Technique of Cluster Validity for Text Mining

Galina Chernyshova
Dept. of Information System in Economics
Saratov Socio-Economic Institute
Saratov, Russia
galacherny@yandex.ru

Gennady Smorodin
EMC Academic Alliance
St. Petersburg, Russia

Alexey Ovchinnikov
Dept. of Information System in Economics
Saratov Socio-Economic Institute
Saratov, Russia

***Abstract*. The purpose of this article is to present the approach to Text Mining using special software. Authors consider possibility of using of k-means with different similarity measures and cluster validity evaluation. The authors have suggested *k*-means method with Bregman divergence for text documents clustering. Results show it is efficient in comparison of similar methods.**

Index Terms — Text Mining, k-means clustering, cluster validity

I. Introduction

Cluster analysis is an element of the exploratory data analysis for different Text Mining applications such as assotiation of similar documents in groups, automated abstracting. The clustering of text-based documents allows simplify process of studying of a large number of documents for studying of certain representatives of clusters. Besides the clustering is an auxiliary element for a formulation of a classification problem. Text clustering is in use in information search and providing information access, business analytics, corporate investigation, national security. Now there is a wide range of cluster algorithms and different modifications. The problem of objects splitting into clusters assumes a set of solutions and the choice of a clustering method is connected with the evaluation of clustering results. Cluster algorithms applicability is carried out in relation to specific data sets. Choosing the suitable algorithm and setting its parameters for text data analysis needs special consideration. In this work the features of k-means algorithm application in document clustering are investigated.

II. Clustering Technique

There are a lot of software of Text Mining now (RapidMiner Studio, IBM Intelligent Miner for Text, SAS Text Miner, Semio Corporation SemioMap, Oracle Text, Knowledge Server, Megaputer Intellidgens TextAnalyst). Such software represents scalable systems with different linguistic and mathematical methods of the text analysis. Similar systems have visualization and data manipulations tools, graphic interfaces, provide access to different data sources.

RapidMiner Studio is the environment for carrying out experiments for data analysis and machine learning, including data loading and data translations (ETL), visualization, modeling. Processes of data analysis are represented randomly by the enclosed operators in created XML files due to the graphic user interface RapidMiner. GUI generates the XML file which contains analytical processes which the user applies to data. The graphic interface can be used for interactive management and check of the started processes. The platform is available both in a cloud, and in the client-server option. For commercial versions an opportunity to work with Big Data is given, connection to different data sources is provided. The platform easily extends by means of libraries, BI platforms and web applications.

Document clustering is a process of detection of natural groups in a collection of documents. Let there X_n is data set $\{x_1,..., x_n\}$ and the function defining degree of similarity of objects in most cases is function of distance between objects $\rho(x_i, x_j)$. It is required to map a sequence of X_n into non-overlapping subsets (clusters) so that each cluster consisted of the objects close on a metrics ρ, and objects of different clusters significantly differed. The cluster algorithm is a function $A: X \rightarrow Y$ which to any object $x \in X$ marks cluster $y_i \in Y$. Y is not known in advance and an additional task is definition of optimum cluster number in terms of metric.

In cluster analysis there are following tasks:

- choice of effective clustering method for a solution of a certain data set;
- choice of characteristics on the basis of which the clustering is carried out (metrics, initial values of the cluster centers, algorithm stop conditions);
- choice of number of clusters. If there are no data of rather possible number of clusters, it is necessary to carry out experiments and to analyze the received results;
- interpretation of clustering results. Specific methods create clusters of certain forms and properties, at the same time there are no similar groups in given set.

978-1-4673-8203-8/16/$31.00 ©2016 IEEE

Clustering algorithms is subdivided into several types such as hierarchical, partitional, density-based, grid-based, fuzzy clustering. Classical hierarchical algorithms allow create full tree of the overlapping clusters. Nonhierarchical algorithms are based on optimization of some objective function which defines splitting objects set [1]. In this group there is special set of clustering algorithms k-means (k-means, fuzzy c-means, Gustafson-Kessel) which use the sum of squares of the weighed deviations of objects coordinates from the centers of required clusters as target function. Clusters are looked for a spherical or ellipsoidal shape. The algorithm of k-means is considered as one of the most effective tools for carrying out clustering of text data, the efficiency of application for this method for similar data types are supported with experiments [2].

On the other hand if clusters are overlapping the application of k-means algorithms is problematic. Also when one cluster is much more the others or clusters have the enclosed structure using of this method is limited. Additional methods are necessary for cluster validity.

