A Comparison of Clustering Techniques in Data Mining

[1] Rahumath Beevi A, [2] Remya R.

[1] Department Of Computer Science and Engineering, Cusat, College Of Engineering Perumon.
Kollam, Kerala, India

[2] Department Of Information Technology ,Cusat, College Of Engineering Perumon
Kollam, Kerala, India

Abstract - Clustering is an important tool in data analysis, as data set grows then their properties and interrelationships will also change. There are different types of cluster model: Connectivity models, Distribution models, Centroid models, Subspace model, Group models and Graph-based models. Clustering algorithms can be categorized based on the models which are using .Traditionally clustering techniques are broadly divided into hierarchical and density based clustering. There are so many clustering methods because the notion of cluster cannot be easily defined. Data mining deals with large data sets and their relationships, while we are imposing clustering to analyze the huge data that needs additional challenges. This leads to an efficient and broadly applicable clustering method. In this paper some of the clustering techniques are discussed.

***Keywords* - Data mining, Hard and soft clustering, fuzzy clustering, sentence level clustering, Concept analysis, Graph centrality.**

1. Introduction

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining is also known as Knowledge Discovery in Data [1]. It is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining is accomplished by building models. A model performs some actions on data based on some algorithm. The notion of automatic discovery refers to the execution of data mining models. Data mining techniques can be divided into supervised or unsupervised. Clustering is one of the unsupervised techniques. Clustering is the process of grouping a set of objects in such a way that object in the same group are more similar to each other than those in other cluster. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Clustering has become an increasingly important topic with the explosion of information available via the Internet. It is an important tool in text mining and knowledge discovery. Representing data by fewer clusters necessarily loses certain fine details, but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters.

From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. The goal of undirected data mining is to discover structure in the data as a whole. There is no target variable to be predicted, thus no distinction is being made between independent and dependent variables. Therefore, clustering is unsupervised learning of a hidden data concept. Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships.

Clustering methods are broadly divided into hierarchical and Partitional [2]. In Partitional method cluster formation is based on portioning the data set into set of disjoint ones. Several extensions to it in form of hard and fuzzy representations have also been proposed in it .In hierarchical technique it produce a nested sequence of partitions; with a single, all inclusive clusters at the top and Singleton clusters of individual points at the bottom. Hard clustering is conventional; here each point of the data set belongs to exactly one cluster. In fuzzy, each point belongs to each cluster to certain degree. Clustering text at the document level is well established and it has been used in a number of different areas of text mining and information retrieval. It has been improved the precession or recall in information retrieval system and has an efficient way of finding the nearest neighbors of a documents. Clustering has been proposed for use in browsing a collection of documents or in organizing the results returned by a search engine in response to users query. In document clustering, documents are represented as data points in high dimensional vector space and each document has its own attribute .document clustering can be done using large range of algorithms. The semantic relationship between documents can be measured in terms of word co-occurrence. However, while the assumption that (semantic) similarity can be measured in terms of word co-occurrence may be valid at the document level, the assumption does not hold for small-sized text fragments such as sentences, since two sentences may be semantically related despite having few, if any, words in common. Many sentence will be related to some degree to no. of sentences, if are able to capture such fuzzy relationship will lead to an increase in the breadth and scope of problems to which sentence clustering can be applied.

2. Literature Survey

2.1 K-means Methods

k- Means [3] is one of the partitioning based clustering methods. The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. In k-means case a cluster is represented by its centroid, which is a mean (usually weighted average) of points within a cluster. Each point is assigned to the cluster with the closest centroid Number of clusters, K, must be specified. This obviously does not work well with a categorical attributes, it has the good geometric and statistical sense for numerical attributes. K-means has problems when clusters are of differing Sizes, Densities, Non-globular shapes and K-means has problems when the data contains outliers.

2.2 k- Medoids Method

When medoids are selected, clusters are defined as subsets of points close to respective medoids, and the objective function is defined as the averaged distance or another dissimilarity measure between a point and its medoid. K-medoid [4] is the most appropriate data point within a cluster that represents it. Representation by k-medoids has two advantages. First, it presents no limitations on attributes types, and, second, the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and, therefore, it is lesser sensitive to the presence of outliers.

