Composite Heuristic Algorithm for Clustering Text Data Sets

Nikita Nikitinsky, Tamara Sokolova and Ekaterina Pshehotskaya
InfoWatch
Nikita.Nikitinsky@infowatch.com, Tamara.Sokolova@infowatch.com,
Ekaterina.Pshehotskaya@infowatch.com

ABSTRACT

Document clustering has become a frequent task in business. Current topic modeling and clustering algorithms can handle this task, but there are some ways to improve the quality of cluster analysis, for example, by introducing some combined algorithms.

In this paper, we will conduct some experiments to define the best clustering algorithm among LSI, LDA and LDA+GS combined with GMM and find heuristics to improve the performance of the best algorithm.

KEYWORDS

Clustering, cluster analysis, topic modeling, LDA, LSI, GMM, Silhouette Coefficient

1 INTRODUCTION

One of the most frequent applications of clustering in business is exploratory data analysis during marketing researches, for example, a customer satisfaction survey.

But there is more than one way to use clustering techniques - cluster analysis of document sets sized up to 50 000 is a task, which may be essential in business.

It might be necessary to cluster, for example, weekly document stream for DLP (Data Leakage Prevention) purposes (e.g. easier categorization of documents).

To cluster small sets of documents we primarily need high clustering quality and may pay little attention to speed or computational complexity of a clustering algorithm – obviously, because modern computer hardware allows the user to perform complex computations in a short time, so small data sets are clustered fast even if an algorithm with high computational complexity is used.

That is why we decided to conduct some experiments on algorithms with high computational complexity in order to combine them in a way, allowing us to maximize quality of clustering.

2 METHODS

Cluster analysis or clustering is a convenient method for identifying homogenous groups of objects called clusters. Objects in a specific cluster share many characteristics, but are very dissimilar to objects not belonging to that cluster. Further in this paper we will discuss clustering algorithms where every object can belong only to one cluster - including cases where an object may belong to no cluster at all. In such cases we will create so called «garbage» cluster and put there all objects not classified by an algorithm.

We will use the following clustering and topic modeling algorithms to create a combination showing highest performance:

LSI (Latent Semantic Indexing) — is an unsupervised machine learning method, which is mostly used for dimensionality reduction. It is an indexing and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. [1]

LDA (Latent Dirichlet Allocation) - is also an unsupervised machine learning method, which is mostly used for object clustering. It is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. Computational complexity of LDA+GS is O(NKW) where N is a number of documents, K is a number of clusters and W is the number of words in vocabulary. [2]

Although mentioned above methods can be used alone, we will conduct experiments, in which we combine them with the following algorithms:

GMM Classifier (Gaussian Mixture Model), which is an unsupervised machine learning method, is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Gaussian mixture models are often used for data clustering. Clusters are assigned by selecting the component that maximizes the posterior probability. Like k-means clustering, Gaussian mixture modeling uses an iterative algorithm that converges to a local optimum. Gaussian mixture modeling may be more appropriate than k-means clustering when clusters have different sizes and correlation within them. Clustering using Gaussian mixture models is sometimes considered a soft clustering method. The posterior probabilities for each point indicate that each data point has some probability of belonging to each cluster.

Gaussian mixture distributions can be used for clustering data, by realizing that the multivariate normal components of the fitted model can represent clusters.

Computational complexity of GMM (using EM-algorithm for convergence) is O(tkmn^3) where k is the number of clusters, n is the number of dimensions in a sample, m is a number of samples and t is a number of iterations. [3]

Our choice fell on GMM because we considered it more robust and fast for cluster analysis compared to, for example, k-means.

When applying GMM we arrange every object only to one cluster (thus, we make it easier to estimate overall performance).

GS (Gibbs Sampling) is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult. GS is widely used to enhance quality of topic modeling algorithms; it is a good algorithm for processing when the dimension of data is not very high. With high dimensional data it may be better to use Variational EM algorithm. [4]

In our experiments we applied faster version of GS algorithm named Collapsed Gibbs Sampling algorithm.

