AN ASYMPTOTIC PERFORMANCE MEASUREMENT OF MACHINE LEARNING APPROACH USING DATAMINING TECHNIQUES

Rajeshwari. M

Dept of Computer Science and Engineering
Mother Terasa College of Engineering and Technology
Illupur, Pudukkottai
rajibtechit90@gmail.com

Harthy Ruby Priya. S

Dept of Computer Science and Engineering
Mother Terasa College of Engineering and Technology
Illupur, Pudukkottai
susaiharthy@gmail.com

Abstract— Data Mining is one of the knowledge discovery steps in database, in which modeling techniques are applied. In this research work, the performance analysis of classification algorithms like K - Means and FCM methods are applied for dealing with medical database for mining. To increase the efficiency of mining process, some preprocessing needs to be done to the data. Medicinal data mining methods are used to analyze the medical data information resources. The effort to develop knowledge and experience of frequent specialists and clinical selection data of patients collected in databases to facilitate the diagnosis process is considered a valuable option. Diagnose the Diabetes, Lung and Liver diseases are a significant and tedious task in medicine. For detecting a disease number of tests should be required from the patient. But using data mining technique the number of test should be reduced. This reduced test plays an important role in time and performance. This research work analyzes and performance study about how data mining techniques are used for predicting the diabetes, lung and liver diseases. In this work, use familiar two data mining algorithms and performance evaluation of different UCI Repository datasets like Diabetes, Liver Disorder and Lung Cancer data on the basis of Performance Measure and Cost Measure. Experimental results showed the good accuracy when applied to the adjust data.

Keywords— KNN, SVM, K-Means, K-Medoids, Fuzzy

I. INTRODUCTION

Data Mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

These include approaches based on splitting and merging such as ISODATA, randomized approaches such as CLARA, CLARANS, and methods based on neural nets, and methods designed to scale to large databases, including DBSCAN, BIRCH and ScaleKM. Among clustering formulations that are based on minimizing a formal objective function, perhaps the most widely used and studied is partition based algorithms like K-Means, and Fuzzy C-Means clustering.

However, the *k*-means algorithm has at least two major theoretic shortcomings:

- First, it has been shown that the worst case running time of the algorithm is super-polynomial in the input size.
- Second, the approximation found can be arbitrarily bad with respect to the objective function compared to the optimal clustering.

The *k*-means++ algorithm addresses the second of these obstacles by specifying a procedure to initialize the cluster centers before proceeding with the standard *k*-means optimization iterations. With the *k*-means++ initialization, the algorithm is guaranteed to find a solution that is O (log *k*) competitive to the optimal *k*-means solution.

Fuzzy c-means is an extension of k-means clustering. The major difference between the fuzzy c-means and k-means is that the later discovers hard clusters where a particular sample can belong to only one cluster while the former discovers soft clusters where a particular sample can belong to more than one cluster with certain probability. This belongingness of a data sample to the cluster is represented using membership values.

II. RELATED WORKS

A. D. Asir Antony Gnana Singh et al.

This paper presents a performance analysis on various clustering algorithm namely K-means, expectation maximization, and density based clustering in order to identify the best clustering algorithm for microarray data. Sum of squared error, log likelihood measures are used to evaluate the performance of these clustering methods. This paper conducted an empirical study on various clustering algorithms in order to observe their performance on gene expression data in terms of sum of squared error and log likelihood. In this empirical study, the performance of the clustering algorithms namely density based clustering, expectation maximization clustering and K-means clustering are evaluated on various gene expression data.

B. Pallavi et al.

This paper analyze the three major clustering algorithms: K-Means, Farthest First and Hierarchical clustering algorithm and compare the performance of these three major clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. The results are tested on three datasets namely Wine, Haberman and Iris dataset using WEKA interface and compute the correctly cluster building instances in proportion with incorrectly formed cluster. The result analysis shows that K-means algorithm performs well without inserting the principle component analysis filter as compared to the Hierarchical clustering algorithm and Farthest first clustering since it have less instances of incorrectly clustered objects on the basis of class clustering. Hierarchical clustering as compared to Farthest fast clustering gives better performance. Also this algorithm performs better after merging principle component analysis filter with it.

