A cosine-based validation measure for Document Clustering

Simona Balbi[1], Michelangelo Misuraca[2], Maria Spano[1]

[1]University "Federico II" + Naples – Italy

[2]University of Calabria + Arcavacata di Rende – Italy

Abstract

Document Clustering is the peculiar application of cluster analysis methods on huge documentary databases. Document Clustering aims at organizing a large quantity of unlabelled documents into a smaller number of meaningful and coherent clusters, similar in content. One of the main unsolved problems in clustering literature is the lack of a reliable methodology to evaluate results, although a wide variety of validation measures has been proposed. If those measures are often unsatisfactory when dealing with numerical databases, they definitely underperform in Document Clustering. This paper proposes a new validation measure. After introducing the most common approaches to Document Clustering, our attention is focused on *Spherical K-means*, do to its strict connection with the Vector Space Model, typical of Information Retrieval. Since *Spherical K-means* adopts a cosine-based similarity measure, we propose a validation measure based on the same criterion. The new measure effectiveness is shown in the frame of a comparative study, by involving 13 different corpora (usually used in literature for comparing different proposals) and 15 validation measures.

Keywords: Cluster Validation, *K-means*, Cosine Similarity, Compactness, Separation.

1. Introduction

Huge documentary bases ever more are increasing in domains where information is traditionally stored in textual bases. However textual data are also used in fields where numerical data bases were most common in the past. In both cases, one of the most challenging issues is to automatically organise the massive amount of information, with or without any prior knowledge.

As we usually aim at finding the topics of the documents, in order to put together documents sharing a similar content, our attention is focused on the so called unsupervised classification process (*clustering*), suitable in the absence of prior information or when information is not consistent to the aims of the organising process. The final result consists in assigning a previously unknown category to each document, related to its main topic.

There is an enormous literature in the field of clustering methods and algorithms. It is not our aim reviewing this literature or going depth in the different questions concerning with the choice of a peculiar technique. Here our attention is focused on the main consequence of this tremendous debate: how to evaluate the quality of the results. The evaluation of results in a clustering process is known as *cluster validation*, and it is related to the existence of a natural grouping in the data, the actual number of groups, and their composition.

Validation methods are usually classified into two groups: external and internal approaches. The difference is whether we refer or not to the presumed real partition, labelling each unit by its own class. The external validation measures quantify how much the identified clusters

correspond to the externally provided labels, as in a supervised classification process. The internal validation measures evaluate the goodness of a clustering structure by examining only the partitioned data, in terms of the internal structure of the obtained partition (compactness, separation, and so on).

In real applications, we normally do not have prior information, and this is the reason why we find internal validation measures more interesting. In the following, a measure for internally evaluating the results of a document clustering process is proposed, taking into account the peculiarity of textual data.

During the years, a wide variety of internal validation measures has been proposed in literature, but even expert researchers experiment problems in choosing a measure that provides an appropriate response to their own aims. The most common measures offer an answer to the question of the number of groups. Spano (2015) shows that poor results are peculiar to Document Clustering, as the meaning of distance in high dimensional spaces lacks its own sense, and the most common validation measures are based on distances.

In this paper, a new validation measure is proposed for the peculiar case of Document Clustering. In particular, after introducing the most common approaches to Document Clustering, our attention is focused on *Spherical K-means*, for its strict connection with the Vector Space Model, typical of Information Retrieval. Since *Spherical K-means* adopts a cosine-based similarity measure, we propose a validation measure based on the same criterion. The effectiveness of the new measure is shown in the frame of a comparative study involving 13 different corpora (usually used in literature for comparing different proposals), and 15 validation measures.

2. Theoretical background

Cluster Analysis techniques are often considered a common exploratory step, preliminary to the proper statistical analysis. As a consequence, the validation of results is frequently held as a costly superfluous addition. This attitude could be dangerous. It is well known that, given a set of data, each clustering algorithm generates a subdivision even when the data do not have a natural grouping. Moreover, different algorithms often lead to different solutions, different choices of the input parameters produce different results with the same algorithm, and the units order in the data set can even affect the final results (Jain and Dubes, 1988).

