International Journal of Knowledge Engineering
ISSN: 0976-5816 & E-ISSN: 0976-5824, Volume 3, Issue 2, 2012, pp.-202-206.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000230

EVALUATION OF PRINCIPAL COMPONENTS ANALYSIS (PCA) AND DATA CLUSTERING TECHNIQUES (DCT) ON MEDICAL DATA

SRIMANI P.K.[1] AND KOTI M.S.[2]*

[1]Dept. of Comp. Science & Maths, Bangalore University, Bangalore-560 056, Karnataka, India.
[2]Dept. of MCA, Dayananda Sagar College of Engineering, Bangalore-560 078, Karnataka, India.
*Corresponding Author: Email- man2san@rediffmail.com

Received: October 25, 2012; Accepted: November 06, 2012

Abstract- The present study investigates the performance analysis of PCA filters and six clustering algorithms on the medical data (Hepatitis) which happens to be multidimensional and of high dimension with complexities much more than the conventional data. By Clustering process data reduction is achieved in order to obtain an efficient processing time to mitigate a curse of dimensionality. Usually, in medical diagnosis, the chief guiding symptoms (rubrics) coupled with the clinical tests help in accurate diagnosis of the diseases/disorders. Hence, the primary factors have maximum impact/influence on the detection of the specific disorders. Therefore, the present study is undertaken and the results predict that farthestfirst clustering algorithm happens to be the best clustering algorithm without PCA filter in general, while cobweb clustering algorithm could be preferred with PCA filter in some other medical datasets.

Keywords- Accuracy, Classifiers, Data Mining, KDD, Learning models, Medical data

Citation: Srimani P.K. and Koti M.S. (2012) Evaluation of Principal Components Analysis (PCA) and Data Clustering Techniques (DCT) on Medical Data. International Journal of Knowledge Engineering, ISSN: 0976-5816 & E-ISSN: 0976-5824, Volume 3, Issue 2, pp.-202-206.

Introduction

Data mining technology provides a user-oriented approach to novel and hidden patterns in data. Data Mining is defined as “the nontrivial extraction of potentially useful, implicit and previously unknown information from data [3]. In data mining, intelligent methods are applied to the data to discover knowledge or patterns. Data mining and statistics both strive towards discovering patterns and structures in data.

Medical Data mining could be thought of as the search for relationships and patterns within the medical data which facilitates the acquisition of useful knowledge for effective diagnosis of the disease. The prediction of the disease becomes more effective and the early detection of disease certainly facilitates an increased exposure to the required patient care and improved cure rates. Usually, in medical diagnosis, the chief guiding symptoms (rubrics) coupled with the clinical tests help in accurate diagnosis of the diseases/disorders. Hence, the primary factors have maximum impact/ influence on the detection of the specific disorders. Actually, in many data mining applications, the choice of data processing methods is restricted by the high dimensionality nature of the data. Some of the major application areas include the studies pertaining to market basket data, text documents, image data etc. In all these cases, the nature of high dimensionality is due to one of the following aspects: wealth of alternative products, a large vocabulary, and the use of large image windows. An optimal statistical approach for dimensionality reduction is to project the data onto a lower dimensional orthogonal subspace that captures as much of the variation of the data as possible. The best (in the mean-square sense) and the most widely used way to do this is the principal component analysis (PCA); unfortunately it is quite expensive to deal with data sets with high-dimension. A dimensionality reduction method would be desirable only when it is computationally simple and does not introduce a significant distortion in the data set. Multidimensional complex problems could be solved by using several clustering algorithms which is done in the present paper.

PCA is a standard technique for visualizing high dimensional data and for data pre-processing. PCA reduces the dimensionality (the number of variables) of a data set by maintaining as much variance as possible. For eg. the three original variables are reduced to a lower number of two new variables which are termed as principal components (PCs). Using PCA, we can identify the two-dimensional plane that optimally describes the highest data variance. Thus, the two-dimensional subspace could be rotated and presented as a two-dimensional component space. Such two-dimensional visualization of the samples allow us to draw qualitative conclusions about the separability of experimental conditions.

