ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 07 - 18 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp02.pdf

Heart Disease Prediction Using Data Mining preprocessing and Hierarchical Clustering

Dr.A.V.Senthil Kumar

Director,Department of Computer Applications,
Hindusthan College of Arts & Science,Coimbatore, India
Email:avsenthilkumar@yahoo.com

Abstract

The diagnosis of diseases is a crucial and difficult job in medicine. The recognition of heartdisease from diverse features or signs is a major issue which is not free from false presumptions often accompanied by unpredictable effects. The healthcare industry gathers enormous amounts of heart disease data that unfortunately, are not mined to determine concealed information for effective diagnosing. Due to this rapid growth is the main motivation for researchers to mine useful information from these medical databases. As the volume of stored data increases, data mining techniques play an important role in finding patterns and extracting knowledge to provide better patient care and effective diagnostic capabilities. Heart disease prediction suffers from the problem of missing data, statistical tests will lose power, results may be based, or analysis may not be feasible at all. There are several ways to handle the problem, for example through imputation. To overcome this problem initially, the data set containing 13 medical attributes were obtained from the Cleveland heart disease database missing attributes data is replaced with the help of imputation method. With imputation, missing values are replaced with estimated values according to an imputation method or model. In this paper, preprocessed dataset from EM is given as input to clustering method for heart disease prediction. In this paper, an efficient approach non negative matrix factorization with hierarchical clustering methods (NMF-HC) is proposed for the intelligent heart disease prediction. The dataset is clustered with the aid of NMF-HC clustering algorithm. The NMF-HC is trained using the preprocessed data sets. The proposed NMF-HC works as promising tool for prediction of heart disease.

Keywords:

Cleveland dataset,Non negative matrix factorization (NMF), Hierarchical clustering, Expectation Maximization (EM) algorithm , Multi-Cycle Expectation Conditional Maximization (MCECM).

INTRODUCTION

Medical errors are both costly and harmful [1]. Medical errors cause tens of thousands of deaths in U.S. hospitals each year, more than from highway accidents, breast cancer, and AIDS combined [2]. Based on a study of 37 million patient records, an average of 195,000 people in the U.S. died due to potentially preventable, in-hospital medical errors [3]. Statistics also show that cardiovascular disease is one of the leading causes of death all over the world [4]. Hence reliable and powerful clinical decision support systems (CDSSs) are required to reduce the time of diagnosis and increase diagnosis accuracy especially for heart disease diagnosis [5]. The early decision support systems, also, were based on Bayesian statistical theory [6], probability diagnoses based on essential variables.

Accurate and error-free of diagnosis and treatment given to patients has been a major issue

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 07 - 18 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp02.pdf

highlighted in medical service nowadays. Quality service in health care field implies diagnosing patients correctly and administering treatments that are effective [7].Hospitals can also minimize the cost of clinical tests by employing appropriate computer-based information and/or decision support systems. Most hospitals today use some sort of hospital information systems to manage their healthcare or patient data [8]. These systems generate huge amounts of data which take the form of numbers, text, charts and images. These data may consist a lot of hidden information which can be use in supporting the clinical decision making.

Most hospitals today employ some sort of patient Information systems to manage their healthcare or patient data [3]. When patient is suffering from heart disease minimum of 13 data are collected by the system [4]. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important research problem as how can we turn these data into useful information that can enable heart disease prediction.

Heart disease refers to various ailments that affect the heart and the blood vessels in the heart. Heart attack Coronary artery disease, heart failure, Angina is some examples, which have different symptoms and causes [12]. The detection of heart disease is a complex procedure because of availability of incomplete data and its dependence on several diverse factors. Therefore, intelligent systems using data mining techniques are required for increasing the accuracy of diagnosis. A large number of clinical decision support systems have been built specially for the diagnosis of various kinds of heart diseases. Six of these systems are discussed here to analyze their performance based on types of heart diseases diagnosed, their strengths and shortcomings. This is the main objective for this paper work.

The use of data mining tools has become widely used in clinical applications for disease diagnosis more effectively. Various data mining techniques such as classification methods decision trees, artificial neural networks, Bayesian networks, support vector machines kernel density, bagging algorithm have been actively used in clinical support systems for diagnosis of heart disease [10]. Although there have been promising results in applying data mining techniques in heart disease diagnosis and treatment, the study done in finding out treatment options for patients and particularly heart patients is comparatively elemental. It has been suggested by researchers that application of data mining techniques for proposing suitable treatments options for patients would not only improve patient care but would also reduce investigation time, errors and would also improve the performance of medical practitioners [11].

There has been a lot of investigation for applying different data mining techniques in the diagnosis of heart disease to find out the most accurate technique but there is no study to find out the data mining technique which can increase reliability and accuracy in finding out effective treatment for heart disease patients.

