RESEARCH PAPER | Engineering | Volume : 4 | Issue : 8 | August 2014 | ISSN - 2249-555X

Performance Metrics of Clustering Algorithm

KEYWORDS: clustering, cluster validation, fuzzy clustering, K means clustering.

A. Viji Amutha Mary
Research Scholar, Dept of CSE, Faculty of Computing, Sathyabama University, Chennai-119,TamilNadu, India.

Dr. T. Jebarajan
Professor and Head, Dept of CSE, Rajalakshmi Engineering College, Chennai, TamilNadu, India.

ABSTRACT *This paper discusses the various performance measures for assessing the quality of Hard C Means Clustering algorithm. This assessment of the clustering quality is referred to as cluster validation which is a similarity measure between two different clusters. A detailed survey of the performance metrics based on the internal and external evaluation measures is presented. The two commonly used clustering algorithms such as K means and Fuzzy C means are compared and its applications are discussed.*

I.Introduction

Clustering is considered the most important unsupervised learning problem which is defined as an assignment of a set of observations into subsets so that observations in the same subset are similar in some sense. Cluster analysis or clustering is the task of assigning a set of objects into groups or clusters so that the objects in the same cluster are more similar to each other than to those in other clusters. It has been the subject of wide research in various fields such as engineering, business and medicine. The most widely used clustering methods are hard (crisp) and soft (fuzzy) clustering. In hard clustering, each object either belongs to a cluster or not. In soft (fuzzy) clustering, each object belongs to each cluster to a certain degree. Evaluation measure is used to compare how well different data clustering algorithms perform on a set of data. When a clustering result is evaluated based on the data that was clustered itself, it is called internal evaluation. The clustering results that are evaluated based on the data that was not used for clustering, such as known class labels and external benchmarks is called external evaluation.

II. Hard C Means Clustering ALGORITHM

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Select 'c' cluster centers randomly.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is the minimum of all the cluster centers..
4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in the i^{th} cluster.

5) Recalculate the distance between each data point and the new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

III.Fuzzy C Means Clustering ALGORITHM

Step 1. Initialize $U=[u_{ij}]$ matrix, U(0)

Step 2. At k-step: calculate the center vectorsC(k)=$[c_j]$ with U(k)

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}}{\sum_{i=1}^{N}}$$

Step 3: Update U(k) and U(k+1)

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{|x_i - c_j|}{|x_i - c_k|} \right)^{\frac{2}{m-1}}}$$

IV. COMPARISON OF HCM AND FCM ALGORITHM

1) Hard C-Means clustering is also known as K-Means. Fuzzy C-means Clustering(FCM), is also known as Fuzzy ISODATA.
2) In HCM each data point will be assigned to only one cluster. Fuzzy c-means (FCM) allows one piece of data to belong to two or more clusters[6].
3) In HCM the centroids of c clusters are achieved by randomly selecting c points from among all the data points. With fuzzy c-means, the centroid of a cluster is computed as being the mean of all points, weighted by their degree of belonging to the cluster.
4) Hard k-means algorithm executes a sharp classification, in which each object is either assigned to a class or not. The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1.
5) The aim of the HCM algorithm is to find the cluster centers(centroids) for each group. FCM iteratively updates the cluster centers and the membership grades for data point till the cluster centers are moved to the "right" location within a dataset.
6) Hard k-means algorithm is dependent on initialization and it is sensitive to outliers. These drawbacks are overcome by FCM.

V.EVALUATION MEASURES

These measures can be used to compare how well different data clustering algorithms perform on a set of data. It is also referred to as cluster validation.

A. Internal Evaluation Measures

When a clustering result is evaluated based on the data that was clustered itself, it is called internal evaluation. This evaluation is biased towards algorithms that use the same cluster model. These methods usually assign the

Volume : 4 | Issue : 8 | August 2014 | ISSN - 2249-555X

best score to the algorithm that produces clusters with high similarity within a cluster and low similarity between clusters. The following methods can be used to assess the quality clustering algorithms based on internal criterion:

(i). Davies–Bouldin index

The Davies–Bouldin index can be calculated by the following formula:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \qquad [2]$$

where n is the number of clusters, c_x is the centroid of cluster x, σ_x is the average distance of all elements in cluster x to centroid c_x, and $d(c_i,c_j)$ is the distance between centroids c_i and c_j. Since algorithms that produce clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies–Bouldin index, the clustering algorithm that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best algorithm based on this criteria.

(ii). Dunn index - The Dunn index aims to identify dense and well-separated clusters. It is defined as the ratio between the minimal inter-cluster distance to maximal intra-cluster distance. For each cluster partition, the Dunn index can be calculated by the following formula

$$D = \min_{1 \le i \le n} \left\{ \min_{1 \le j \le n, j \neq i} \left\{ \frac{d(i,j)}{\max_{1 \le k \le n} d'(k)} \right\} \right\}$$

where d(i,j) represents the distance between clusters i and j, and d'(k) measures the intra-cluster distance of cluster k. The inter-cluster distance d(i,j) between two clusters may be any number of distance measures, such as the distance between the centroids of the clusters. Similarly, the intra-cluster distance d'(k) may be measured in a variety ways, such as the maximal distance between any pair of elements in cluster k. Since internal criterion seek clusters with high intra-cluster similarity and low inter-cluster similarity, algorithms that produce clusters with high Dunn index are more desirable.