Therefore, evaluation criteria are important in order to provide the reliability to results. Suitable measures can be useful in identifying the number of clusters in the data, to assess whether the obtained clusters are the actual ones or they are just induced by the algorithm, or to decide which of the different algorithms is better to use.

Given a *corpus* of n documents d_i $(i=1,\ldots,n)$, classifying its documents on the basis of their content is one of the most important tasks of Text Mining.

The usual assumption is that each d_i, is represented as a vector, in the space spanned by the p terms in our vocabulary (*vector space model*):

$$d_i = \left(t_1,\ldots,t_m,\ldots,t_p\right) \tag{1}$$

where t_m is the importance of the *m-th* term in d_i. It is usually measured by the frequency (for a discussion on the choice of t_m see: Balbi and Misuraca, 2005).

The goal of grouping similar documents in distinguishable subsets can be achieved by referring to methods of supervised or unsupervised classification. In the first case, some *a priori* information is available, related to the number of groups, their peculiarities and their composition. The prior information is based on an expert knowledge, and it is usually related to the topic. Therefore, a given label/category is assigned to a subset of documents, in order to automatically attribute the same category to similar documents. An unsupervised approach aims at grouping documents trying to bring out the natural structure of the categories, without external knowledge. In this case the process, named clustering, is based solely on the available data.

In the following, our attention is focused on the validation of clustering procedures. In other words, we assume the absence of prior information. We start with the assumption that documents of different categories have a different frequency distribution of terms. In the most extreme form, each documents' category uses almost exclusively a portion of the vocabulary, which properly constitutes a peculiar vocabulary consisting of specific terms. In practical problems, the specific vocabularies of different categories are overlaid, so that a given document may use terms peculiar to a document category to which it does not belong.

2.1. Document Clustering

The strategies usually adopted are common with the clustering of numerical data. The two main families of algorithms are: agglomerative hierarchical algorithms and partitive centre-based algorithms.

Hierarchical algorithms allow the visualisation of the association structure at different levels of granularity. One of the main interesting consequences is that the number of clusters is not an input of the algorithm. The different solutions are sequentially nested and displayed in a tree structure. This is of special interest in the frame of content analysis. However, once two (sets of) documents have been aggregated, they will not be separated in the subsequent steps. Furthermore, the hierarchical algorithms are less scalable in the case of huge collections of documents.

The partitional algorithms create a one-level solution, given as input parameter the number k of desired clusters. Initially, k documents are (randomly) selected as initial centroids. Then, for each document, its proximity to these k centroids is computed (by a suitable measure), and the document is assigned to the cluster according with the highest proximity measure. This forms the initial clustering in k groups. The clustering is then repeatedly refined in order to optimise the chosen clustering criterion function.

The *K-means* algorithm (MacQueen, 1967) is still the most important reference in literature and it is widely used for the clustering of documents (Jain and Dubes, 1988), although many new algorithms have been proposed throughout fifty years of scientific research (Wang *et al*., 1986; Iezzi, 2012).

In classical *K-means*, the proximities are measured in terms of Euclidean distance. In the framework of Document Clustering we deal with high dimensional data sets. It is proved that Euclidean distance loses its readability and interpretability at the increasing of dimensionality (Aggarwal *et al*., 2001). This condition – known as the *curse of dimensionality* – has peculiar effects on clustering methods, as it has already shown that the relative difference of the distances of the closest and farthest data points of an independently selected point tends to 0 as dimensionality increases (for a review see: Balbi, 2010).

One of the most interesting variation in *K-means* family, is the so-called *Spherical K-means* (Dhillon and Modha, 2001). This algorithm is based on the cosine similarity, suggested by Salton and McGill (1983) in Information Retrieval. Given two documents d_i and d_j in a *corpus*, the so-called *cosine similarity* is given by:

$$cos(d_i, d_j) = \frac{d_i^T d_j}{\|d_i\| \cdot \|d_j\|} \tag{2}$$

The cosine is 1 if the documents use the same words, and 0 if they have no terms in common. The effect of the different length of documents is mitigated by the normalisation, in order to represent the documents in a high dimensional unit sphere.

