IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

An Efficient K-Means Clustering Algorithm Using Euclidean Distance Techniques

N.Suresh[1], Dr.K.Arulanandam[2]

Research Scholar, Dept. Of Computer Applications, Govt. Thirumagal Mills College, Gudiyattam[1]

Research Supervisor, Asst. Prof. & Head, Dept. Of Computer Applications, Govt. Thirumagal Mills College, Gudiyattam[2]

Abstract: The quality of the healthcare and treatment outcome relies heavily on Data Mining field to exchange information for accurate detection of the life threatening causes. The quality of health care system can be improved by employing an intelligent system via the Data Mining Techniques. Data Mining Techniques are used to reveal the hidden patterns from the vast collection of patient's data. Analysis of the data uses techniques and statistical measures to get insight into vast patient's information and predict the possible causes for the health issues and its impact on individual patients. Earlier analysis of the health related issues will reduce psychological stress and gives enormous time to identify the specialist in the respective field to acquire pre-determined treatment.

Keywords: Data Mining, Agglomerative, Clustering, K-Means, K-Medoids, Dataset in Excel.

I. INTRODUCTION

1.1 DATA MINING

Data Mining is the abstraction of useful data from concealed knowledge of valuable information from large databases. It is the relevant method of searching legitimate, novel, potentially beneficial and in the end understandable templates in records. It uses a variance of technique to identify chunks of information for determination in the available dataset and deriving these in such way that they can be used in various areas. The term is contradicted because the target is to extract the knowledge from enormous amounts of data. It can be applied to any form of information processing system as well as any decision support system. The real mining venture is the partially automatic or computerized evaluation of huge quantities of information to extract unknown, exciting styles consisting of agencies of records, unusual statistics and dependencies. This generally includes using database strategies. The output patterns are a kind of summary of the input data, and may be used in future analysis. The statistics mining step may become aware of multiple groups within the records that could be used to attain greater correct prediction results with the assistance of a support machine. The related terms "Data degrading, Data fishing and Data snooping" refers to the use of data mining statistics to a large populated data. These methods can be used in creating new axioms to test against the larger data. This testing finds a model which explains the data by creating a hypothesis and then test the hypothesis against the data. It is usually verified by searching the data sample. As the dataset has grown in size and complexity manual process becomes complicated and ineffective. So the implementation of automated process aided by the mining techniques is put forth.

1.2 DATA MINING TECHNIQUES

Key Techniques are:

Association

Association or relation is a transparent operation which co-relates more than one that belongs to the same type to identify the patterns. For example, Associating the people with their buying habits. People who buy Bread also buy Jam and vice versa.

The rule is in the form,

X=>Y
Where X and Y are the sets of items

----- (1)

Where X is 'antecedent' and Y is 'consequent' .

To select the rules from the set, some conditions on various measures are used. The existing measures are 'support and confidence'.

Support

Support indicates the frequency of the item set that resides in the set.

The support of X with respect to T is explained as the proportion of transactions t in the dataset that consist the item set X.

Support is denoted as:

$$\text{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|} \quad \text{----- (2)}$$

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

Confidence

Confidence indicates how frequent the rule is supposed to be positive.

The confidence of a rule X=> Y , with respect to T, is the proportion of the transactions that includes X which also includes Y.

Confidence is defined as:

$$\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X) \quad \text{--- (3)}$$

Process

These rules satisfies minimum support and minimum confidence specified by the user at the similar time.

The two steps in Association rule,

(i) Minimum support is activated to find all frequent item sets in a database.

(ii) Minimum confidence is activated to the frequent item sets to form rules.

Some other measures

- All-confidence
- Collective strength
- Conviction
- Leverage
- Lift

II. LITERATURE REVIEW

2.1 INTRODUCTION

Data Mining is applied in many areas for its automated process. There exist a vast and variant techniques used now-a-days. The techniques should handle big data also. Some algorithms only work on the small dataset. So before using it one should analyse the algorithms and its Pros and Cons. This chapter gives the detailed assess of the paper which demonstrates the techniques elaborately.

2.2 REVIEW OF LITERATURE

Ashfaq Ahmed . K *et al* [1] , compares the existing classification technique for better performance. Classification technique comes under the machine learning is used to sort based on the label. These technique is also used to analyse the life threatening disease. It has also been noticed that classification when combined with learning procedures can be used effectively to gain the accuracy of prediction.