3 EVALUATION METRICS

To evaluate algorithm performance we used two types of metrics often utilized for cluster analysis purposes:

3.1 External Evaluation Metrics

In external evaluation, clustering results are evaluated based on data that was not used for clustering, such as known class labels and external benchmarks. Such benchmarks consist of a set of pre-classified items, and these sets are often created by human (experts). Thus, the benchmark sets can be thought of as a gold standard for evaluation. [5]

We used the following external measurements:

Jaccard index - also known as the Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. [6]

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

V-measure score - is an entropy-based measure which explicitly measures how successfully the criteria of homogeneity and completeness have been satisfied. V-measure is computed as the harmonic mean of distinct homogeneity and completeness scores, just as precision and recall are commonly combined into F-measure.

$$V = \frac{2 * (H * C)}{(H + C)} \quad (2),$$

where:
H is homogeneity
C is completeness

This metric is independent of the absolute values of the labels: a permutation of the class or cluster label values won’t change the score value in any way. [7]

Adjusted Rand score - is a measure of the similarity between two data clusterings.[8]

Adjusted mutual information score - a variation of mutual information (which is a measure of the variables' mutual dependence) may be used for comparing clusterings. [9][10]

3.2 Internal Evaluation Metrics

In internal evaluation clustering result is evaluated based on the data that was clustered itself. These methods usually assign the best score to the algorithm that produces clusters with high similarity within a cluster and low similarity between clusters. [11]

We used the following internal measurement:

Silhouette Coefficient — is a measure of how appropriately the data has been clustered and how well each object lies within its cluster.
The Silhouette Coefficient is defined for each sample and is composed of two scores:

1. The mean distance between a sample and all other points in the same class.
2. The mean distance between a sample and all other points in the next nearest cluster.

Which can be written as:

$$S(\mathrm{i}) = \frac{(b(\mathrm{i}) - a(\mathrm{i}))}{\max(a(\mathrm{i}), b(\mathrm{i}))} \quad (3),$$

where:
i is the sample
a(i) is the average dissimilarity of i with all other data within the same cluster (i.e. the mean distance between a sample and all other points in the same class)
b(i) is the lowest average dissimilarity of i to any other cluster which i is not a member (i.e. the mean distance between a sample and all other points in the next nearest cluster.)

We used cosine metric as the most common for measuring the distances for Silhouette Coefficient. When we have higher value of Silhouette Coefficient, it means that we have better distribution of documents to topics. [12]

Based on Silhouette Coefficient measurements we apply Elbow method to define the number of clusters. This method assumes a choice of a number of clusters so that adding another cluster doesn't give much better modeling of the data (so called “Knee of a curve”). This method was originally designed to make predictions based on the percentage of variance explained and in some cases may appear unsuitable; in such cases we will choose the number of clusters where Silhouette Coefficient reaches maximum value [13]

Since, in real conditions, we are unable to use external metrics for evaluation of algorithms (because we usually don’t know the true number of clusters), we will evaluate quality of our models basing mostly on Silhouette Coefficient, applying external metrics as supplementary. The main external metric will be V-measure score as the most appropriate, because it is based upon two main criteria for clustering usefulness, homogeneity and completeness, which capture a clustering solution’s success in including all and only data-points from a given class in a given cluster. For some cases we will utilize Jaccard index to let the reader better understand the situation.

4 DATA SETS

We used some different data sets to check and validate the results:

1. Data set containing 600 documents, distributed to 5 topics – a «good» collection (distribution of documents: 83 to 163 documents per topic). Topics are easily distinguishable by human expert.
2. Data set containing 157 documents, distributed to 14 topics - «bad» collection (distribution of documents: 3 to 21 documents per topic). Topics are not distinguishable by human expert.
3. Data set containing 1000 documents, randomly assigned from the real document stream of the company; topic distribution is not predetermined; human experts considered the number of topics between 3 and 5 (including 3 and 5).
4. Data set containing 35000 documents, randomly assigned from the real document stream of the company; topic distribution is not predetermined. Human experts then estimated quality of the best algorithm performance on this data set.

5 EXPERIMENTS

We tested all these algorithms on the «good» collection to find out the best one and then evaluated the best algorithm performance on other collections

5.1 Choosing the Best Algorithm

5.1.1 LSI+GMM

Data preprocessing:
All words with length less than 3 symbols were deleted as well as all non-alphabetic characters.
To obtain better results we preprocessed input data with TF-IDF algorithm.

