Comparative Study of Different Data Mining Prediction Algorithms

M Krishna Satya Varma
Asst. Professor, Dept of IT,SRKR Engineering College, Bhimavaram A P, India

N K Kameswara Rao
Assoc. Professor, Dept of IT,SRKR Engineering College Bhimavaram, A P, India

Abstract-: The main objective of this research paper is to prove the efficiency of high dimensional data analysis and different algorithms in the prediction process of Data mining. The approach made for this survey includes, an extensive literature search on published papers in the application of Data mining in prediction. The main algorithms which were involved in this study include classification using decision tree, clustering algorithm, Apriori algorithm and association rules. All the algorithms were analyzed with different data sets which were collected from real life applications.

Keywords- Classification, Clustering, Data mining, Decision tree, High Dimensional data analysis. Prediction.

I INTRODUCTION

Data mining is an interdisciplinary field of computer science. It is the progression that results in the innovation of new patterns in large database. It utilizes methods at the juncture of artificial intelligence, machine learning, statistics, and database systems. The overall objective of the data mining process is to mine knowledge from an accessible data set and transform it into a human explicable structure for auxiliary use. Besides the rare analysis step, it involves database and data management aspects, data processing model and presumption considerations, interestingness metrics, complication, considerations, post-processing of found structures, visualization, and online updating [1].

There are different techniques for multidimensional data analysis. The models included in the predictive data mining consist of clustering, decision tree, association rules, pattern matching, classification rules, statistical analysis etc. Through this paper we tried to analyze the advantages and disadvantages of main three algorithms such as association rules, clustering algorithm and classification and prediction methods. In clustering algorithm, we have selected the k-means clustering algorithm for sample study, similarly for classification and prediction methods, decision tree have used which is used C4.5 algorithm. Association rule algorithm also verified with sample data. Data set used for this study is UCI repository.

A. Multidimensional Analysis

Multidimensional data is a type of data that records facts related to variable entities called dimensions. Technologies such as Integrated Data Analysis & Simulation module (IDASM) provide an environment where multiple data sets can be integrated to conduct analysis across different cases Dimensions are entities which are used to analyze data and may represent the sides of a multidimensional cube. Selecting proper dimensions in data analysis is indeed very crucial for multidimensional analysis and gaining greater insights from the data. Dimension modeling is very important in data analysis because queries can be satisfied only if dimensions are properly defined. If one of the dimensional values changes, then output also will change. The multi-dimensional analysis tools are capable of handling large data sets. These tools provide facilities to define convenient hierarchies and dimensions. But they are unable to predict trends and find the patterns and guess the hidden behavior [2].

B. Related Works in this Area

Many works related in this area have been going on. “In A New Approach for Evaluation of Data Mining Techniques” by Moawia Elfaki Yahia[19],the authors tried to put a new direction for the evaluation of some techniques for solving data mining tasks such as: Statistics, Visualization, Clustering, Decision Trees, Association Rules and Neural Networks. This paper also provides an summary of techniques that are used to

improvise the efficiency of Association Rule Mining (ARM) from huge databases. In another article " K-means v/s K-medoids: A Comparative Study" Shalini S Singh explained that portioned based clustering methods are suitable for spherical shaped clusters in medium sized datasets and also proved that K-means are not sensitive to noisy or outliers.[21]. In an article "Predicting School Failure Using Data Mining C". MÁRQUEZ-VERA explained the prediction methods and the application of classification rule in decision tree for predicting the school failures.[22].There are many research works carrying out related with data mining technology in conjecture such as financial stock market predict, rainfall forecasting, appliance of data mining technique in health care, base oils biodegradability predicting with data mining technique etc,[23].

II. THE ROLE OF DATA MINING IN HIGH DIMENSIONAL ANALYSIS:

Due to the advancement in algorithm and changing scenario, new techniques have emerged in data analysis, which are used to predict and generate data patterns and to classify entities having multivariate attributes. These techniques are used to identify the pre-existing relationship in the data that are not readily available. Predictive Data mining deals with impact patterns of data [4].