As large data sets have become more common in biological and data mining applications, missing data imputation and clustering is a significant challenge. Motivate missing data estimation in matrix data with the example of the Cleveland dataset. Large number of missing data in the dataset themselves might not correctly characterize the datasets. This problem known as "data imputation" creates a challenge for various clustering and classification of for decision support. This can increase the risk of taking into account correlated or redundant attributes which can lead to lower classification accuracy. Therefore, the process of finding missing attributes is a vital phase for designing decision support systems with high accuracy. Therefore, the key objective of this paper is to design an Multi-Cycle Expectation Conditional Maximization (MCECM) approach to finds the

ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 07 - 18 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp02.pdf

missed attribute information and to obtain higher accuracy classification rates. The proposed MCECM consists of two-step process: In the first step for calculating conditional distributions. Accordingly, after the new data is applied for calculating conditional distributions. Then clustering is performed based on the NMF-hierarchical clustering (HC) methods is also compared with existing clustering methods such as k-means clustering, Fuzzy k means clustering, the proposed algorithm attain higher clustering accuracy with other clustering Algorithms.

The rest of the paper is planned as follows. The next section introduces a literature survey for existing data mining methods for heart disease prediction problem. Section 3 describes proposed Multi-Cycle Expectation Conditional Maximization (MCECM) algorithms for missing attributes imputation and hierarchical clustering methods for clustering the heart disease into two class's positive and negative classes. In Section 4, proposed NMF- **hierarchical** clustering (HC) results and existing clustering methods experimental results are discussed and the conclusion is presented in Sections 5.At end of the chapter the scope of future work is also discussed.

RELATED WORK

Intelligent and Effective Heart Attack Prediction (IEHPS) [12] presents methodology for the extraction of significant patterns from the heart disease warehouses for heart attack prediction. Initially, the data warehouse is preprocessed to make the mining process more efficient. The preprocessed data warehouse is then clustered using the K-means clustering algorithm. Frequent Itemset Mining (FIM) is performed using MAFIA (MAximal Frequent Itemset Algorithm) for the extraction of association rules from the clustered dataset. Weightage is then calculated and the patterns vital to heart attackprediction are selected according to the weightage. The neural network is trained with the selected patterns. Multi-layer Perceptron model is used withBack-propagation as the training algorithm. The significant patterns are extracted with the aid of thesignificance weightage greater than the pre-defined threshold.

M A. Jabbar et al. proposed Association Rule mining based on the sequence number and clustering for heart attack prediction [13]. The entire database is divided into partitions of equal size. The dataset with 14 attributes was used in that work and also each cluster is considered one at a time for calculating frequent item sets. This approach reduces main memory requirement. To predict the heart attack in an efficient way the patterns are extracted from the database with significant weight calculation. The frequent patterns having a value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Three mining goals were defined based on data exploration and all those models could answer complex queries in predicting heart attack.

Mai Shouman, et al. [14] proposed k-means clustering with the decision tree method to predict the heart disease. In their work they suggested several centroid selection methods for k-means clustering to increase efficiency. The 13 input attributes were collected from Cleveland Clinic Foundation Heart disease data set. The sensitivity, specificity, and accuracy are calculated with different initial centroids selection methods and different numbers of clusters. For the random attribute and random row methods, ten runs were executed and the average and best for each method were calculated. When comparing integrating k-means clustering and decision tree with traditional decision tree applied previously on the same data set in diagnosing heart disease patients. In Addition, integrating k-means clustering and decision tree could achieve higher accuracy than the paging algorithm in the diagnosis of heart disease patients. The accuracy achieved was 83.9% by the enabler method with two clusters.

Frequently the proximity with regard to some defined distance measure [15] is known as Clustering. The clustering problem has been

ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 07 - 18 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp02.pdf

identified in numerous contexts and addressed being proven beneficial in many medical applications. Clustering the medical data into small with meaningful data can aid in the discovery of patterns by supporting the extraction of numerous appropriate features from each of the clusters thereby introducing structure into the data and aiding the application of conventional data mining techniques. In [16], association rule mining technique is used for predicting heart attack. In this work proposed a novel method CBARBSN, for association rule mining based on sequence numbers and clustering the transactional database for predicting heart attack. The two important steps of this process are, first the medical data is transformed into binary and the proposed method is applied to the binary transactional data. The data is collected from Cleveland database.

In [17], the author proposed enhanced K means clustering algorithm for predicting coronary heart disease. There are two strategies are used for enhancing K-means clustering algorithm. First the author proposed weighted ranking algorithm to overcome the problem of random selection of initial centroids. Second the attributes associated with weights concerned by the physicians are taken into account in both ranking and the K-means algorithm instead of assigning unit weight to all the attributes. The heart dataset was collected from UCI machine learning repository [2]. Moreover 35 conditions are carried out to assign weights to attributes. From an experiment the author concluded that the proposed algorithm improves the consistency and quality of the final clusters. The unique clusters generates in turns of consistency.