In this algorithm we may vary two main parameters: number of LSI topics and number of GMM clusters.
The LSI algorithm takes as input the collection of documents, processes it and then documents-topic matrix is returned. This matrix is then given to an input of GMM classifier, which processes the input matrix assembling documents to final categories (this is likely to increase the quality of clustering).

We tested two heuristics:
1. Number of LSI topics is equal to number of output GMM clusters
2. Number of LSI topics is equal to number of output GMM clusters plus one, such as number of LSI topics is n+1, while number o GMM clusters is n (one of the topics becomes so called «garbage» topic — it accumulates objects, which could not be unambiguously arranged to other «real» topics)

Table 1 contains evaluation metrics estimated on the «good» collection for LSI+GMM algorithm with 5 output categories:

Table 1.

	Heuristic 1	Heuristic 2
Jaccard index	0.575	0.57
Adjusted mutual information score	0.75	0.735
Adjusted Rand score	0.66	0.66
V measure score	0.74	0.74
Silhouette Coefficient	0.61	0.5

We can see that both heuristics showed comparable results when tested on a real number of categories; Heuristic 2 showed a decrease in Silhouette Coefficient value.
But, more generally, if we vary the number of output categories and estimate Silhouette Coefficient and V-measure for them we will get the following results (Figures 1, 2, where green (upper) line is V-measure and blue (lower) line is Silhouette score):

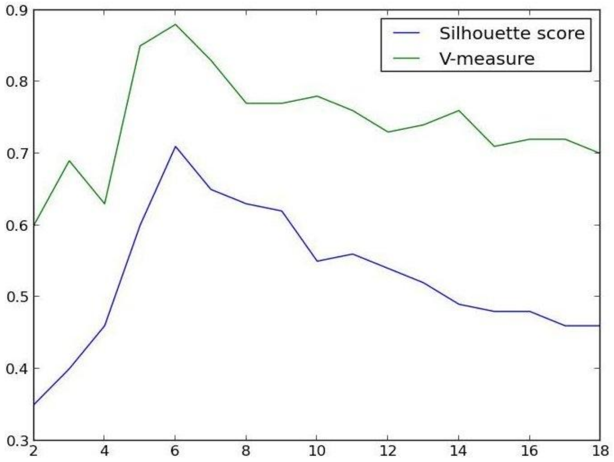

Figure 1. LSI+GMM, Heuristic 1

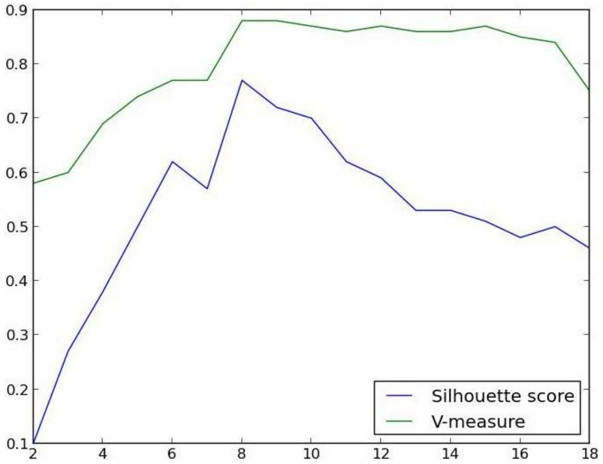

Figure 2. LSI+GMM, Heuristic 2

According to the results, the Silhouette Coefficient reached higher levels when we implemented Heuristic 2 (Figure 2). Nevertheless, both pikes indicated incorrect number of output clusters (6 and 8 correspondingly), V-measure showed comparable and also incorrect results (pikes at 6 and 8 correspondingly).

5.1.2. LDA+GMM

Data preprocessing:
All words with length less than 3 symbols were deleted as well as all non-alphabetic characters
Words occurring only once (hapax legomena) were deleted

The LDA algorithm takes as input the collection of documents processes it and then documents-topic matrix is returned. This matrix is then given to an input of GMM classifier which processes the input matrix assembling documents to final clusters (this must increase the quality of clustering).

We tested the same two heuristics.

Table 2 contains metrics estimated on the «good» collection for LDA+GMM algorithm with 5 output categories:

Table 2.