From a statistical viewpoint, it can be interpreted in terms of linear correlation (if documents are centred with respect to the vector means). As the distribution of correlation between random vectors becomes narrowly focused around zero as the dimensionality grows, the significance of small correlations increases with growing dimensionality. It is good at capturing the similarity of patterns of feature changes, disregarding at the same time the absolute amplitude of the compared feature vectors. Therefore, the algorithm exploits the sparsity of textual data and, although quickly converges to local maxima, in experimental results shows a good performance, according to Dhillon and Modha.

As the dissimilarity is intuitively related to a distance, and clustering algorithms often deal with distances, the complement of the corresponding similarity measure, known as the *cosine dissimilarity*, is widely used.

2.2. Cluster Validation

The assessment of clustering quality and the selection of the most appropriate method are still open challenges. Since clustering defines clusters that are not known ahead, irrespective of grouping criterion, the final partition of the data requires an evaluation. The procedure of evaluating the results of a clustering algorithm is known in literature as *Cluster Validation*. The two basic approaches to validate a partition are *external* and *internal* validation. The difference is whether or not prior information is used in the validation process. As external measures assess how much the identified clusters correspond to the externally provided labels, they are mainly used for choosing an optimal clustering algorithm on a specific dataset. Internal validation measures evaluate the goodness of a clustering structure without any additional information, and they can be used to choose the best clustering algorithm as well as the optimal number of clusters.

In the majority of applicative scenarios, information regarding the true number of clusters, or either the real composition of the corresponding groups, are not available. In this case internal measures are the only option to evaluate clustering results.

The *overall validity* of a clustering solution can be expressed as a weighted sum of the validity of each cluster:

$$overall\ validity = \sum_{h=1}^{k} w_h \cdot validity\,(C_h) \tag{3}$$

where w_h is the weight assigned to a generic cluster C_h. As the aim of clustering is to obtain compact and well-separated groups, a better solution will consider a higher proximity in each group (*compactness*) as well as a lower proximity among the groups (*separation*).

An overall validity function can consider just one aspect, or some combination of both, and the optimisation depends on what is taken into account. At the same time it is important to specify how all the elements contribute in the overall evaluation, because it is possible to measure the proximity of all the elements each other (*graph-based view*), or of each element with respect to a reference one (*prototype-based view*). The following Table 1 shows the main internal measures with respect to the way compactness and separation are computed:

graph-based view	**prototype-based view**
C Index (C) Dunn (D) Gamma (G) G^+ (G_p) McClain-Rao (MCR) Point Biserial (PB) Silhouette (S) Tau (T)	Calinski-Harabasz (CH) Davies-Bouldin (DB) PBM Ratkowski-Lance (RL) Ray-Turi (RT) Wemmert-Gancarski (WG) Xie-Beni (XB)

Table 1 – Overview of the main internal validation measures

Since it is not possible to present in detail all the above mentioned indices, a critic discussion can be found elsewhere (Halkidi *et al*., 2001; Liu *et al.*, 2010).

3. Preliminary experimental evaluation

The aim of the following experiment is to compare the behaviour of the validation measures in Table 1. A large number of configurations is considered, in order to identify which of them provides a more appropriate response with respect to the (known) natural partition in the data. The experiment has been planned considering different factors. Due to the explosion of combinations, each factor is necessarily limited to few levels. The adopted comparative methodology is widely shared in literature (Milligan, 1981; Legány *et al.*, 2006). For evaluating the quality of a clustering solution an algorithm is ran by setting different parameters, aiming at obtaining a set of different partitions. The value assumed by the internal measure of validity is then calculated for each partition. The number of clusters in the partition corresponding to the best results is considered a prediction of the validation index. More specifically, this prediction will be satisfactory if the number of groups identified by the measure coincides with the true number of classes in the analysed dataset.

