Precise Data Cluster Validation using K- Means Algorithm

T. P. Latchoumi
Assistant Professor, Department of
Computer Scienceand Engineering,
Christ College of Engineering and
Technology, Pondicherry, India.

K. Kaarguzhali
Final year M.Tech, Department of Computer
Science and Engineering,
Christ College of Engineering and Technology,
Pondicherry, India

***Abstract*—Data mining refers to extracting or mining knowledge from large amounts of data.Clustering algorithm and cluster validity are two highly correlated parts in cluster analysis. A novel idea for cluster validity and a clustering algorithm based on the validity index ACentroid Ratio is firstly introduced to compare two clustering results. 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. We 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. The centroid ratio is shown to be highly correlated to the mean square error (MSE) and other external indices. Likewise, it is fast and simple to calculate. The term vector is an algebraic model for representing text documents as vectors of identifiers. It is used in information sieving, information recovery, indexing and relevancy levels. Clustering is performed by means of K-Mean algorithm. The k-Mean clustering is distance threshold based clustering. Clusters formed by similarity distance threshold value. It is used ensure a good clustering quality.**

Keywords—Data clustering;Cosine similarity; Clustering evaluation; Mean Square Error;k-means.

I INTRODUCTION

Data Mining is an systematic process designed to discover data (usually large amounts of data - typically business or market related - also known as "big data") in search of reliable patterns and/or systematic relationships among variables, and then to authenticate the findings by relating the detected patterns .Data mining is the process of analysing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs these are the Data mining software is one of a number of analytical tools for analysing data. It use to analyse data from many different dimensions or angles, categorize it, and summarize the interactions identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data are any facts, numbers, or text that can be handled by a computer. Today, organizations are accruing vast and growing amounts of data in different formats and different databases. The patterns, associations, or relationship are providing the information. Information can be renewed into knowledge about historical patterns and future trends. The actual data mining task is the automatic or semi-automatic analysis of large quantities of

Datato extract previously unknown interesting patterns such as groups of data records, unusual records and dependencies.Clusteranalysis is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.Clustering can be deliberated the most important unsupervisedlearning so, as every other problem of this kind; it deal with finding a structure in a group of unlabelled data.The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a worthy clustering. There is no absolute “best” criterion which would be independent of the final aim of the clustering. Accordingly, it is the user which must resource this standard, in such a way that the result of the clustering will suit their needs.

II EXISTING SYSTEM

Several other methods have been developed, which are based on stochastic global optimization such as simulated annealing and genetic algorithms. These methods have not gained wide acceptance because of their great time complexity. A global k-means algorithm (GKM) [2] is an incremental approach that dynamically adds onecluster center at a time through a deterministic global search procedure. Clustering algorithm and cluster validity are two highly correlated parts in cluster analysis. A common way to address the initialization problem is to run k-means multiple times with a different set of randomly chosen initial parameters and to choose the bestsolution as a result. A cost function is used to evaluate the quality of the clustering. There is no universal function for all clustering problems, and the choice of the function depends on the application. We consider the clustering as an optimization problem. The replacement location of the swapped centroid can be chosen by considering the locations of all possible data points. First, people have studied the impact of the high dimensionality on the performance of K-means [3] and found that the traditional Euclidean notion of proximity is not very effective for K-means on high-dimensional data sets, such as gene expression data sets and document data sets.

There are some other internal validations measures in Literature some have poor performance while some are designed for data sets with specific structures.Take Composed Density between and within clusters index (*CDbw*)[10] and Symmetry distance-based index. Internal indexes such as the Bayesian information criterion (BIC) and sum-of-squares have difficulties in finding a knee point of the indexes, so detection methods through BIC in partition-based clustering are proposed in the present study.

Most of the currently existing parametric clustering methods partition data into a predefined number of clusters, with a cluster representative corresponding to each cluster, so that a well-defined cost function involving the data and its representatives is minimized. There are many different ways to express and formulate the clustering delinquent, as both clustering algorithm may deliver a different grouping for a data set depending on the cost function used. The categorisation of clustering methods is neither straightforward nor canonical, but one option is to classify the methods as hierarchical methods, partitional methods, density-based methods, graph-based methods, grid-based methods and methods for high-dimensional space data.

III DRAWBACKS OF EXISTING SYSTEM

It has low computational and memory space requirements. These methods have not gained wide acceptance because of their great time complexity. There is no universal function for all clustering problems. It has not gained wide acceptance because of their great time complexity. In a high dimensional space, the data converts sparse, and modern indexing and algorithmic techniques fail to be efficient and/or effective.

IV PROPOSED SYSTEM

A Centroid Ratio is firstly introduced to compare two clustering consequences. This centroid proportion is then used in prototype-based clustering by introducing a Pairwise Random Swap clustering algorithm to avoid the local optimum problem of k-means. We propose a cluster validity index called the centroid ratio, which can be used to compare two clusterings and find unstable and incorrectly located centroids in them. As the centroid ratio can find incorrectly located centroids in two clusterings, we use this property and propose a novel clustering algorithm called the Pairwise Random Swap (PRS) [1] clustering algorithm.

