Analysis of K-Means Algorithm Using Classification Techniques in Mammographic Dataset

[1]Mrs.K.K.Kavitha, [2]Ms.S.Kiruthiga

[1] MCA. M.Phil., SET., Head of the Department of Computer Science, Selvamm Arts and Science College (Autonomous), Namakkal
[2]Selvamm Arts and Science College (Autonomous), Namakkal

Abstract: **In today's world, gigantic amount of data is available in science, industry, business and many other areas. This data can provide valuable information which can be used by management for making important decisions. But problem is that how can find valuable information. The answer is data mining. The work focuses on the fundamental concept of the Data mining i.e. Clustering and Classification Techniques. Clustering is done by analyzing k-means algorithm and classification techniques such as J48 and NAÏVE BAYES. Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. Classification is an important data mining technique with broad applications. It classifies data of various kinds. Classification is used to classify each item in a set of data into one of predefined set of classes or groups. This work has been carried out to make a performance evaluation of Naïve Bayes and j48 classification algorithm. Naive Bayes algorithm is based on probability and j48 algorithm is based on decision tree, that to make comparative evaluation of classifiers J48 and NAÏVE BAYES.**

Keywords: **Classification, Clustering, J48, Naïve Bayes, k-means, unsupervised learning.**

I. INTRODUCTION

Data mining involves the use of various sophisticated data analysis tools for discovering previously unknown, valid patterns and relationships in huge data set. These tools are nothing but the machine learning methods, statistical models and mathematical algorithm. Data mining consists of more than collection and managing the data, it also includes analysis and prediction. Classification technique in data mining is capable of processing a wider variety of data than regression and is growing in popularity. There are several data mining techniques are preprocessing, association, classification, pattern recognition and clustering. In our work performs by clustering and classification techniques.

CLUSTERING:

Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

CLASSIFICATION:

Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

II. K-MEANS CLUSTERING ALGORITHM

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bar center of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function knows as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

Where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and $v_{j.}$

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

III. NAIVE BAYES CLASSIFIER

Naïve Bayes Classifier:

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Algorithm:

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(c \mid x) = \frac{P(x \mid c) P(c)}{P(x)}$$

$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

IV. J48 DECISION TREE

Decision trees are very popular tools for classification as they represent rules. A decision tree is in the form of a tree, where each node is either a leaf node (it indicates value of target class ofexamples) or a decision node (it specifies some test tobe carried out on a single feature-value), with two ormore than two branches and each branch has a sub-tree. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple.

Algorithm

```
INPUT:
D //Training data
OUTPUT
T //Decision tree
DTBUILD (*D)
{
T=φ;
T= Create root node and label with splitting attribute;
T= Add arc to root node for each split predicate and
label;
for each arc do
D= Database created by applying splitting
Predicate to D;
if stopping point reached for this path, then
T'= create leaf node and label with appropriate class;
else
T'= DTBUILD(D);
T= add T' to arc;
}
```

V. EXPERIMENTAL RESULTS

Data Set Acquisition:

Breast cancer is the most common form of cancer amongst women. Early and accurate detection of breast cancer results in long survival of patients. Machinelearning techniques are being used to improve diagnostic capability for breast cancer. Various classificationtechniques (naïve bayes, decision trees, support vector machines,fuzzy- genetic algorithmic.) have been used to studybreast cancer dataset. Breast cancer data was taken from UCI machine learning data repository [6].

WEKA Tool:

WEKA [9] is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes a visualization tools.

VI. PERFORMANCE EVALUATION

i) Comparison between J48 and Naïve classifier:

Decision Trees are very flexible, easy to understand, and easy to debug. They will work with classification problems and regression problems. So if you are trying to predict a categorical value like (red, green, up, down) or if you are trying to predict a continuous value like 2.9, 3.4 etc Decision Trees will handle both problems. Probably one of the coolest things about Decision Trees is they only need a table of data and they will build a classifier directly from that data without needing any up front design work to take place[8].

Naive bayes will answer as a continuous classifier. There are techniques to adapt it to categorical prediction however they will answer in terms of probabilities like (A 90%, B 5%, C 2.5% D 2.5%) Decision trees work better with lots of data compared to Naive Bayes. Decision trees are neat because they tell you what inputs are the best predicators of the outputs so often decision trees can guide you to find if there is a statistical relationship between a given input to the output and how strong that relationship is. Often the resulting decision tree is less important than relationships it describes. So decision trees can be used a research tool as you learn about your data so you can build other classifiers.