Principal component analysis (PCA) rotates the original data space such that the axes of the new coordinate system point into the directions of highest data variance. The new variables or the axes would be referred to as the principal components (PCs) and are ordered by variance: The first component, PC1, represents the direction of the highest variance of the data. The direction of the second component, PC2, represents the highest of the remaining variance orthogonal to the first component. Accordingly, one can naturally extend this to obtain the required number of components which together span a component space covering the desired amount of variance. Since components describe specific directions in the data space, there is a certain amount of dependency of each component on each of the original variables. Certainly each component can be mathematically expressed as a linear combination of all the original variables. Low variance can often be assumed to represent undesired background noise. Without the loss of relevant information, the dimensionality reduction of the data could be achieved by retrieving a lower dimensional component space covering the highest variance. The usage of a subset of the principal components instead of the high-dimensional original data is a common pre-processing step that often improves results of subsequent analysis such as classification. For effective visualization, the first two components can be plotted against each other to obtain a two-dimensional representation of the data that captures most of the variance, useful to analyze and interpret the structure of a data set.

In cluster analysis, dimension reduction is an essential step which not only makes the high dimensional data addressable and of less computational cost, but also can provide the users with a crystal clear view of the data set of interest[2]. Contributing areas of research include data mining, statistics, machine learning, special database technology, biology and marketing. Typical requirements of clustering in data mining are: scalability, ability to deal with different types of attributes, discovery of clusters with different shapes, minimal requirement for domain knowledge to determine input parameters, capacity to handle noisy data, input records of any order and incremental or constraint based clustering, high dimensionality and usability.

Related Work

A practical tool for visualizing and data mining medical time series is stated by [7] who has concluded that increasing interest in time series data mining had surprisingly little impact on real world medical applications. Clustering technique is applied when there is no class to predict but rather when the instances divide into natural groups [11]. Clustering for multidimensional data has many challenges like noise, complexity and redundancy in data. In order to overcome these problems dimensionality reduction is required. In statistics, dimension reduction is the process of reducing the number of random variables. The process is classified into feature selection, feature extraction [8], and the taxonomy of dimension reduction problems. The principal components analysis (PCA) and partial least squares (PLS) which are dimension reduction techniques can be used to reduce the dimension of the microarray data before certain classifier is used [4]. Some of the recent works in medical mining include [12-15]. A survey of the papers where the authors have made a detailed study of the various classifiers on medical data is made. No work with regard to the present topic is available. Therefore the present investigation is carried out to study the performance of the different clustering algorithms in the presence and absence of PCA which would facilitate the early detection and efficient diagnosis of the diseases.

Methodology

The important functions of data mining are association, classification, prediction, correlation, clustering, analysis of trends, outliers and deviation, similarity and dissimilarity analyses. The algorithms such as characterization attribute subset selection and classification involve clustering as a pre-processing step that operates on the detected clusters and the selected attributes or features. Unlike classification, clustering does not rely on the predefined classes, class labels and training examples and thus it is an unsupervised learning. Accordingly, clustering is learning by observation rather than learning by examples. Grouping the data into classes or clusters is essentially the process of clustering so that objects within a cluster have high similarity while objects in different clusters have high dissimilarity. The various clustering techniques are: partitioning methods, hierarchical methods, density based methods, model based methods, grid based methods, methods for data with high dimension and constraint based clustering. Clustering is also called as data segmentation because clustering partitions large data sets into groups according to their similarity. The clustering algorithms employed in this paper are:

A. K-Means Algorithm

K-Means [1] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. Initially we determine the number of clusters K and assume the centroid as the center of these clusters. The first K objects or any random objects can also serve as the initial centroids. Finally, the K means algorithm will perform the following three steps for convergence and performs iterations until *stable* (= no object move group):

- Determine the centroid coordinate
- Compute the distance of each object from the centroids

Based on the criterion of minimum distance, group the objects (Identify the closest centroid).

The algorithm comprises the following computational steps:

Step 1: Introduce K points in the space represented by the objects that are being clustered. The initial group centroids are represented by these points.

Step 2: Each object is assigned to the group having the closest centroid.