C. Neha D et al.

This paper mainly presents an overview of types of clustering techniques and some of the applications of data mining where clustering techniques can be applied. The main goal of clustering is to produce a good and high quality clusters that depends mainly on the similarity measure which has the ability to discover some or all hidden patterns and also make the analysis of data easy. he quality of clusters produced by clustering method is measured by its ability to discover some or all of the hidden patterns. It has been observed that, the most common type of clustering technique that has been used by different applications of data mining is the k-means clustering technique.

D. G.G.Gokilam et al.

In this paper we take diabetes and heart datasets relate with their matching fields then apply the classification algorithm in diabetes heart dataset in software tool finding weather people affected by diabetes are getting chance to get heart disease or not, output are evaluated as Tested Negative (No Diabetes), Tested Normal(Not affected), Tested High(affected). a new approach for efficiently predicting the diabetes_heart disease from some medical records of patients. Dataset has designed with matching attributes applied in classification algorithms like J48, Random Tree, Random Forest, REP, Naïve Bayesian algorithm. [4]

E. Gopala Krishna Murthy Nookala et al.

In this study, we have made a comprehensive comparative analysis of 14 different classification algorithms and their performance has been evaluated by using 3 different cancer data sets. The results indicate that none of the classifiers outperformed all others in terms of the accuracy when applied on all the 3 data sets. Most of the algorithms performed better as the size of the data set is increased. We recommend the users not to stick to a particular classification method and should evaluate different classification algorithms and select the better algorithm.

III. PROBLEM DESCRIPTION

A. *Cluster Analysis*

The objective of cluster analysis is the classification of objects according to similarities among them, and organizing of data into groups. Clustering techniques are among the unsupervised methods, they do not use prior class identifiers. The main potential of clustering is to detect the underlying structure in data, not only for classification and pattern recognition, but for model reduction and optimization. Different classifications can be related to the algorithmic approach of the clustering techniques.

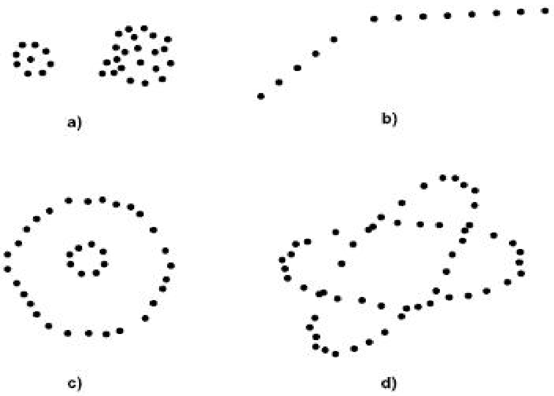

Fig3. 1 Clusters of different shapes and dimensions in R2.

$$\mathbf{U}=\begin{bmatrix} \mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,c} \\ \mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,c} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{N,1} & \mu_{N,2} & \cdots & \mu_{N,c} \end{bmatrix}$$

K-means algorithm aims at minimizing an objective function, namely sum of squared error (SSE). SSE is defined as

$$\mathbf{E} = \sum_{i=0}^{k} \sum_{p \in C_i} |p - m_i|^2 \quad \text{--------}\rightarrow (1)$$

Where E is sum of the square error of objects with cluster means for k cluster. p is the object belong to a cluster C_i and mi is the mean of cluster C_i .The time complexity of K-means is O(t*k*n) where t is the number of iterations, k is number of clusters and n is the total number of records in dataset.

Input is k is the number of clusters, D is input data set
Output is k clusters.

1. Randomly choose k objects from D as the initial cluster centers.
2. Repeat
3. Assign each object from D to one of k clusters to which the object is most similar based on the mean value of the objects in the cluster.
4. Update the cluster means by taking the mean value of the objects for each of k cluster.
5. Until no change in cluster means/ min error E is reached.