Madhuri V. Joseph *et al* [2] did a relative study on various techniques. Normally knowledge Mining is used to get the patterns to search the underlying matter for analytical process. There are vast amount of techniques and all those cannot solve the problem so it is risk job to decide which algorithm should be used where. So this paper analyse all the popular algorithm with the dataset and list the results. Since the business economy grows very faster it's duty to select the robust algorithm which is suitable for particular field.

This paper analyse the most effective techniques and its Pros and Cons are listed.

III. METHODLOLOGY

3.1 EXISTING SYSTEM

Existing work

The existing work categorize the Tuberculosis affected patients from the normal patients and get the most desirable factors which influence the disease.

Data set

The work was conducted on 600 patients from "Masih -e-Daneshvari Tuberculosis Research Centre" during 2015-2016.

Flow of Work

It consists of two main Phases.

Pre-processing (Clustering);

In the pre-processing steps are performed in order to find most important factors by applying 'K- means' algorithm to cluster features.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

Flow Chart

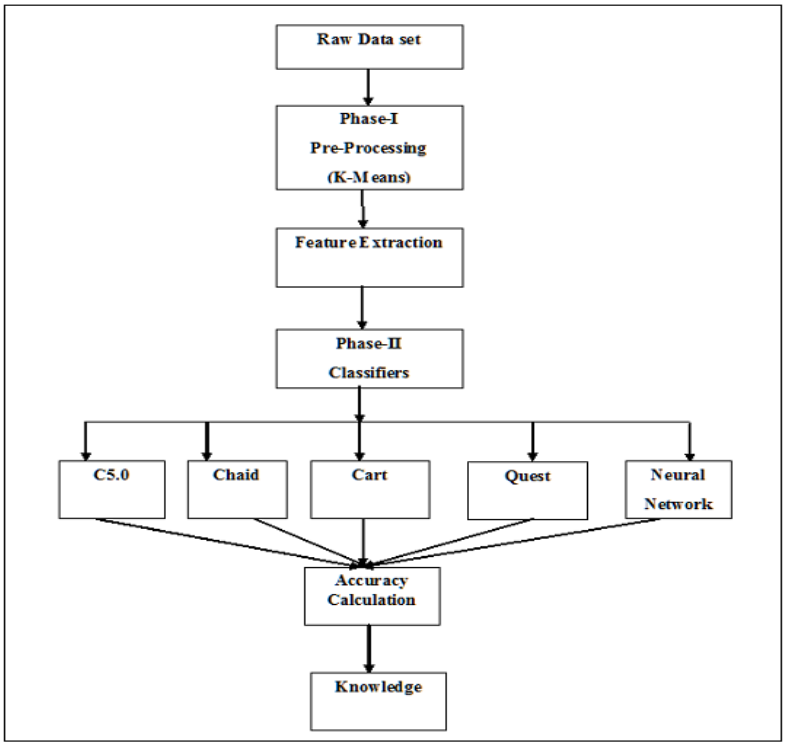

Figure 3.1 Flowchart of Existing system

Techniques used in Existing work

Pre-Processing

The data processing is applied to remove the inconsistencies and incompleteness with the record. There are numerous techniques in the Pre –Processing such as " Remove the undesirable attribute, Picking the most desirable attribute, Filling the Missed Values, Discretization, Normalization" etc.

Technique used for Pre-processing

The aim is to remove items which has no value for laboratory and demographic characteristics. For this, 'K-means' algorithm is used. The aim of the indicator is to maximize the intra object distance while minimizing inter object distance.

The main objective is to get the common factors between tuberculosis patients. Initially 22 factors are considered before Pre-Processing.

Methods Used

Clustering

It was designed and introduced in anthropology by 'Kroeber and Driver' in 1932 and in psychology by 'Zubin' in 1938. And 'Robert Tryon' initiated in 1939 and was used by 'Cattell' in 1943 for classification in personality psychology.

'Cluster analysis or clustering' is to make the objects in the way that objects in the same group are more identical than in other . It is a useful technique of distribution and finding co-relation in the data. The aim is to uncover the dense and scarce regions in a set. It is necessary to analyse the principle of minimizing the operations.

Numerical and Categorical data type

Clustering can handle numerical & categorical type. For clustering numerical type the geometric properties or distance function are applied to define the interval between the points.

For Clustering categorical type, the distance functions are not used.