Step 3: Recalculate the positions of the K centroids after all the objects have been assigned.

Repeat Steps 2 and 3 until the centroids remain stationary. This results in a segregation of the objects into groups from which the metric to be minimized can be calculated.

The main objective of this algorithm is the minimization of the *objective function*- which in this case is a squared error function. The objective function is,

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 ,$$

Here, $\left\| x_i^{(j)} - c_j \right\|^2$ represents a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j and is an indicator of the distance of the *n* data points from their respective cluster centres.

B. Farthestfirst Algorithm

This algorithm [5,10] is a variant of K-Means that places each cluster centre in turn at the point farthermost from the existing cluster centre. For accelerating the clustering process in most of the cases this point should lie within the data area because it facilitates less reassignment and adjustment.

C. Expectation Maximization Algorithm

EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters [9]. EM can decide how many clusters are to be created by cross validation, or one may specify apriori how many clusters could be generated.

In order to determine the number of clusters, the cross validation is performed which comprises the following steps:

Step 1: Set no. of clusters=1

Step 2: Split the training set randomly into 10 folds

Step 3: Perform the algorithm 10 times using the 10 folds

Step 4: Perform the loglikelihood average over the 10 results.

Step 5: If loglikelihood has increased the number of clusters by 1 then go to step 2

As long as the number of instances in the training set is greater than 10, the number of folds is 10(fixed).

D. Hierarchical clustering Algorithm

Hierarchical clustering is based on the core idea of objects being more related to nearby objects than to objects which are far away. Thus, these algorithms connect "objects" to form "clusters" based on their distance criterion. Large clusters can be formed on the basis of the maximum distance needed to connect parts of the clusters. A dendogram, represents different clusters that are formed at different distances and also explains the source for "hierarchical clustering". These algorithms provide an extensive hierarchy of clusters that merge with each other at certain distances but do not make a single partitioning of the data set. In a dendogram, the objects are placed along the x-axis while the distance at which the clusters merge are represented along the y-axis. This clearly avoids the mixing of clusters.

The algorithmic steps involved in hierarchical clustering defined by [6] are:

Step 1: Consider N items to be clustered and an N*N distance (or similarity) matrix

Step 2: Start by assigning each item to a cluster, so that if there are N items, then there will be N clusters such that, each contains exactly one item. Suppose the distances (similarities) between the clusters be the same as the distances (similarities) between the items they contain.

Step 3: Find the closest (most similar) pair of clusters and merge them into a single cluster, thereby reducing the number of clusters by one.

Step 4: Compute distances (similarities) between the new cluster and each of the old clusters.

Step 5: Repeat steps 2 and 3 until all items are clustered into a single cluster of size N (*)

Step 3 can be done in different ways, which distinguishes single-linkage from complete-linkage and average-linkage clustering.

E. Make Density Based Clusterer Algorithm

MakeDensityBasedClusterer Algorithm is a class for wrapping a Cluster and returns a distribution and density. The normal and discrete distributions produced by the wrapped clusterer will be fitted within each cluster.

The algorithmic steps are:

Step 1: Consider the set of elements D, no. of clusters K, minimum number of points and max distance for density measure

Step 2: Initialize k=1

Step 3: loop

Step 4: if t_i not in cluster then X={ t_j | t_j is density-reachable from t_i

Step 5: If X is a valid cluster then k=k+1; K_k=X

Step 6: until i=n

F. Cobweb Algorithm

Cobweb is an incremental system for hierarchical conceptual clustering, in the sense that it organizes the observations as a classification tree in which each node represents a class (concept) and is labeled by a probabilistic concept that summarizes the attribute-value distributions of objects classified under the node.

Database Description

Standard structure of database has to be followed by the clinics and hospitals which will help in application of data mining techniques and analysis.Practical care should be taken for the effective and efficient clinical tests and their analysis. Too many tests and trials may confuse and affect the reliability of the diagnosis of the diseases and directly contribute to the escalation of the cost. In these days of sophisticated medical electronics and computer-based tools, optimization of the process is the primary key. The database used in our experiment is Hepatitis which is available from UCI repository. This data set has 20 attributes with 155 instances with the class distribution of 2 with missing values.

