Effective and Efficient Correlation Based Subset Selection Algorithm for Fast Data Retrieval

R. Jayasri[1], Dr. E. George Dharma Prakash Raj[2]
[1, 2] Department of Computer Science and Engineering
[1, 2] Bharathidasan University, Tiruchirappalli, Tamil Nadu, India

Abstract- *Many feature subset selection methods have been studied for machine learning applications. In filter feature selection method applied the application of cluster analysis has been verified to more effective than traditional feature selection algorithms. They can be divided into different broad categories, A fast clustering-based feature subset selection algorithm workings in, 1. Features are divided into cluster by using graph-theoretic clustering methods.2. The most representative feature strongly correlated to target classes and each cluster to form a subset of features. The FAST algorithm was tested 35 widely available image, microarray, and text data sets. Hence, each cluster is treated as a single feature. Feature selection algorithm composed to the two connected components of irrelevant features reduced and redundant features elimination. We proposed, A Fast clustering bAsed feature subset Selection algoriThm (FAST) for effective and efficient data retrieval, with the different Correlation measures for accurate and valuable data in the retrieved dataset. In this research paper, Wehave increased the Correlation measures and Classification accuracy from the original data set has been improved by FAST algorithm.*

Keywords- FAST algorithm, Feature Subset Selection Methods, Classification accuracy, Correlation measures, Breast Cancer Dataset.

I. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD) an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.The actual data mining task is theautomatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining). Data mining consists the more than collection and managing data; it includes analysis and prediction people are frequently do mistakes while analysing or, possibly, when trying to set up relationships between multiple features. Clustering is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.It is a main task of exploratory data mining, and a common technique for statisticaldata analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Many clustering algorithm and prediction methods have been proposed by researchers in machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Most algorithm are memory resident, typically assuming a small data size. Recent data mining research has been built on such work, developing scalable clustering and prediction of handling large disk-resident. Clustering can therefore be formulated as a multi-objective optimization problem.

Correlation clustering provides a method for clustering a set of objects into the optimum number of clusters without specifying that number in advance.Correlation clustering also relates to a different task, where correlations among attributes of feature vectors in a high-dimensional space are assumed to exist guiding the clustering process. These correlations may be different in different clusters, thus a global decorrelation cannot reduce this to traditional (uncorrelated) clustering.

Correlations among subsets of attributes result in different spatial shapes of clusters. Hence, the similarity between cluster objects is defined by taking into account the local correlation patterns. With this notion, the term has been introduced in simultaneously with the notion discussed above. Different methods for correlation clustering of this type are discussed in the relationship to different types of clustering is discussed in, see also clustering high-dimensional data.

This Research focuses on the clustering based fast data retrieval into effective and efficient through the use of fast clustering based feature subset selection algorithm. The aim of this paper is to apply the Fast clustering bAsed feature subset Selection algorithm (FAST) on some publicly available datasets of the Breast Cancer repository in order to clustering the retrieve the effective and efficient data retrieval. The selected datasets are: WBCD, WDBC, UMC-USA, and IBCDB.

II. RELATED WORK

Sherin Mary Varghese, M.N.Sushmitha **"Efficient Feature Subset Selection Techniques for High Dimensional Data"**International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 3, March 2014.In this research is novel clustering- based feature subset selection algorithm for high dimensional data.A database can contain several dimensions or attributes. When the dimensionality increases, data in the irrelevant dimensions may produce much noise and mask the real clusters to be discovered. To deal with these problems, an efficient feature subset selection technique for high dimensional data has been proposed. Feature subset selection reduces the data size by removing irrelevant or redundant attributes. The algorithm involves removing the irrelevant features, constructs the minimum spanning tree from relative ones, and partitioning the MST and selects representative features. In the proposed algorithm, a cluster based feature selection is done and thus dimensionality is drastically reduced. For efficiency Pearson correlation is used for removing the redundant data. The advantages of algorithm Low Time Consuming process, Effective search is achieved based on feature search, no outliers in the data, Easy to cluster the values.

R.Pavithra, J.Vinitha Grace, A.Arun Sethupathy Raja, V.Stalin, M.Ramakrishnan, **"Analysing the Efficiency of Data by Using Fast Clustering and Subset Selection Algorithm"**International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 3, March 2014.