The problem of the analysis for text in a natural language is complexity of selection useful information except for its size and metadata. To make possible using of traditional cluster algorithms is necessary to transform an array of text documents to numerical form. There are two common models for representation of text collections: treelike and vector models. The treelike model is sets of the chains following one after another words. Such way allows create similar chains among different documents and reveal their similarity.

The vector model is a matrix with frequencies of the words occurrence in a document. Let *T* is an array of text data, *N* – total quantity of terms in all arrays. Let *T* is a set of terms in an array of text documents. Then document presents as a vector wiht length N in which each element corresponds term from a set *T*. The coefficients specifying the frequency of occurrence of the term in the document can be values of elements.

Texts in vector model are considered as set of the terms constituting them. This approach is called a bag of words. Application of vector model assumes the choice of a method of terms weighing. There are several standard methods for numerical estimation of the document term. Term frequency (TF) defines term weight depending on the number of occurrences in the document. Thus importance of the word in the document is estimated.

Inverse frequency of the document (IDF) represents the return frequency of the document with which some word occurs in documents of a collection promotes reduction of weight of the most common words:

$$IDF_t = log\frac{|D|}{|\ d_i \supseteq t_i\ |},$$

where $|D|$ – number of documents in a collection *D*, $d_i \supseteq t_i$ – number of documents in which is available t_i.

TF-IDF (term frequency - inverse document frequency) is the statistical measure used for an assessment of importance of a term in the context of all set of documents. TF-IDF is calculated as product of the number of word entries into the document and functions from value reciprocalof the documents number:

$$TF - IDF(t,d,D) = TF(t,d)\times IDF(t,D)\ .$$

TF-IDF value increases for terms which often occur in the specific document and seldom used in other documents of a collection.

The vector space model allows define quickly with a small error a key word in the text and document subject. The vector model despite of the shortcomings remains the most often used in text analysis.

In the software intended for text analysis additional types of metrics of distances are used:

- quantitative measures (Euclidean measure, Chebyshev measure, cosine measure, Jacquard measure, Manhattan distance);
- categorial measures (Kulchinsky, Rogers-Tanimoto, Russell-Rao);
- measures applying Bregman divergence (Itakura, Kulbaka-Leybler, Makhalanobis distance, square Euclidean measure).

Bregman divergence represents following function:

$D\varphi(x,y)=\varphi(x)- \varphi(y) - \nabla\varphi(y)\ (x-y)$,

where φ (x) — any strictly convex real-valued function, $\nabla\varphi(y)$ — its derivative on y.

For example if φ (x)=$\|x\|^2$, Dφ(x,y)=$\|x - y\|^2$ is a square Euclidean measure. Similarly other types of function φ allow to define other types of divergence of Bregman such as Itakura, Kulbaka-Leyblera, Makhalanobis distance [3].

For a solution of a problem of a rubrication it makes sense to construct cluster models using different types of metrics and to compare results by means of criteria of accuracy. In such a way it is possible to define the most suitable way of calculation of distance for this type of the text data.

Important element of clustering is the criterion for cluster model evaluation. The criterion for evaluation is a numerical indicator which is calculated by results of a clustering, the essence of this criterion consists in quantitative expression of clustering quality.

Cluster validation includes external, internal and relative methods. Metrics is used any already famous information on the structure of clusters existing in the considered set belong to external. As a rule such metrics are applied at an assessment overall performance of clustering algorithm when as a test set with the known structure is given. Metrics belong to internal (index of Rand, Jacquard, Folkes-Mallows index, F-measure) which at an assessment are guided only by that information which can be received on data set. Relative methods (Devies-Bouldin index, Dunn index, silhouette index, Maulik-Bandyopadhyay index, Calinski-Harabasz index) estimate quality comparing several cluster designs among themselves not having the priori information and taking into consideration only cluster type and data set [4].

Relative methods are widely applied including in cluster analysis software. However it should be noted that they are useful only in certain situations not general purpose [5]. For example Dunn index has high computational complexity and badly applicable for the analysis of noisy data. The area of its application is limited to identification of pure clusters in the data sets containing small amount of elements. Davies-Bouldin index yields good results for different data. Nevertheless it is not intended for detection of the blocked clusters. The silhouette index cannot be applied to the enclosed clusters. The Maulik-Bandyopadhyay index substantially depends on the parameters set by the user.

