Text Mining and Clustering Analysis

Shobha S. Raskar[†], **D. M. Thakore**[††]

[†]Bharati Vidyapeeth College of Engineering , BVUCOE, Dhanakwadi, pune
[††]Bharati Vidyapeeth University, College of Engineering BVUCOE, Dhankawadi ,Pune

Abstract:
Cluster analysis is required in text mining for grouping objects. Cluster analysis consists of different algorithms and methods for grouping objects of similar kinds into respective categories. Cluster analysis is exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal, if they belong to same group and minimal otherwise. It can be used to discover structure in data without providing an explanation or interpretation. Cluster analysis simply discover structure in data without explaining, why they exist. Aim of text mining, text clustering is to divide collection of text document into different category group should be of little similarity. Cluster is comprised of number of similar object collected or grouped together. Cluster analysis is tool for exploring structure of data clustering is subjective or problem dependent. Basic objective in cluster analysis is to discover natural grouping of items. Quantitative scale is developing which measure association between object, these scales are referred as similarity measure.

Keywords:
Clustering, K-mean, Expectation Maximization(EM) , Distance measure

1. Introduction

The term cluster analysis which is first used by Tryon, in 1939, consists of number of different algorithm and methods for grouping

objects of similar kind into respective categories. Data clustering is a method in which cluster of objects made that are somehow similar in characteristics. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre defined classes, whereas in clustering the classes are also to be defined. Precisely, Data Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency in the database systems the number of disk accesses is to be minimized. In clustering the objects of similar properties are placed in one class of objects and a single access to the disk makes the entire class available.

A cluster is an ordered list of objects, which have some common characteristics. The distance between two clusters involves some or all elements of the two clusters. A similarity measure SIMILAR (Di, Dj) can be used to represent the similarity between the documents. There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. Clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap. These methods are sometimes divided into partitioning methods, in which the classes are mutually exclusive, and the less common clumping method, in which overlap is allowed. Each object is a member of the cluster with which it is most similar; however the threshold of similarity has to be defined. The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or divisive methods. In agglomerative methods, the hierarchy is build up in a series of N-1 agglomerations beginning with the un-clustered dataset. The less common divisive methods begin with all objects in a single cluster and at each of N-1 steps divide some clusters into two smaller clusters, until each object resides in its own cluster.

The main distinction between clustering and the nearest neighbor technique is that clustering is what is called an unsupervised learning technique and nearest neighbor is generally used for prediction or a supervised learning technique. Unsupervised learning techniques are unsupervised in the sense that when they are run there is not particular reason for the creation of the models the way there is for supervised learning techniques that are trying to perform prediction. In prediction, the patterns that are found in the database and presented in the model are always the most important patterns in the database for performing some particular prediction. In clustering there is no particular sense of why certain records are near to each other or why they all fall into the same cluster. Some of the differences between clustering and nearest neighbor prediction can be summarized in Table 1.1.

Manuscript received June 5, 2011
Manuscript revised June 20, 2011

Nearest Neighbor	*Clustering*
Used for prediction as well as consolidation.	Used mostly for consolidating data into a high-level view and general grouping of records into like behaviours.
Space is defined by the problem to be solved (supervised learning).	Space is defined as default n-dimensional space, or is defined by the user, or is a predefined space driven by past experience (unsupervised learning).
Generally only uses distance metrics to determine nearness	Can use other metrics besides distance to determine nearness of two records - for example linking two points together

Similarities are a set of rules that serve as criteria for grouping or separating items. Cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Given the above, cluster analysis can be used to discover structures in data without providing an explanation/interpretation.

Consider example, a group of diners sharing the same table in a restaurant may be regarded as a cluster of people. In food stores items of similar nature, such as different types of meat or vegetables are displayed in the same or nearby locations. There are a countless number of examples in which clustering plays an important role. General categories of cluster analysis methods are Tree clustering, Block clustering, EM clustering and k-means clustering,.

2. Uses of Distance Measures in tree clustering:

The joining or tree clustering method uses the dissimilarities (similarities) or distances between objects when forming the clusters. Similarities are a set of rules that serve as criteria for grouping or separating items. In the previous example the rule for grouping a number of dinners was whether they shared the same table or not. These distances (similarities) can be based on a single dimension or multiple dimensions, with each dimension representing a rule or condition for grouping objects. For example, if cluster fast foods, we could take into account the number of calories they contain, their price. The Microsoft Clustering algorithm provides two methods for creating clusters and assigning data points to the clusters. The first, the K-means algorithm, is a hard clustering method. This means that a data point can belong to only one cluster, and that a single probability is calculated for the membership of each data point in that cluster. The second method, the Expectation Maximization (EM) method, is a soft clustering method. This means that a data point always belongs to multiple clusters, and that a probability is calculated for each combination of data point and cluster.

