Data Mining Project Report

Document Clustering

Meryem Uzun-Per

504112506

Table of Content

1.Project Definition

There is a rapidly growing amount of available electronic information such as online newspapers, journals, conference proceedings, Web sites, e-mails, etc. Using all these electronic information, controlling, indexing or searching is not feasible especially for human and also for search engines. Thus, automatic document organization is an important issue. By using document clustering methods we can insight into data distribution or we can preprocess data for other applications [1]. For example, if a search engine uses clustered documents in order to search an item, it can produce results more effectively and efficiently.

Document clustering is an automatic clustering operation of text documents so that similar or related documents are presented in same cluster, dissimilar or unrelated documents are presented in different clusters [1]. Clustering is an unsupervised learning method which does not need any training step; pre-defined categories and labeled documents. So, there is no need for a training set while applying the clustering algorithms. It just uses the input data in order to find regularities in it.

In this project, we aim to cluster documents into clusters by using some clustering methods and make a comparison between them.

2.Literature Survey

There are several clustering approaches. These are partitioning (eg. K-means, k-medoids), hierarchical (eg. DIANA, AGNES, BIRCH), density-based (eg. DBSACN, OPTICS), grid-based (eg. STING, CLIQUE), model based (eg. EM, COBWEB), frequent pattern-based (eg. p-Cluster), user-quided or constraint-based (eg. COD), and link-based (eg. SimRank, LinkClus) clustering approaches [1]. Most of these are explained and some of them firstly proposed in the book of Kaufman and Rousseeuw in 1990 which are partitioning, hierarchical and fuzzy clustering approaches [2].

The most frequent method which is applied to documents is hierarchical clustering method. In 1988, Willett applied agglomerative clustering methods to documents by changing the calculation method of distance between clusters [3]. These algorithms have several problems with clusters that finding stopping point is very difficult and they run too slowly for thousands of documents. Hierarchical clustering algorithms are applied to documents for several times by Zhao and Karypis [4,5] and in 2005 they tried to improve agglomerative clustering algorithm by adding constrains [6].

K-means and its variants, which are partitioning clustering algorithms that create a non-hierarchical clustering consisting of k clusters, are applied to documents [2]. These algorithms are more efficient and scalable, and their complexity is linear to the number of documents. A disadvantage of k-means is that estimating the value of *k* wrongly leads worse accuracy. Moreover, k-means can stuck on a local maximum because of randomly chosen initial centroids. In order to solve this problem, Kaufman and Rousseeuw proposed k-medoids algorithm but this algorithm is computationally much more expensive and does not scale well large document sets [2].

A comparison of document clustering techniques is done by Steinbach and et al. in 2000 [7]. In their study, they applied k-means, its variant bisecting k-means, and hierarchical clustering algorithms to documents. It shows that average-link algorithm generally performs better than single-link and complete-link algorithms among hierarchical clustering methods for the document data sets used in the experiments. On the other hand, average-link algorithm is compared with k-means and bisecting k-means and it has been concluded that bisecting k-means performs better than average-link agglomerative hierarchical clustering algorithm and k-means algorithm in most cases for the data sets used in the experiments. The most recent study on document clustering is done by Liu and Xiong in 2011 [8]. They introduce common text clustering algorithms which are hierarchical clustering, partitioned clustering, density-based algorithm and self-organizing maps algorithm, analyze and compare some aspects of clustering algorithms such as the applicable scope, the initial parameters, termination conditions and noise sensitivity.

3.Methods

In this project, we want to apply and compare some partitioning and hierarchical approaches. The methods that we are going to handle are k-means, AGNES (Agglomerative Nesting) with different calculation methods for updating similarity matrix.

