P- ISSN 1991-8941 E-ISSN 2706-6703 *Journal of University of Anbar for Pure Science (JUAPS)* Open Access
2008,(2), (1) :179-185

EFFICIENCY OF SOME CLASSIFICATION METHODS ADOPTED IN CLINICAL REMINDER SYSTEMS IN CASES OF MIXED DATA

Murtadhah M. H.* **Imad H.A. AL-Iathary****

*College of Computers, Anbar University.
** College of Science , Anbar University

ARTICLE INFO

Received: 1 / 4 /2008
Accepted: 24 / 4 /2008
Available online: 30/4/2008
DOI: 10.37652/juaps.2008.15442

Keywords:
Cluster analysis,
inferential statistics,
CDSS,
PCA.

ABSTRACT

Cluster analysis techniques are widely used in medical researches. Clustering techniques are not unique and hence users must be extremely conscious about what to use in order to analyze their data. The choice of unsuitable technique will resulted directly in a misleading output that cannot be interpreted or even give hints for further investigations. Understanding data variability by the use of inferential methods will help us adopt the most appropriate classification technique and accordingly enhance both building more robust CRS and CDSS's.

Introduction

Physicians require reliable information to understand, diagnose, treat, or establish the prognosis for any disease[1,2]. However, the methods for acquiring and interpreting relevant evidence in clinical research are not universal and well defined. The methods that can be used for assimilating relevant clinical information for rational clinical decision making according to the principles of evidence-based medicine are therefore affected in way or another by the art and science of data mining.

Several factors may be responsible for the increasing use of investigations, such as the increasing demand for care (due to ageing of the population and increasing numbers of chronically ill people and the urge to make use of new technology. Once an abnormal test result is found, doctors may order further investigations, not realizing that on average 5% of test results are outside their reference ranges, and a cascade of testing may result.

* Corresponding author at: College of Computers, Anbar University, Iraq.E-mail address: mortadha61@yahoo.com

Furthermore, higher standards of care, the guidelines for which often recommend additional testing, and defensive behaviour have led to more investigations. Unfortunately, when guidelines on selective and rational ordering of investigations are introduced, numerous motives for ignoring evidence based recommendations, such as fear of litigation, or procrastination on the part of the doctor, come into play in daily practice and are difficult to influence. involves a range of activities to stimulate the use of guidelines, such as communication and information about their contents and relevance, providing insight into the problem of inappropriate ordering of tests and the need to change, and, most importantly, interventions to achieve actual behavioural changes[3].

There is significant evidence from research studies that Clinical Decision Support Systems (CDSS) can enhance clinical performance in drug dosing, preventive care, and other aspects of medical care[4].

The Clinical Reminder System is designed to access a patient's treatment history by integrating the

hospital's administrative, laboratory, and clinical records systems into a single database. This information is made available to the physicians via desktop computers installed in every examination room of the clinic. Thus the patient's latest medical status is used to provide just-in-time reminders to clinicians at the point of care that reflect evidence-based medicine guidelines. Reminders generated by CRS take the form of recommendations to have tests scheduled and performed, receive vaccinations, alert clinicians to review abnormal test results, or follow up on patients with medical conditions that require unscheduled intervention[5].

In this context, data analysis will be of a high magnitude to both patient and physician. Since the data in the clinical reminder system will be restricted to a single patient, then potential variability will not be as uncommon practice, and so, revealing the hidden pattern in the set of data related to the patient will be the first step in the process of diagnosis and understanding the progress of patient's status.

Cluster analysis

Cluster analysis identifies and classifies objects individuals or variables on the basis of the similarity of the characteristics they possess. It seeks to minimize within-group variance and maximize between-group variance. The result of cluster analysis is a number of heterogeneous groups with homogeneous contents. There are substantial differences between the groups, but the individuals within a single group are similar.

k-means cluster analysis

The K-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster[6].

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster.

These centroids should be placed in a cunning way because of different locations causes different results. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function

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

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

The algorithm is composed of the following steps:

1. *Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.*
2. *Assign each object to the group that has the closest centroid.*
3. *When all objects have been assigned, recalculate the positions of the K centroids.*
4. *Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.*

Hierarchical clustering: centroid method

Hierarchical clustering follows one of two approaches: Agglomerative methods start with each observation as a cluster and with each step combine observations to form clusters until there is only one large cluster. Divisive methods begin with one large cluster and proceed to split into smaller clusters items that are most dissimilar [6].

In centroid linkage, the distance between two clusters is the distance between the cluster centroids (a vector containing means of the considered variables).