In [18], the heart attack symptoms are predicted using biomedical data mining techniques. The author used data classification which is based on supervised machine learning algorithms. For data classification the Tanagra tool is used. Using entropy based cross validations and partitioned techniques, the data is evaluated and the results are compared. The algorithms used in these techniques are K- nearest neighbors, K-means and Mean Clustering Algorithm (EMC) is the extension of the K-mean algorithm for clustering process which reduces the number of iterations. As a result the author analyzed that the mean clustering algorithm performs well when compared to other algorithms. To run the data the time taken is very fast and it gives the result of accuracy about 82.90%. Further this work will enhanced by applying unsupervised machine learning algorithm.

Cleaning and filtering of the data might be necessarily carried out with respect to the data and data mining algorithm employed so as to avoid the creation of deceptive or inappropriate rules or patterns [19]. The steps involved in the pre-processing of a dataset are the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points and removing unneeded data fields. To make data appropriate for the mining process it needs to be transformed. The raw data is changed into data sets with a few appropriate characteristics. Moreover it might be essential to combine the data so as to reduce the number of data sets besides minimizing the memory and processing resources required by the data mining algorithm. This leads to removal of duplicate records and supplying the missing values in the heart disease data warehouse. In addition, it is also transformed to a new form which is appropriate for clustering [20]. Most of the existing methods use a k means and fuzzy k means clustering methods to group similar Cleveland dataset which is studied in literature.

These clustering method initial number of selection of k number to predict different risk levels becomes difficult and missing attributes is simply replaced by zero ,so it reduce the clustering accuracy or the prediction results of heart disease. To address limitations of existing approaches a new technique based on NMF-HC and missing attribute imputation is done based on the Multi-Cycle Expectation Conditional Maximization (MCECM) for heart disease prediction is proposed in this paper.

ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 07 - 18 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp02.pdf

PROPOSED EXPECTATION MAXIMIZATION FOR PREPROCESSING AND NMF- HIERARCHICAL CLUSTERING (HC) ALGORITHM

Propose an efficient missing attribute imputation method for Cleveland database with 13 attributes. With imputation, missing values are replaced with estimated values according to an imputation matrix-variate normal, the mean-restricted matrix-variate normal, in which the rows and columns each have a separate mean vector and covariance matrix from Multi-Cycle Expectation Conditional Maximization (MCECM). Then proposed a novel hierarchical clustering method based on a new rank-2 NMF. When the two block coordinate descent framework of nonnegative least squares is applied to computing rank-2 NMF, each sub problem requires a solution for nonnegative least squares with only two columns in the matrix for data imputation results . In addition, design a measure based on the results of rank-2 NMF for determining which leaf node should be further split to group similar data or not from pre-processing methods.

Multi-Cycle Expectation Conditional Maximization (MCECM) for missing attribute imputation :

MCECM algorithm maximizing the observed penalized log-likelihood for input data from Cleveland database . The algorithm exploits the structure by maximizing with respect to one block of coordinates at a time, saving considerable mathematical and computational time. As previously mentioned, use i to denote the row index and j the column index. The observed and missing parts of row i are o_i and m_i , respectively, and o_j and m_j are the analogous parts of column j. Let m and o denote the totality of missing and observed elements, respectively. Since with transposable data there is no natural orientation, set n to always be the larger dimension of $\boldsymbol{X}$ and p number of feature in the dataset smaller. Develop the ECM-type algorithm for imputation mathematically, beginning with the observed data log-likelihood which seek to maximize.

$$x^*_{oj,j} = \sum_{oj,oj}^{-\frac{1}{2}} (x_{oj,j} - v_{oj}), \tag{1}$$

$$l(v,\mu,\Sigma,\Delta) = \frac{1}{2}\left[\sum_{j=1}^{p} \log\left|\sum 1^{-1}_{oj,oj}\right| + 1\left|\sum \Delta^{-1}_{oj,oj}\right| + \right] - \frac{1}{2} tr\Big(\sum_{i=1}^{n} \big(x^*_{i,oi} - \mu_{o_i}\big)^T \big(x^*_{i,oi} - \mu_{o_i}\big) \Delta^{-1}_{oi,oi} - \rho_r \left|\left|\sum 1^{-1}\right|\right|^{q_r} - \rho_c \left|\left|\sum 1^{-1}\right|\right|^{q_c} \tag{2}$$

