Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 11, November 2014, pg.629 – 632

RESEARCH ARTICLE

To Determine Document Clustering By using Centroid Ratio for Pair wise Random Swap Clustering Algorithm

R.B. RAMADEVI, S.SUBASHINI

PG STUDENT, ASSISTANT PROFESSOR

NPR COLLEGE OF ENGINEERING AND TECHNOLOGY, TAMILNADU, INDIA

EMAIL: devirama364@gmail.com, balasundar1957@gmail.com

Abstract: Clustering algorithm and cluster validity are two highly correlated parts in cluster analysis. In this paper, a novel idea for cluster validity and a clustering algorithm based on the validity index are introduced. A Centroid Ratio is firstly introduced to compare two clustering results. This centroid ratio is then used in prototype-based clustering by introducing a Pair wise Random Swap clustering algorithm to avoid the local optimum problem of k-means. Before clustering, the number of clusters is an essential parameter for the clustering algorithm, while after clustering; the validity of the clustering is performed. The similarity value for comparing two clusterings from the centroid ratio can be used as a stopping criterion in the algorithm.

I propose a cluster-level validity criterion called a centroid ratio. It has low time complexity and is applicable for detecting unstable or incorrectly located centroids. Employing

the centroid ratio in swap-based clustering, I further suggest a pair wise random swap clustering algorithm, for which no stopping criterion is required. The centroid ratio is shown to be highly correlated to the mean square error (MSE) and other external indices. Moreover, it is fast and simple to calculate. An empirical study of several different datasets indicates that the proposed algorithm works more efficiently than Random Swap, Deterministic Random Swap, Repeated k-means or k-means++. The algorithm is successfully applied to document clustering.

2 LITERATURE SURVEY

2.1 Data clustering: 50 years beyond K-means:

Organizing data into sensible groupings is one of the most fundamental modes of understanding and learning. Cluster analysis is the formal study of methods and algorithms for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity. The aim of clustering is to find structure in data and is therefore exploratory in nature. Clustering has a long and rich history in a variety of scientific fields. One of the most popular and simple clustering algorithms, K-means, was first published in 1955. In spite of the fact that K-means was proposed over 50 years ago and thousands of clustering

2.2 **Understanding of Internal Clustering Validation Measures**

Clustering validation can be categorized into two classes, external clustering validation and internal clustering validation. In this paper, we focus on internal clustering validation and present a detailed study widely used internal clustering validation measures for crisp clustering. The idea of SD index ($\mathcal{SD}$) is based on the concepts of the average scattering and the total separation of clusters. The first term evaluates compactness based on variances of cluster objects, and the second term evaluates based on distances between cluster centers. The value of this index is the summation of these two terms. Experiment results show that $SDbw$is the only internal validation measure which performs well in all five aspects, while other measures have certain limitations in different application scenarios.

2.3 K-means Clustering versus Validation Measures: A Data Distribution Perspective

K-means is a widely used partition clustering method. While there are considerable research efforts to characterize the key features of K-means clustering. We show that the entropy measure, an external clustering validation measure, has the favorite on the clustering algorithms which tend to reduce high variation on the cluster sizes. Finally, our experimental results indicate that K-means tends to produce the clusters in which the variation of the cluster sizes, as measured by the Coefficient of Variation (CV), is in a specific range, approximately from 0.3 to 1.0. We have conducted extensive experiments on a number of real-world data sets from various different application domains. The larger the CV value is, the greater the variability in the data. K-means [15] is a prototype-based, simple partition clustering technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids.

2.4 Adapting the Right Measures for K-means Clustering

While many validation measures have been developed for evaluating the performance of clustering algorithms, these measures often provide inconsistent information about the clustering performance and the best suitable measures to use in practice remain unknown. This paper thus fills this crucial void by giving an organized study of 16external validation measures for K-means clustering. we first introduce the importance of measure normalization in the evaluation of the clustering performance on data with imbalanced class distributions. We also provide normalization solutions for several measures. We provide a guide line to select the most suitable validation measures for K-means clustering.

2.5.:Extending external validity measures for determining the number of cluster:

External validity measures in cluster analysis evaluate how well the clustering results match to a prior knowledge about the data. In this paper, we extend the external validity measures for both hard and soft partitions bya resampling method, where no prior information is needed. The proposed method is then applied and reviewed in determining the number of clusters for the problem of unsupervised learning, cluster analysis. Experimental results have demonstrated the proposed method is very effective in solving the number of clusters. External validity measures are preferable for evaluating the goodness of clustering's when ground truth labels are available. External measures are mainly designed for hard partitions. Researchers shed light on extensions of external measures for fuzzy results.

3. CONCLUSION

We proposed a novel evaluation criterion called the centroid ratio, based on the centroids in prototype-based clustering, which compares two clustering's Cluster analysis is one of the most widely used techniques for exploratory data analysis, with applications ranging from image processing, speech processing, information retrieval and Web applications clustering has been developed and modified for different application fields, providing many clustering algorithms . The cost function in clustering algorithms is used to decide whether the clustering result is suitable for certain kinds of data structures. The Mean squared error (MSE)-based cost function. The cluster validity is an important issue in cluster analysis, as evaluating different clustering algorithms helps the user to gain a better understanding on the properties and efficiency on different algorithms.

REFERENCES

- K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit. Lett. vol. 31, no. 8,pp. 651–666, 2010.
- Y. Liu, Z. Li, H. Xian, X. Gao, and J.Wu, "Understanding of internal clustering validation measures," in Proc. 10th ICDM, Sydney,NSW, Australia, 2010, pp. 911–916.
- H. Xiong, J. Wu, and J. Chen, "K-means clustering versus validation measures: A data-distribution perspective," IEEE Trans. Syst., Man, Cybern., vol. 39, no. 2, pp. 318–331, Apr. 2009.
- J. Wu, H. Xiong, and J. Chen, "Adapting the right measures fork-means clustering," in Proc. 15th ACM SIGKDD Int. Conf. KDD, Paris, France, 2009, pp. 877–886.
- P. Fränti and J. Kivijärvi, "Randomized local search algorithm for the clustering problem," Pattern Anal. Applicat., vol. 3, no. 4, pp. 358–369, 2000.