Davies-Bouldin criterion is based on a ratio of intra cluster and intercluster distance. Davies-Bouldin index is defined:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j}\{ D_{i,j} \},$$

where $D_{i,j} = \frac{(\overline{d}_i + \overline{d}_j)}{d_{i,j}}$, $\overline{d}_i$ – average distance between points of a cluster to *i* and centrodes of a cluster of *i*, $\overline{d_j}$ - average distance between points of a cluster to i and centrodes of a cluster of *j*, $d_{i,j}$ – Euclidean distance between centrodes of clusters of *i* and *j*.

Davies-Bouldin index defines average similarity between a cluster of c_i and the closest cluster to it. It is meant that clusters structure considerably differs from each other, the structure with the minimum of index value will be optimal [6].

The external valuation methods based on reference sets can be considered as the standard for an assessment. This approach allows define how cluster model corresponds to the test set classes.

As the simplest characteristics at an assessment of cluster model measures of similarity are precision and recall borrowed from information search.

Let the control sample consist of the *M* objects, from them *m* of objects were correctly distributed in clusters. Document clustering accuracy is calculated as the attitude of the documents which are correctly attributed to a cluster towards total quantity of the documents attributed to a cluster:

$$p(u) = \frac{|u \cap v|}{|u|},$$

where $|u \cap v|$ – correctly attributed documents; $|u|$ – total quantity of the documents attributed to a cluster.

Completeness of document clustering is calculated as the attitude of correctly attributed documents to total quantity of the documents carried to a cluster:

$$r(u) = \frac{|u \cap v|}{|v|},$$

where $|u \cap v|$ – correctly attributed documents; $|v|$ – total quantity of the documents carried to a cluster.

The *F*-measure represents a harmonic average between the accuracy and completeness.

III. Results and analysis

As the tool for document clustering we selected RapidMiner Studio 6.002, at the last version there is a set of modern algorithms, tools and approaches for text analysis [7].

For document rubrication we constructed cluster models using different types of metrics and compared results by means of criteria of accuracy. In such a way it is possible to define the most suitable way of calculation of distance for text data type.

Initial data are presented by a set of text documents from a news line goarticles.com. Text documents contain from 420 to 650 words. For an assessment of cluster model documents were grouped by expert way in four thematic categories (education, web design, real estate, cars).

The special operator of loading Loop Files is applied to import a collection of text documents. During the analysis of text data it is necessary to transform contents of all documents for separate words. The operator Process Documents carries out preprocessing of the text, creating a bag of words, and also calculates the frequency of each word presenting the models of a vector space.

In this process the operator Process Documents consists of 6 subprocesses which are consistently connected (Tokenize Non-letters; Tokenize Linguistic; Filter Stopwords; Filter Tokens (by Length); Word stemming (Stem); Transform Cases).

The operators Tokenize Non-letters and Tokenize Linguistic are created by adding in subprocess of the operator Tokenize with the choice of different parameters. The first operator breaks into the lexemes based not on letters whereas the second breaks into lexemes based on linguistic sentences within this or that language.

The operator Filter Stopwords deletes all words which have length less than 3 signs or more than 25. Stem carries out process of finding of a word stem. Transform Cases will transform all characters in selection to the lower register.

Often there is an issue of attributes setting before application of some operators especially for big and difficult data sets. The operator Select Attributes allows select the necessary attributes thanks to different types of filters. The selected attributes will be on the operator's output. This conversion is necessary for the following operator *k*-means who carries out a clustering only for numerical values.

For a clustering of text collections the algorithm of *k*-means is used. In RapidMiner there are different operators capable to give help in selection of best value of parameter *k* and an assessment of clustering quality. For a solution of this problem by the *k*-means method new process for cluster validity is constructed. The operator Cluster Distance Performance is used for an assessment of efficiency of a clustering. This operator provides the list of values of efficiency criteria including Davies-Bouldin index.

Application of different distances was followed by the analysis of model on a test set. Results of this process with the Davies-Bouldin index for different numbers of clusters are given in table 1.

TABLE 1. Davies-Bouldin index

Quantity k clusters	**Metrics type**			
	Divergence Bregman	***Euclidean Measure***	***Chebyshev Measure***	***Manhattan distance***
2	4,77	4,74	4,95	4,83
3	4,38	4,40	4,46	4,41
4	**3,91**	3,91	4,24	4,51
5	4,51	4,40	4,29	4,45
6	4,34	**3,66**	4,53	4,54
7	4,50	4,55	4,43	4,62
8	4,29	3,91	4,04	4,48
9	3,94	4,14	**3,92**	**4,18**
10	4,23	4,07	4,03	4,21

The optimum quantity of clusters corresponds to the smallest value of Davies-Bouldin index bolded in table 1. However the correct number of clusters k=4 defined for an expert way for test set was received when using Bregman divergence as a metrics.