In this paper, mainly focus about the micro array images by analyzing the efficiency of the proposed work. The efficiency consider the time taken to retrieve the data will be better in the proposed by removing all the irrelevant features which are gets analyzed. Using Subset Selection Algorithm for Analyzingthe Efficiency of Data. The overall function leads to the subset selection and FAST algorithm which involves, removing irrelevant features, constructing a minimum spanning tree from relative ones (clustering) and reducing data redundancy and also it reduces time consumption during data retrieval. It supports the microarray data in database; we can upload and download the data set from the database easily. Images can be downloaded from the database. Thus we have presented a FAST algorithm which involves removal of relevant features and selection of datasets along with the less time to retrieve the data from the databases. The identification of relevant data's is also very easy by using subset selection algorithm.

Qinbao Song, Jingjie Ni and Guangtao Wang **"A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data"** IEEE Transactions on Knowledge and Data Engineering vol:25 no:1 year 2013.Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm evaluated from both the efficiency and effectiveness points of view. The efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. In this paper, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. We have compared the performance of the proposed algorithm with those of the five well-known feature selection algorithms FCBF, Relief-F, CFS, Consist, and FOCUS-SF on the 35 publicly available image, microarray, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record. Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime, and the best classification accuracy for Naive Bayes, and RIPPER, and the second best classification accuracy for IB1. The Win/Draw/Loss records confirmed the conclusions. We also found that FAST obtains the rank of 1 for microarray data, the rank of 2 for text data, and the rank of 3 for image data.

K. Revathi and T. Kalai Selvi**"Efficient and Effective Subset Selection Process based on Clustering Algorithm"** International Journal of Innovation and Scientific Research ISSN 2351-8014 Vol. 6 No. 1 Aug. 2014.Process with high dimensional data is enormous issue in data mining and machine learning applications. Feature selection is the mode of recognize the good number of features that produce well-suited outcome as the unique entire set of features. Feature selection process constructs a pathway to reduce the dimensionality and time complexity and also improve the accuracy level of classifier. In this paper, we use an alternative approach, called affinity propagation algorithm for effective and efficient feature selection and clustering process. The endeavor is to improve the performance in terms accuracy and time complexity. Main benefit of proposed algorithm is high speed and low error. The motive of this process is increasing the accuracy level of classifiers and reducing the runtime of the algorithm.

J.K.Madhavi, G.Venkatesh Yadav, **"An Improved Fast Clustering method for Feature Subset Selection on High-Dimensional Data clustering"** International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 10, October 2014.In this paper, we proposed Feature extraction as the process of eliminating the irrelevant information and features during Data Mining. Feature subset selection can be analyzed as the practice of identifying and removing as lot of inappropriate and unnecessary features as achievable. This if for the reason that, irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to receiving a better analysis for that they provide typically information which is previously present in other features of all the existing feature subset selection algorithms, most of them can effectively eliminate irrelevant features but fail to handle redundant features. The improved FAST algorithm is evaluated using various types of data like text data, micro-array data and image data to represent its performance. The identification of relevant data's is also very easy by using subset selection algorithm. The Improved FAST algorithm is proposed to provide an efficient method for feature subset selection for different categories of data. This work is done is four phases to remove irrelevant feature, clustering similar features, removing redundant features and subset selection. Thus our Improved FAST algorithm works efficiently and shows higher performance than FAST in terms of search time.

Dr. K.Sakthivel, R.Abinaya, I.Nivetha, R.Arun Kumar, **"Region Based Image Retrieval using k-means and Hierarchical Clustering Algorithms"**International Journal of Innovative Research in Science, Engineering and Technology ,ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014.In this Research Region Based Image Retrieval (RBIR) is an image retrieval approach which focuses on contents from regions of images. This approach applies image segmentation to divide an image into discrete regions, which if the segmentation is ideal, it corresponds to objects. Thus the capture of region is improved so as to enhance the indexing and retrieval performance and also to provide a better similarity distance computation. During image segmentation, a modified k-means algorithm for image retrieval is developed where hierarchical clustering algorithm is used to generate the initial number of clusters and the cluster centers. In addition, during similarity distance computation, object weight based on object's uniqueness is introduced. Therefore considering images based on regions using RBIR allows the users to pay more attention to regional properties that may better characterize objects which are also made up of local regions. This strategy is able to better reflect the characteristics of the images from the perspective of image regions and objects. The advantages of the proposed method, 1)An improvement in image segmentation accuracy, especially for simple images 2) An improvement during similarity distance computation by using the parameter of object uniqueness into consideration.