3.1. Data description and pre-processing

In order to perform the comparative study 13 corpora have been selected among the most used in the literature of Text Categorisation and Information Retrieval. The collections are real-world data sets obtained from different applicative domains. The *fbis* corpus has been obtained from the *Foreign Broadcast Information Service* data in the *TREC-5* collection, with classes corresponding to the different categories used in the collection. The corpora *k1a*, *k1b* and *wap* are part of the project *WebACE*, with documents corresponding to the web pages listed in the *Yahoo!* subject directory, categorised according to different levels of granularity. The corpora *la1* and *la2* have been obtained from the *Los Angeles Times* articles listed in

TREC-5. The corpora *re0* and *re1* are from the *Reuters-21578* collection. The corpora *tr11*, *tr12*, *tr23*, *tr41* and *tr45* have been built from the *TREC-5*, *TREC-6* and *TREC-7* collections, and the used classes correspond to documents that have been judged relevant in some particular queries. On each corpus, the same pre-treatment has been carried out. The terms of each collection have been stemmed using Porter's suffix-stripping algorithm, stop-words have been then completely removed from the documents.

In Table 2, it is possible to read some characteristics of the data sets, in terms of number of documents, number of terms in each collection and number of classes in which the different collections are categorised. The *cv* index represents the *coefficient of variation*, used to characterize the class imbalance, while *density* is the ratio of non-zero terms in a data set. A large *cv* indicates a severe class imbalance, a small density indicates instead a high sparsity.

dataset	# documents	# terms	# classes	cv	density
fbis	2463	2000	17	0.961	0.0799
k1a	2340	21839	20	1.004	0.0068
k1b	2340	21839	6	1.316	0.0068
la1	3204	21604	6	1.022	0.0048
la2	3075	31472	6	0.516	0.0048
re0	1504	2886	13	1.502	0.0179
re1	1657	3578	25	1.385	0.0140
tr11	414	6429	9	0.882	0.0438
tr12	313	5804	8	0.638	0.0471
tr23	204	5832	6	0.935	0.0661
tr41	878	7454	10	0.913	0.0262
tr45	690	8261	10	0.669	0.0340
wap	1560	8460	20	1.040	0.0167

Table 2 – Characteristics of the analysed corpora

3.2. Performance of the validation measures

On each dataset *Spherical K-means* has been performed by setting the number of seeds *k* between *2* and *2g*, where *g* is the "true" number of classes in the specific dataset. For each partition the value assumed by each of the 15 validation measures has been calculated.

dataset	*C*	*D*	*G*	*G_p*	*MCR*	*PB*	*S*	*T*	*CH*	*DB*	*PBM*	*RL*	*RT*	*WG*	*XB*
fbis	18	6	17	18	21	20	15	19	16	7	18	15	22	30	17
k1a	24	37	23	17	24	30	32	30	23	30	37	20	24	34	38
k1b	**2**	**2**	**2**	7	**1**	**2**	7	4	5	**3**	6	5	**2**	**3**	**3**
la1	8	**2**	9	6	9	8	11	8	5	8	5	5	8	10	6
la2	3	**1**	**1**	7	**1**	**3**	5	**3**	5	8	5	5	4	5	**2**
re0	17	**1**	17	15	17	19	14	19	12	12	12	14	8	9	**1**
re1	12	30	6	26	6	7	24	4	22	28	23	15	28	25	37
tr11	8	**1**	11	9	**2**	**3**	17	9	6	7	8	8	13	16	10
tr12	9	**2**	14	9	8	6	13	14	**3**	13	**1**	**1**	4	14	10
tr23	**2**	**3**	**3**	8	**2**	**2**	5	4	**2**	8	**1**	**1**	**3**	**3**	4
tr41	14	10	14	10	14	14	14	18	18	10	14	14	18	17	15
tr45	10	**1**	**2**	11	9	6	7	**3**	8	6	14	6	9	**1**	7
wap	21	32	17	23	17	18	15	20	19	29	22	19	16	15	36

Table 3 – Ranks of the optimal solution according to the internal validation measures

In Table 3, it is possible to observe each result computed for the different partitions, in terms of rank. A value equal to 1 means that the index has identified the real partition as the best possible solution among all the *2g-1* partitions, while a higher value (in ascending order)

implies a greater shift from the true classification in the data. It is possible to see the better results in bold and different shades of grey.