A. Models used in Predictive Data Mining

The models mainly used in predictive data mining includes Regression, Time series, neural networks, statistical mining tools, pattern matching, association rules, clustering, classification trees etc. [5].
Regression model is used to express relationship between dependent and independent variables using an expression. It is used when the relationship is linear in nature. If there is a nonlinear relationship, then it cannot be expressed using any expression, but the relationship can be built using neural networks. In time series models, historic data is used to generate trends for the future. Statistical mining models are used to determine the statistical validity of test parameters and can be utilized to test hypothesis undertake correlation studies and transform and prepare data for further analysis. Pattern matching are used to find hidden characteristics within data and the methods used to find patterns with the data includes association rules. [16]
Association rules allows the analysts to identify the behavior pattern with respect to a particular event where as frequent items are used to find how groups are segmented for a specific set. Clustering is used to find the similarity between entities having multiple attributes and grouping similar entities and classification rules are used to categorize data using multiple attributes.

III. C4.5

"C4.5" is an algorithm used to initiate a decision tree developed by Ross Quinlan [1]. It is an lean-to Quinlan's prior ID3 algorithm. The decision trees generated with "C4.5" can be used for classification, and for this reason, "C4.5" is frequently referred to as a statistical classifier [3]. This algorithm builds decision trees from a set of training data with the concept of information entropy.
It is handling both constant and discrete attributes, handling training data with missing attribute values and also handling attributes with differing costs. In construction a decision tree we can compact with training sets that contain records with unknown attribute standards by evaluating the gain, or else the gain ratio, for an aspect by allowing for only the records where that aspect is defined. In using a decision tree, we can organize records that have unknown aspect values by estimating the probability of the various possible results [14].
Follows the algorithms employed in "C4.5" using decision tree.

A. Decision trees

Given a set S of cases, "C4.5" first grows an primary tree using the divide-and-conquer algorithm as follows [7]. If all the cases in S belong to the similar class or S is small, the tree is a leaf labeled with the majority frequent class in S. Otherwise, choose a test based on a single attribute among two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1, S2 . . . according to the outcome for each case, and apply the similar procedure recursively to each subset. Use either information gain or gain ratio to rank the possible tests. Check the estimation error [12].

Advantages	**Limitations**
Error rate is less	Decision trees typically require certain knowledge of quantitative or statistical experience on the way to complete the process precisely. Fault to precisely understand decision trees can lead to a distorted outcome of business opportunities or decision possibilities.

Decomposition is easier as compared with other techniques	It can also be intricate to include variables on the decision tree, eliminate replica information or express information in a logical, steady manner. The incapability to complete the decision tree using only one set of information can be somewhat complicated.
Represent the knowledge in the form of IF-THEN rules. Rules are easier for humans to understand.	While partial information can create difficulties within the decision-tree process, excessively much information can also be an issue.

IV. CLUSTERING ALGORITHM

Clustering is the task of conveying a set of objects into groups so that the objects in the similar cluster are more parallel to each other than to those in further clusters. Clustering is a core job of explorative data mining, and a common technique for data analysis used in many fields including information retrieval. .Cluster analysis groups objects based on their similarity. The measure of similarity can be computed for various types of data. [5]
Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods, k-means algorithm, graph based model etc.

A. K means algorithm

Within data mining K-means clustering is a system of cluster analysis which aims to partition n observations into k clusters in which each study belongs to the cluster with the nearby mean. The k-means algorithm[also referred as Lloyd's algorithm] is a easy iterative method to partition a given dataset into a user specified number of clusters [8]. The algorithm operates on a set of d-dimensional vectors, D = {xi | i = 1, . . . , N}, where xi ∈ _d denote the i^{th} data point. The algorithm is initialized by picking k points in _d as the initial k cluster representatives or "centroids".
Techniques for selecting these original seeds embrace sampling at unsystematic from the dataset, setting them as the outcome of clustering a small subset of the data or perturbing the comprehensive mean of the data k times. Subsequently the algorithm iterates between two steps till convergence: [18]
Step 1: Data Assignment. Each data point is assigned headed for its closest centroid, with ties broken arbitrarily. This results in a partitioning of the data.
Step 2: Replacement of "means". Every cluster representative is relocated to the core (mean) of all data points assigned to it. If the data points come with a prospect measure (weights), then the replacement is to the expectations (weighted mean) of the data partitions.
The algorithm converges when the assignments (and hence the 'cj' values) no longer change. The algorithm execution is visually depicted in Fig. 1. Note that all iteration needs N x k comparisons, which determines the time intricacy of one iteration. The number of iterations required for union varies and may depend on N, but as a first cut, this algorithm can be considered linear in the dataset size.