Paradigms

The two main strategies in clustering are: (i) Hierarchical clustering,
(ii) Partitioning Clustering.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

The strategies differ among themselves in handling different attributes, accuracy and the capability to use inconsistent data.

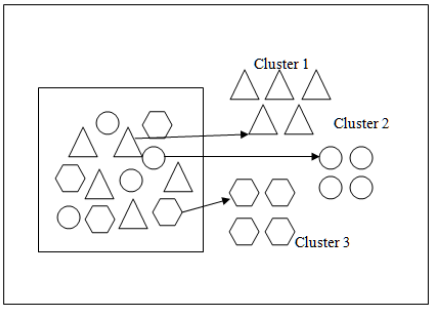

Figure 3.2 Clustering heterogeneous data

Figure 3.2 shows how the data are clustered. Three clusters are fixed on the basis of the shape.

(i) Hierarchical clustering

This method do a consecutive partitions and each is build into the next one in the order. So it creates a hierarchy of groups.

Categories: It is of the types: (a) Agglomerative,
(b) Divisive.

Agglomerative - Agglomerative clustering starts with many clusters and terminates in a group which is built under a single which contains all the records at the top. At each stage, this method joins together the two groups that are closest together. It is referred as a 'Bottom-Up' approach.

Divisive - Divisive Clustering starts with all the objects in one cluster and it splits into small pieces. It is referreded as a 'Top-Down' approach.

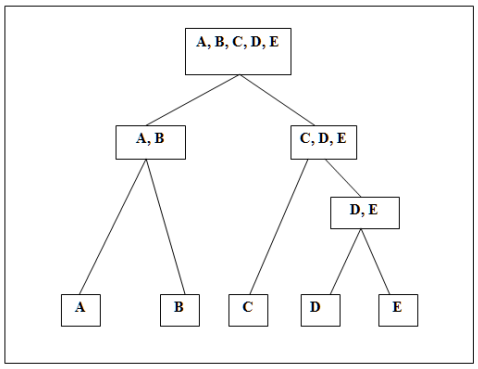

Figure 3.3 Hierarchical Clustering

Figure 3.3 shows the Hierarchical technique. It proceeds from root to leaf.

(ii) Partition Clustering

This method splits the set into a set of disjoint array. For given a set with M points, the procedure constructs N partitions, and each represents a set.

It splits into N groups by the following constraints:
(1) each contains at least single object,
(2) each belongs to exactly one set.

Categories: The two main categories are,
(a) K-Means- Each package is represented by the centre object,
(b) K-Medoids - Every package is defined by one of the objects that lies closest to the core.

K-Means algorithm

- The standard algorithm was first put forward d by 'Stuart Lloyd in 1957'.
- The term "k-means" was first utilized by James 'Mac Queen in 1967'.

K-means clustering aims to partition the given data points into k clusters in which each one belongs to the cluster with the nearest mean.This results leads to 'Voronoi cells'.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

Voronoi cell - There is a region contains all observations closer than to any other. These are called 'Voronoi cells'.
The Voronoi cell R_k , with the site P_k is the set of all points in dataset X, whose distance to P_k is not exceed than the other sites P_j, where j is any index from k.

$$R_k = \{x \in X \mid d(x, P_k) \leq d(x, P_j) \text{ for all } j \neq k\} \quad \text{-----(5)}$$

This is similar to the expectation-maximization for Gaussian distributions by an iterative refinement which is employed by both . But it differs from the expectation-maximization by that k-means clustering tends to get clusters with the almost equal shape, while the other mechanism allows to have variant shapes.

Also this algorithm has a loose relations with the k-nearest neighbor classifier, because of the k in the name. If 1-nearest neighbor is applied on the centers obtained by k-means then it is referred as "Nearest centroid classifier or Rocchio algorithm".

Choosing Number of Clusters

The algorithm is not designed to fix the number of sets so it relies upon the user to fix this in prior. If a group of people is taken based on the gender , the number should be defined as two. Similarly if a group of individuals were taken based upon home state the number should be defined based on all the possible way to be effective. So it is suggested to experiment with different values to define the value that best suits .

Distance Function

K-Means algorithm use only Euclidean distance.The 'Euclidean distance' works in straight-line to calculate the distance between two points in Euclidean space. In this distance, it becomes a metric space. This metric is referred as 'Pythagorean metric' in earlier . The norm is 'L2 norm or L2 distance'.