In EM clustering, the algorithm iteratively refines an initial cluster model to fit the data and determines the probability that a data point exists in a cluster. The algorithm ends the process when the probabilistic model fits the data. The function used to determine the fit is the log-likelihood of the data given the model. If empty clusters are generated during the process, or if the membership of one or more of the clusters falls below a given threshold, the clusters with low populations are reseeded at new points and the EM algorithm is rerun. The results of the EM clustering method are probabilistic. This means that every data point belongs to all clusters, but each assignment of a data point to a cluster has a different probability. Because the method allows for clusters to overlap, the sum of items in all the clusters may exceed the total items in the training set. In the mining model results, scores that indicate support are adjusted to account for this EM algorithm is the default algorithm used in Microsoft clustering models. This algorithm is used as the default because it offers multiple advantages in comparison to k-means clustering:

- Requires one database scan, at most.
- Will work despite limited memory (RAM).
- Has the ability to use a forward-only cursor.
- Outperforms sampling approaches.

The Microsoft implementation provides two options: scalable and non-scalable EM. By default, in scalable EM, the first 50,000 records are used to seed the initial scan. If this is successful, the model uses this data only. If the model cannot be fit using 50,000 records, an additional 50,000 records are read. In non-scalable EM, the entire dataset is read regardless of its size. This method might create more accurate clusters, but the memory requirements can be significant. Because scalable EM operates on a local buffer, iterating through the data is much faster, and the algorithm makes much better use of the CPU memory cache than non-scalable EM. Moreover, scalable EM is three times faster than non-scalable EM, even if all the data can fit in main memory. In the majority of cases, the performance improvement does not lead to lower quality of the complete model.

The EM (expectation maximization) algorithm extends this basic approach to clustering in two important ways:

1. Instead of assigning cases or observations to clusters to maximize the differences in means for continuous variables, the EM clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions. The goal of the clustering algorithm then is to maximize the overall probability or likelihood of the data, given the final clusters.

2. Unlike the classic implementation of k-means clustering, the general EM algorithm can be applied to both continuous and categorical variables.

The goal of EM clustering is to estimate the means and standard deviations for each cluster so as to maximize the likelihood of the observed data or distribution

2.1 Term frequency and Inverse document frequency weighting:

This scheme overcomes the drawback of term frequency model by including the global weight. To understand the global weight and the local weight following table can be used.

Table 2.1: Term document Matrix

	Doc1	***Doc2***	***Doc3***	***Doc4***
Term1	2	1		3
Term2		2		1

The term1 occurs in total of 3 documents while term2 occurs in total of 2 documents. Now,

Global weight of term1 = total documents / number of document(s) with the term1 =log(4/3)

Similarly, the global weight of term2 = log(4/2)

Thus, the global weight is the overall importance of the term which decreases as the number of document containing the tem increases. The tfidf scheme aims at balancing the local and the global term occurrences in the documents and can be defined as,

$$aij=tfij \cdot \log(N/dfi)$$

where ,

tfij : is the term frequency in document dj,

dfi :denotes the number of documents in which term i appears, and N represents the total number of documents in the collection.

3. K-Means Clustering

K-means clustering is a well-known method of assigning cluster membership by minimizing the differences among items in a cluster while maximizing the distance between clusters. The "means" in k-means refers to the centroid of the cluster, which is a data point that is chosen arbitrarily and then refined iteratively until it represents the true mean of all data points in the cluster. The "k" refers to an arbitrary number of points that are used to seed the clustering process. The k-means algorithm calculates the squared Euclidean distances between data records in a cluster and the vector that represents the cluster mean, and converges on a final set of k clusters when that sum reaches its minimum value. The k-means algorithm assigns each data point to exactly one cluster, and does not allow for uncertainty in membership. Membership in a cluster is expressed as a distance from the centroid. Typically, the k-means algorithm is used for creating clusters of continuous attributes, where calculating distance to a mean is straightforward. However, the Microsoft implementation adapts the k-means method to cluster discrete attributes, by using probabilities. For discrete attributes, the distance of a data point from a particular cluster is calculated using function. In contrast to clustering, where groups are unknown at the beginning, classification tries to put specific documents into groups known in advance. The same basic means can be used as in clustering, like bag-of-words representation as a way to formalize unstructured text. The real-world examples are spam classification of e-mails or classifying news articles into topics. Examples of a very simple classifier (k-nearest neighbor), and more advanced method (Support Vector Machines).

k-means method will produce exactly k different clusters of greatest possible distinction. It should be mentioned that the best number of clusters k leading to the greatest separation (distance) is not known as a *priori* and must be computed from the data.

n K-Means clustering algorithm values of k is given and k-means algorithm is implemented in 4 steps:

1. Partition objects into k nonempty subsets.
2. Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
3. Assign each object to the cluster with the nearest seed point. Go back to Step 2, stop when no more new assignment.

The k-means algorithm provides two methods of sampling the data set: non-scalable K-means, which loads the entire data set and makes one clustering pass, or scalable k-means, where the algorithm uses the first 50,000 cases and reads more cases only if it needs more data to achieve a good fit of model to data. The most straightforward way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances.