P.Abinaya, IIDr.J.Sutha, **"Effective Feature Selection For High Dimensional Data using Fast Algorithm",**International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014), Vol. 2 Issue Special.1 Jan-March 2014.Feature subset clustering is a powerful technique to reduce the dimensionality of feature vectors for text classification. In this paper, we propose a similarity-based self-constructing algorithm for feature clustering with the help of K-Means strategy. The words in the feature vector of a document set are grouped into clusters, based on similarity test. Words that are similar to each other are grouped into the same cluster, and make a head to each cluster data sets. By the FAST algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data. The user need not specify the number of extracted features in advance, and trial-and-error for determining the appropriate number of extracted features can then be avoided. Experimental results show that our FAST algorithm implementation can run faster and obtain better-extracted features than other methods. The advantage of the FAST algorithm Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with each other. 2. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

K.vidyasagar, A.Bhujangarao, T.Madhu, **"Brain Tumor Detection and its Area Estimation in MRI Images using Pillar k-means algorithm"**, International Journal of Engineering Sciences Research-IJESR ,Vol 05, Article 03345; March 2014.In this research is detect the brain tumour from CT or MRI scan. CT or MRI scan, when directed into intracranial cavity, produces a complete image of the brain. This image can be examined visually by the physician or fed to a computer for detection & diagnosis of brain tumour. A new approach is proposed to detect the Brain tumour based on segmentation using Pillar K-means algorithm. At the end of the process the tumour is extracted from the MR image and its exact position and shape are determined. The stage of the tumour is displayed based on the amount of area (size and shape) calculated from the cluster. The proposed Pillar K-Means algorithm has shown better results than the other methods and its able to optimize computation time, improved the precision, enhanced quality of image segmentation.

Kaijian XIA ,Yue WU, Xiaogang REN, Yong JIN, **"Research in Clustering Algorithm for Diseases Analysis"**, International Journal of Advanced Research in Computer

Science and Software Engineering. Volume 48 Issue 7, July 2013.In this paper is analysis Cluster Algorithm-Fuzzy C-Means Algorithm (FCM).The clustering algorithm plays a very important role in the applications of medical analysis, it can effective analysis the log of disease. It can accurately analyse the characteristics of various diseases, thus providing accurate basis for the doctor's diagnosis. Fuzzy C-Means cluster Algorithm (FCM) produce the better result on the traditional algorithm. Experimentally, Parkinson's disease is effectively analysis, analysis accurate result and save the more time in the analytical process.

S.Saravanakumar, S.Rinesh, **"Effective Heart Disease Prediction using Frequent Feature Selection Method"**International Journal of Innovative Research in Computer and Communication Engineering ,Vol.2, Special Issue 1, March 2014.This research paper proposed a frequent feature selection method for Heart Disease Prediction. Good performance of this method comes from the use of the fuzzy measure and the relevant nonlinear integral. The none additively of the fuzzy measure reflects the importance of the feature attributes as well as their interactions. The main purpose of frequent feature selection is to reduce the number of features used in classification while maintaining acceptable classification accuracy. The main advantage of this method is that it produces a hierarchy of feature subsets with the best selection for each dimension. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. Clustering the objects which have similar meaning, the proposed approach improves the accuracy and reduces the computational time.

EshaSarkar, C.H Sekhar, **"Organizing Data in Cloud using Clustering Approach"** International Journal of Scientific & Engineering Research, Volume 5, Issue 5, May-2014.In this research is Organizing Data in Cloud Environments using Clustering approaches and effectively retrieved data in Cloud Environments. Cloud computing is the latest technology that delivers computing resources as a service such as infrastructure, storage, application development platforms, software etc. Cloud computing is gaining popularity and now-a-days it is on the boom. Huge amount of data is stored in the cloud which needs to be retrieved efficiently. The retrieval of information from cloud takes a lot of time as the data is not stored in an organized way. Data mining is thus important in cloud computing. We can integrate data mining and cloud computing (Integrated Data Mining and Cloud Computing– IDMCC) which will provide agility and quick access to the technology. The integration should be so strong that it will be able to deal with increasing production of data and will help in efficient mining of massive amount of data. In this paper, we provide brief description about cloud computing and clustering techniques.