The distance matrix $D = \{d_{ij}\}$ represents the distances between airs of observation with regard to a certain similarity measure. In this paper the Euclidian distance was adopted:

$$d_{ij} = \sqrt{\sum_{k} (x_{ik} - x_{jk})^2}$$

The linkage method determines how the distance between two clusters is defined. At each stage there is a distance matrix. The entry, d(m,j), in row m and column j of this matrix is the distance from cluster m to cluster j. At the beginning, when each observation constitutes a cluster, the distance from cluster m to cluster j is the corresponding value in D, giving the distance from observation m to observation j. On each step of the amalgamation algorithm, the two rows (and columns) of the distance matrix corresponding to the two clusters to be joined are replaced by a new row (and column) corresponding to the new cluster created by joining the two clusters. The linkage method determines how the elements, d(m,j), of the new row, m, are calculated from the elements, d(k,j) and d(l,j), of the deleted rows, k and l.

Measures of similarity

There are many measures of similarities as well as dissimilarities that accompany hierarchical amalgamation of clusters. The most common measures of similarities are Euclidian distance, squared Euclidian distance, Manhattan, Pearson, and squared Pearson. Each of these measures serve a particular situation as described in details by Everitte, B., 1980[7] and Gordon, A., 1999[8].

Clustering methods

Accompanied with the hierarchical clustering different methods that well stated in Everitte, B., 1980[7] and Gordon, A., 1999[8]. the most common clustering methods are the single linkage, complete linkage, centroid and Ward's method.

Principal component analysis (PCA)

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA is theoretically the optimum transform for a given data in least square terms.

Typically speaking a data matrix of *n* observations on *p* correlated variables can be put as $x_1, x_2, x_3, \ldots, x_p$.

PCA looks for a transformation of the *xi* into *p* new variables *yi* that are uncorrelated

Looking for a transformation of the data matrix X (*n*x*p*) such that

$$Y = \delta^T X = \delta_1 X_1 + \delta_2 X_2 + \ldots + \delta_p X_p$$

Where $\delta = (\delta_1, \delta_2, \ldots, \delta_p)^T$ is a column vector of wheights with

$$\delta_1^2 + \delta_2^2 + \ldots + \delta_p^2$$

Maximize the variance of the projection of the observations on the Y variables

Find δ *so that*

$$Var(\delta^T X) = \delta^T Var(X) \delta \text{ is maximal}$$

The matrix C = Var(X) is the covariance matrix of the X_i variables

$$\mathbf{C} = \begin{pmatrix} v(x_1)\, c(x_1,x_2) \ldots\ldots c(x_1,x_p) \\ c(x_1,x_2)\, v(x_2) \ldots\ldots c(x_2,x_p) \\ \\ c(x_1,x_p)\, c(x_2,x_p) \ldots\ldots v(x_p) \end{pmatrix}$$

The direction of δ is given by the eigenvector γ_1 corresponding to the largest eigenvalue of matrix C. The second vector that is orthogonal (uncorrelated) to the first is the one that has the second highest variance which comes to be the eignevector corresponding to the second eigenvalue, and so forth.

New variables Y_i that are linear combination of the original variables (x_i):

$$Y_i = a_{i1}x_1 + a_{i2}x_2 + \ldots a_{ip}x_p \; ; i=1..p$$

The new variables Y_i are derived in decreasing order of importance; they are called 'principal components'

PCA can be used for dimensionality reduction in a data set by retaining those characteristics of the data set that contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones. Such low-order components often contain the "most important" aspects of the data. However, depending on the application this may not always be the case.

Testing Data

Since this research work is restricted to the cases of clinical reminder, the data should be related to one individual. The data was simulated using Monte Carlo method and mimic an example of real life from Everitt, B.S., 1989[9]. The data set contained 9 variables measured on 20 weeks of follow-up.

X1: Pulse rate

X2: Anxiety: 0=none, 1=slight, 2=moderate, 3=severe.

X3: Depression: 0=none, 1=slight, 2=moderate, 3=severe.

X4: Sleeping normally: 0=no, 1=yes.

X5: Loss of appetite: 0=no, 1=yes.

X6: Suicide attempt: 0=no, 1=yes.

X7: Walking alone: 0=no, 1=yes.

X8: Weight change.

X9: Week.

Experimental Results

The use of cluster analysis for observations (cases) by adopting the centroid method and the use of Euclidian distance as a measure of similarity, the following dendrogram (fig. 1) has been obtained. In this figure three main clusters can be detected. Each cluster contained different number of observations. The first cluster contained 15 observations, the second contained 4 and the third cluster contained only 1 observation.