Observed log-likelihood by starting with the multivariate observed log-likelihood and using $vec(X)$ and the corresponding $vec(M)$ and .Maximize (2) via an EM-type algorithm which, similarly to the multivariate case, gives the imputed values as a part of the Expectation step. In imputation work presents two forms of the E-step ,one which leads to simple maximization with the respect to $\sum^{-1}$ and other with respect to Δ^{-1} .Letting the $\theta = \{v,\mu,\Delta,\sum\}$, the parameters of the mean-restricted matrix-variate normal, and letting o be the indices of the observed values, the E step, denoted by $Q(\theta|\theta^{'}, X_o)$, has the following form. Here, assume that X is centered:

$$Q(\theta|\theta^{'}, X_o) = \mathrm{E}(l(v,\mu,\Sigma,\Delta)|X_o,\theta^{'}) \alpha E(tr(X^T \sum 1^{-1} \mathrm{X}\Delta^{-1})|X_o,\theta^{'}) \alpha\, E(tr(X^T\sum^{-1} \mathrm{X}|X_o,\theta^{'}\Delta^{-1})) \alpha(tr[E(\mathrm{X}\Delta^{-1})|X_o,\theta^{'})\sum 1^{-1}] \tag{3}$$

ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 07 - 18 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp02.pdf

Let $X_i \sim N(0, \Delta)$ for $i = 1, .. n$ be the number of input samples from Cleveland database with p features and their covariance matrix is represented as $\Delta \in R^{n \times p}$, row mean $v \in R^n$,the column mean $\mu \in R^p$,the row covariance $\sum \in R^{n \times n}$ and column covariance $\sum \in R^{p \times p}$

Algorithm 1: MCECM for missing attribute imputation

1.Initialization
   a. Estimate $\hat{v}$ & $\hat{\mu}$ from the observed Cleveland database
   b. if x_{ij} is missing then set $x_{ij} = \hat{v}_i + \hat{\mu}_j$
   c. start with non singular estimates $\hat{\Sigma}$&$\hat{\Delta}$

2.E step (Δ):Calculate $\hat{X}^T \hat{\sum}^{-1} \hat{X} + G(\hat{\sum}^{-1})$

3.M step (Δ)
   a. Update estimates of $\hat{v}$&$\hat{\mu}$
   b. Maximize Q with respect to Δ^{-1} to obtain $\hat{\Delta}$

4.E step (Δ):Calculate $\hat{X}^T \hat{\Delta}^{-1} \hat{X} + F(\hat{\Delta}^{-1})$

5. E step ($\sum$)
   a. Update estimates of $\hat{v}$&$\hat{\mu}$
   b. Maximize Q with respect to $\sum^{-1}$ to obtain $\hat{\sum}$

6.Repeat steps 2-5 until missing data is completed

NMF HIERARCHICAL CLUSTERING method for disease prediction :ank-2 NMF to the recursive splitting of a Cleveland dataset with preprocessed samples from data imputation or missing attribute imputation method , empirical results reveal that if a balanced hierarchical tree structured is constructed, its clustering quality is often worse than that of a at partitioning. Thus need to adaptively determine the next node for imputated Cleveland dataset to split. To perform this splitting process for preprocessed data need to compute a score for each leaf node to evaluate whether it is composed of two well-separated clusters based on the two basis vectors generated by rank-2 NMF before deciding which one to split. Rank-2 NMF can be recursively applied to a preprocessed data, generating a hierarchical tree structure. In particular, propose **NMF Hierarchical Clustering** of choosing an existing leaf node at each splitting step mainly relying on cluster labels induced by the current tree structure. In the context of NMF, however, additional information about the clusters: Each column of C is a cluster representative for preprocessed dataset samples from Cleveland datasets and the largest elements in the column correspond to the disease dataset for preprocessed dataset. **NMF Hierarchical Clustering** is to compute a score for each leaf node by running rank-2 NMF two columns of W. Then select the current leaf node with the highest score as the next node to split. Initially split a leaf node N if at least two well-separated classes can be discovered within the node. Thus expect that N receives a high score of the positive class in the preprocessed dataset samples then it is splitted as left or else right in the tree which is determined based on the normalized discounted cumulative gain (NDCG) between 0 and 1. A leaf node N in tree is associated with a attribute distribution a_N, given by a column of 'a' from the rank-2 NMF run that generates the node N. Then can obtain a ranked list of possible data matrix in the positive and negative class by sorting the elements in a_N, in descending order, denoted by dpr_N, Similarly, can obtain ranked lists of terms for its two potential children, L and R, denoted by cdr_L,and cdr_R. Assuming cdr_N is a perfect ranked list, compute a modified NDCG (mNDCG) score for each of dpr_L,and dpr_R. Suppose the perfectly ordered clustered data points cdr_N is $dp_1, dp_2, \ldots dp_m$ and the shuffled orderings in cdr_L,and cdr_R are respectively $dp_{l_1}, \ldots . dp_{l_m}$ $dp_{r_1}, \ldots . dp_{r_m}$First define a position discount factor $p(dp_i)$ (4) and a gain $g(dp_i)$ (5) for each cluster data points ,

$$p(dp_i) = \log(m - \max\{i_1, i_2\} + 1) \quad (4)$$

$$g(dp_i) = \frac{\log(m - i + 1)}{p(dp_i)} \quad (5)$$

Where $l_{i_1} = r_{i_2} = i$ .In other words, for each datapoints dp_i find its positions i_1, i_2 in the two shuffled orderings, and place a large discount in the gain of preprocessed datapoint dp_i if this datapoint is high-ranked in both shuffled

ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 07 - 18 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp02.pdf

$\{g(dp_i)\}_{i=1}^{m}$orderings. The sequence of gain $\{\hat{g}(dp_i)\}_{i=1}^{m}$ is sorted in descending order. Then, for shuffled ordering $dp_s(dp_s = dp_L or\ dp_R)$, mNDCG (6) is defined as:

$$mDCG(dp_s) = g(dp_{n_1}) + \sum_{i=2}^{m} \frac{g(dp_{n_1})}{\log_2(i)} \quad (6)$$

$$mIDCG = \hat{g} + \sum_{i=2}^{m} \hat{g}/\log_2(i) \quad (7)$$

$$mNDCG(f_s) = \frac{mDCG(dp_s)}{mIDCG} \quad (8)$$

Finally, the score of the leaf node N is computed (8) as:

$$score(N) = mNDCG(dp_L) \times mNDCG(dp_R) \quad (9)$$

- When the two potential children L;R describe well separated different disease , a selected attribute N is high-ranked in one of the two shuffled orderings dp_L; $dp_R$and low-ranked in the other. Thus the top words will not suffer from a large discount and both $mNDCG(f_L)$ and $mNDCG(f_R)$ will be large.
- When both L and R describe the same diseases from preprocessed dataset samples of N, a selected attribute with data matrix for N is high-ranked in both the shuffled orderings. Thus the top words will get a large discount, and both $mNDCG(dp_L)$and $mNDCG(dp_R)$ will be small.
- When L describes the same diseases from preprocessed dataset samples of N, and R describes a totally unrelated class(e.g. outliers in N), then $mNDCG(dp_L)$ is large and $mNDCG(dp_R)$is small, and score(N) is small.

EXPERIMENTATION RESULT

The data set is taken from the Data Mining Repository of the University of alifornia, Irvineerg(UCI) [9]. To end with the system is tested using Cleveland data sets. Attributes such as Age, ex, chest pain type, resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar, resting electrocardiographic results, and maximum heart rate achieved, exercise induced angina, ST depression, and slope of the peak exercise ST segment, number of major vessels, that and the diagnosis of heart disease are presented. In experimentation work we have used a total of 909 records with 13 medical attributes. This dataset is taken from Cleveland Heart Disease database [9].Have split this record into two categories: one is training dataset (455 records) and second is testing dataset (454 records). The records for each category are selected randomly. “Diagnosis” attribute is the target predictable attribute. Value “1” of this attribute for patients with heart disease and value “0” for patients with no heart disease. “PatientID” is used as the key; the rest are input attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved. Table 1shows the attribute information of Cleveland Heart Disease database. There are two classes to be predicted: Absence or presence of heart disease in patients.

Table 1: Attribute information of Cleveland Heart Disease database

Attributes	Description	Type
Age	age in years	Numerical
sex	sex (1 = male; 0 = female)	Categorical
cp	chest pain type • Value 1: typical angina • Value 2: atypical angina • Value 3: non-anginal pain • Value 4: asymptomatic	Categorical
restbps	resting blood pressure (in mm Hg on admission to the hospital	Numerical
chol	serum cholestoral in	Numerical

ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 07 - 18 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp02.pdf

	mg/dl #10 (trestbps)	
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	Categorical
restecg:	resting electrocardiographic results Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria	Categorical
thalach	maximum heart rate achieved	Numerical
exang	exercise induced angina (1 = yes; 0 = no)	Categorical
Oldpeak	ST depression induced by exercise relative to rest	Numerical
slope	the slope of the peak exercise ST segment Value 1: upsloping, Value 2: flat and Value 3: downsloping	Categorical
ca:	number of major vessels (0-3) colored by flourosopy	Categorical
thal	3 = normal; 6 = fixed defect; 7 = reversable defect	Categorical
num:	diagnosis of heart disease Value 1: present Value 0: not_present	Categorical

The clustering accuracy for measuring the clustering results is computed as follows .Suppose that the final number of cluster is k ,clustering accuracy r is defined as ,

$$r = \frac{\sum_{i=1}^{k} a_i}{n} \qquad (10)$$

a_i be the number of instance occurs in both cluster and its corresponding class ,which has maximum value and error of the cluster is determined by e=1-r.

