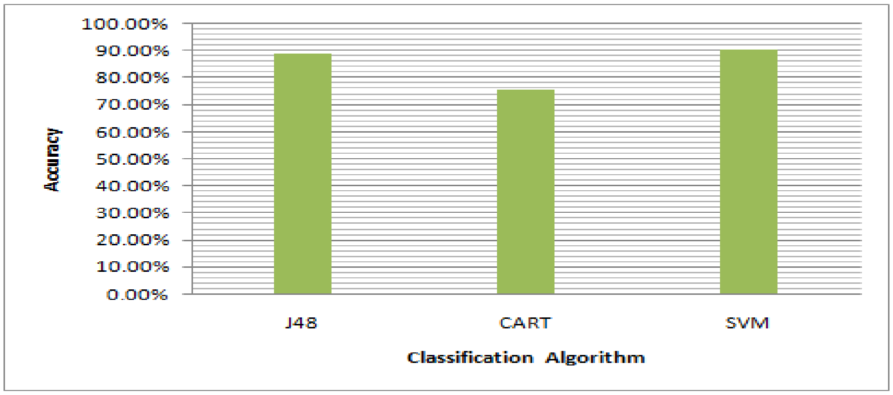

Figure 8. Comparison chart of breast cancer dataset Classification Algorithms

The figure 8 shows the accuracy of classification algorithms. Here SVM provided better results than J48 and CART.

4. CONCLUSION

Nobody says a specific technique is the excellent method for prognosis using a real-world data that is usually set up for certain functions. But, this probably refers to the efficacy of the selected data. The concept is based on this performance of the two partitioned clustering algorithms, namely k-Means and Expectation-Maximization (EM) algorithms. The decision was examined by a lot executions of the programs.Generally, the run time difference from one activity to other applications, it depends on the speed and the kind of the system. Partition based methods worked well for discovery results in many domains, including the medical field.machine techniques based Clustering and classification data mining can be used to obtain some information.It basically uses clustering and classification to classify each object in a set of data. In this research work, clustering and classification algorithms were methods have been studied and their attribute values are based on the dataset of breast cancer. The results show that the comparison of breast cancer data with the highest accuracy of the mean is used at this time and the K-means method performs betterfrom all machine learning classification and clustering methods.

REFERENCES

1. Siva Priya and Ashok Ku.E., "HMPFIM-B: Hybrid Markov Penalized FCM in Mammograms for Breast Cancer", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 2 (10), pp. 3033-3037, 2014.
2. Venkatesan.E and Velmurugan.T.," Performance Analysis of Decision Tree Algorithms for Breast Cancer Classiication", Indian Journal of Science and Technology, Vol 8 (29), pp 1-8, 2015.
3. Venkatesan. E,." Performance Analysis of classification Algorithms using Clinical Dataset", Journal of Information and Computational Science, Vol.9 (9), pp. 395- 401, 2019.
4. Venkatesan.E and Velmurugan.T.," Performance Analysis of Decision Tree Algorithms for Breast Cancer Classiication", Indian Journal of Science and Technology, Vol 9(30), pp 1-10, 2016.
5. Venkatesan. E and Velmurugan. T.," Prediction of Tumor in Classifying Mammogram images by k-Means, J48 and CART Algorithms", International Journal of Data Mining Techniques and Applications,Vol 04 (02), pp 29-34, 2015.
6. Velmurugan. T and Venkatesan. E.," A Hybrid Multifarious Clustering Algorithm for the Analysis of Mammogram Images", International Journal of Computer and Communications,Vol 07 (12), pp 136-151, 2019.
7. Kaur. G and Chhabra. A., "Improved J48 classification algorithm for the prediction of diabetes", International Journal of Computer Applications.Vol.98 (22), pp. 13–7, 2014.
8. Kambo., Rubi and AmitYerpude.,"Classification of basmati rice grain variety using image processing and principal component analysis", International Journal of Computer Trends and Technology, Vol. 11(2), pp. 80-85, 2014.
9. Sujata Joshi and Priyanka Shetty.S.R.,"Performance Analysis of Different Classification Methods in Data Mining for Diabetes Dataset Using WEKA Tool", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 3(3), pp. 1168 – 1173, 2015.
10. Ramani. R., Valarmathy. S and SuthanthiraVanitha. N., "Breast Cancer Detection in Mammograms based on Clustering Techniques-A Survey", International Journal of Computer Applications, Vol. 62 (11), pp. 17-21, 2013.
11. ZainulAbdinJaffery., Zaheeruddin and Laxman Singh.," Performance Analysis of Image Segmentation Methods for the Detection of Masses in Mammograms", Performance Analysis of Image Segmentation Methods for the Detection of Masses in Mammograms, Vol. 82 (2). pp. 44-50, 2013.
12. Karmilasari., Suryarini Widodo., Matrissya Hermita and Louisiana ETP.,"Sample K-Means Clustering Method for Determining the Stage of Breast Cancer Malignancy Based on Cancer Size on Mammogram Image Basis", Vol.5 (3), pp.86-90, 2014.
13. Hiral N. Pokarand. Poorvi H. Patel.,"Survey of Different Techniques Used For Detection of Malignancy in Mammograms of Breast Cancer", International Journal of Advance Engineering and Research Development, Vol.2 (12), pp.656-664, 2011.
14. Akila.K and Sumathy.P., "Early Breast Cancer Tumor Detection on Mammogram Images", International Journal Computer Science, Engineering and Technology, Vol. 5 (9), pp. 334-336, 2015.

15. Sujata Joshi and Priyanka Shetty. S. R.,"Performance Analysis of Different Classification Methods in Data Mining for Diabetes Dataset Using WEKA Tool", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 3 (3), pp. 1168 – 1173, 2015.
16. Venkatesan. E and Velmurugan. T.," Role of Classification Algorithms in Medical domain: A Survey", International Conference on Information, System and Convergence Applications, June 24-27, 2015 in Kuala Lumpur, Malaysia.
17. Saravananathan.K and Velmurugan.T.," Analyzing Diabetic Data using Classification Algorithms in Data Mining", Indian Journal of Science and Technology, Vol 9 (43), PP.1-6, 2016
18. Govindasamy. K and Velmurugan. T.," A Study on Classification and Clustering Data Mining Algorithms based on Students Academic Performance Prediction", International Journal of Control Theory and Applications, Vol. 10 (23), pp. 147 – 159, 2017.
19. Dharmarajan. A, Velmurugan. T.," Efficiency of k-Means and k-Medoids Clustering Algorithms using Lung Cancer Dataset", International Journal of Data Mining Techniques and Applications, Vol. 05 (2), pp. 150156, 2016.
20. Dharmarajan. A, Velmurugan. T.," Lung Cancer Data Analysis by k-means and Farthest First Clustering Algorithms", Indian Journal of Science and Technology, Vol. 08 (15), pp. 1-8, 2015.
21. Velmurugan.T and Venkatesan.E.," Effective Fuzzy C Means Algorithm for the Segmentation of Mammogram images of Identify Breast Cancer", International Journal of Control Theory and Applications , Vol 09(10), pp 4647-4660, 2016.