3.1. K-means algorithm

K-means algorithm is first applied to an N-dimensional population for clustering them into *k* sets on the basis of a sample by MacQueen in 1967 [9]. The algorithm is based on the input parameter *k*. First of all, *k* centroid point is selected randomly. These *k* centroids are the means of *k* clusters. Then, each item in the dataset is assigned to a cluster which is nearest to them. Then, means of all clusters are calculated again with new points added to them, until values of means do not change. In his book, Alpaydin symbolizes this algorithm like below where $\boldsymbol{m}$ is sequence of means, $\boldsymbol{x^t}$ is sequence of samples, and b is sequence of estimated labels [10].

```
Initialize $\boldsymbol{m}_i$, $i = 1,\ldots,k$, for example, to $k$ random $\boldsymbol{x}^t$
Repeat
    For all $\boldsymbol{x}^t \in X$
        $b_i^t \leftarrow 1$ if $\| \boldsymbol{x}^t - \boldsymbol{m}_i \| = \min_j \| \boldsymbol{x}^t - \boldsymbol{m}_j \|$
        $b_i^t \leftarrow 0$ otherwise
    For all $\boldsymbol{m}_i$, $i = 1,\ldots,k$
        $\boldsymbol{m}_i \leftarrow \sum_t b_i^t x^t \ / \sum_t b_i^t$
Until $\boldsymbol{m}_i$ converge
```

3.2. Hierarchical Clustering

The hierarchical clustering is a commonly used text clustering method, which can generate hierarchical nested classes. It clusters similar instances in a group by using similarities of them [10]. This requires the use of a similarity (distance) measure which is generally Euclidean measure in general, and cosine similarity for documents. Therefore, a similarity (distance) matrix of instances has to be created before running the method. Hierarchical clustering can be categorized into two; agglomerative (bottom-up) and divisive (top-down) clustering which are explained in [2] with the names AGNES, and DIANA.

3.2.1. AGNES: An agglomerative clustering algorithm starts with clusters which each of them contain only one instance and for each iteration merges the most similar clusters until the stopping criterion is met such as a requested number *k* of clusters is achieved [10]. The algorithm of agglomerative clustering [11]:

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the similarities between the clusters equal to the similarities between the items they contain.

2. Find the most similar pair of clusters and merge them into a single cluster, so that now you have one less cluster.

3. Compute similarities between the new cluster and each of the old clusters.

4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

At third step, the similarity (or distance) matrix is updated after merging two clusters. Although AGNES uses single-link method, this update can be done by different approaches [1,10, 11]. General definitions of these approaches are based on distance matrix but we can adapt them for similarity matrix like below where *G* symbolizes clusters:

Single-link: The similarity between two clusters is defined as the maximum similarity value from any member of one cluster to any member of the other cluster.

$$sim\left(G_i,\ G_j\right) = \max_{\mathrm{x^r \in G_i, x^s \in G_j}} sim(x^r, x^s)$$

Complete-link: The similarity between two clusters is defined as the minimum similarity value from any member of one cluster to any member of the other cluster.

$$sim\left(G_i,\ G_j\right) = \min_{\mathrm{x^r \in G_i, x^s \in G_j}} sim(x^r, x^s)$$

Centroid: The similarity between two clusters is defined as the similarity between centroids of two clusters.

$$sim\left(G_i,\ G_j\right) = sim(m^i, m^j)$$

3.2.2. DIANA: A divisive algorithm can be considered as the reverse form of an agglomerative algorithm that starts with one cluster containing all instances and at each iteration split the most appropriate cluster until a stopping criterion such as a requested number *k* of clusters is achieved [10, 11]. Since the algorithm is exactly does what AGNES algorithm does by top-down approach, the results obtained for AGNES is going to be valid for DIANA. So, coding and obtaining the results for AGNES will be enough.

3.3. Evaluation method

The most significant part of most projects is evaluation part since the value of the study can be assessed in this part. Error rate and accuracy measures are widely used metrics to evaluate correctness of results of data mining projects. **Error rate** is the proportion of count of wrong estimation over the count of items in the dataset. **Accuracy** is the proportion of count of the correct estimation over the count of items in the dataset. These can be formulized by using confusion matrix of predicted and real classes table below.

		Predicted Classes		
		1	2	3
Real Classes	1	T1: Correct prediction for class 1	F12: Predicted as 2 for the item of class 1	F13: Predicted as 3 for the item of class 1
	2	F21: Predicted as 1 for the item of class 2	T2: Correct prediction for class 2	F23: Predicted as 3 for the item of class 2
	3	F31: Predicted as 1 for the item of class 3	F32: Predicted as 2 for the item of class 3	T3: Correct prediction for class 3

$$Error\ rate = \frac{\mathrm{F12 + F\ 13 + F21 + F23 + F31 + F32}}{\mathrm{F12 + F\ 13 + F21 + F23 + F31 + F32 + T1 + T2 + T3}}$$

$$Accuracy = \frac{\mathrm{T1 + T2 + T3}}{\mathrm{F12 + F\ 13 + F21 + F23 + F31 + F32 + T1 + T2 + T3}}$$