The experimental results indicate that the proposed algorithm requires 26% to 96% less processing time than the second fastest algorithm (RS) and avoids the local optimality problem better than the other swap strategies. We proposed a novel evaluation criterion called the centroid ratio, based on the centroids in prototype-based clustering, which compares two clusterings and detects unstable centroids [10] and incorrectly located centroids. The design of the internal indices is based on three elements: the data set, the point level partitions, and centroids. A criterion such as MSE uses quantities and features inherent in the dataset, which gives a global level of evaluation. Since it relates to points and clusters, its time complexity is at least *O(MN)*.

The efficiency of a swap-based clustering algorithm depends on two issues: how many iterations (swaps) are needed and how much time each iteration consumes. Swap-based clustering can be categorized into four types in terms of the swap strategy: *RR*, *RD*, *DR* and *DD*.The centroid ratio is highly correlated with external indices and MSE values. The applications of the proposed algorithm to document clustering and color image quantization indicate that the algorithm is useful and is not restricted by the distance function of *k*-means.

For high-dimensional data [7], the running time of the proposed algorithm has high variance.Centroids represent a global structure of prototypes, and utilizing only centroids in the evaluation reduces the time complexity.Swapping iterations are needed in RS and DRS and repetitions are needed for RKM and KM++ to guarantee good performance.The proposed algorithm has the best performance in its running time. Most efficient method among these algorithms according to the experimental results. PRS requires 26% to 96% ess processing time than the others on different data sets.

V MODULES

- Term Vector Calculation
- Vector Similarity Calculation
- Clustering
- Pairwise Random swap

A) Term Vector Calculation

The term vector is an algebraic model for representing text documents as vectors of identifiers.Itis used in information straining, information recovery; indexing and relevancy rankings. Each dimension corresponds to a separate term. If a term occurs in the text, its worth in the path is non-zero otherwise nothing.

B) Vector Similarity Calculation

In this work, vector similarity calculation accomplished by cosine similarity. Cosine connection is a quantity of correspondence between two vectors of an inner product space that measures the cosine of the angle among them.Cosine similarity is mostly used in optimistic space, where the outcome is neatly bounded in [0,1].Cosine similarity is most commonly used in high-dimensional positive spaces. The technique is also used to measure cohesion within clusters in the field of data mining.
The Cosine Similarity of two vectors (d1 and d2) is definedas:
cos(d1, d2) = dot(d1,d2) / ||d1|| ||d2||

Where dot(d1,d2) = d1[0]*d2[0] + d1[1]*d2[1] …
And Where ||d1|| = sqrt(d1[0]^2 + d1[1]^2 …)

C) Clustering

Clustering is performed by means of K-Mean algorithm.The k-Mean clustering is distance threshold based clustering.Clusters formed by similarity distance threshold value.

K-Mean algorithm

K-means clustering is a method of vector quantization, originally from indication processing, that is current for cluster examination in data mining. K-means clustering [3] aims to partition observations into k clusters in which each observation belongs to the cluster with the adjacentmean, ration as a prototype of the collection.

Given a set of observations (x_1, x_2…x_n), where each observation is a d-dimensional actualpath, k-means clustering purposes to barrier the n observations into k sets (k ≤ n) S = {S_1, S_2… S_k} so as to minimize the within-cluster sum of squares whereμ_i is the mean of points in S_i.Given an initial set of k means $m_1^{(1)}$,$m_k^{(1)}$

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2 \qquad (1)$$

Assignment step: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclideandistance, this is intuitively the bordering mean.

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \le \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \le j \le k\}, \qquad (2)$$

Where each $x_p$is assigned to exactly one$S^{(t)}$, even if it could be is assigned to two or more of them.

Update step: Calculate the new means to be the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \qquad (3)$$

Since the arithmetic mean is a least-squaresestimator, this also minimizes the within-cluster sum of squares (WCSS) objective.

The algorithm has converged when the projects no extended change. Since both stages optimize the WCSS independent, and there only occurs a limited number of such partitioning’s, the algorithm must converge to anoptimal. There is no assurance that the global peak is found using this algorithm.

D) Pairwise Random swap

In this Module, calculate the mean square error value and initializing cluster centroid randomly. Determine the dislocated centroids by means of MSE value.This swapping continued until best fit centroids are found.

Mean Square Error

If $\hat{Y}$ is a vector of n calculations, and Y is the route of the true standards, then the

$$\operatorname{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2. \qquad (4)$$

This is a recognized, computed quantity given a particular sample (and hence is sample-dependent).The MSE of an estimator$\hat{\theta}$with respect to the unknown parameter $\theta$is defined as

$$\operatorname{MSE}(\hat{\theta}) = \operatorname{E}\left[(\hat{\theta} - \theta)^2\right]. \qquad (5)$$

This definition depends on the unidentifiedlimitation, and the MSE in this sense is a property of an estimator (of a method of obtaining an estimate).

$$\operatorname{MSE}(\hat{\theta}) = \operatorname{Var}(\hat{\theta}) + \left(\operatorname{Bias}(\hat{\theta}, \theta)\right)^2. \qquad (6)$$

The MSE is equal to the sum of the variance and the squared bias of the estimator or of the estimates. In the circumstance of the MSE of an estimator.