Euclidean Distance working principle

It is determined from the middle of the home cell to the middle of each surrounding cells. The interval between two points is computed by taking square root for the sum of the squares between the differences .

The distance between the two points x and y is taken as the measurement of the line segment which links them .

In Cartesian coordinates, if $x = (x_1, x_2,..., x_n)$ and $y = (y_1, y_2,..., y_n)$ are the points in Euclidean space, then the interval d from x to y or from y to x is given by the Pythagorean formula as,

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad \text{-----(6)}$$

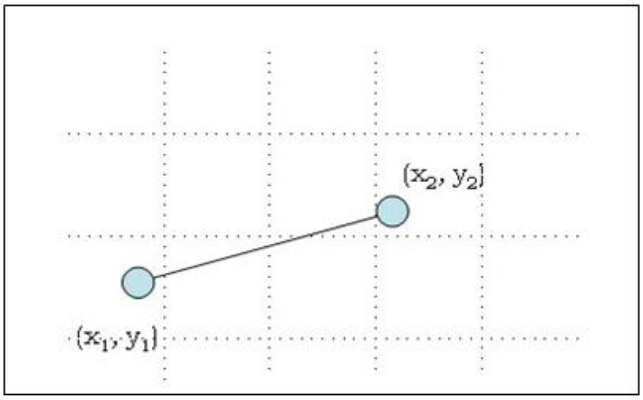

Figure 3.4 Euclidean Distance

Figure 3.4 shows the measurement of Euclidean distance in the space.
As an example, the measurement between points (2, -1) and (-2, 2) is given as,

$$\begin{aligned} \text{Dist}((2,-1),(-2,2)) &= \text{sqrt}\,(2 - (-2))^2 + ((-1) - 2)^2 \\ &= \text{sqrt}\,((2 + 2)^2 + (-1 - 2)^2 \\ &= \text{sqrt}\,((4)^2 + (-3)^2) \\ &= \text{sqrt}\,(16 + 9) \\ &= \text{sqrt}\,(25) \\ &= 5. \end{aligned}$$

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

Working Principle

The idea is to choose centroid , one for each set. This should be fixed in a well planned manner because different centre will give different results. The most usual method for initializing the centres is "Forgy and Random Partition" . The optimum choice is to place them as far as possible away. Then take each point and fix it to the closest till there is none. Now early group is done. Then re-calculate new centres and binding is done between the same set points and the nearest new centre leading to the number of loop. The centres change their point until no changes .

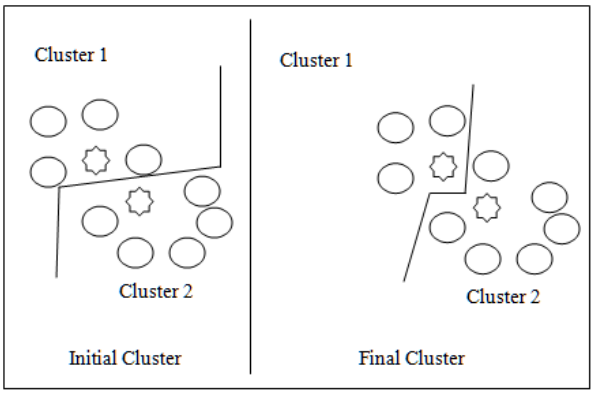

Figure 3.5 Initial and Final Cluster

Figure 3.5 Shows the assignment of first and final clusters.

Algorithm

1. Clusters the points into 'k' disjoint groups where k is the number and is predefined.
2. Select 'k' points at random as centres.
3. Cumulate the remaining objects to their closest 'k' based on the distance function.
4. Calculate the mean for all objects.
5. Repeat steps from 2 to 4 until the points are assigned correctly and there is none.

Advantages

- Speed – Due to its simplexes of the algorithm, the iterations runs fast.
- Simplicity and reliability – Solves the problem with a optimum solution for large ones.
- Complexity – Presents a good result with a less computational complexity.
- Optimum solution – Provides the good result even the points are located far.

Disadvantages

- No Categorical Data – It won't work in categorical type.
- Number of Clusters –It must be pre-defined.
- Initialization Method –Different results will gain for different K.
- Outliers – Can't handle outlier and noisy data.
- Non-linear data set- Fails to work for non-linear data.

IV. RESULTS AND DISCUSSION

4.1 INTRODUCTION

This chapter displays the results of the research work in detail. The results for the "Pre-processing, Classification and Clustering" are layout in this chapter. Excel sheet is used for creating database. The proposed research is implemented in WEKA tool. It gives the clear visual and clarification about the research work.