3.1 Euclidean distance.

This is probably the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space. It is computed as:

$$\text{distance}(x,y) = \{ \sum_i (x_i - y_i)^2 \}^{1/2}$$

Euclidean and squared Euclidean distances are usually computed from raw data, and not from standardized data. This method has certain advantages e.g., the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers. However, the distances can be greatly affected by differences in scale among the dimensions from which the distances are computed. For example, if one of the

dimensions denotes a measured length in centimetres, and you then convert it to millimetres by multiplying the values by 10, the resulting Euclidean or squared Euclidean distances which are computed from multiple dimensions can be greatly affected i.e., biased by those dimensions which have a larger scale, and consequently, the results of cluster analyses may be very different. Generally, it is good practice to transform the dimensions so they have similar scales.

3.2 Squared Euclidean distance.

This square the standard Euclidean distance in order to place progressively greater weight on objects that are further apart. This distance is computed as :

$$\text{distance}(x,y) = \quad i (x_i - y_i)^2$$

3.3 City-block (Manhattan) distance.

This distance is simply the average difference across dimensions. In most cases, this distance measure yields results similar to the simple Euclidean distance. However, note that in this measure, the effect of single large differences (outliers) is dampened (since they are not squared). The city-block distance is computed as:

$$\text{distance}(x,y) = \quad i |x_i - y_i|$$

3.4 Chebychev distance.

This distance measure may be appropriate to define two objects as "different" if they are different on any one of the dimensions. The Chebychev distance is computed as:

$$\text{distance}(x,y) = \text{Maximum}|x_i - y_i|$$

3.5 Power distance.

To increase or decrease the progressive weight that is placed on dimensions on which the respective objects are very different. This can be accomplished via the power distance. The power distance is computed as:

$$\text{distance}(x,y) = \left(\quad i |x_i - y_i|^p \right)^{1/r}$$

where r and p are user-defined parameters. Parameter p controls the progressive weight that is placed on differences on individual dimensions, parameter r controls the progressive weight that is placed on larger differences between objects. If r and p are equal to 2, then this distance is equal to the Euclidean distance.

3.6 Percent disagreement.

This measure is particularly useful if the data for the dimensions included in the analysis are categorical in nature. This distance is computed as:

$$\text{distance}(x,y) = (\text{Number of } x_i \quad y_i)/ i$$

3.7 Advantage K-mean :

Rapidity, simplicity, high scalability, easy realization, fast speed, improves efficiency.

3.8 Disadvantage:

Numbers of cluster should known in advance. In K-means, if pre-specified number of clusters modified, the precision of each result is also modified.

4 Conclusions:

The Microsoft Clustering algorithm does not expose the distance function used in computing k-means, and measures of distance are not available in the completed model. Prediction function use to return a value that corresponds to distance, where distance is computed as the probability of a data point belonging to the cluster. The basic operation of k mean algorithm is relatively simple: Given a fixed number of k clusters, assign observations to those clusters so that the means across clusters are as different from each other as possible.

References:

[1] Athman Bouguettaya "On Line Clustering", IEEE Transaction on Knowledge and Data Engineering Vol 8, No. 2, Apr 1996.

[2] Euripides G.M. Petrakis and Christos Faloutsos "Similarity Searching in Medical Image Databases", IEEE Transaction on Knowledge and Data Engineering Vol 9, No. 3, MAY/JUNE 1997.

[3] Rob Short, Rod Gamache, John Vert and Mike Massa "Windows NT Clusters for Availability and Scalability" Microsoft Online Research Papers, Microsoft Corporation.

[4] Simona Balbi, Emilio Di meglio, Dip. Di mothe statistica university, Federico ildi Napoli, ‘ Text mining strategy based on local context of words.’

[5] Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, university of Waikato, New Zealand, ‘KEA: Practical Automatic Keyphrase Extraction’.

[6] Milos Radovanovic, Mirjana Ivanovic, ‘Text mining: Approaches and Applications’, Novi Sad J. Math , Vol. 38, No. 3 , 2008

Mrs. Shobha S. Raskar, is a student of M. Tech in the Department of Computer Engineering, Bharati Vidyapeeth University College of Engineering, Pune. She obtained her B.E. Computer Engineering in 1999 from Cummins College of Engineering for Women, Pune University. Her research interest is, Text mining/Data mining.

Prof. D. M. Thakore graduated (B.E–Computer Engineering) from Walchand College of Engineering, Sangli and State–Maharashtra in 1990. He pursued his M.E. (Computer) from Bharati
Vidyapeeth University College of Engineering, Pune in 2004. He also completed MBA (Marketing) from Pune University in 1996. He is currently working and pursuing his Ph.D. with subject Data Mining/Text Mining from Bharati Vidyapeeth Deemed University College of Engineering, Pune.