Entropy is a measure of discretization [1]. It commonly explores class distributions in data and able to find split-points. Here we can use entropy in order to find if clusters distributed smoothly or not. If entropy is equal to 0, this means samples are in same class; if entropy is 1, this means samples are distributed to the classes equally; otherwise samples are distributed randomly [12]. The formulation of entropy is below.

$$H(p_1, p_2, \ldots .\, p_s) = -\sum_{i=1}^{s} p_i \log(p_i)$$

4.Data Set

In this project, we decided to work on Classic3 dataset which is used in several papers in order to do experiments and give results. Classic3 dataset contains 3891 document consist of 1460 documents from information retrieval papers, 1398 document aeronautical system papers, and 1033 documents from medical journals.

The documents do not require any natural language processing operations because they are obtained clean that all stop words are removed and stemming operations are done. As a result, 10930 distinct words are obtained from documents. The data is represented in a matrix (3891 * 10930) in which rows represent documents, columns represent terms, and the intersection of a row and a column gives the normalization of the multiplication of Term Frequency (TF) per document and Inverse Document Frequency (IDF) value of term among documents.

5.Implementation

The algorithms explained in Section 3 coded in MATLAB R2009b. The details which are paid attention while coding each algorithm are below.

5.1. K-means

While implementing k-means algorithm the most important details are selecting the value of *k* and initializing means of *k* clusters. Since the dataset contains three different types of documents *k* is selected as 3. Means of clusters could be initialized by random values. However, in order to get values of means approximate to the values of items, initially the items are sorted randomly, and then first *k* of them selected as the means of *k* clusters. The other items are assigned to the clusters according to cosine similarity measure to the means of clusters.

Another important detail about k-means is avoiding from local maximum values. For this aim, k-means algorithm is run several times. Then items are clustered with the items which are frequently clustered as together.

5.2. Hierarchical Clustering

In order to run hierarchical clustering algorithm, we need to have a similarity matrix. Rows and columns of this matrix represent documents, and the intersection of a row and a column gives the similarity of two documents. The similarity of items could be calculated by several distance or similarity measures such as euclidean, cityblock, hamming, and chebychev. However, for this dataset cosine similarity measure is selected as similarity calculation method because cosine similarity is known as the most suitable measure for documents. Even so, all similarity measure methods are also tested for several different datasets.

A problem that we have to struggle is realized with complete link similarity matrix update method. Single link and centroid approaches are implemented without any problem. However in complete link approach, since the matrix is updated according to the furthest documents, after a while all documents contain some documents whose similarity to each other is zero, in other words the documents do not contain any common term. Namely, all non-diagonals become zero. Therefore, the most similar clusters could not be found in order to be merged. In order to solve this problem, if the similarity of the furthest documents of clusters is zero, the similarity matrix is updated by average of the similarity of the furthest documents and nearer documents.

For each approach the program is stopped after reaching the wanted number of clusters which is three for this dataset.

5.3. Testing

Clustering is an unsupervised learning method and there are no pre-defined labels or categories for items that are clustered. As a result, the outputs of algorithms are unlabeled. So, clusters cannot be compared with known classes of documents. In order to compare documents whether they are clustered correctly or not, labels have to be assigned to the clusters. Label 1 is assigned to the first large document group symbolized by same character, label 2 to second large group symbolized by the same character, label 3 to the third large group symbolized by the same character. For example, in the table below first column is converted to the second column.

A	1
A	1
A	1
A	1
B	3
A	1
A	1
C	2
C	2
C	2
C	2
C	2
A	1
B	3
B	3
B	3
C	2

After preparing outputs, the results are compared with already known classes of documents and confusion matrixes are calculated for each algorithm. Then we evaluated our estimations by error rate, accuracy and entropy.

6.Results and Evaluation

After doing several experiments like explained before, we got the results below.