Experiments and Results

In this section a detailed study of principal components analysis filters (PCA filters) and the clustering algorithms viz., K-Means, Farthestfirst, Expectation Maximization, Hierarchical, Makedensitybased and Cobweb which have their implemented source code in WEKA 3.7 version on Hepatitis medical data that contains 20 attributes and 155 instances is made. The simulated results for the clustering algorithms are presented in [Table-1], [Table-2] and [Table-3].

From [Table-1] and [Table-2] it is found that (i) All the clustering algorithms except Cobweb perform well in the absence of PCA filter. This is because the incorrectly clustered instances are less when compared to those with PCA filter (ii) Farthestfirst is considered to be the best clustering algorithm which has the least value of incorrectly clustered instances among others (iii) In the case of Cobweb clustering algorithm the performance is better for the case with PCA filter. In this case the incorrectly clustered instances are 96(61.9%) (iv) Although Farthestfirst algorithm performs very fast analysis, Hierarchical clustering could be considered as an equally well performer (v) In the case of K-means algorithm, the values of "squared error" are less with respect to PCA filter which is in accordance with our results predicted above (vi) Of all the algorithms Farthestfirst performs extremely well without the PCA filter and (vii) Finally, it is concluded that PCA filter need not be recommended for medical data in general in the case of clustering analysis but for some medical data it may prove to be useful.

[Table-3] presents the number of clusters formed in the clustering analysis by applying the six algorithms for the Hepatitis medical dataset considered above. From [Table-3] it is found that the maximum number of clusters is in the case of EM, while in all the other cases except Hierarchical clustering it is 2. In the case of Hierarchical clustering only one cluster is formed with 53 instances. However, in the case of Cobweb algorithm, the number of clusters formed happens to be 21.

Table 1- Evaluation of Clusters for Various Algorithms

Algorithms	Without PCA filter						With PCA filter					
	No. Cluster instances	Incorrectly cluster instances	% Incorrectly cluster	Time	SSE	Loglikelihood	No. Cluster instances	Incorrectly cluster instances	% Incorrectly cluster	Time	MSE	Loglikelihood
K-means	2	57	36.77	0.02	288.86	----	2	70	45.16	0.03	73.96	-------
Farthestfirst	2	50	32.25	0	----	-----	2	69	44.51	0	------	-----
Expectation Maximisation	5	93	60	19.44	------	-19.76	4	80	51.61	17.44	------	-19.96
Hierarchial	1	70	45.16	0.004	-----	------	2	69	44.51	0.2	---	-----
Make density based	2	59	38	0.05	288.86	-23.01	2	70	45.16	0.06	73.96	-21.97

Table 2- Results of Cobweb Algorithm

	Cobweb	
Without PCA	Incorrect	145
	%	93
	Time	1.95
With PCA	Incorrect	96
	%	61.9
	Time	0.2