Table 1: Comparison between J48 and Naïve classifier

Accuracy	J48	Naïve
Correctly classified instances	75.52%	71.68%
Incorrectly classified instances	24.48%	28.32%
Kappa static	0.2826	0.2857
Mean absolute error	0.3676	0.3272
Root Mean Squared Error	0.4324	0.4534
Relative Absolute Error	87.8635	78.2086
Root Relative Squared Error	94.61%	99%

ii) Accuracy level of J48 and Naïve classifier:

Naive-Bayesian classifiers are very robust to irrelevant attributes, and classification takes into account evidence from many attributes to make the final prediction. On the downside, Naive- Bayes classifiers require making strong independence assumptions and when these are violated, the achievable accuracy may asymptote early and will not improve much as the database size increases. Decision-tree classifiers are also fast and comprehensible, but current induction methods based on recursive partitioning. As each split is made, the data is split based on the test and after two dozen levels there are usually base decisions.

In this work describe attempts to utilize the advantages of decision trees (i.e., segmentation) compared to Naive-Bayes (evidence accumulation from multiple attributes). A decision tree is built with univariate splits at each node, but with Naive-Bayes classifiers at the leaves. The induction process is very different and geared toward larger datasets. The resulting classifier is as easy to interpret as Decision -trees and Naive-Bayes. The decision-tree segments the data, a task that is consider an essential part of the data mining process in large databases.

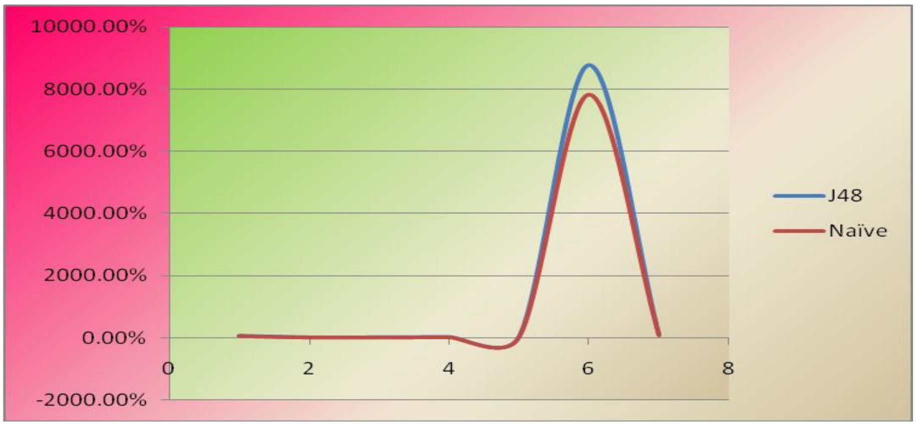

Figure 1:Accuracy level between J48 and Naïve

iii) Performance of J48 and Naïve classifier:

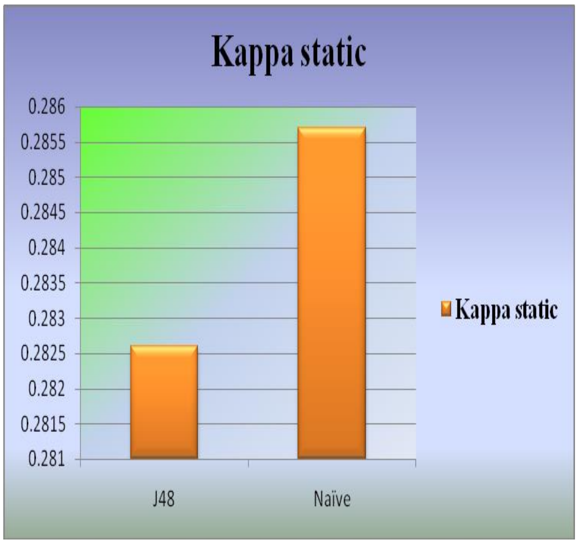

Figure 2:Kappa Static

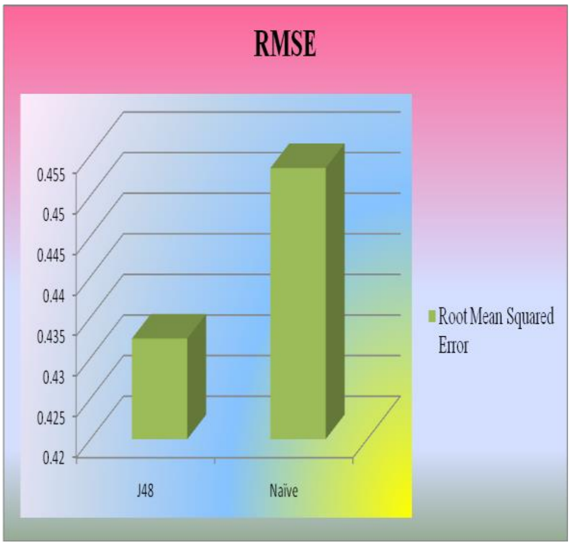

Figure 3:RMSE

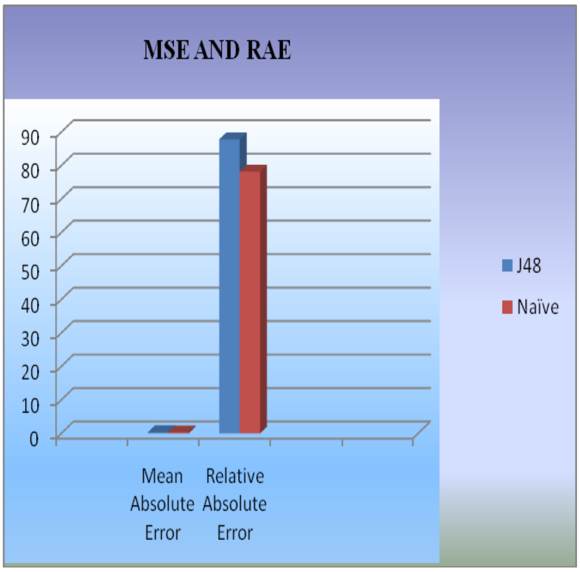

Figure 4: MSE and RAE

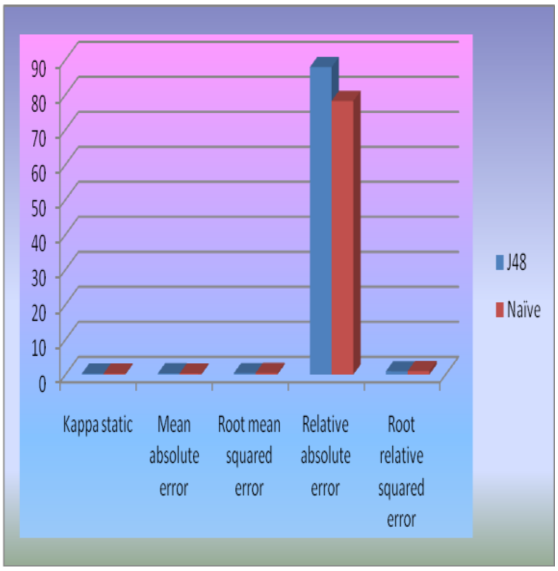

Figure 5: Comparitive Analysis of J48 and Naïve Bayes

VII. CONCLUSION

From the above analysis method, concluded that J48 Decision tree Classifier provides high accuracy than the Naïve Bayes. K-means clustering algorithm taken as minimum time to processed the dataset. This proves that the, J48 is a simple classifier technique to make a decision tree. Efficient result has been taken from mammographic dataset using weka tool in the experiment. Naïve Bayes classifier is also showing good results. The experiments results shown in the study are about classification accuracy. J48 gives more classification accuracy for class nominal in mammographic dataset having two values recurrence event and no recurrence event.

Future Work:

In future J48 classified to other decision tree classifiers like AD tree, NB tree and Random tree structure.J48 takes more time for classifying large dataset to compare other classification but J48 gives accurate result. Further implementation work can be concentrated on improving reduction of the time delay.

REFERENCES

[1] Anshl Goyal and Rajnimehta "Performance Comparison Naïve Bayes and J48 Classification Algorithms using Bank Dataset"ISSN 0973-4562, Vol-7, No.11, 2012.

[2] Gangnjot Kawrand Amit Chabra "Improved J48 Classification Algorithm for the Prediction of Diabetes", Vol-98, No-22, July 2014.

[3] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques",Second Edition. University of Illinois at Urbana-Champaign.

[4] Michel Steinbach,George Karypis and Vipin Kumar,"A Comparison of Document Clustering Techniques", Computer Science Engineering,Technical Report #00-034.

[5] Mahmutkaya and Oktay Yildiz," Breast Cancer Diagnosis Based on Naïve Bayes Machine Learning Classifier with KNN Missing Data Imputation", Awer Proceeding Information Technology and Computer science, Vol-4(401-407) 2013.

[6] Machine learning repository, https://archive.ics.uci.edu/ml/datasets.html

[7] Shiv Shakti Shrivastara and Anjali sant," An Overview on Data Mining Approach on Breast Cancer Data" Vol-3, No-4, Issue13, Dec 2013.

[8] www.stackflow.com/decision tree - vs. - naive bayes-classifier.

[9] www.cs.waikato.ac.nz/ml/weka/documentation.html.