K-means++ (David Arthur et. Al., 2007) is another variation of k-means; a new approach to select initial cluster centers by random starting centers with specific probabilities is used.

The steps used in this algorithm are described below:

1. Step 1: Choose first initial cluster center c_1 randomly from the given dataset X.
2. Step 2: choose next cluster center $c_i = x_j \in X$ with probability p_i where; denote the shortest distance from x to the closest center already chosen.
3. Step 3: Repeat step2 until k cluster centers are chosen.
4. Step 4: After initial selection of k cluster centers, apply k-means algorithm to get final k clusters.

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In **fuzzy clustering** (also referred to as **soft clustering**), data elements can belong to more than one cluster, and associated with each element is a set of membership levels.

$$JU = u\ i, j \in [0,1],\ i = 1,\ldots.,n,\ j = 1,\ldots,c \text{-----}\rightarrow (2)$$

where each element u_{ij} tells the degree to which element x_i belongs to cluster c_j . Like the k-means algorithm, the FCM aims to minimize an objective function. The standard function is:

$$u_k(x) = \frac{1}{\Sigma_j \left(\frac{d(center_k, x)}{d(center_k, x} \right)^{2/(m-1)}}$$

which differs from the k-means objective function by the addition of the membership values u_{ij} and the fuzzifier m. The fuzzifier m determines the level of cluster fuzziness. A large m results in smaller memberships u_{ij} and hence, fuzzier clusters. In the limit m = 1, the memberships u_{ij} converge to 0 or 1, which implies a crisp partitioning. Any point *x* has a set of coefficients giving the degree of being in the *k*th cluster $w_k(x)$. With fuzzy *c*-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$C_k = \boldsymbol{\Sigma}_x\, w_k\,(x)x \,/\, \boldsymbol{\Sigma}_x\, w_k\,(x) \quad \text{-------------}\rightarrow (3)$$

The degree of belonging, $w_k(x)$, is related inversely to the distance from *x* to the cluster center as calculated on the previous pass.

IV. METHODOLOGY

A. K-means Algorithm

The *k*-means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. Cluster similarity is measured in regard to the *mean* value of the objects in a cluster, which can be viewed as the cluster's *centroid* or *center of gravity*.

B. Algorithm

Given the data set X, choose the number of clusters 1 < *c* < *N*.
Initialize with random cluster centers chosen from the data set.
Repeat for *l* = 1*;* 2*;*
Step 1 Compute the distances

$$D_{ik}^2 = (x_k - v_i)^T (x_k - v_i), \qquad 1 < i < c, \qquad 1 < k < N.$$

Step 2 Select the points for a cluster with the minimal distances, they belong to that cluster.
Step 3 Calculate cluster centers

$$v_i^{(l)} = \frac{\Sigma_{j=1}^{N_i} x_i}{N_i}$$

Until

$$\prod_{k=1}^{n} max \left| v^{(l)} - v^{(l-1)} \right| \neq 0$$

Ending Calculate the partition matrix

C. Euclidean Distance

The Euclidean distance, data vector p and centroid q is computed as

$$d(p,q) = \sqrt{\sum_{k=1}^{n} (q_{ik} - p_{ik})^2}$$

------------→ (4)

E. Cluster Validity Measure

Cluster validity refers to the problem whether a given fuzzy partition fits to the data all. The clustering algorithm always tries to find the best fit for a fixed number of clusters and the parameterized cluster shapes.

- Starting with a sufficiently large number of clusters, and successively reducing this number by merging clusters that are similar (compatible) with respect to some predefined criteria. This approach is called *compatible cluster merging*.
- Clustering data for different values of *c*, and using *validity measures* to assess the goodness of the obtained partitions. This can be done in two ways:
    - The first approach is to define a validity function which evaluates a complete partition. An upper bound for the number of clusters must be estimated (c_{max}), and the algorithms have to be run with each $c \in \{2; 3; : c_{max}\}$. for each partition, the validity function provides a value such that the results of the analysis can be compared indirectly.
    - The second approach consists of the definition of a validity function that evaluates individual clusters of a

Different scalar validity measures have been proposed in the literature, none of them is perfect by oneself, and therefore we used several indexes in our Toolbox, which are described below:

1. ***Partition Coefficient (PC):*** measures the amount of "overlapping" between clusters.