2.3 Vector Space Model

This model is more suitable for document clustering .Hence it can be further applicable in document compression. Here the documents are represented using the vector space model. In this model, each document, d, is considered to be a vector, d, in the term-space. In its simplest form, each document is represented by the (TF) vector [5],

$$d_{tf} = (tf_1,\ tf_2,\ \ldots,\ tf_n), \qquad (1)$$

where tfi is the frequency of the ith term in the document.

In addition, here use version of this model that weights each term based on its inverse document frequency (IDF) in the document collection. (This discounts frequent words with little discriminating power.) Finally, in order to account for documents of different lengths, each document vector is normalized so that it is of unit length. Here the similarity can be found out using one of the similarity measures such as cosine similarity, multi viewpoint similarity. This approach is does not hold for sentence level clustering.

2.4 Fuzzy Clustering

The fuzzy set, first proposed by Zadeh [6] in 1965, is an extension to allow $p_i(x)$ to be a function (called a membership function) assuming values in the interval [0, 1].Traditional clustering approaches generate partitions; in a partition, each pattern belongs to one and only one cluster. Hence, the clusters in a hard clustering are disjoint. Fuzzy clustering extends this notion to associate each pattern with every cluster using a membership function. The output of such algorithms is a clustering, but not a partition.

2.5 Fuzzy C Means

Ruspini [7] introduced a fuzzy c-partition p = $(p_1,p_2 \ldots. p_c)$ by the extension to allow $p_i(x)$ to be functions assuming values in the interval (O ,l] such that $P_1(x) + \ldots + p_c(x) = 1$ since he first applied the fuzzy set in cluster. In fuzzy object data clustering, on the other hand, the problem of classifying N objects into C types is typically solved by, first, finding C prototypes, which best represent the characteristics of as many groups of objects, and then building a cluster around each such prototype, by assigning each object a membership degree that is as much higher as greater its similarity degree with the prototype is. A prototype may be either a cluster center, or the most centrally located object in a cluster, or a probability distribution, etc., depending on the type of available data and the specific algorithm adopted. It should be noted that the knowledge of prototypes, which are a condensed representation of the key characteristics of the corresponding clusters, is also an important factor. Here the distance calculations for stable clusters in the iterative process, when the number of proceeding iterations increases the cluster center number will also increases.

2.6 ARCA Algorithm

P .Corsini et al introduced Any Relational Clustering Algorithm (ARCA) [8] which is based on the Fuzzy C-means (FCM) algorithm. ARCA is very stable and it represents clusters with high membership value in terms of the mutual relationship between the objects. It uses an attribute based representation. Each object is represented by the vector of its relation strengths with other objects in the data Set. It takes relational data as input, thus FCM to be applied.

The algorithm initially choose a partition $U(0)$ and at the step *l, l=0,1,2,*etc it will calculate the prototype vector $V(l)$ .After that $U(l)$ will be updated to $U(l+1)$ and compare its value in a suitable matrix. Depending upon the value of the predetermined threshold, algorithm will either stop or return to a particular point. Prototypes in this system are therefore objects (not necessarily present in the original data set) whose relationship with all objects in the data set is representative of the mutual relationships of a group of similar objects. A limitation of this approach is the high dimensionality introduced by representing objects in terms of their similarity with all other objects.

2.7 FRECCA (fuzzy relational eigenvector centrality based clustering algorithm)

Clustering text at the document level is well established in the Information Retrieval (IR). Because it is able to adequately capture much of the semantic content of document-level text. This is because documents that are semantically related are likely to contain many words in common, and thus are found to be similar according to popular vector space measures such as cosine similarity, which are based on word co-occurrence. This is valid only at the document level only not hold for small sized text fragments such as sentences .Fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. This is important in domains such as sentence clustering, since a sentence is likely to be related to more than one theme or topic present within a document or set of documents. However, because most sentence similarity measures do not represent sentences in a common metric space, conventional fuzzy clustering approaches based on prototypes or mixtures of Gaussians are generally not applicable to sentence clustering.