Then, it also describes about cloud data mining. This paper proposes a model that applies hierarchical clustering algorithm in the data storage.

Monika Jain and Dr. S.K.Singh, **"An Experimental Study on Content Based Image Retrieval Based On Number of Clusters Using Hierarchical Clustering Algorithm"** International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.7, No.4 (2014).This paper discusses the image retrieval based on Number of Cluster(NC) which is evaluated using hierarchical agglomerative clustering algorithm (HAC). In this paper, we determine the optimal number of clusters using HAC applied on RGB images and validate them using some validity indices. Based on number of clusters, we retrieve set of images. These cluster values can be further used for divide and conquer technology and indexing for large image dataset. An experimental study is presented on real data sets.Nowadays the content based image retrieval (CBIR) is becoming a source of exact and fast retrieval. CBIR presents challenges in indexing, accessing of image data and how end systems are evaluated. Data clustering is an unsupervised method for extraction hidden pattern from huge data sets. Many clustering and segmentation algorithms both suffer from the limitation of the number of clusters specified by a human user. The hierarchical agglomerative clustering algorithm will become good approach for efficient content based image retrieval system for very large dataset.

Mugdha Jain, Chakradar verma, **"Adapting k-means for clustering in Big Data"** International Journal of Computer Science,vol-101,no-1,2014.Big data if used properly can bring huge benefits to the business, science and humanity. The various properties of big data like volume, velocity, variety, variation and veracity render the existing techniques of data analysis ineffective. Big data analysis needs fusion of techniques for data mining with those of machine learning. The k-means algorithm is one such algorithm which has presence in both the fields. This paper describes an approximate algorithm based on k-means. It is a novel method for big data analysis which is very fast, scalable and has high accuracy. It overcomes the drawback of k-means of uncertain number of iterations by fixing the number of iterations, without losing the precision.

Akshay S. Chavan,**"Survey on Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data",**International Journal of Advance Research inComputer Science and Management StudiesVolume 2, Issue 12, December 2014.Feature selection

involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. In this paper, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data.

The algorithm involves

(i) removing irrelevant features,

(ii) Constructing a minimum spanning tree from relative ones, and

(iii) Partitioning the MST and selecting representative features.

In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. Extensive experiments are carried out to compare FAST and several representative feature selection algorithms, namely, FCBF, Relief-F, CFS, Consist, and FOCUS-SF.

III. METHODOLOGY

WBCD, WDBC, UMC-USA, IBCDB dataset are available from UCI Breast Cancer dataset repository. All these dataset are used for effective and efficient data retrieval. This paper used these datasets. This dataset contains several attributes. The Breast Cancer attributes is represent many patient details such as Patient ID, Diagnosis, radius, texture, area, smoothness, compactness, concavity, concavity points etc….

In the FAST algorithm i mainly focused on four Breast Cancer data sets attributes, because these typeof attributes are very important in Effective and Efficient fast data retrieval.

The subset selection algorithm identifying accuracy and correlation measures from selected attributes. The attributes are necessary to increase prediction analysis for fast data retrieval research.

The subset selection filter methods are very important in effective and efficient data retrieval; any defects in these kinds of attributes will affect data efficiency, difficulty measure.

The difficulty measure is related to the difficulty of the program to write or understand the result. Discussion part describes the overall accuracy and correlation measure this attributes.

A Fast Clustering based subset Selection algorithm (FAST)

Many feature subset selection methods have been studied for machine learning applications. Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible.

Ineffective at removing and redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multi-class problems, but still cannot identify redundant features. Irrelevant features do not contribute to the predictive accuracy.

Some subset selection methods can effectively eliminate irrelevant features but fail to handle redundant features. Irrelevant features, redundant features also affect the speed and accuracy of learning algorithms.

My Research focus on this type of problem. The problem addressed in this research is the Effective and Efficient Correlation Based Subset Selection Algorithm for Fast Data Retrieval.

The proposed FAST algorithms reduce the time complexity and select the best features in the high dimensional data sets. In this algorithm was designed to be fast and effective.