B. Advantages and Limitations

Table 2: Advantages and Limitations of k-means algorithm

Advantages	Limitations
Relatively efficient and easy to implement.	Sensitive to initialization
Terminates at local optimum.	Limiting case of fixed data.
Apply even large data sets	Difficult to compare with different numbers of clusters
The clusters are non-hierarchical and they do not overlap	Needs to specify the number of clusters in advance.
With a large number of variables, K-Means may be computationally faster than hierarchical clustering	Unable to handle noisy data or outliers.
K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular	Not suitable to discover clusters with non-convex shapes

V. ASSOCIATION RULES ALGORITHM

Association rule mining searches for interesting relationships among items in a given set. Here the main rule interestingness is rule support and confidence which reflect the usefulness and certainty of exposed rules. Association rule mining algorithm is a two step process where we must have to find all the numerous item sets and generate strapping association rules from the regular item sets [9].
If a rule concern association between the incidence or absence of items, it is a Boolean association rule. If a rule describes association between quantitative items or attributes, then it is known as quantitative association rules. Here the quantitative values for items are partitioned into intervals. The algorithm can be formed based on dimensions, based on level of abstractions involved in the rule set and also based on various extensions to association mining such as correlation analysis [17].

A. Multi dimensional Association Rules:

In multi dimensional databases, each distinct predicate in a rule as a dimension. Association rule that involve two or more dimensions each of which occurs only once in the rule can be referred as multidimensional association rules.
Multi dimension association rules with no persistent predicates are called inter dimension association rules and may with repeated predicates which can contain multiple occurrences of some predicates are called hybrid dimension association rules. For example
Age(X,50…..70)^ FAMILYHISTORY(X,DISEASE)=> DISEASEHIT(X,"TYPHOID").
Here the database attributes can be categorical or quantitative with no ordering among the values The basic definition of association rule states that, Let A={l1,l2,……..lm} be a set of items, and Let T, the transaction database, be a set of transaction, where each transaction t is a set of items and thus t is a subset of A. An association rule tells us about the association between two or more items. For example, if we are given a set of items where items can be referred as disease hit in an area and a large collection of patients who are subsets of some inhabitants in the area. . The task is to find relationship between the presences of disease hit within these group. In order for the rules to be useful there are two pieces of information that must be supplied as well as the actual rule: Support is how often does the rule apply? and confidence is how often is the rule is correct. [19]
In fact association rule mining is a two-step process: Find all frequent item sets / disease hit-by definition, each of these item sets will occur at least as frequently as a predetermined minimum support count, and then generate strong association rules from the frequent item sets by definition, these rules must satisfy minimum support and minimum confidence. In this study predicting the chances of disease hit an area, by correlating the parameters or attributes such as climate, environmental condition, heredity, education with the inhabitants. And also finding how these parameters are associated with the chances of disease hit.

Table 3: Association Rules-Advantages and Disadvantages

Advantages	Limitations
Association rule algorithms can be formulated to look for sequential patterns.	Association rules do not show reasonable patterns with dependent variable and cannot reduce the number of independent variables by removing.
The methods of data acquisition and integration, and integrity checks are the most relevant to association rules.	Association rules cannot be useful if the information do not provide support and confidence of rule are correct.

VI. METRICS USED TO VALIDATE THE ALGORITHMS

The accuracy of the algorithms can be finding out by the validation. Validation of the algorithms can be done by using the metrics Precision (P), Recall (R), F-measure, Error Rate and Accuracy Rate.

A. Precision:

Precision is also called as positive predictive value. In pattern recognition and information retrieval with binary classification, precision is the fraction of retrieved instances that are appropriate. Documents referred as instances and a set of specified documents returned by a search engine referred as tasks in an information retrieval system or equivalently, "relevant" and "not relevant" are two categories assigned to each document. In this situation the "relevant" documents are belong to the "relevant" category. Then the relevant documents retrieved divided by total number of documents retrieved in this search is called the precision. Every result retrieved is relevant (but it may not retrieved all the relevant documents) by as search engine then precision becomes prefect with the score 1.0 in the information retrieval.

$$\text{Precision} = \frac{|\{\text{relevant documents }\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

i.e.Precision (P) = (Correct Values/ Prediction Values)*100

The division between the number of true positives and the total number of positive class's items labeled is called as the precision in classification (i.e. the number of items properly labeled as belongs to the positive class divided by the total number of positives (total number of items labeled in class) (true positives + false positives).