	Heuristic 1	Heuristic 2
Jaccard index	0.51	0.85
Adjusted mutual information score	0.57	0.83
Adjusted Rand score	0.53	0.76
V measure score	0.6	0.84
Silhouette Coefficient	0.45	0.52

We can see that Heuristic 2 showed far better results for external metrics, but insignificantly better result for Silhouette Coefficient.
If we vary the number of output categories and estimate both Silhouette Coefficient score and V-measure score for them we will get the following results (Figures 3, 4 where green (upper) line is V-measure and blue (lower) line is Silhouette score):

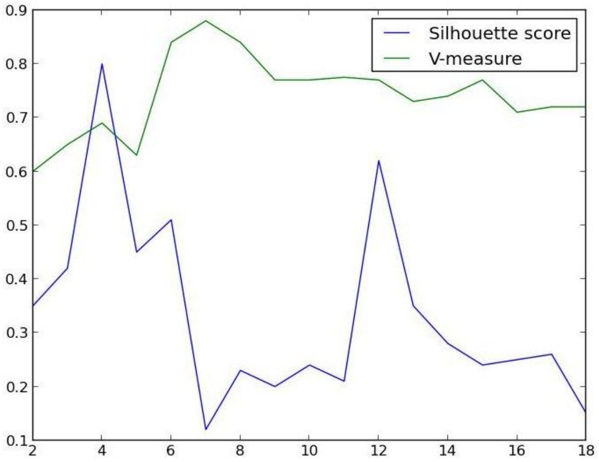

Figure 3. LDA+GMM, Heuristic 1

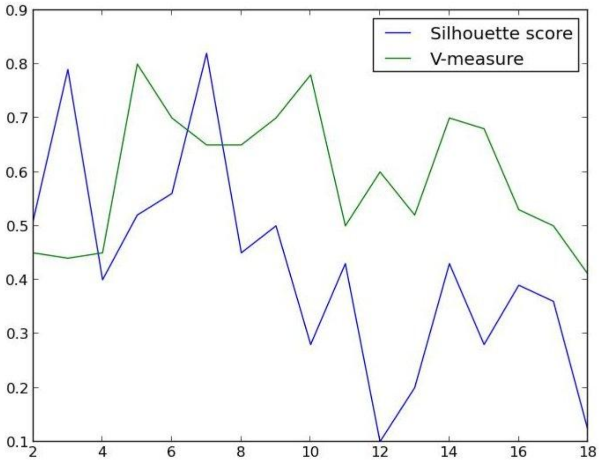

Figure 4. LDA+GMM, Heuristic 2

According to the results, Silhouette Coefficient reached a bit higher levels when we implemented Heuristic 2 (Figure 4). Nevertheless, both pikes indicated incorrect number of output clusters (4 and 7 correspondingly). V-measure, on the contrary, reached higher levels at Figure 3, however, indicating wrong number of categories (7 clusters), but on the Figure 4 this measure correctly identified correct number of clusters (5 clusters) reaching at the same time lower levels.

5.1.3 LDA+GS+GMM

Data preprocessing:
All words with length less than 3 symbols were deleted as well as all non-alphabetic characters
Words occurring only once (hapax legomena) were deleted

In this algorithm we may vary three main parameters: number of LDA topics, number of Gibbs Samples and number of GMM clusters.

For given quantity of LDA topics there are n iterations of Gibbs Sampling (where n is number of Gibbs Samples) and then documents-topic matrix is returned. This matrix is then given to an input of GMM classifier which processes the input matrix assembling documents to final clusters.

Choosing proper number of Gibbs Samples:

Knowing the real quantity of output categories we iteratively start the algorithm changing the number of samples and keeping other parameters the same.

The best number of Gibbs Samples is considered the number of samples when metric (e,g, Silhouette Coefficient) reaches highest values and then doesn't fluctuate much.