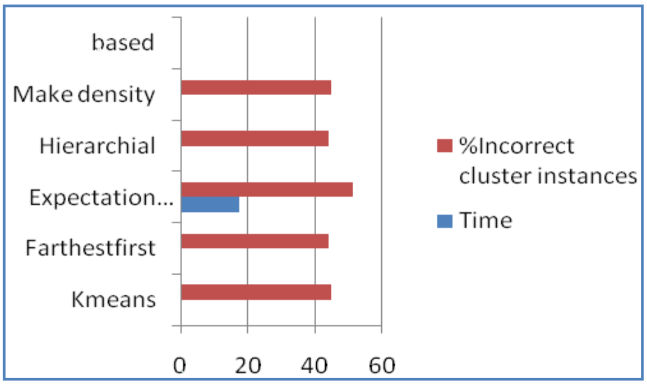

Fig. 1- Cluster evaluation with PCA filters

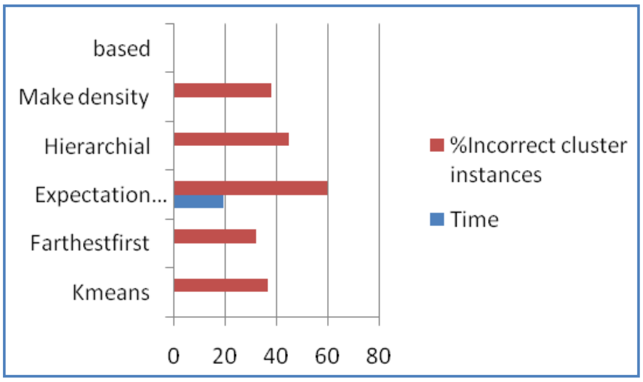

Fig. 2- Cluster evaluation without PCA filters

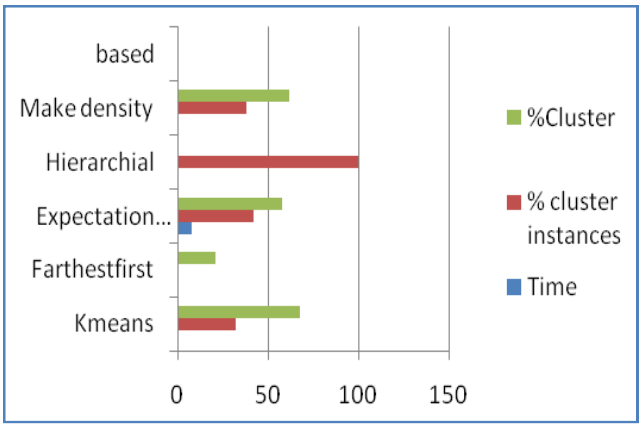

Fig. 3- Cross Validation with PCA filters

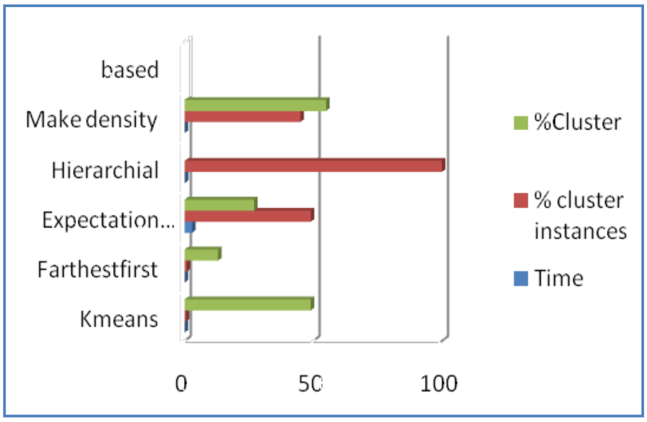

Fig. 4- Cross Validation without PCA filters

In [Table-3] and [Fig-1], [Fig-2], [Fig-3], [Fig-4], [Fig-5] the graphs of Cluster evaluation with and without PCA filters, Crossvalidation with and without PCA filters and performance of cobweb algorithm with and without PCA filters are presented, which are self explanatory and as per our above predicted results.

Table 3- Results for Cross Validation

Algorithms	Without PCA filter						With PCA filter					
	No. Cluster instances	Cluster instances	% cluster	Time	SSE	Loglikelihood	No. Cluster instances	cluster instances	% cluster	Time	MSE	Loglikelihood
K-means	2	17 36	32 68	0.02	221.29	----	2	27 26	51 49	0.03	57.27	------
Farthestfirst	2	42 11	79 21	0	----	----	2	46 7	87 13	0.01	-----	-------
Expectation Maximisation	2	22 31	42 58	7.8	----	-28.32	2	26 27	49 27	2.86	------	-21.31
Hierarchial	1	53	100	0.13	------	-------	1	53	100	0.11	-----	-----
Make density based	2	20 33	38 62	0.02	-------	-27.36	2	24 29	45 55	0.02	--------	-23.01