Andrew Skabar et al. proposed the Fuzzy Relational Eigen Vector Centrality Based clustering algorithm (FRECCA) [9], motivated by the mixture model approach in which the data is represented as a combination of components. A graph representation of data objects is used here along with the PageRank algorithm. It operates within an Expectation–Maximization (EM) [10] framework. Each sentence in a document is represented by node in the directed graph and the weighted objects will indicate the object similarity. In order to measure the relative importance of a hyperlinked set of documents PageRank will assigns numerical weighting to each element. And by using this importance we can easily determine the centrality of the graph.

Cluster membership values for each node indicate the contribution of each data object into a particular cluster and the mixing coefficients will point out the probability of an object being generated from a component. These two parameters are needed to determine to start on with the FRECCA algorithm and will be optimized by Expectation Maximization. The input to the algorithm is pairwise similarities between the sentences and the required number of output of clusters. The semantic similarity between sentences can be measured by using cosine similarity.

3. Performance Analysis

The performance evaluation of the above discussed approaches is based on the external cluster evaluation measures such as Partition Entropy Coefficient (PE), Purity and Entropy, V-Measure, Rand Index and F-Measure. The experimental comparison is carried out for 5 numbers of clusters.

We apply the algorithm to clustering famous quotations from [11]. Table 1 shows the result of applying the FRECCA, ARCA, vector space model and k-Medoids algorithms to the quotations data set and evaluating using the external measures.

Table 1: Clustering Performance Evaluation

Techniques	Purity	Entropy	V-meas	F-meas
K-medoids	0.608	0.396	0.560	0.482
Fuzzy medoids	0.699	0.365	0.598	0.549
Vector space	0.678	0.380	0.624	0.689
ARCA	0.699	0.375	0.611	0.481
FRECCA	0.713	0.335	0.634	0.531