The *Dunn index* (*D*, reported in the second column) generally shows better performances, even if in some cases (*re1*, *k1a*, and *wap*) the optimal solution is very far from the one proposed by the validation measure. It is important to state, however, that none of the other measures performs well with the *re1*, *k1a*, and *wap* data sets.

4. A cosine-based measure for validating Document Clustering

The internal measures proposed in literature usually consider a metric based on Euclidean distances as a criterion for evaluating the compactness and the separation. This choice may be inconsistent with the criterion beneath the algorithm used for obtaining the partitions. With this motivation, if the cosine similarity is adopted in the clustering algorithm, it should be introduced also in the validation measure. Furthermore, the joint use of both compactness and separation has to be considered, because it is desirable to have clusters not only with a high cohesion but also well distinguished. In Document Clustering this means to consider groups of documents that share one (or few) topic(s).

Let $X=\{d_1,\ldots,d_n\}\subset\Re^p$ be a set of n document vectors in the term space of dimension p. By the clustering procedure, k groups of documents C_h (with $h=1,\ldots,k$) have been identified, so that each document has one of the labels identifying the k different groups.

Let us start by defining *compactness* and *separation*. From a geometric viewpoint the aim of a clustering algorithm is to maximise intra-cluster proximities whilst minimise inter-cluster proximities. Let d_i, $d_{i'}$ and d_j be three generic documents included in X, with d_i, and $d_{i'}$ belonging to the same cluster. If all the documents contribute to the evaluation (*graph-based view*), the *compactness* and *separation* can be calculated as in the follow:

$$compactness(C_h)=\sum_{d_i,d_{i'}\in C_h} proximity(d_i,d_{i'}) \tag{4a}$$

$$separation(C_h,C_{h'})=\sum_{\substack{d_i\in C_h\\ d_j\in C_{h'}}} proximity(d_i,d_j) \tag{4b}$$

where *proximity* (.) is usually the Euclidean distance. Starting from the experimental results previously shown, we adopt *Dunn index* which seems to be a more suitable measure for evaluating the clustering solutions in the peculiar field of Document Clustering. In this measure, a ratio between separation and compactness is used. Maximising this ratio leads to the best possible solution. The *Dunn index* of each clustering solution is given by:

$$D=\frac{\min\limits_{1\le h<h'\le k}\{separation(C_h,C_{h'})\}}{\max\limits_{1\le h\le k}\{compactness(C_h)\}} \tag{5}$$

The *separation* between two generic clusters C_h and $C_{h'}$ is measured by the minimum Euclidean distance observable between the closer documents belonging to any pair of clusters. The *compactness* of a generic cluster C_h is measured by the distance between the furthest documents belonging to the cluster. This means to calculate the diameter of the

cluster itself. In this viewpoint, both the worst observable cohesion and the worst observable separation are considered for evaluating the "quality" of each cluster of documents.

By considering the cosine dissimilarity, (5) can be rewritten as our new index *BMS*:

$$BMS = \frac{\min\limits_{\substack{d_i \in C_h \\ d_j \in C_{h'}}} \{1 - \cos(d_i, d_j)\}}{\max\limits_{d_i, d_{i'} \in C_h} \{1 - \cos(d_i, d_{i'})\}} \tag{6}$$

The new *BMS* measure has been computed for all the collections. In Table 4 it is possible to see the results in comparison with the original *Dunn Index* performances, already shown in Table 3. Again, better results are highlighted in bold and different shades of grey.

dataset	*D*	*BMS*
fbis	6	15
k1a	37	31
k1b	**2**	7
la1	**2**	**1**
la2	**1**	**1**
re0	**1**	**1**
re1	30	**1**
tr11	**1**	16
tr12	**2**	**1**
tr23	**3**	**1**
tr41	10	**3**
tr45	**1**	5
wap	32	34