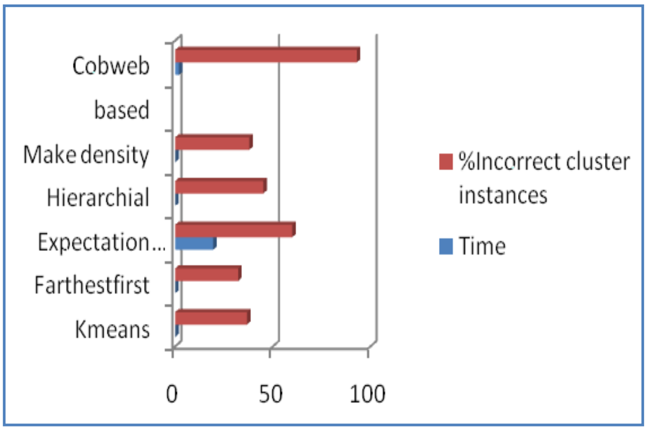

Fig. 5- Performance of Cobweb algorithm with and without PCA

Conclusion

The most common difficulties that arise during the statistical analysis of medical data come from the following facts: In most cases we encounter with a very large number of parameters, and the parameters are often very different in their nature- because, in order to establish a correct and accurate diagnosis, the physician needs to make many analyses, to observe many parameters that characterize the patient's condition and to get information using all the possible sources.

A useful computational tool for this purpose is the present data clustering approach: (i) first we record all the medical parameters that characterize a disease or a class of diseases, and try to classify them in a number of clusters equal with the number of possible diagnosis and (ii) knowing the right diagnosis for each record. In this way we find the percentage accuracy of clustering. The algorithm can be changed if the accuracy is poor or we can change the set of analysed parameters by adding or deleting them, until the desired accuracy is achieved. Next, the procedure can be used in order to establish automatically the diagnosis for new patients, by using the previously selected parameters and the clustering algorithm. Finally, it is concluded that (i) All the clustering algorithms except Cobweb perform well in the absence of PCA filter. (ii) Farthestfirst is considered to be the best clustering algorithm which has the least value of incorrectly clustered instances among others and (iii) PCA filter need not be recommended for medical data in general in the case of clustering analysis but for some medical data it may prove to be useful. Finally, it is concluded that the results of the present investigation would be effective for the early prediction of the diseases so that the survival rate could be drastically enhanced.

Acknowledgment

Mrs. Manjula Sanjay Koti is grateful to Bharathiar University, Tamil Nadu and Dayananda Sagar College of Engineering, Bangalore for providing the facilities to carry out the research work.

References

[1] Andrew Moore, Daniel B. Neill (2004) *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 256-265.

[2] Fodor I.K. (2002) *LLNL Technical Report*, *UCRL-ID*-148494, 1-18.

[3] Frawley and Piatetsky-Shapiro (1996) *Knowledge Discovery in Databases*, AAAI/MIT Press, 1-27.

[4] Gennari J.H., Langley P., Fisher D. (1990) *Artificial Intelligence*, 40, 11-61.

[5] Hochbaum Shmoys (1985) *Mathematics of Operations Research,* 10(2), 180-184.

[6] Johnson S.C. (1967) *Psychometrika*, 2, 241-254.

[7] Li Wei, Nitin Kumar, Venkata Lolla and Helga Van Herle (2005) 18*th IEEE Symposium on Computer-Based Medical Systems,* 106-125.

[8] Nisbet, Robert, John Elder, Gary Miner (2009) *Statistical Analysis & Data Mining Application*, Elsevier Inc., 111-269.

[9] Osmar R. Zaïane (2010) *Journal of Information and Data Management,* 1(1), 37-51.

[10] Sanjoy Dasgupta (2002) 15*th Annual Conference on Computational Learning Theory*, 351-363.

[11] Sembiring, Rahmat Widia, Jasni Mohamad Zain, Abdullah Embong (2010) *International Journal Of Computer Science & Information Technology*, 2(4), 162-170.

[12] Srimani P.K. and Manjula Sanjay Koti (2011) *International Journal Current Research*, 33(6), 402-407.

[13] Srimani P.K. and Manjula Sanjay Koti (2011) *AIP Conf. Proc.* 1414, 51-55.

[14] Srimani P.K. and Manjula Sanjay Koti (2012) *World Academy of Science, Engineering and Technology,* 61, 1641-1647.

[15] Srimani P.K. and Manjula Sanjay Koti (2012) *International Journal of Engineering Science and Technology*, 4, 239-246.