VISYSTEM ARCHITECTURE

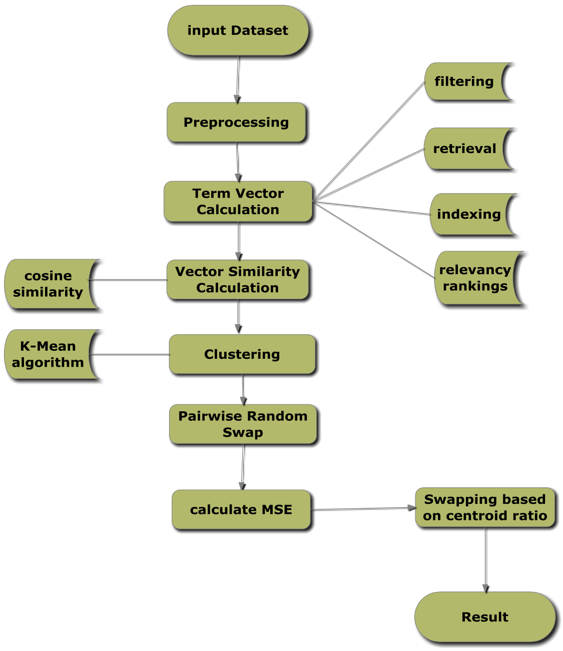

Fig 6.1 system architecture

In the fig 6.1 the input datasets are pre-processed by the term vector calculation by representing as a text documents. It is used to filtering,retrieval,indexing,and relevancyrankings. The vector similarity calculation accomplished by cosine similarity. Clustering is performed K-Means algorithm it is based on threshold based clustering. It is formed by similarity distance threshold value. The mean square error is calculated by initializing cluster to determine the dislocated centroids by MSE values. The swapping continues until best fit centroids are found.

VII CONCLUSION

We proposed a novel evaluation criterion called the centroid ratio, based on the centroids in prototype-based clustering, which compares two clusterings. Cluster analysisis 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 filesarrangements. The Mean squared errorcreated 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.

The experimental results indicate that the proposed algorithm requires 26% to 96% less processing time than the second fastest algorithm (RS) and avoids the local optimality problem. To ensure a good clustering quality, the number of iterations for random swap should be set large enough to find successful swap setter than the other swap strategies. The similarity value obtained from the centroid ratio is employed as a stopping criterion in the algorithm. Random swap needs a large number of iterations to provide a good quality of clustering.

REFERENCES

[1] C.Veenman, M. Reinders, and E. Backer, “A maximum variance cluster algorithm,” *IEEE* Trans. Pattern Anal. Mach. Intell., vol. 24, no. 9, pp. 1273–1280, Sept. 2002

[2] H. Xiong, J. Wu, and J. Chen, “K-means clustering versus validation measures: A data-distribution perspective,”IEEE Trans. Syst., Man, Cybern. B, Cybern.,.vol. 39, no. 2,pp. 318–331, Apr. 2009.

[3] I. Dhillon, Y. Guan, and J. Kogan, “Iterative clustering of high dimensional text data augmented by local search,” *in Proc. IEEE ICDM*, Washington, DC, USA, 2002,pp. 131–138.

[4] Jiasi Chen, Video Clustering Using Compression Quality and Motion Features Princeton University,2011

[5] K. Krishna and M. NarasimhaMurty, Genetic K-Means Algorithm,IEEETrans. Syst., Man, Cybern. B, Cybern., vol. 29, no. 3, pp. 433–439,Jun. 1999.

[6] M. Steinbach, G. Karypis, and V. Kumar, “A comparison of document clustering techniques,” in Proc. Knowl. Discov.Databases Workshop Text Mining, 2000.

[7] P. Franti, M. Rezaei, and Q. Zhao, Centroid index: Cluster level similarity measure, Pattern Recognition.

[8] Qinpei Zhao and Pasi Fränti, Centroid Ratio for a Pairwise Random Swap Clustering Algorithm SeniorMember, IEEE transactions on knowledge and data engineering, vol. 26, no. 5, may 2014.

[9] T. Zhang, R. Ramakrishnan, and M.Livny, “BIRCH: A new data clustering algorithm and its applications,” Data Min. Knowl. Discov, vol. 1, no. 2, pp. 141–182, 1997.

[10] T. Hasan, Y. Lei, A. Chandrasekaran, and J. Hansen, “A novel feature sub-sampling method for efficient universal background model training in speaker verification,” in *Proc. IEEE ICASSP*,Dallas, TX, USA, 2010, pp. 4494–4497.