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^2$$

-------→ (5)

Where μ_{ij} is the membership of data point *j* in cluster *i*. The disadvantage of PC is lack of direct connection to some property of the data themselves. The optimal number of cluster is at the maximum value.

2. Classification Entropy (CE)**:** it measures the fuzzyness of the cluster partition only, which is similar to the Partition Coefficient.

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}log(\mu_{ij}),$$

3. Partition Index (SC): is the ratio of the sum of compactness and separation of the clusters. It is a sum of individual cluster validity measures normalized through division by the fuzzy cardinality of each cluster.

$$SC(c) = \sum_{i=1}^{c}\frac{\sum_{j=1}^{N}(\mu_{ij})^m||x_j - v_i||^2}{N_i\sum_{k=1}^{c}||v_k - v_i||^2}$$

SC is useful when comparing different partitions having equal number of clusters. A lower value of *SC* indicates a better partition.

4. Separation Index (S): on the contrary of partition index (SC), the separation index uses a minimum-distance separation for partition validity.

$$S(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^2||x_j - v_i||^2}{N min_{i,k}||v_k - v_i||^2}$$

5. Xie and Beni's Index (XB): it aims to quantify the ratio of the total variation within clusters and the separation of clusters.

$$XB(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^m||x_j - v_i||^2}{N min_{i,j}||x_j - v_i||^2}$$

The optimal number of clusters should minimize the value of the index.

6. Dunn's Index (DI): this index is originally proposed to use at the identification of "compact and well separated clusters". So the result of the clustering has to

$$DI(c) = min_{i\in c}\{min_{j\in c, i\neq j}\{\frac{min_{x\in C_i, y\in C_j}d(x,y)}{max_{k\in c}\{max_{x,y\in C}d(x,y)\}}\}\}$$

7. Alternative Dunn Index (ADI)**:** the aim of modifying the original Dunn's index was that the calculation becomes more simple, when the dissimilarity function between two clusters ($min_{x\epsilon\ Ci, y\epsilon\ Cj}$ $d(x, y)$) is rated in value from beneath by the triangle-non equality:

$$d(x,y) \geq |d(y,v_j) - d(x,v_j)|$$

$$ADI(c) = min_{i\in c}\{min_{j\in c, i\neq j}\{\frac{min_{x_i\in C_i, x_j\in C_j}|d(y,v_j) - d(x_i,v_j)|}{max_{k\in c}\{max_{x,y\in C}d(x,y)\}}\}\}$$

Dataset	Algorithm	PC	CE	SC	S	XB	DI	ADI
Diabetes	K-Means	1	NaN	0.726	0.001	3.17	0.648	0.563
	K-Means++	1	NaN	0.723	0.001	Inf	0.648	0.550
	FCM	0.80	0.331	0.954	0.001	2.61	0.648	0.547

Table 3.1 Validity measure

F. K-Medoids

K-medoid is a classical partitioning technique of clustering that clusters the data set of *n* objects into *k* clusters known a priori. A useful tool for determining *k* is the silhouette. It is more robust to noise and outliers as compared to *k*-means because it minimizes a sum of pair wise dissimilarities instead of a sum of squared Euclidean distances. A medoid can be defined as the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal. i.e. it is a most centrally located point in the cluster.

Steps:

1: Arbitrarily choose k data items as the initial medoids.

2: Assign each remaining data item to a cluster with the nearest medoid.

3. Randomly select a non-medoid data item and compute the total cost of swapping old medoid data item with the currently selected non-medoid data item.

4. If the total cost of swapping is less than zero, then perform the swap operation to generate the new set of k-medoids.

