Acta HealthMedica (ISSN: 2414-6528) http://www.ActaHealthMedica.com
Volume: 2, Issue: 1, January-March 2017, Pages: 116, DOI: http://dx.doi.org/10.19082/ah116

VALIDATION OF CLUSTERING METHODS FOR MEDICAL DATA SETS

Azam Orooji[1], Farzaneh Kermani[2,*]

1: PhD Student of Medical Informatics, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran
2: PhD Student of Medical Informatics, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

Correspondence:
Tel: +98-2188794301, E-mail: Kermani.f@tak.iums.ac.ir

TYPE OF ARTICLE: CONFERENCE ABSTRACT

ABSTRACT

Introduction: Data mining techniques have been increasingly applied to medical data in the past decade and are divided into two categories: predictive algorithms such as classification and descriptive algorithms such as clustering and association rule mining (ARM). Clustering means partitioning a data set into a set of clusters in such a way that the samples belonging to the same clusters are similar and those belonging to different clusters are dissimilar. These algorithms are currently used as a preprocessing technique prior to medical data mining and analyzing. However, clustering algorithms have different behaviors depending on the features of the data set. Therefore, in most applications, results of clustering are evaluated in terms of some validity clustering measures. In addition, most of the medical data sets are usually complex, with nonlinear patterns, which are extremely large and difficult to cluster. The aim of this study is to performance analyze several clustering techniques based on the basis of two different classes of clustering quality measures, named internal and stability, over the medical data sets.

Methods: In this study, the performance of six common clustering algorithms, including hierarchal clustering, K-means, fuzzy C-means (FCM), self-organizing tree algorithm (SOTA), Diana and partitioning around medoids (PAM), has been examined by using four different medical data sets, which are available on UCI repositories, including the Indian liver patient data set (ILPD), Pima Indians diabetes, breast cancer, and statlog heart disease. The results of clustering methods have been evaluated based on two different classes of clustering quality measures called “internal” and “stability.” Stability measures include average proportion of non-overlap (APN), average distance (AD), average distance between means (ADM), and figure of merit (FOM) and internal measures consist of connectivity, Silhouette and Dunn index.

Results: Due to given data sets consisting of two classes of samples, the number of cluster is assumed to two. The evaluation results showed that hierarchical clustering is the best algorithm to cluster all of the current data sets on the basis of internal validity measures. While based on stability measures, a Diana algorithm for ILPD and breast data sets, hierarchical clustering method for heart data set and SOTA for Pima data set outperformed other clustering algorithms.

Conclusion: In this study, performance of several clustering techniques on the basis of internal and stability clustering measures over the medical data sets were investigated. The hierarchical clustering and Diana methods revealed better results. The number of clusters is assumed to two, while this parameter can have an effect on performance of different clustering algorithms. Future work will mainly cover the optimal estimation number of clusters and also consider more clustering methods for evaluation.

KEYWORDS: Clustering, Health care data, Validity assessment, Data mining