The FAST algorithm mainly designed for improving the accuracy in fast data retrieval. In this algorithm use attribute selection method which is the process of selecting a subset of relevant features for use best cluster formation in Breast Cancer data sets.

IV. IMPLEMENTATION

Implemented algorithm mainly focused on four Breast Cancer data sets attributes, because these types of attributes are very important in analyzing efficient and effective fast data retrieval and identifying accurate value, measure correlation for valid attributes feature selection.

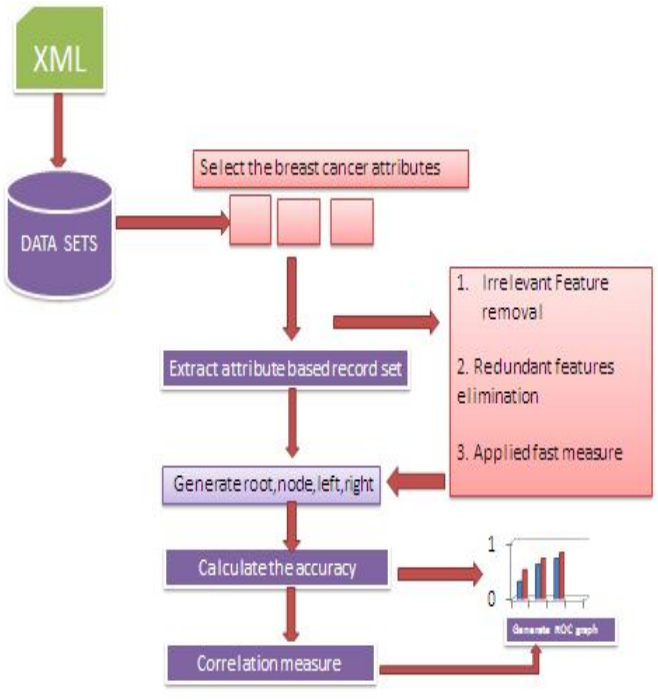
These four kinds of data set attributes are selected through filtering process.

The attribute are thirty three are classified three categories level like Mean, Standard Error and worst.

Which attributes I want select these attribute are fast clustering and predict the accuracy and correlation measure. The implementation steps are given below.

Step 1: Load the Brest Cancer Data set from UCI data repository.
Step 2: Select the attributes.
Step 3: upload the data set in SQL data base and save the data set.
Step 4: Then Fast -Measure is applying, first load the data set.
Step 5: Irrelevant Features are removal.
Step 6: Apply redundant features elimination.
Step 7: Then ApplyFast Clustering (Minimum Spanning Tree) Method. The selected attributes result is display.
Step 8: Then accuracy is calculated.
Step 9: next correlation is measured.
Step 10: Compare the attribute selection result with overall results of the attributes.
Step 11: Then plot the ROC accuracy Graph and ROC correlation graph.
Step 12: The ROC accuracy graph showed the average result.
Step 13: The ROC correlation measure graph showed correlated ranges in selected attributes.

System Architecture



3.2 Figure System Architecture

PROPOSED MODEL PERFORMANCE EVALUATION

The performance of proposed subset selection algorithm was mainly designed for improving the accuracy in high dimensional data; In this algorithm I reduced noisy data from high dimensional data and selected best feature in target class using the Filter subset selection method. We propose, A Fast clustering bAsed feature Selection algoriThm (FAST) for an effective and efficient Data retrieval. Where we perform a different Correlation measures on the retrieved Dataset in order to have an accurate and valuable Data. Fast Correlation-based Filter method addresses explicitly the correlation between features. It first ranks the features according to their mutual information with the class to predict, and remove those which mutual information is lesser than a threshold. And this proposed algorithm solve the low speed and given better accuracy of fast data retrieval.

Table 1: selected attribute's Description

S.NO	ATTRIBUTE	DESCRIPTION
1	**Patient id**	Patient ID is describe the identification number of patient.
2	**Radius**	It is represent range of radius in breast cancer.
3	**Perimeter**	The Perimeter represent the outside enclosing a shape measure.
4	**Area**	This type attribute represent the size of the surface. And the tumour extract position is identified.
5	**Texture-SE**	It represents range of texture standard error.
6	**Perimeter-SE**	It represents range of perimeter standard error.
7	**Worst texture**	It represents the range of worst texture and tumor shape bad.
8	**Worst area**	It represents the range of worst area size.