The received distribution of documents in clusters is compared to expert objects distribution in classes. Results are evaluated by means of standard characteristics of precision, recall and F-measure (Tabel 2). Coincidence of results of a clustering to these subjects can serve as objective criterion of estimation constructed model.

TABLE 2. Assessment of cluster validity evaluation

Metrics	Precision	Recall	*F*-measure
Bregman divergence	0,98	0,99	0,98
Euclidean measure	0,98	0,99	0,98
Chebyshev measure	0,77	0,74	0,75
Manhattan distance	0,76	0,82	0,79

The algorithm of k-meams with Bregman divergence and standard Euclidean measure yields good results for clustering of documents. However an essential lack of the method is selection of optimum number of clusters. For selection of optimum number of clusters by the most applicable there was an approach with Bregman divergence (Fig. 1, Fig. 2).

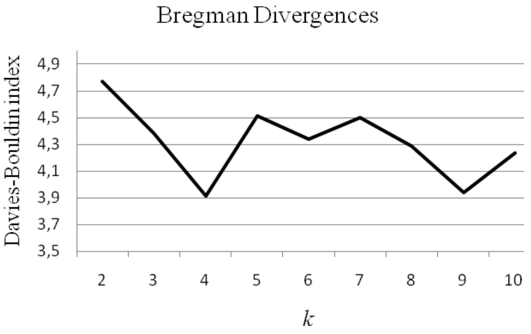

Fig. 1. Clustering with Bregman divergence.

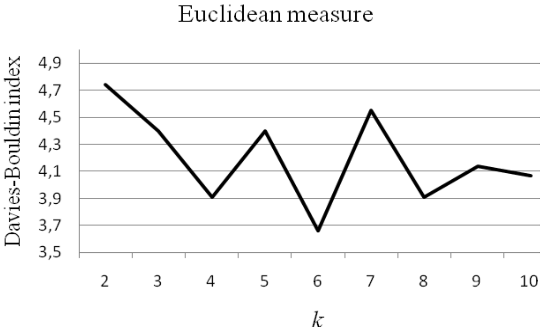

Fig. 2. Clustering with Euclidean measure.

The correct number of clusters is four. This is not the case with Eucledean measure which identifies six clusters.

IV. Conclusions

In order to solve practical problems in economy apart from the traditional task of forecasting the need for analysis of text documents increasingly arises [8]. In case of the text analysis is able to find the correct number of clusters for noisy, unbalanced, overlapped and mixed clusters. The offered technique of carrying out a document clustering using RapidMiner includes application for cluster validation. We use as a metric Bregman divergence which allowed to define optimum number of clusters by Davies-Bouldin index. Clustering by k-means algorithm with using Bregman divergence as a measure of distance has allowed adequate accuracy for expert classes.

Using RapidMiner Studio on the basis of this technique allow the analyst quickly without labor-consuming development to carry out document clustering and selection of number of clusters. Besides results of the received distribution of documents on clusters can be used for creation of decision tree and solution of classification task.

References

[1] S. Koteeswaran, P. Visu, J. Janet, “A review on clustering and outlier analysis techniques in Data Mining”, Am. J. Applied Sci., 9: 254-258. DOI:10.3844/ajassp.2012.254.258, 2012.

[2] N. Andrews, E. Fox, “Recent Developments in Document Clustering”, Virginia Tech, Blacksburg, VA 24060, October 16, 2007.

[3] A. Banerjee, S.Merugu, I. Dhillon, J. Ghosh, “Clustering with Bregman divergences”, JMLR, 6:1705–1749, 2005.

[4] S. Saitta, B. Raphael, I. F. C. Smith, “A bounded index for Cluster validity”, in Proc. of Int. Conf. on Machine Learning and Data Mining in Pattern Recognition, pp. 174–187, Springer, 2007.

[5] U. Maulik, S. Bandyopadhyay, “Performance evaluation of some clustering algorithms and validity indices”, IEEE Transactions Pattern Analysis Machine Intelligence, vol. 24(12), pp 1650–1654, 2002.

[6] D. Davies, D. Bouldin, “A Cluster Separation Measure”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, No. 2, 1979, pp. 224–227.

[7] M. Hofmann, R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2013.

[8] G. Chernyshova, V. Gusyatnikov, “Application of Forecasting Technique for Economic Indicators”, in Proc. Int. Conf. on Cloud, Big Data and Trust, pp. 23-24, Bhopal, India, 2013.