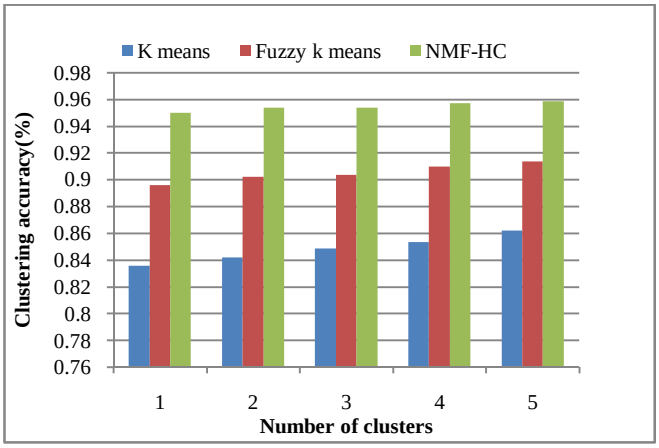

Fig 1: Clustering accuracy comparison for methods vs number of clusters

From Fig 1 shows the clustering results of the existing k means (KM) ,fuzzy k means(FKM) and NMF-HC clustering for Cleveland database with 13 attributes . It shows that the clustering accuracy results of the proposed schema is increases when compare to existing methods since the proposed NMF-HC the missing attribute data is replaced with the help of preprocessing methods ,similarly the error rate of the proposed schema is less when compare to traditional clustering methods is shown in table 2.

Table 2 .Performance error rate comparison for clustering methods

Number of clusters	Error rate (%)		
	K means	Fuzzy k	NMF-HC
1	0.164	0.104	0.052
2	0.158	0.098	0.0464
3	0.151	0.0964	0.0462
4	0.1464	0.0912	0.0423
5	0.1673	0.0865	0.0416

Clustering accuracy of k means, fuzzy k means and NMF-HC clustering is also measured and evaluated using the following metrics such as PE, V-Measure [21] and RI.

ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 07 - 18 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp02.pdf

Partition Entropy Coefficient (PE): Many unsupervised evaluation measures have been defined, but most are only applicable to clusters represented using prototypes. Two exceptions are the Partition Coefficient (PC) and the closely related Partition Entropy Coefficient (11) , the latter of which is defined as,

$$PE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|L|}(u_{ij}\log_a u_{ij}) \quad (11)$$

where $u_{ij}$is the membership of instance i to cluster j. The value of this index ranges from 0 to $\log_a|L|$. The closer the value is to 0, the crisper the clustering is. The highest value is obtained when all of the $u_{ij}$is are equal. The remainder of the criteria that we describe is all supervised.

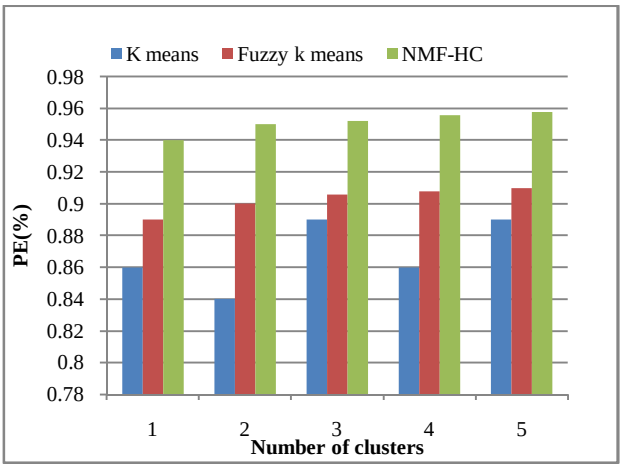

Fig 2: PE comparison for methods vs number of clusters

From Fig 2 shows the PE clustering results of the existing k means (KM) ,fuzzy k means(FKM) and NMF-HC clustering for Cleveland database with 13 attributes . It shows that the PE results of the proposed schema is increases when compare to existing methods since the proposed NMF-HC the missing attribute data is replaced with the help of preprocessing methods The PE results is measured between clustering based on the number of clusters .

V-Measure [22]: This problem with purity and entropy is overcome by the V –measure (12) , also known as the Normalized Mutual Information (NMI) , which is defined as the harmonic mean of homogeneity (h) and completeness (c); i.e.,

$$V = \frac{hc}{h+c} \quad (13)$$

where h (14) and c are defined as

$$h = 1 - \frac{H(C|L)}{H(C)} \,\&\, c = 1 - \frac{H(L|C)}{H(L)} \quad (14)$$

$$H(C) = -\sum_{i=1}^{|C|}\frac{|c_i|}{N}log\frac{|c_i|}{N} \quad (15)$$

$$H(L) = -\sum_{i=1}^{|L|}\frac{|w_i|}{N}log\frac{|w_j|}{N} \quad (16)$$

$$H(C|L) = -\sum_{j=1}^{|L|}\sum_{i=1}^{|C|}\frac{|W_j \cap c_i|}{N}\log\frac{|W_j \cap c_i|}{|c_i|} \quad (17)$$

$$H(L|C) = -\sum_{i=1}^{|C|}\sum_{j=1}^{|L|}\frac{|w_j \cap c_i|}{N}\log\frac{|w_j \cap c_i|}{|c_i|} \quad (18)$$

Because it takes into account both homogeneity and completeness, V -measure is more reliable than purity or entropy when comparing clusterings with different numbers of clusters.