4.2 IMPLEMENTATAION

Microsoft Excel

Excel could be a program developed by Microsoft for 'Windows, mac OS, golem and iOS'. Its options are unit calculation, graph tools, tables, and a macro artificial language known as Visual Basic for applications .It's been a really wide applied program for these platforms , particularly since version five in 1993, and it's replaced Lotus 1-2-3 because the business customary for spreadsheets. Excel forms a part of workplace.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 11, November 2017

4.3 RESULTS

Dataset in Excel

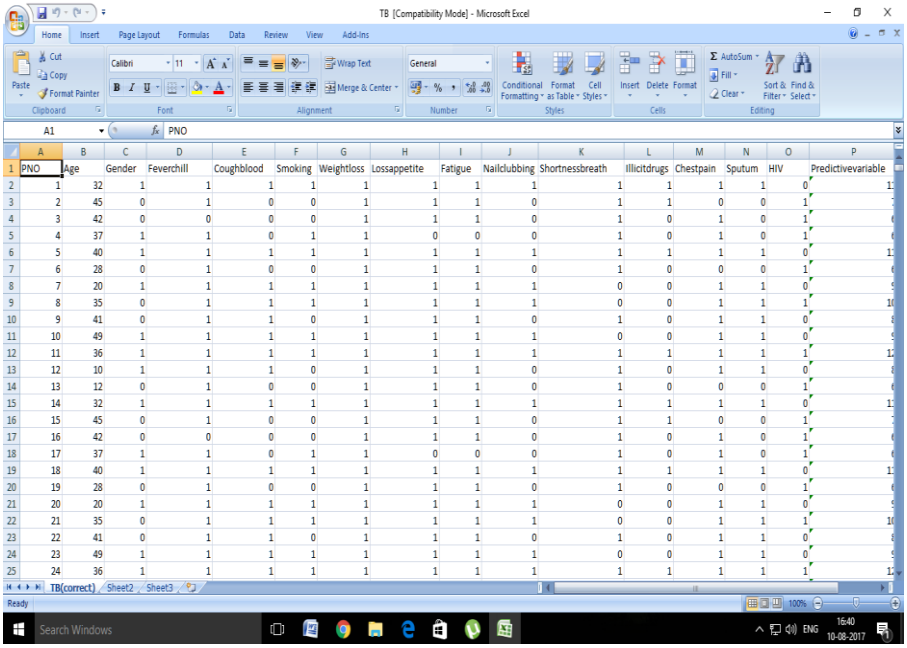

Figure 4.1 Dataset with attributes

Figure 4.1 shows 14 attributes in Microsoft Excel.

The attributes are:

Age,
Gender,
Cough with Blood,
Smoking,
Weight Loss,
Loss of Appetite,
Fatigue,
Nail Clubbing,
Shortness breath,
Illicit Drugs,
Chest pain, Sputum and HIV.