$\text{Precision} = \frac{tp}{tp+fp}$ Where tp is true positives, fp is false positives,

B. Recall:

Recall is also called as sensitivity. In pattern recognition and information retrieval with binary classification, Recall is the fraction of significant instances that are retrieved. Documents referred as instances and a set of specified documents returned by a search engine referred as tasks in an information retrieval system or equivalently, "relevant" and "not relevant" are two categories assigned to each document. In this situation the "relevant" documents are belong to the "relevant" category. Then the relevant documents retrieved divided by total number of existing relevant documents in this search is called the Recall. Every relevant document is retrieved (but it won't say about number of irrelevant documents are retrieved) by as search engine then recall becomes prefect with the score 1.0 in the information retrieval.

$$\text{Recall} = \frac{|\{\text{relevant documents }\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

i.e. Recall (R)= (Correct Values/Actual Values)*100

The division between the number of true positives and the total number of elements that initially belong to the positive class is called as the recall in classification (i.e. the number of items properly labeled as belongs to the positive class divided by the summation of true positives and false negatives, which are items which were not labeled as belonging to the positive class, but should have been).

$\text{Precision} = \frac{tp}{tp+fn}$ Where tp is true positives, fn is false negatives

Both precision and recall are then based on a perceptive and determine of relevance.

C. F-measure:

F-measure is also called as F-score or F_1 Score. In statistical analysis of binary classification, the F1 score is a measure of a test's accuracy. It contains both the precision p and the recall r of the test to calculate the score: p is the number of exact results divided by the number of all arrived results and r is the number of exact results divided by the number of results that must have been arrived. The F_1 score can be measured as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0.

The conventional F-measure or balanced F-score (F_1 score) is the harmonic mean of precision and recall.

$F_1 = 2.\frac{\text{precision .recall}}{\text{precision+ recall}}$ i.e. F-measure= 2*((P*R)/(P+R))

There are numerous reasons that the F-score can be disapproved in Specific circumstances due to its unfairness as an assessment metric. This is also known as the F_1 measure, because recall and precision are consistently weighted. The general formula for positive real β is:

$$F_\beta = (1+\beta^2).\frac{\text{precision . recall}}{(\beta^2.\text{precision}) + \text{recall}}$$

The formula in terms of Type I and type II errors:

$$F_\beta = \frac{(1+\beta^2).\text{true positive}}{(1+\beta^2).\text{true positive} + \beta^2.\text{flase negative} + \text{false positive}}$$

D. Accuracy & Error rates:

The accuracy of a measurement system is the magnitude of proximity of measurements of a quantity to that quantity's real (true) value. The precision of a measurement system, associated to reproducibility and repeatability, is the magnitude to which recurring measurements under unchanged circumstances show the same outcomes. Even though the two words precision and accuracy can be identical in colloquial use, they are intentionally contrasted in the situation of the scientific method.

A measurement system can be exact but not precise, precise but not exact, neither, or both. For example, if an experiment having a systematic error, then rising the sample size normally increases precision but does not get

better accuracy. The effect would be a reliable yet imprecise string of results from the defective experiment. Eliminating the methodical error improves accurateness but does not modify precision.

A measurement scheme is considered suitable if it is both precise and accurate. Interrelated terms include unfairness (non-random or directed effects caused by an issue or issues unrelated to the autonomous variable) and error (random inconsistency). The terminology is applied to tortuous measurements—that is, values acquired by a computational process from observed data. That is, the proportion of true results is the accuracy (both true positives and true negatives) in the inhabitants. To make the situation clear by the semantics, it is frequently referred to as the "Rand Accuracy". It is a stricture of the test.

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{flase negatives} + \text{true negatives}}$$

i.e. Accuracy Rate= Correct Values/Actual Values*100

Error Rate= (Actual Values-Correct Values)/Actual Values*100

Table-4 Metrics - Classification Algorithm

Parameter	Male	Female	Total
Precision(P)	86.08	85.42	85.76
Recall(R)	74.73	75.93	75.29
F-measure	80.00	80.39	80.19
Error Rate	25.27	24.07	24.71
Accuracy Rate	74.73	75.93	75.29