P- ISSN 1991-8941 E-ISSN 2706-6703 *Journal of University of Anbar for Pure Science (JUAPS)* Open Access
2008,(2), (1) :179-185

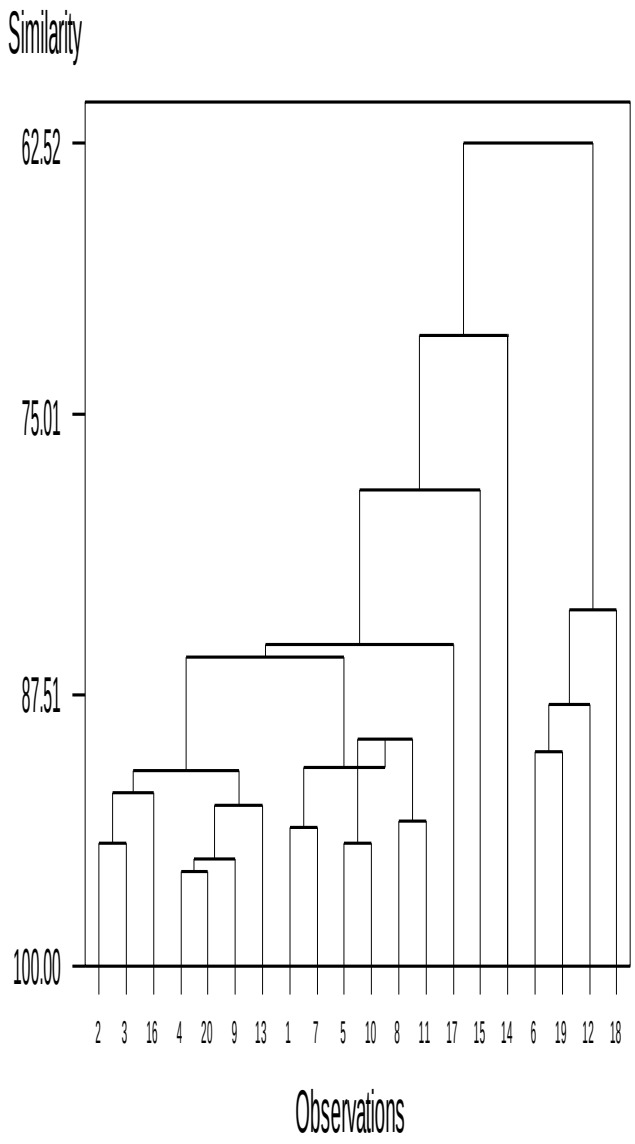

Fig. 1. Dendrogram showing clusters obtained by the use of centroid method and Euclidian distance.

The k-means cluster analysis revealed three main clusters but with different size. The first cluster contained 4 observations, the second contained 12 and the third cluster contained 4 observations.

The use of principal component analysis for the same data showed more than three clusters. Figure 2, showed the plot of the first two scored components.

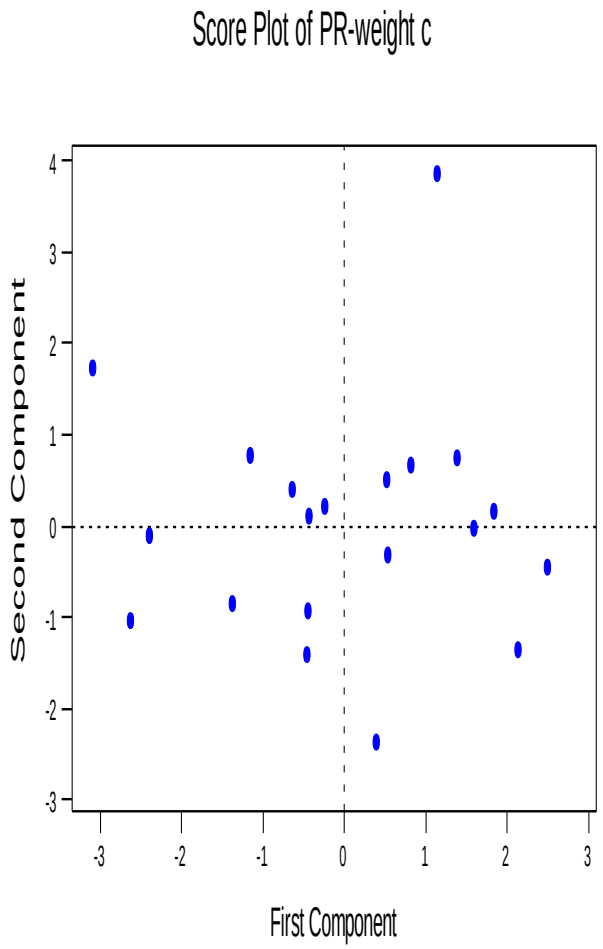

Fig. 2. Plot of the first two scored components.