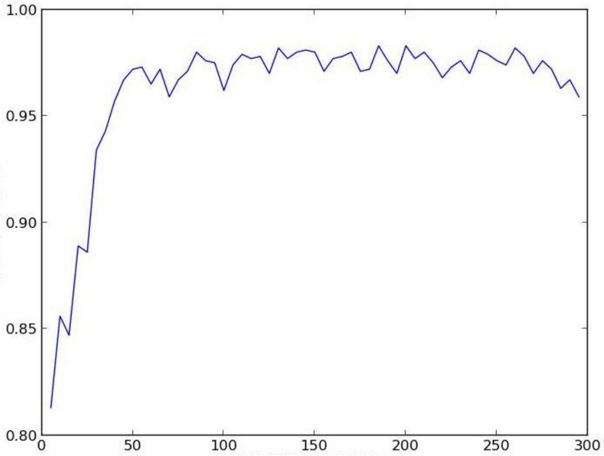

Figure 5.

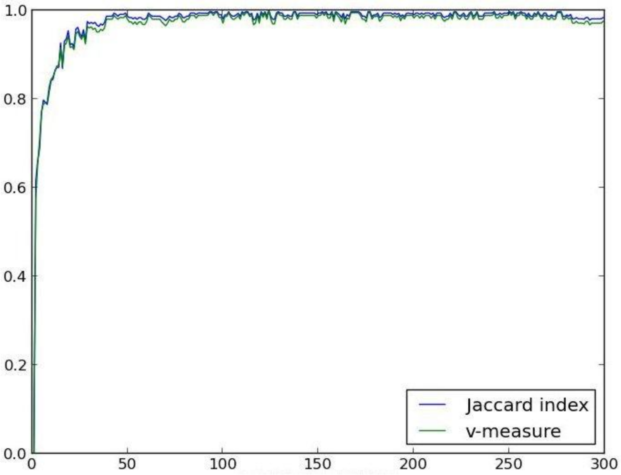

Figure 6.

We selected the best number of GS samples on the “good” collection. The unchanged parameters were the number of LDA topics and the number of GMM clusters (as in Heuristic 2). As we can see from the picture (Figure 5), the plotted line reaches highest values at 50 samples and then don’t fluctuate much, so we assume that we can choose any quantity of samples above 50. Figure 6 verifies this assumption: plotted lines of Jaccard index and V-measure also reach highest levels and don’t fluctuate much at approximately 50 samples.
Thus we will then use 100 samples as optimal

and versatile number of samples.

We tested the same two heuristics.

Table 3 contains metrics estimated on the «good» collection for LDA+GS+GMM algorithm with 5 output categories:

Table 3.

	Heuristic 1	Heuristic 2
Jaccard index	0.66	0.99
Adjusted mutual information score	0.77	0.99
Adjusted Rand score	0.72	0.99
V measure score	0.79	0.99
Silhouette Coefficient	0.82	0.98

We can see that Heuristic 2 showed far better results for all metrics. It means that documents are better distributed to said number of output categories with Heuristic 2 implemented for this algorithm.
If we vary the number of output categories and estimate both Silhouette Coefficient score and V-measure score for them we will get the following results (Figure 7, where green (upper) line is V-measure and blue (lower) line is Silhouette score) and Figure 7 where green (lower) line is V-measure and blue (upper) line is Silhouette score):