5. Repeat steps 2, 3 and 4 till the medoids stabilize their locations.

G. Fuzzy C-means (FCM)

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. Straightly speaking, this algorithm works by assigning membership to each data point correspoinding to each cluster center on the basis of distance between the cluster and the data point. The algorithm is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad , \quad 1 \le m < \infty$$

where *m* (the Fuzziness Exponent) is any real number greater than 1, *N* is the number of data, *C* is the number of clusters, u_{ij} is the degree of membership of x_i in the cluster *j*, x_i is the *i*th of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}$$

$$= \frac{1}{\left(\frac{\|x_i - c_j\|}{\|x_i - c_1\|}\right)^{2/(m-1)} + \left(\frac{\|x_i - c_j\|}{\|x_i - c_2\|}\right)^{2/(m-1)} + \cdots + \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{2/(m-1)}}$$

centers *k*.

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

The iteration will stop when $\max_{ij}\left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas *k* are the iteration steps. This procedure converges to a local minimum or a saddle point J_m

V. EXPERIMENTAL RESULTS

A. Pima Indian Diabetes Dataset

World Health Organization (WHO) report had shown a marked increase in the number of diabetics and this trend is expected to grow in the next couple of decades. In the International Diabetes Federation Conference 2003 held in Paris, India was labeled, as "Diabetes Capital of the World," as of about 190 million diabetics worldwide, more than 33 million are Indians.

The Pima Indian diabetes data set is taken from the UCI machine learning repository [18].The data set has 768 instances with two class problems to test.

Dataset	Number of Objects	Number of Attributes	Number of Clusters
Pima Indian Diabetes	768	8	2

Table 5.1 Pima Indian dataset

Class Distribution: Class value 1 is interpreted as "tested positive for diabetes"
Class Value: 0 - Number of instances - 500
Class Value: 1 - Number of instances – 268

No	Attribute	Description	Missing Values
1	pregnant	Number of times pregnant	110
2	glucose	Plasma glucose concentration (glucose tolerance test)	5
3	pressure	Diastolic blood pressure (mm Hg)	35
4	triceps	Triceps skin fold thickness (mm)	227
5	insulin	2-Hour serum insulin (mu U/ml)	374
6	mass	Body mass index (weight in kg/(height in m)^2)	11
7	pedigree	Diabetes pedigree function	0
8	age	Age (years)	0
9	diabetes	Class variable (test for diabetes)	0

Table 5.2 Description of Dataset

VI. CONCLUSION

Cluster analysis is one of the major tasks in various research areas. The clustering aims at identifying and extract significant groups in underlying data. Thus based on a certain clustering criterion the data are grouped so that data points in a cluster are more similar to each other than points in different clusters. Since clustering is applied in many fields, a number of clustering techniques and algorithms have been proposed and are available in literature. In the proposed system to analysis the major clustering algorithms such as K-Means, K-Medoids and Fuzzy C-Means with Euclidean distance measure by using UCI dataset.

It illustrates the efficiency of clustering algorithm with its validity measures. It shows the Fuzzy C-Means clustering algorithm had better than other clustering algorithms. The experimental result shows the performance of the Fuzzy C-Means algorithm was improved significantly.

VII. REFERENCES

[1]D. Asir Antony Gnana Singh, A. Escalin Fernando,, E. Jebamalar Leavline by Performance Analysis on Clustering Approaches for Gene Expression Data.

[2] Pallavi , Sunila Godara by A Comparative Performance Analysis of Clustering Algorithms, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 1, Issue 3, pp.441-445.

[3] Neha D, Ballari, B.M. Vidyavathi by A Survey on Applications of Data Mining using Clustering Techniques. International Journal of Computer Applications (0975 – 8887)Volume 126 – No.2, September 2015

[4]G.G.Gokilam, Dr K.Shanthi by Performance Analysis of Various Data mining Classification Algorithms on Diabetes Heart dataset. An international journal of advanced computer technology March - 2016 (Volume-V, Issue-III) ISSN:2320-0790.

[5] Tanvi Sharma, Anand Sharma, Prof. Vibhakar Mansotra by "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data", International Journal of Innovative Research in Computer and Communication Engineering Certified Organization) Vol. 4, Issue 6, June 2016.