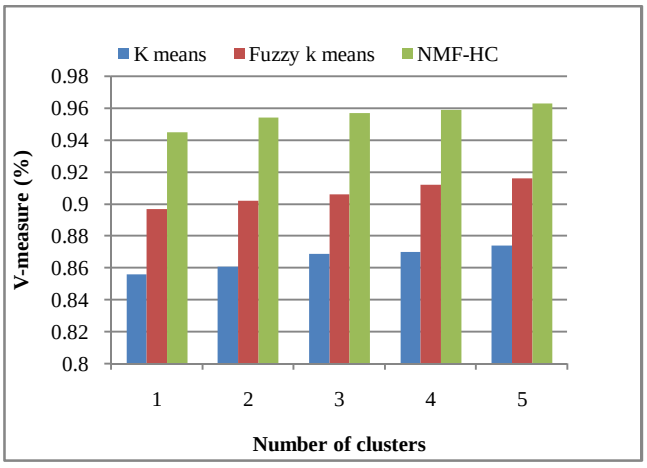

Fig 3: V-Measure comparison for methods vs number of clusters

From Fig 3 shows the **V-Measure** clustering results of the existing k means (KM)

ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 07 - 18 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp02.pdf

,fuzzy k means(FKM) and NMF-HC clustering for Cleveland database with 13 attributes . It shows that the **V-Measure** results of the proposed schema is increases when compare to existing methods since the proposed NMF-HC the missing attribute data is replaced with the help of preprocessing methods The homogeneity and completeness of the clustering results is measured between methods with number of clusters .Accuracy is the typically used measure to evaluate the efficacy of clustering methods ; it is used to reckon how the test was worthy and consistent. In order to calculate these metric, we first compute some of the terms like, True positive (TP), True negative (TN), False negative (FN)and False positive (FP) based on Table 3.

Table 3.Confusion matrix

Result of the diagnostic test		Physician diagnosis	
		Positive	Negative
Clustering results	Positive	TP	FP
	Negative	FN	TN

Rand Index : Unlike V-measure, which are based on statistics, Rand Index based on a combinatorial approach which considers each possible pair of objects. Each pair can fall into one of four groups: if both objects belong to the same class and same cluster then the pair is a true positive (TP); if objects belong to the same cluster but different classes the pair is a false positive (FP); if objects belong to the same class but different clusters the pair is a false negative (FN); otherwise the objects belong to different classes and different clusters, and the pair is a true negative (TN)

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \qquad (19)$$

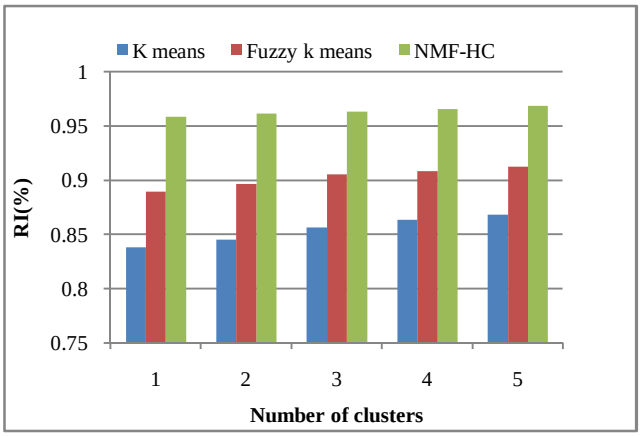

Fig 4: Rand Index comparison for methods vs number of clusters

From Fig 4 shows the **rand index** clustering results of the existing k means (KM) ,fuzzy k means(FKM) and NMF-HC clustering for Cleveland database with 13 attributes . It shows that the **rand index** results of the proposed schema is increases when compare to existing methods since the proposed NMF-HC the missing attribute data is replaced with the help of preprocessing methods In order to evaluate rand index first compute some of the terms like, TP, TN, FN and FP . From this confusion matrix the results shows that the clustering rand index results of the NMF-HC achieves higher results .