Table-5 Metrics - Clustering Algorithm

Parameter	Male	Female	Total
Precision(P)	91.08	84.21	87.70
Recall(R)	78.57	79.01	78.78
F-measure	84.37	81.53	83.00
Error Rate	21.43	20.99	21.22
Accuracy Rate	78.57	79.01	78.78

Table-6 Metrics - Apriori Algorithm

Parameter	Male	Female	Total
Precision(P)	88.69	88.00	88.36
Recall(R)	81.87	81.48	81.69
F-measure	85.14	84.62	84.89
Error Rate	18.13	18.52	18.31
Accuracy Rate	81.87	81.48	81.69

Table-7 Metrics - Association Rules (Decision Tree)

Parameter	Male	Female	Total
Precision(P)	90.64	88.46	89.60
Recall(R)	85.16	85.19	85.17
F-measure	87.82	86.79	87.33
Error Rate	14.84	14.81	14.83
Accuracy Rate	85.16	85.19	85.17

Table-8 Metrics - Hybrid Algorithm

Parameter	Male	Female	Total
Precision(P)	96.05	95.54	95.81
Recall(R)	93.41	92.59	93.02
F-measure	94.71	94.04	94.40
Error Rate	06.59	07.41	06.98
Accuracy Rate	93.41	92.59	93.02

Table-9 Comparison of Algorithms with No. of people Effected

Algorithm	Actual	Classificatior	Clusterin	Apriori	Decision Tre	Hybrid
Count	344	302	309	318	327	334

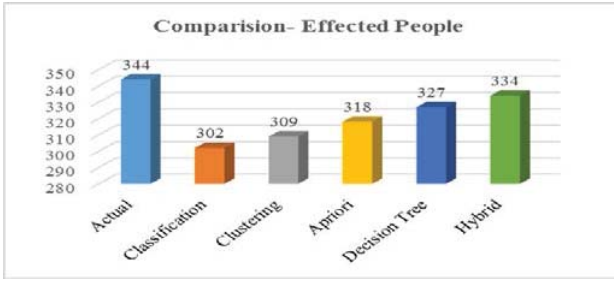

Figure 1. Comparison of Algorithms- victims

Table-10 Comparison of Algorithms -Gender wise

Algorithm	Actual		Classification		Clustering		Apriori		Decision Tree		Hybrid	
Gender	M	F	M	F	M	F	M	F	M	F	M	F
Count	182	162	158	144	157	152	168	150	171	156	177	157

Table-11 Comparison of Algorithms Area wise

Area	Actual	Classification	Clustering	Apriori	Decision Tree	Hybrid
Tribal	49.1	49.0	49.2	50.0	48.9	49.4
Hill	25.9	24.8	26.2	26.1	25.4	25.4
Rural	15.7	16.2	16.2	14.5	16.2	15.6
Urban	9.3	9.9	8.4	9.4	9.5	9.6

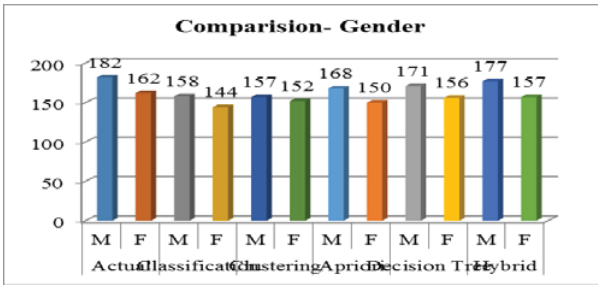

Figure 2. Comparison of Algorithms -Gender wise

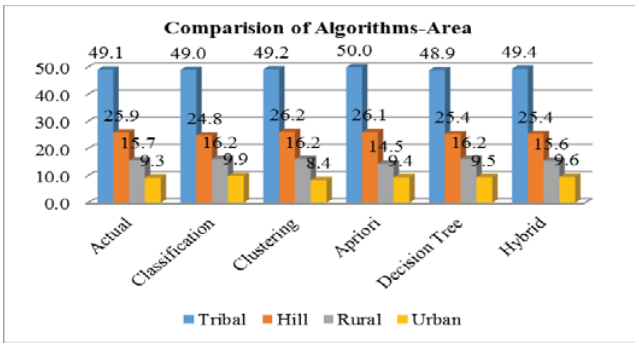

Figure 3. Comparison of Algorithms -Area wise

Table-12 Comparison of Algorithms- Validation Parameters

	Classification	Clustering	Apriori	Decision Tree	Hybrid
Precision	84.75	85.12	83.87	84.38	94.78
Recall	70.42	72.54	73.24	76.06	89.44
F-measure	76.92	78.33	78.20	80.00	92.03
Error Rate	29.58	27.46	26.76	23.94	10.56
Accuracy	70.42	72.54	73.24	76.06	89.44