B. External Evaluation Measures

In external evaluation, clustering results are evaluated based on known class labels and external benchmarks which consist of a set of pre-classified items, and these sets are often created by human experts. These types of evaluation methods measure how close the clustering is to the predetermined benchmark classes. Since classes can contain internal structure, the attributes present may not allow separation of clusters or the classes may contain anomalies. Additionally, from a knowledge discovery point of view, the reproduction of known knowledge may not necessarily be the intended result.

Some of the measures of quality of a clustering algorithm using external criterion include:

(i). Rand measure

This was developed by William M. Rand in 1971. The Rand index computes how similar the clusters resulted by the clustering algorithm are to the benchmark classifications. The Rand index is also viewed as a measure of the percentage of correct decisions made by the algorithm. It can be computed using the following formula:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. One issue with the Rand index is that false positives and false negatives are equally weighted. This may be an undesirable characteristic for some clustering applications. This concern is addressed by the F-measure.

(ii). F-measure

The F-measure can be used to balance the contribution of false negatives by weighting recall through a parameter $\beta \geq 0$. Let precision and recall be defined as follows:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

where P is the precision rate and R is the recall rate. We can calculate the F-measure by using the following formula [21]:

$$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

When $\beta = 0$, $F_0 = P$. Therefore, recall has no impact on the F-measure when $\beta = 0$, and increasing β allocates an increasing amount of weight to recall in the final F-measure.

(iii). Pair-counting F-Measure

This is the F-Measure applied to the set of object pairs, where objects are paired with each other when they are part of the same cluster. This measure is able to compare clusterings with different numbers of clusters.

(iv). Jaccard index

The Jaccard index is used to quantify the similarity between two datasets. The Jaccard index takes on a value between 0 and 1. An index of 1 means that the two datasets are identical, and an index of 0 indicates that the datasets have no common elements. The Jaccard index is defined by the following formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad [1].$$

This is the number of unique elements common to both sets divided by the total number of unique elements in both the sets.

(v) Confusion matrix

A confusion matrix can be used to quickly visualize the results of a clustering algorithm. It shows how different a cluster is different from the gold standard cluster.

(vi). Mutual Information

This is an information theoretic measure of how much information is shared between a clustering and a ground-truth classification that can detect a non-linear similarity between two clusterings. Adjusted mutual information is the corrected-for-chance variant of this that has a reduced bias for varying cluster numbers.

VII. APPLICATIONS

1. **Sequence Analysis :** The homologous sequences are grouped into gene families.
2. **Human genetic clustering:** The similarity of genetic

Volume : 4 | Issue : 8 | August 2014 | ISSN - 2249-555X

data is used in clustering to infer population structures.

3. **Medical imaging :** Different types of tissue and blood are differentiated in a three dimensional image.
4. **Market Research:** The general population of consumers are partitioned into market segments and to better understand the relationships between different groups of consumers.
5. **Software evolution:** The legacy properties in code are reduced by reforming functionality that has become dispersed.
6. **Image segmentation:** A digital image is divided into distinct regions for border detection or object recognition.
7. **Crime Analysis:** Areas of greater incidences of particular types of crime are identified.
8. **Educational data mining:** Groups of schools or students with similar properties are identified.
9. **Climatology:** Weather regimes or preferred sea level pressure atmospheric patterns are found.
10. **Petroleum Geology:** To reconstruct missing bottom hole core data or missing log curves inorder to evaluate reservoir properties.

VIII. conclusion

The role of validity index is very important in clustering. The cluster validity index provides a measure of the quality of the partition that was found and finds out whether there exists a better partition. Thus, it has been used to search for the optimal number of clusters when the number of clusters is not known a priori. The number of clusters that present an image may be determined automatically with the help of these indices. Such indices based on internal and external measures are given a detailed study in this paper.

REFERENCE

[1] Lin, Jiang & Lee. (2013, Jan). Similarity Measure for Text Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering. | [2] El-Melegy, Zanaty E. A. (2007). On Cluster Validity Indexes in Fuzzy and Hard Clustering Algorithms for Image Segmentation. | [3] Maulik U. & Bandyopadhyay S. (2002). Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. 24, no. 12. | [4] Likas A, Vlassis N & Verbeek J. (2001). The global k- means clustering algorithm (Technical Report). Computer Science Institute, University of Amsterdam, The Netherlands ISA-UVA-01-02. | [5] Carl G. L. (1999). A fuzzy clustering and fuzzy merging algorithm, Technical Report CS- UNR-101. | [6] Kumar & Sirohi. (2010 August). Comparative Analysis of FCM and HCM Algorithm. International Journal of Computer Applications (0975 – 8887). Volume 5– No.2. | [7] Bezdek J. C, et al. (1999) Fuzzy Models and Algorithms for | Pattern Recognition and Image Processing. Kluwer Academy Publishers, Boston9. | [8] Yanp M. S, Wu K. L & Yub J. (2003). A novel fuzzy clustering algorithm. IEEE International Symposium on Computational Intelligence in Robotics and Automation, Vol. 2, pp. 647- 652. | [9] Xu Y, Richard G. & Brereton A. (2005). A comparative study of cluster validation indices applied to genotyping data. Chemometrics and Intelligent Laboratory Systems, Vol. 78, pp. 30–40. | [10] Jain A. K & Dubes R. C. (1998) Algorithms for Clustering, Prentice- Hall, Englewood Cliffs, NJ. |