CONCLUSION AND FUTURE WORK

Data mining techniques play an important role in finding patterns and extracting knowledge from large volume of data. It is very helpful to provide better patient care and effective diagnostic capabilities. Many clinical diagnosis, examined the data mining techniques for prediction of heart disease. In this paper we have presented an efficient non negative matrix with hierarchical clustering method for extracting significant patterns from Cleveland heart disease data for the efficient prediction of heart attack. The main application of this paper has been to missing value imputation by proposing MCECM find that the algorithm converges faster if we start with the

MLE estimates with the missing values fixed and set to the estimated mean. Transposable regularized covariance models may be of potential mathematical and practical interest in numerous fields for nonsingular estimation of the covariances of the rows and columns, which is essential for any application. The preprocessed heart disease data was clustered to extract data most relevant to heart attack using NMF-HC algorithm for the valuable prediction of heart attack. The proposed NMF-HC algorithm validated against a test Cleveland heart disease dataset. The most effective model to predict patients with heart disease appeared to be the new proposed technique NMF-HC algorithm. The proposed NMF-HC algorithm classifies the group of the objects based on attributes into n number of groups in hierarchical manner. In addition, design a measure based on the results of rank-2 NMF for determining which leaf node of data should be further split. The proposed NMF-HC clustering methods shows that the compactness and connectedness is found that the efficiency and effectiveness of the method for predicting Heart Disease is better than the other two techniques. In future work, we have planned to design and develop an efficient heart attack prediction system with the aid of these selected significant patterns using artificial intelligence techniques. It can be further extended to huge amount of unstructured data in which include images such as Electro Cardio Gram (ECG) scanned images.

REFERENCES

1. Hall, J., First, make no mistakes. The New York Times. (2009).
2. SoRelle, R., Reducing the rate of medical errors in the United States. (2000).
3. Patient Safety in American Hospitals Study Survey by HealthGrades, 2004.
4. CDC‟s report, http://www.cdc.gov /nccdphp/ overview. html.
5. Yan, H.-M., Jiang, Y.-T., Zheng, J., Peng, C.-L., & Li, Q.-H. A multilayer perceptron-based medical decision support system for heart disease diagnosis. (2006)
6. http:/astrosun.tn.cornell.edu/staff/loredo/bayes/
7. Herbert Diamond, Michael P. Johnson, Rema Padman, Kai Zheng, “Clinical Reminder System: A Relational Database Application for Evidence-Based Medicine Practice ” INFORMS Spring National Conference, Salt Lake City, , 2004
8. Sellappan Palaniappan , Rafiah Awang “Web-Based Heart Disease Decision Support System using Data Mining Classification Modeling Techniques” Proceedings of iiWAS2007
9. Blake, C.L., Mertz, C.J.: “UCI Machine Learning Databases”, http://mlearn.ics.uci.edu/databases/heart-disease/, 2004
10. Sitar-Taut, V.A., et al., Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics, 2009.
11. Garg, A.X., et al., Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: a Systematic Review. Journal of the American Medical Association, 2005.
12. Patil, S. & Kumaraswamy, Y., “Intelligent and effective heart attack prediction system using data mining and artificial neural network, “European Journal of Science Research. Vol 31, No. 4.2009.
13. M A. Jabbar, Priti Chandra and B. L. Deekshatulu, “Cluster based association rule mining for heart attack prediction”, Journal of Theoretical and Applied Information Technology, Vol. 32, No.2, pp. 197 - 201, 2011.
14. Mai Shouman, Tim Turner and Rob Stocker, “Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients”, Proceedings of the International Conference on Data Mining, 2012

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol. 4 , No.6 Pages : 07 - 18 (2015)
Special Issue of ICAEECC 2015 - Held during November 16-17, 2015 in PARKROYAL on Kitchener Road, Singapore
http://warse.org/IJATCSE/static/pdf/Issue/icaeecc2015sp02.pdf

15. Zakaria Nouir, Berna Sayrac, Benoît Fourestié, Walid Tabbara, and Françoise Brouaye, "Generalization
16. Capabilities Enhancement of a Learning System by Fuzzy Space Clustering," Journal of Communications, Vol. 2, No. 6, pp. 30-37, November 2007.
17. Ma.Jabbar, Dr.Priti Chandra, B.L.Deekshatulu "Cluster Based Association Rule Mining For Heart Attack Prediction" JTAIT Vol. 32 No.2 October 2011.
18. R. Sumathi, E. Kirubakaran "Enhanced Weighted K-Means Clustering Based Risk Level Prediction for Coronary Heart Disease" European Journal of Scientific Research ISSN 1450- 216X Vol.71 No.4 (2012), pp. 490-500.
19. V.V.Jaya Rama krishniah, D.V.Chandra Sekar, Dr.K.Ramchand H Rao, "Predicting the Heart Attack Symptoms using Biomedical Data Mining Techniques" Volume 1, No. 3, The TIJCSA, 2012 .
20. Gerhard Münz, Sa Li, and Georg Carle, "Traffic anomaly detection using k-means clustering", In Proc. Of performance, reliability and dependability evaluation of communication networks and distributed systems, 4 GI / ITG Workshop MMBnet, 2007..
21. A. Budanitsky and G. Hirst, "Evaluating WordNet-Based Measures of Lexical Semantic Relatedness," Computational Linguistics, vol. 32, no. 1, pp. 13-47, 2006.
22. A. Rosenberg and J. Hirschberg, "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure," Proc Conf. Empirical Methods in Natural Language Processing (EMNLP '07), pp. 410-420, 2007.