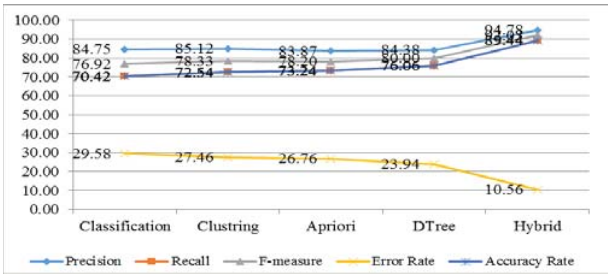

Figure4 Comparison of Algorithms- Validation Parameters

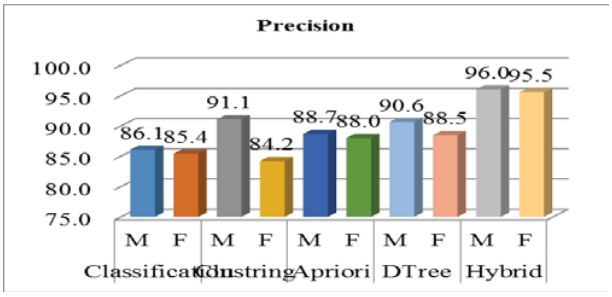

Figure-5 Comparison of Algorithms- Precision

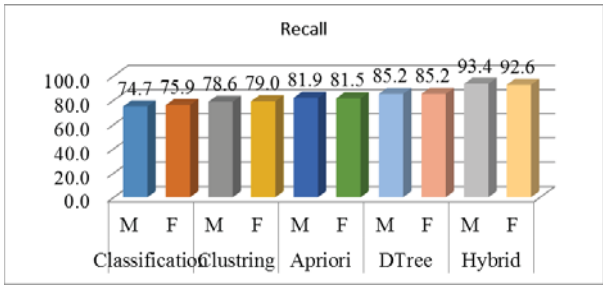

Figure-6 Comparison of Algorithms- Recall

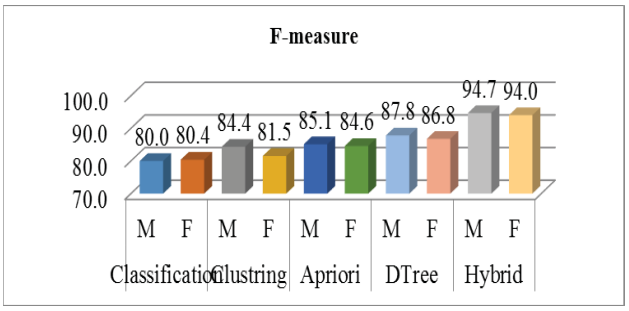

Figure-7 Comparison of Algorithms- F-Measure

Figure-8 Comparison of Algorithms- Error Rate

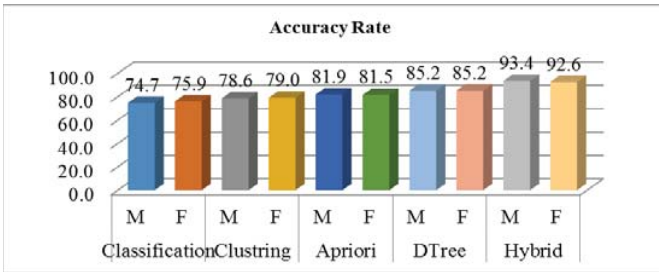

Figure-9 Comparison of Algorithms- Accuracy Rate

VII. CONCLUSION

In this research work, I have made a study to make a comparison of the some of the existing data mining algorithm for high dimensional data clusters to estimate prediction in data mining technique. The main techniques included in the survey are decision tree, clustering algorithm, k-means algorithm and association rule algorithm. Studied each algorithm with the help of high dimensional data set with UCI repository and find the advantages and disadvantages of each. By comparing the advantages and disadvantages of each algorithm, I am trying to develop a hybrid algorithm for multidimensional data analysis. The efficiency was calculated on the basis of time complexity, space complexity, space requirements etc. The sample used in this study includes UCI repository. The efficiency of new algorithm can be checked with real time data.