K-means:

	Estimation		
Real Classes	1455	2	3
	14	1383	1
	14	2	1017

Misclustering number: 36 → error rate: 0,935%

Correct clustering number: 3855 → accuracy: 99,075%

Entropy:

0,15399	0,03137	0,04014

Hierarchical clustering:

Single link:

	Estimation		
Real Classes	1460	0	0
	1398	0	0
	1031	1	1

Misclustering number: 2430 → error rate: 62,45%

Correct clustering number: 1461 → accuracy: 37,55%

Entropy:

1,56900	0,00000	0,00000

Complete link:

	Estimation		
Real Classes	1358	76	26
	58	853	487
	70	26	937

Misclustering number: 743 → error rate: 19,10%

Correct clustering number: 3148 → accuracy: 80,9%

Entropy:

0,50903	0,57768	1,03976

Complete link: stop if similarity matrix values become zero except diagonal

It generates 271 clusters. If we take 3 clusters to compare with real classes:

	Estimation		
Real Classes	89	0	1
	0	70	0
	0	0	58

Misclustering number: 3674 → error rate: 94,42%

Correct clustering number: 217 → accuracy: 5,58%

Centroid:

	Estimation		
Real Classes	1425	0	35
	1392	0	6
	29	2	1002

Misclustering number: 1464 → error rate: 37,63%

Correct clustering number: 2427 → accuracy: 62,37%

Entropy:

1,07176	0,00000	0,26273

The results show that k-means algorithm outperforms the hierarchical clustering algorithm for each similarity matrix update method. Hierarchical clustering methods face with difficulties for cluster number 3. Since, in each step two clusters are merged, the algorithm is more suitable for times of two. Even so, the results of AGNES with modified complete link method and centroid method are not so bad. The worst method for this dataset is AGNES with single link method. Since the single link method update the similarity matrix with the maximum similarity value from a cluster to another cluster, it leads to seem all clusters are very similar to each other and so, it merges many clusters to a cluster.

Entropy value is significant for k-means algorithm that it shows clusters are distributed evenly since the entropy values are very close to zero. However, for hierarchical clustering the entropy values are not very significant. For example, in the single link clustering the entropy of second class is zero which means items are clustered perfectly. However, only one item is clustered for that sample (it had to be 1398), moreover it is wrongly clustered. So, entropy measures do not give very valuable measure or it needs an interpretation by looking the confusion matrix to decide if items are clustered perfectly wrong or perfectly right.

7.Conclusion

In this project, we searched a solution for document organization which is a need due to growing vast amount of electronic information. We used k-means and hierarchical clustering algorithms in order to cluster documents. Classic3 corpus is used as a dataset. As a result, it has seen that k-means algorithm is superior to hierarchical clustering algorithm for the document dataset used in the experiments. AGNES with modified complete link approach is the best among hierarchical clustering similarity matrix update approaches for the dataset used.

8.References

[1] Han, J. and Kamber, M., *Data Mining: Concepts and Techniques, 2nd ed.*, Morgan Kaufmann Publishers, 2006.

[2] Kaufman, L. and Rousseeuw, P., *Finding Groups in Data*, Wiley, New York, NY, 1990.

[3] P. Willett, "Recent Trends in Hierarchic Document Clustering: A Critical Review", *Information Processing and Management*, 24(5), pp.577-597, 1988.

[4] Zhao, Y., and Karypis, G., “Criterion functions for document clustering: Experiments and analysis”, *Technical Report*, Department of Computer Science, University of Minnesota, 2001.

[5] Zhao, Y., and Karypis, G., “Evaluation of hierarchical clustering algorithms for document datasets”, *International Conference on Information and Knowledge Management*, McLean, Virginia, United States, pp.515-524, 2002.

[6] Zhao, Y. and Karypis, G., “Hierarchical Clustering Algorithms for Document Datasets”, *Data Mining and Knowledge Discovery* [C].10(2), pp.141-168, 2005.

[7] Steinbach, M., Karypis, G. and Kumar, V., “A Comparison of Document Clustering Techniques”, *KDD Workshop on Text Mining*, 2000.

[8] Liu, F. and Xiong, L., “Survey on text clustering algorithm -Research present situation of text clustering algorithm” *IEEE 2nd International Conference on Software Engineering and Service Science (ICSESS)*, pp.196-199, 2011**.**

[9] MacQueen, J. B., "Some Methods for classification and Analysis of Multivariate Observations”, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297, 1967.

[10] Alpaydın, E., *Introduction to Machine Learning, 2e,* The MIT Press, London, England, February, 2010.

[11] Borgatti, S.P., “How to Explain Hierarchical Clustering”, *Connections*, 17(2):81-84, 1994.

[12] Gündüz-Öğüdücü, Ş., *Data Mining Lecturer Notes*, Istanbul Technical University, 2011.