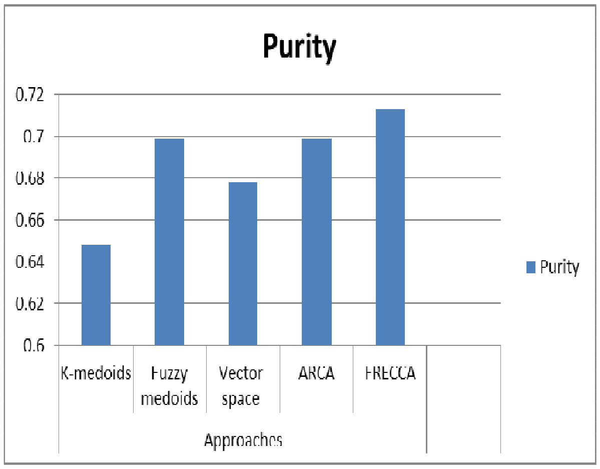

Fig. 1 Purity Comparison

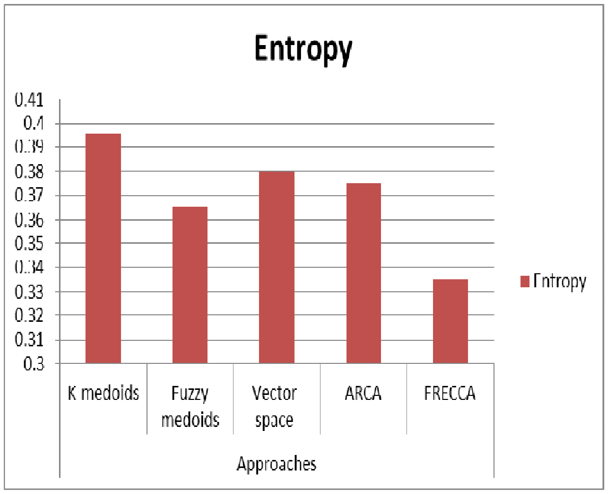

Fig 2 Entropy Comparison

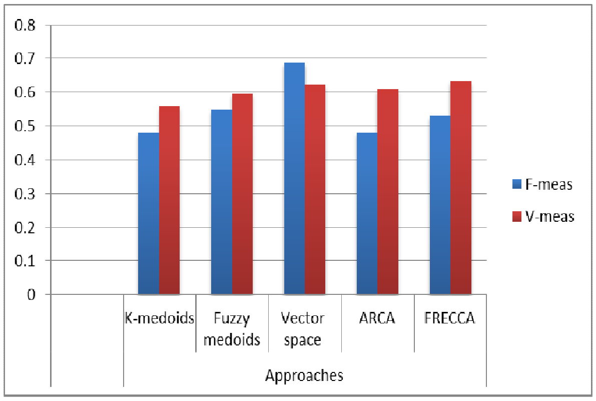

Fig.3 Performance Comparison

4. Conclusions

We have already reviewed numerous clustering algorithms. But it is necessary to pre assume the number c of clusters for all these algorithms. Therefore, the method to find optimal c is very important. By analyzing various methods it is clear that each of them have their own advantages and disadvantages. The quality of clusters depends on the particular application. When the inter-object relationship has no metric characteristics then ARCA is a better choice. Among the different fuzzy clustering techniques FRECCA algorithm is superior to others. It is able to overcome the problems in sentence level clustering. But when time is critical factor then we cannot adopt fuzzy based approaches. A good clustering of text requires effective feature selection and a proper choice of the algorithm for the task at hand. It is observed from the above analysis that fuzzy based clustering approaches provide significant performance. But, fuzzy approaches do have certain drawbacks which have to be eliminated.

Acknowledgments

We express my sincere gratitude to all the staff of Computer Science and Engineering Department in our college and beloved family members who helped me with their timely suggestions and support. I also express my sincere thanks to all my friends who helped me throughout the successful completion of the work. All glory and honor be to the Almighty God, who showered His abundant grace on me to make this work a success.

References

[1] Oded Maimon, Lior Rokach, “Data Mining AND Knowlwdge Discovery Handbook”, Springer Science+Business Media.Inc, pp.321-352, 2005.

[2] P. Berkhin, “A Survey of Clustering Data Mining Techniques” Kogan,Jacob; Nicholas, Charles; Teboulle, Marc (Eds) Grouping Multidimensional Data, Springer Press (2006) 25-72

[3] J.B MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations,” Proc. Fifth Berkeley Symp. Math. Statistics and Probability, pp. 281-297, 1967.

[4] R. Krishnapuram, A. Joshi, and Y. Liyu, “A Fuzzy Relative of them k-Medoids Algorithm with Application to Web Document and Snippet Clustering,” Proc. IEEE Fuzzy Systems Conf., pp. 1281-1286,1999.

[5] Gerald Kowalski, Information Retrieval Systems – Theory and Implementation, Kluwer Academic Publishers, 1997.

[6] L.A. Zadeh, Fuzzy sets, Inform. and Control 8, 338-353 (1965).

[7] E.H. Ruspini, A new approach to clustering, Inform. and Control 15, 22-32 (1969).

[8] P. Corsini, F. Lazzerini, and F. Marcelloni, “A New Fuzzy Relational Clustering Algorithm Based on the Fuzzy C-Means Algorithm,” Soft Computing, vol. 9, pp. 439-447, 2005.

[9] Andrew Skabar, Member, IEEE, and Khaled Abdalgader “Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm” 1041-4347/13/$31.00 2013 IEEE Published by the IEEE Computer Society.

[10] A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” J. the Royal Statistical Soc. Series B (Methodological), vol. 39, no. 1, pp. 1-38, 1977.

[11] http://www.famousquotesandauthors.com, 2012.

Author Details

Rahumath Beevi A Received the bachelor’s degree in Computer Science and Engineering from Cochin University of Science and Technology, Kerala in 2012. Presently she is pursuing her Mtech in the department of Computer Science and Engineering from Cochin University of Science and Technology, Kerala. Her research interests include Data Mining.

Remya R Received the bachelor’s degree in Information Technology from University college Trivandrum, Kerala in 2004 and master’s degree in Computer Science and Engineering from Anna University, Coimbatore in 2008. Currently working as an Assistant Professor in Information Technology department, of College of Engineering Perumon,under Cochin university of Science and Technology. She has teaching experience of eight years. Research interests Includes Data Mining.

.