Dataset in Excel

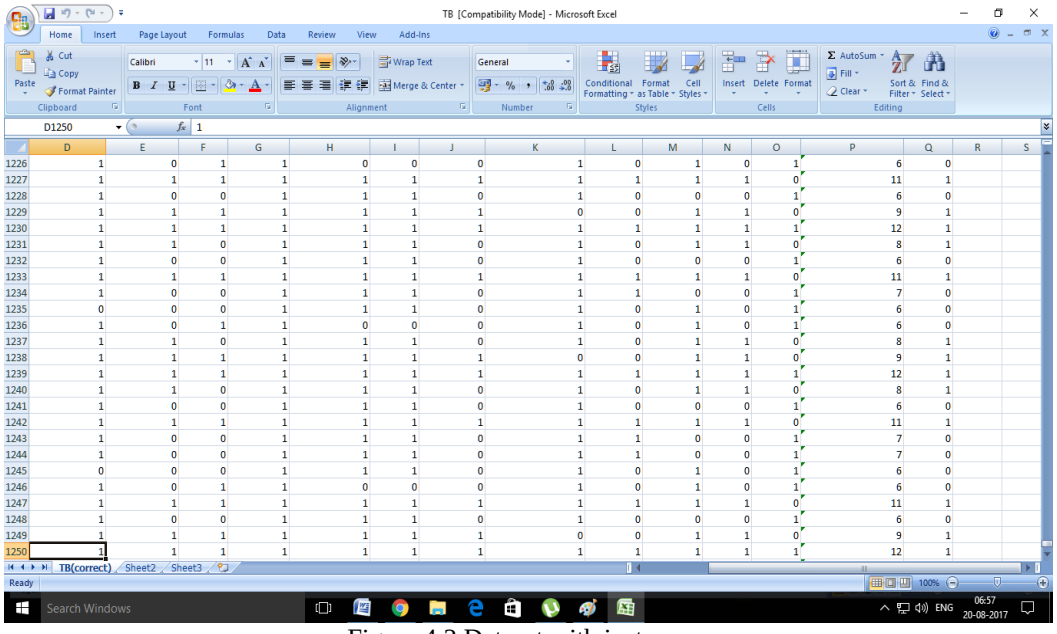

Figure 4.2 Dataset with instances

Figure 4.2 shows 1250 instances in Microsoft Excel.

V. CONCLUSION

There will be many prosperity in the aspect of life lead by the patient as the use of analysis on the patient's acquired data increases. The motivation of this work is to deploy an intellectual systematic process to assess the disease flawlessly. Data Mining routines are used to reveal the hidden norms from the vast collection of patient's data. Classification and Clustering are evaluated for Predictive and Descriptive analysis respectively. When the both routines are allied it produces high precision and also takes part in detecting the Outliers.

VI. FUTURE ENHANCEMENT

- This system can be enhanced by embodying variant Data Mining analytics.
- More Signs and Risk feature are also added for analysing the disease deeply.

REFERENCES

[1] Ashfaq Ahmed . K, Sultan Aljahdali and Syed Naimatullah Hussain, "Comparative Prediction performance with support Vector Machine and Random Forest Classification Techniques " , International Journal of Computer Applications, Volume 69-No.11, pp no 12-16. 2013.
[2] Madhuri V. Joseph Data Mining : " A Comparative study in various techniques and methods", IJARCSSE, Volume 3,Issue 2,Feb 2013.
[3] Manish Shukla and Sonali Agarwal, "Hybrid approach for tuberculosis data classification using optimal centroid selection based clustering" DOI: 10.1109/SCES.2014.6880115 Conference: Students Conference on Engineering and Systems (SCES) 2014.
[4] K. R. Lakshmi, M. Veera Krishna, S. Prem Kumar, " Utilization of Data Mining Techniques for Prediction and Diagnosis of Tuberculosis Disease Survivability" DOI: 10.5815/IJMECS, 02.08.2013 .
[5] Orhan Er, Feyzullah Temurtas and A.C. Tantrikulu, "Tuberculosis disease diagnosis using Artificial Neural networks ", Journal of Medical Systems, Springer, DOI 10.1007/s10916-008-9241, 2008.
[6] Wai Yan Nyein Naing , Zaw Z. Htike, IIUM, Malaysia, "Advances in Automatic Tuberculosis Detection in Chest X-Ray Images " volume 5, number 6, SIPIJ, December 2014.
[7] Collins K. Ahorlu, Frank Bonsu, "Factors affecting TB case detection and treatment in the Sisala East District, Ghana", Journal of Tuberculosis Research, 1 , 29-36. DOI: 10.4236/jtr.2013.13006.
[8] Tamer, "Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches", World Conference on Information Technology, Vol.3, pp.1404-1411, 2011.
[9] Asha.T, S. Natarajan and K.N.B. Murthy, "Diagnosis of Tuberculosis using Ensemble methods", IEEE, 978-1-4244-5539-3/10, 2010.
[10] K.S. Al-Sultan, "A Tabu Search Approach to the Clustering Problem," Pattern Recognition, vol.28, no.9, pp. 1443-1451, 1995.
[11] K. Rajalakshmi Dr. S. S. Dhenakaran, "Analysis of Data mining Prediction Techniques in Healthcare Management System" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, April 2015.
[12] Khajehei, M. and Etemady, F. (2010) "Data Mining and Medical Research Studies".Cimsim. 2nd International Conference on Computational Intelligence , Modelling and Simulation, 28-30 , 119-122, September 2010.
[13] Ameri, H., Alizadeh, S. and Hadizadeh, M. " Assessing the Effects of Infertility Treatment Drugs Using Clustering Algorithms and Data Mining Techniques", Journal of Mazandaran University of Medical Sciences , 24, 26-35. (Persian), 2014.