In this section we present the experimental results in terms of the proportion of selected features, the time to obtain the feature subset, the classification accuracy, and the selected attributes record.

For the purpose of exploring the statistical significance of the results, we performed a generally, all the subset selection algorithms achieve significant reduction of dimensionality by selecting only a small portion of the original features. FAST on average obtains the best proportion of selected features.

Accuracy

Accuracy is also referred as "correct classification rate" and is measured by taking the selected attributes separate accuracy value is correctly prediction to the total prediction made by the effective and efficient retrieval accuracy is predict and is formulated.

$$\text{Accuracy} = \frac{\text{Accepted value + my measurement}}{\text{Accepted value}} * 100$$

= The percentage of prediction that is correct

Fast Correlation Measure

The fast correlation measure is identifying relationship between selected data attributes. Calculated formula as:

x,y is data set attributes.

$$\text{Correlation } (r) = \frac{n\,(\sum xy) - (\sum x)(\sum y)}{\sqrt{[(\sum x^2) - (\sum x)^2][(\sum y) - (\sum y)^2}}$$

Table 2: Result of A Fast clustering based Feature Subset Selection Algorithm for Selected Attributes in WBCD, WDBC, UMC-USA and IBCDB

Dataset	Accuracy	F-Correlation
IBCDB	88%	0.315
UMC-USA	88.75%	0.493
WBCD	93.25%	0.598
WDBC	96.6%	0. 605

Chart Representation of Selected Attribute's Result

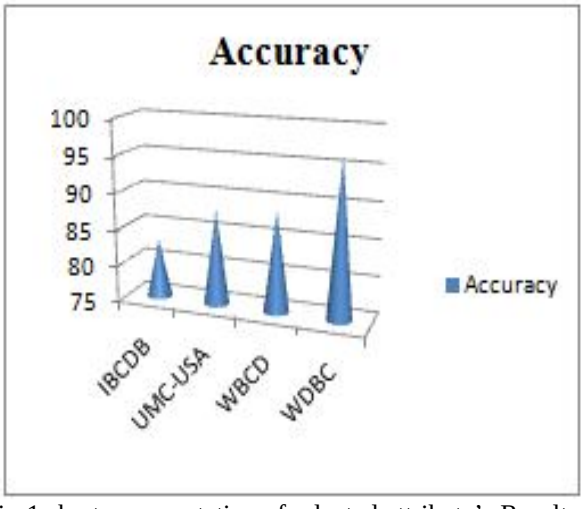


Fig:1 chart representation of selected attribute's Result of Accuracy

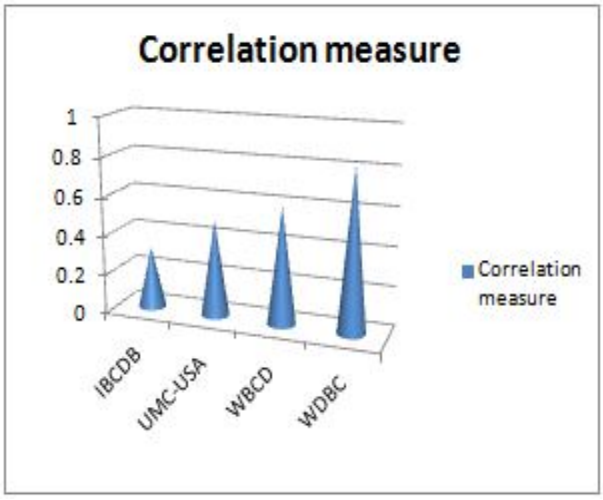


Fig: 2 chart representation of overall attributes of Correlation Result

V. CONCLUSION & FUTURE ENHANCEMENT

In this Research paper, we have presented Effective and Efficient Correlation based subset Selection Algorithm for fast data retrieval. In this research work mainly focused on the different types correlation measure for fast data retrieval, the correlation measurement is represent the relationship between selected features. The algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, (iii)partitioning the MST and selecting representative features, (iv)Get subset of features in the data, (v) Create a cluster and Remove the duplicates, and (vi)Collect the Valid features for the effective and efficient data retrieval. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. Performance is increase, increasing learning accuracy, and improving result comprehensibility for the fast retrieval process. Main benefit of proposed algorithm is high speed and low error. The motive of this process is increasing the accuracy level of classifiers and increased correlation measure.