Discussion

It has been shown recently[10,11] that the relaxed solution of K-means clustering, specified by the cluster indicators, are given by the PCA (principal component analysis) principal components, and the PCA subspace spanned by the principal directions is identical to the cluster centroid subspace specified by the between-class scatter matrix. The results that achieved in this paper contradict the above argument and this probably due to the clear difference in the distributions of the data. Mixture data suggests more sophisticated manipulation and techniques. Probabilistic models are one of the many suggestions in this context although its not necessary that such

techniques will resulted in a perfect variability capturing and hence well performed models.

The k-means clustering and hierarchical clustering by the use of centroid method and Euclidian distance as a measure of similarity thought to give the same results in the case of mixed set of data. In this research work the adoption of these two techniques resulted in two different clustering which indicates the huge effects of mixed data set on the performance of these techniques in data from real life.

Conclusion

In the cases of mixed data set, clustering techniques are not working identically, thus adoption of data manipulation techniques such as methods of transformation and weighing the given set of data will be very necessary. Weighing criteria can play the role of prior probabilities and hence probabilistic models will be easily performed.

References

1. JAMA patient page. Medical research. Finding the best information. JAMA. 2000;284:1336.
2. Djulbegovic B, Soares HP, Kumar A. What kind of evidence do patients and practitioners need: evidence profiles based on 5 key evidence-based principles to summarize data on benefits and harms. Cancer Treat Rev. 2006;32:572-576. Epub 2006 Aug 17.
3. Ron Winkens and Geert-Jan Dinant. Evidence base of clinical diagnosis: Rational, cost effective use of investigations in clinical practice. BMJ, 2002, 324; 783-785.
4. M. Weiner, C. M. Callahan, W. M. Tierney, J M. Overhage, B. Mamlin, P. R. Dexter, and C. J. McDonald Using Information Technology To Improve the Health Care of Older Adults. *Ann Intern Me*d, September 2, 2003; 139(5): 430 -436.
5. Kai Zheng a , Rema Padman a , Michael P. Johnson a , John Engberg a, Herbert H. Diamond b. An Adoption Study of a Clinical Reminder System in Ambulatory Care Using A Developmental Trajectory Approach. MEDINFO 2004, M. Fieschi et al. (Eds), Amsterdam: IOS Press, © 2004 IMIA. All rights reserved
6. MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297.
7. Everitt, B., 1980: Cluster analysis. Second edition. New York.
8. Gordon, A. D., 1999: Classification. 2nd edition. London-New York.
9. Everitt, B. S, 1989: Statistical methods for medical investigations. First edition. New York.
10. H. Zha, C. Ding, M. Gu, X. He and H.D. Simon. "Spectral Relaxation for K-means Clustering", Neural Information Processing Systems vol.14 (NIPS 2001). pp. 1057-1064, Vancouver, Canada. Dec. 2001.
11. Chris Ding and Xiaofeng He. "K-means Clustering via Principal Component Analysis". Proc. of Int'l Conf. Machine Learning (ICML 2004), pp 225-232. July 2004.

P- ISSN 1991-8941 E-ISSN 2706-6703 *Journal of University of Anbar for Pure Science (JUAPS)* Open Access
2008,(2), (1) :179-185

كفاءة بعض طرق التصنيف لنظام الرسائل التذكيرية السريري ذو البيانات الغير متجانسة

مرتضى محمد حمد **عماد هجول عبود**

Email : mortadha61@yahoo.com

الخلاصة

تقنيات تحليل العنقدة تستخدم بشكل واسع في البحوث الطبية. وهذه التقنيات ليست أحادية الاستخدام لذلك تتطلب من المستخدم تحديد نوع الاستخدام لتحليل البيانات على ضوء ذلك.

أن اختيار التقنية الغير مناسبة يؤدي الى حصول مخرجات ناقصة لاتتمكن من اعطاء تفسير او ملاحظات عن التنبؤات المستقبلية.

ان اسلوب فهم تغير البيانات بأستخدام الطرق الاستدلالية يساعد في اعتماد تقنية عنقدة مناسبة، وهذا له الدور في تحسين الاداء لبناء (CRS) و (CDSS) أكثر متانة.