REFERENCES

[1] en.wikipedia.org/wiki/Data mining
[2] Multidimensional Data Analysis and data mining, Black Book, Arijay Chaudhry and Dr. P .S. Deshpande.
[3] R. Agarwal, T. Imielinski and A.Swamy “Mining association Rules between Set of Items in Large Database”.In ACM SIGMO international conference on Management of Data .
[4] Goulbourene G, Coenen F and Leng P, Algorithms for Computing Association Rules using a Partial support Tree” j. Knowledge Based System 13(2000) pp-141-149.
[5] David Hand, Heikki Mannila, Padhraic Smyth,” principles of Data Mining”.
[6] “Designing Association Models for disease Prediction using Apriori”, N. K. Kameswara Rao, Dr. G. P. Saradhi Varma, Elixir International Journal, Elixir Inform. Tech. 71 (2014), pp. 24666-24669 June -2014.
[7] “Classification Rules Using Decision Tree for Dengue Disease”, N. K. Kameswara Rao, Dr. G. P. Saradhi Varma, Dr M. Nagabhushana Rao. International Journal of Research in Computer and Communication Technology, Vol.No. 3, Issue.3, pp.340-343 March -2014.
[8] “Knowledge Discovery from Realtime Data using Data Mining”, N. K. Kameswara Rao, Dr. G. P. Saradhi Varma, International Journal for Research in Science & Advanced Technologies, Volume No. 2, Issue No. 3, pp.034-037, Mar-2014.
[9] “Another Look at Measures of Forecast Accuracy” Hyndman R and Koehler A (2005).
[10] “Mining the structural knowledge of high-dimensional medical data using Isomap” S. Weng I C. Zhang I Z. Lin I X. Zhang 2
[11] Bhattachariee, A., Richards, W. G., Staunton, J., Li, C., Monti,S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M.,Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W.,Johnson, B. E., Golub, T. R., Sugarbaker, D. J., And Meyerson, M. (2001): 'Classification of human lung carcinomasby mRNA expression profiling reveals distinct adenocarcinoma sub classes', Proc. Nat. Acad. Sci. USA, 98, pp. 13790 13795 BLAKE, C. L. and Merz, C. J. (1998):
[12] BORG, I., and GROENEN, P. (1997): 'Modern multidimensional scaling: theory and application' (Springer-Verlag, New York, Berlin, Heidelberg, 1997).
[13] Adomavicius G, TuzhilinA2001 Expert-driven validation of rule-based user models in personalization applications. Data Mining Knowledge Discovery 5(1/2): 33–58.
[14] Shekar B, Natarajan R 2004b A transaction-based neighbourhood-driven approach to quantifying interestingness of association rules. Proc. Fourth IEEE Int. Conf. on Data Mining (ICDM 2004) (Washington, DC: IEEE Comput. Soc. Press) pp 194–201

[15] Mohammed J. Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and WeiLi. Parallel algorithms for discovery of association rules. Data Mining and Knowledge Discovery: An International Journal, special issue on Scalable High-Performance Computing for KDD, 1(4):343–373, Dec, 2001.
[16] Refaat, M. Data Preparation for Data Mining Using SAS,Elsevier, 2007.
[17] El-taher, M. Evaluation of Data Mining Techniques, M.Sc thesis (partial-fulfillment), University of Khartoum, Sudan, 2009.
[18] Lee, S and Siau, K. A review of data mining techniques, Journal of Industrial Management & Data Systems, vol 101, no 1, 2001, pp.41-46.
[19] "A New Approach for Evaluation of Data Mining Techniques", Moawia Elfaki Yahia1, Murtada El-mukashfi El-taher2, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.
[20] "A study on effective mining of association rules from huge database" V.Umarani et. al.International Journal of Computer Science and Research, Vol. 1 Issue 1, 2010.
[21] "K-means v/s K-medoids: A Comparative Study" Shalini S Singh, National Conference on Recent Trends in Engineering & Technology, May 2011.
[22] Predicting School Failure Using Data Mining"C. MÁRQUEZ-VERA
[23] Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" K.Srinivas et al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255.
[24] "A hybrid Algorithm for Dengue Disease Prediction with Multi Dimensional Data" N. K. Kameswara Rao, Dr. G. P. Saradhi Varma, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4, Issue 3, pages: 1033-1037. March 2014