Advantages

- Effectively remove irrelevant features and redundant feature.
- Easy to retrieve data.
- Easy to access the data or record set
- Can classify new instances rapidly
- Performance is good better than traditional feature subset selection algorithm.
- It produce the accurate result And measure correlation it's specify relationship between of data set attributes feature.

FUTURE ENHANCEMENT

- ✓ In this study future enhancement are (i) several of data set are used and increase the performance level, and study some formal properties of feature space. In future we can consider this external tool for fast retrieved data accuracy and that correlation measures.

REFERENCES

[1] Sherin Mary Varghese, M.N.Sushmitha **"Efficient Feature Subset Selection Techniques for High Dimensional Data"**International Journal of Innovative Research in Computer and Communication Engineering,Vol. 2, Issue 3, March 2014.

[2] R.Pavithra, J.Vinitha Grace, A.Arun Sethupathy Raja, V.Stalin, M.Ramakrishnan, **"Analysing the Efficiency of Data by Using Fast Clustering and Subset Selection Algorithm"**International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 3, March 2014.

[3] Qinbao Song, Jingjie Ni and Guangtao Wang **"A Fast Clustering-Based Feature SubsetSelection Algorithm for High Dimensional Data"** IEEE Transactions on Knowledge and Data Engineering vol: 25 no:1 year 2013.

[4] K. Revathi and T. Kalai Selvi**"Efficient and Effective Subset Selection Process based on Clustering Algorithm"**International Journal of Innovation and Scientific Research ISSN 2351-8014 Vol. 6 No. 1 Aug. 2014.

[5] J.K.Madhavi, G.Venkatesh Yadav, **"An Improved Fast Clustering method for FeatureSubset Selection on High-Dimensional Data clustering"**International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 10, October 2014.

[6] Dr. K.Sakthivel, R.Abinaya, I.Nivetha, R.Arun Kumar, **"Region Based Image Retrieval using k-means and Hierarchical Clustering Algorithms"**International Journal of Innovative Research in Science, Engineering and Technology ,ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014.

[7] P.Abinaya, IIDr.J.Sutha, **"Effective Feature Selection for High Dimensional Data using Fast Algorithm",** International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014), Vol. 2 Issue Special 1 Jan-March 2014.

[8] K.vidyasagar, A.Bhujangarao, T.Madhu, **"Brain Tumor Detection and its Area Estimation in MRI Images using Pillar k-means algorithm"**, International Journal of Engineering Sciences Research-IJESR ,Vol 05, Article 03345; March 2014 .

[9] Kaijian XIA ,Yue WU, Xiaogang REN, Yong JIN, **"Research in Clustering Algorithm for Diseases Analysis"**, International Journal of Advanced Research in Computer Science and Software Engineering. Vol: 48,Issue 7, July 2013.

[10] S.Saravanakumar, S.Rinesh, **"Effective Heart Disease Prediction using Frequent Feature Selection Method"**International Journal of Innovative Research in Computer and Communication Engineering ,Vol.2, Special Issue 1, March 2014.

[11] Esha Sarkar, C.H Sekhar, **"Organizing Data in Cloud using Clustering Approach",** International Journal of Scientific & Engineering Research, Volume 5, Issue 5, May-2014.

[12] Monika Jain and Dr. S.K.Singh, **"An Experimental Study on Content Based Image Retrieval Based On Number of Clusters Using Hierarchical Clustering Algorithm"** International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.7, No.4 (2014).

[13] Mugdha Jain,Chakradar verma, **"Adapting k-means for clustering in Big Data"** International Journal of Computer Science,vol-101,no-1,2014.

[14] Akshay S. Chavan,**"Survey on Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data",**International Journal of Advance Research inComputer Science and Management StudiesVolume 2, Issue 12, December 2014.

[15] Alan Jose, S.Ravi, M.Sambath, **"Brain Tumor Segmentation Using K-Means Clustering And Fuzzy C-Means Algorithms And Its Area Calculation",** International Journal of Innovative Research in Computer and Communication Engineering,Vol. 2, Issue 3, March 2014.