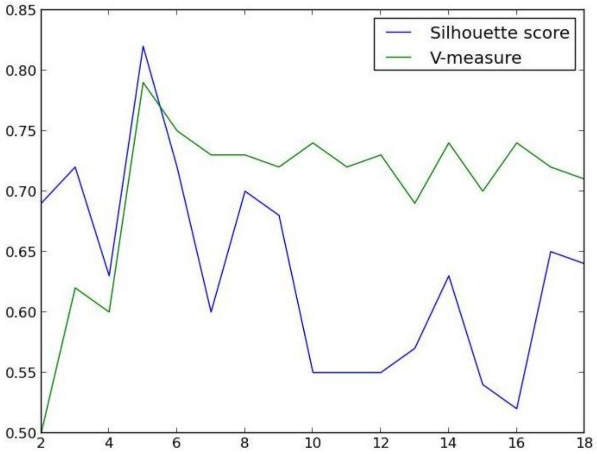

Figure 7. LDA+GS+GMM, Heuristic 1

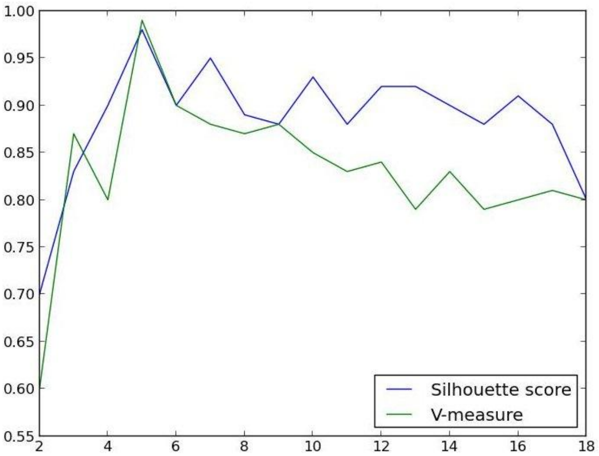

Figure 8. LDA+GS+GMM, Heuristic 2

According to the results, while both pikes indicated the same true number of clusters, Silhouette Coefficient reached higher levels when we implemented Heuristic 2 (Figure 8).
V-measure also reached higher levels on Figure 7 while indicating true number of categories on both figures. We can suggest that Heuristic 2 improves the performance of LDA+GS+GMM and intensifies the results making it easier to determine the number of output categories

5.2 Estimating the Best Algorithm on Other Data Sets

We tested LDA+GS+GMM algorithm on other collections using the parameters that we considered the best testing the algorithm on the “good” collection:

Number of GS samples is equal to 100
Number of LDA topics is equal to number of GMM clusters plus one (e.g. while number of GMM clusters is 5, number of LDA topics is 6)

Data preprocessing:
All words with length less than 3 symbols were deleted as well as all non-alphabetic characters
Words occurring only once (hapax legomena) were deleted

5.2.1 Data Set №2

We tested LDA+GS+GMM algorithm on the «bad» collection, estimated Silhouette coefficient score, V-measure score and Jaccard coefficient on it and had the following results (see Silhouette score on Figure 9, Jaccard score and V-measure score on Figure 10, where green (upper) line is V-measure score and blue (lower) line is Jaccard coefficient):

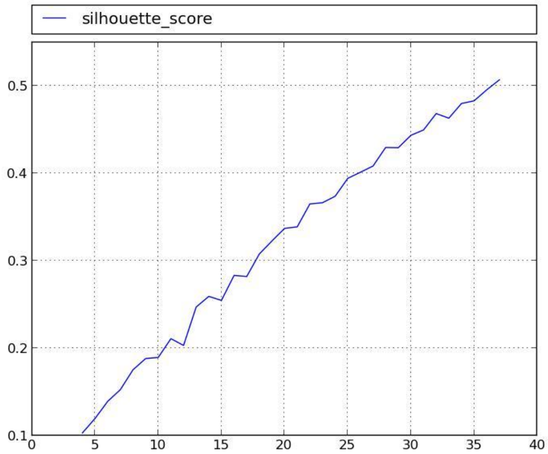

Figure 9.

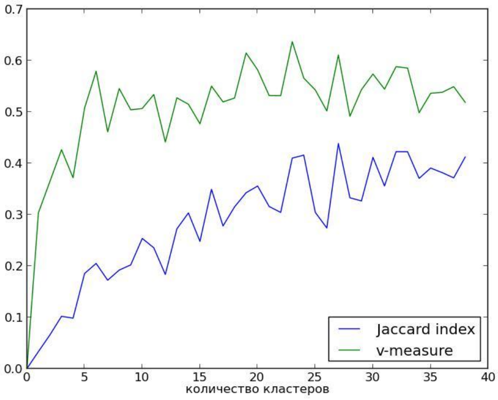

Figure 10.

Assuming that we selected the optimal parameters and using Elbow method based on Silhouette Coefficient plot we found it impossible to define (even approximately) the best number of output categories. Jaccard coefficient and V-measure score also showed contradictory results.

We obtained such results because of two main factors:

1. The distribution of documents to topics is conventional (in such cases there are either no much difference in vocabulary between documents of different categories or difference between all documents is too high to group at least some of them into one definite cluster)
2. Number of documents is small. For a collection with such conventional distribution of documents to topics the decently large number of documents is a need. But for a collection, where a difference in vocabulary between documents of different categories is significant this is likely not to be an issue. For example, a group of articles about cars and another group of articles about vegetables will be easily clustered even if we have size of each group of articles less than 50 items.

5.2.2. Data Set №3

We tested LDA+GS+GMM algorithm on the data set №3 containing 1000 documents and had the following results (Figure 11):

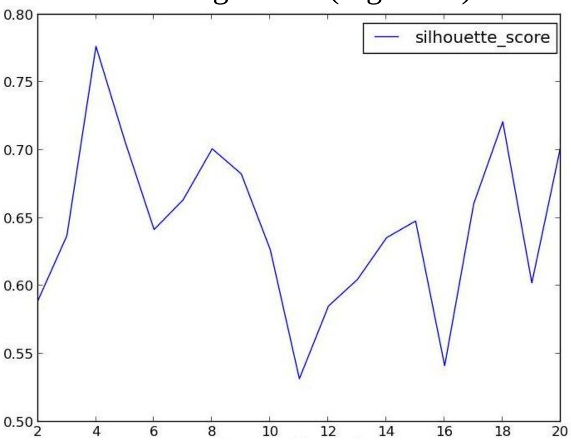

Figure 11.

As soon as we can't use external evaluation metrics to estimate quality of clustering on this data set (because we don't know the true number of clusters), here we will utilize only Silhouette Coefficient.

Basing on Silhouette Coefficient plot we decided that 4 categories is the best number of clusters for this data set. Human experts considered the result of the algorithm good. Documents in four categories could easily be defined as contracts, financial documents, application forms and information letters + instructions.

5.2.3. Data Set №4

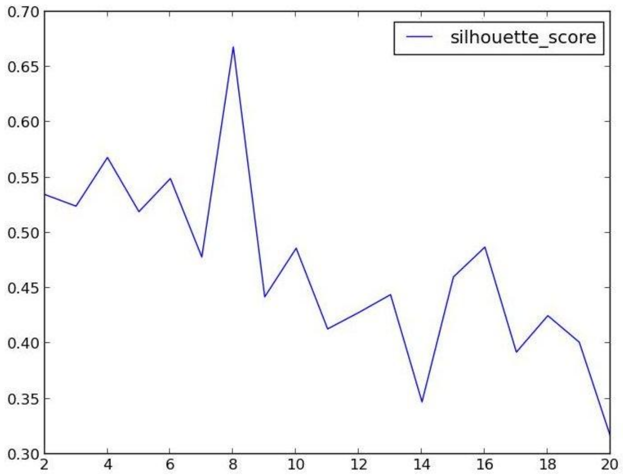

Figure 12.

As soon as we can’t use external evaluation metrics to estimate quality of clustering on this data set (because we don’t know the true number of clusters), here we will utilize only Silhouette Coefficient.

Basing on Silhouette Coefficient plot (Figure 12) we decided that 8 categories were the best quantity for this data set.

Human experts defined documents in 8 categories as contracts, financial documents, documents in other languages, information letters, instructions, application forms and other internal documents.

6 AUTOMATING NUMBER OF CLUSTERS DETECTION

The question about the most versatile algorithm to define proper number of clusters is still open-ended, but there are some general methods that may help to find a so called Knee of a curve. They are:

1. The largest magnitude difference between two points
2. The largest ratio difference between two points
3. The first data point with a second derivative above some threshold value
4. The data point with the largest second derivative
5. The point on the curve that is furthest from a line fitted to the entire curve.

This list is ordered from the methods that make a decision about the knee locally, to the methods that locate the knee globally by considering more points of the curve. The first two methods use only single pairs of adjacent points to determine where the knee is located. The third and fourth methods use more than one pair of points, but still only consider local trends in the graph. The last method considers all data points at the same time. Local methods may work well for smooth, monotonically increasing/decreasing curves. However, they are very sensitive to outliers and local trends, which may not be globally significant. The fifth method takes every point into account, but only works well for continuous functions, and not curves where the knee is a sharp jump. [14]

When we use other evaluation metrics – not Silhouette Coefficient score - these simple number of cluster detection methods may help.

But, nevertheless, in most cases it may be enough simply to choose a number of clusters, where Silhouette Coefficient score reaches the highest value.

7 CONCLUSION

According to the experiments we conducted, the best algorithm for processing relatively small set of documents (up to 50 000) with relatively small quantity of topics (up to 20) is LDA+GS+GMM. The Heuristic 2 may help to improve quality of LDA+GS+GMM and make it easier to determine number of output categories. Usage of Silhouette Coefficient is considered appropriate for determining best number of output clusters.

The data set should not be too small in order to provide the clustering algorithm with processable data: data sets containing less than 500 documents are likely to be incorrectly classified provided the data set contains documents with no much difference in vocabulary between them. In cases, when we have data set with significant differences in vocabulary between its items this is not an issue.

8 FURTHER READING

There are some papers on automated number of clusters detection algorithms, such as [14], proposing state-of-the-art algorithms that may be useful for cluster analysis.

Hierarchical Dirichlet Process (HDP) is also a generative probabilistic topic modeling algorithm for text clustering, showing the performance comparable to Latent Dirichlet Allocation topic modeling algorithm [15]

Although Latent Dirichlet Allocation works well for topic modeling there are now conducted multiple researches on more advanced topic modeling algorithms such as Higher-order Latent Dirichlet Allocation and other Higher-order topic modeling algorithms [16].

For processing large collection of documents different algorithms will be helpful, where the data is partitioned across separate processors and inference is done in a parallel, distributed fashion. These algorithms are Approximate Distributed Latent Dirichlet Allocation (AD-LDA), Hierarchical Distributed Latent Dirichlet Allocation (HD-LDA) and Approximate Distributed Hierarchical Dirichlet Processes (AD-HDP). The easiest to implement algorithm among these three is AD-LDA, but it has no formal convergence guarantee. HD-LDA is more complicated than AD-LDA, but it inherits the usual convergence properties of Markov chain Monte Carlo (MCMC). AD-HDP algorithm followed the same approach as AD-LDA, but with an additional step to merge newly instantiated topics. [17]

9 REFERENCES

[1] Deerwester, S., et al, Improving Information Retrieval with Latent Semantic Indexing, Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, 1988, pp. 36–40.

[2] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). "Latent Dirichlet allocation". In Lafferty, John. Journal of Machine Learning Research 3 (4–5): pp. 993–1022

[3] Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin

[4] Casella, George; George, Edward I. (1992). "Explaining the Gibbs sampler". The American Statistician 46 (3): 167–174

[5] Kaufman L, Rousseeuw PJ (2005) Finding groups in data. An introduction to cluster analysis. Wiley, Hoboken, NY

[6] Tan, Pang-Ning; Steinbach, Michael; Kumar, Vipin (2005), Introduction to Data Mining

[7] Rosenberg, Andrew and Hirschberg , Julia. V-Measure: A conditional entropy-based external cluster evaluation measure. Columbia University, New York.

[8] Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods". Journal of the American Statistical Association (American Statistical Association) 66 (336): 846–850

[9] Meila, M. (2007). "Comparing clusterings—an information based distance". Journal of Multivariate Analysis 98 (5): 873–895.

[10] Vinh, N. X.; Epps, J.; Bailey, J. (2009). "Information theoretic measures for clusterings comparison". Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09. p. 1.

[11] Manning, Christopher D, Raghavan, Prabhakar & Schütze, Hinrich. Introduction to Information Retrieval. Cambridge University Press.

[12] Rousseeuw, Peter J. (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65

[13] Ketchen, David J., Jr & Shook, Christopher L. (1996). "The application of cluster analysis in Strategic Management Research: An analysis and critique". Strategic Management Journal 17 (6): 441–458.

[14] Salvador, Stan and Chan, Philip. Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms, Dept. of Computer Sciences Florida Institute of Technology, Melbourne

[15] The, Yee Whye; Jordan, Michael I.; Beal, Matthew J.; Blei, David M.. “Hierarchical Dirichlet Processes”, Journal of the American Statistical Association. Vol. 101, No. 476, Dec., 2006, pp. 1566-1581

[16] Nelson, Christie, Pottenger, William M., Keiler, Hannah, and Grinberg, Nir. "Nuclear Detection Using Higher-Order Topic Modeling." 2012 IEEE International Conference on Technologies for Homeland Security. Waltham, MA. 13-15 Nov 2012.

[17] Newman, David; Asuncion, Arthur; Smyth, Padhraic; Welling, Max. “Distributed Algorithms for Topic Models”, Journal of Machine Learning Research 10 (2009), pp. 1801-1828