Table 4 – Ranks of the best solutions for the cosine-based measure

It is interesting to notice that in some cases the proposal seems to perform better, while in other cases the change in the proximity measure leads to worse results. In particular, with the collection *la1*, *re1*, *tr12* and *tr23* it is possible to obtain a higher accuracy with *BMS* measure. On the other hand, with *tr11* and *tr45* the original measure provides a more appropriate response. It seems that cosine dissimilarity does not improve the effectiveness of *Dunn index* when a huge number of classes has to be considered (e.g., *fbis*, *k1a*, *wap*), where no other measure offers appropriate results.

It could be fruitful going in depth into the data structure, by analysing the balance and the sparseness and their effects on the validation process.

5. Remarks and future development

The measure proposed in this paper has the advantage of considering the same optimisation criterion both in the clustering algorithm and in the validation measure. The cosine similarity quantifies the proximity among documents in terms of shared vocabulary. In this way, it evaluates the closeness between two different documents represented in a vector space, as well as their lexical similarity. In this viewpoint, dealing with Document Clustering, this

ability is more desirable than simply considering a distance between two elements. Aiming at improving the effectiveness of the clustering process and its consequent validation, a soft-cosine could be considered (Sidorov *et al*., 2014). This measure includes in the classical cosine formula a weight for taking into account the semantic similarity (synonymy), by using external linguistic resources (e.g., WordNet).

In the validity concept two characteristics - compactness and separation - are usually taken into account, as suitable indicators of the clustering quality. However, other aspects should be considered. For example, the density of the different clusters in a given partition has to be evaluated, because high sparsity affects the accuracy of clustering. A density measure has to be included in the validation process.

Another challenging issue in a clustering validation process is the evaluation of the quality of a solution in terms of "composition" of groups. The different measures discussed above, as well as our proposal, consider the optimal solutions only from the "number of clusters" viewpoint. It is important to see if the elements in a cluster are actually well classified. Comparative analyses usually do not take into account this aspect. As widely discussed by Spano (2015), an optimal solution has to be evaluated also in terms of semantic similarity. An ex-post analysis is necessary, and some refinements of the solution have to be considered in order to obtain a partition that effectively represents the grouping structure in the data.

References

Aggarwal C.C., Hinneburg A. and Keim D.A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. *Proceeding ICDT '01 Proceedings of the 8th International Conference on Database Theory*: 420-434.

Balbi S. (2010). Beyond the curse of multidimensionality: high dimensional clustering in Text Mining. *Statistica Applicata - Italian Journal of Applied Statistics*, 22 (1): 53-63.

Balbi S. and Misuraca M. (2005). Visualization Techniques in Non Symmetrical Relationships. In Sirmakessis S., editor, *Knowledge Mining*, Springer-Verlag, Heidelberg: 23-29.

Dhillon I.S. and Modha D.S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42 (1-2): 143-175.

Halkidi M., Batistakis Y. and Vazirgiannis M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17 (2-3): 107-145.

Iezzi D.F. (2012). A new method for adapting the k-means algorithm to text mining. *Statistica Applicata*, 22: 69-80.

Jain A.K. and Dubes R.C. (1988). *Algorithms for clustering data*. Prentice Hall.

Legány C., Juhász S. and Babos A. (2006). Cluster Validity Measurement Techniques. In *Proceeding AIKED'06 Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*: 388-393.

Liu Y., Li Z., Xiong H., Gao X. and Wu J. (2010). Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*: 911-916.

MacQueen J. (1967). Some methods for classification and analysis of multivariate observations, In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1 (14): 281-297.

Milligan G.W. (1981). A monte carlo study of thirty internal criterion measures for cluster analysis. *Psycometrika*, 46 (2): 187-199.

Salton G. and McGill M.J. (1983). *Introduction to Modern Retrieval*. McGraw-Hill.

Sidorov G., Gelbukh A., Gómez-Adorno H. and Pinto D. (2014). Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Computación y Sistemasi*, 18 (3): 491-504.

Spano M. (2015). *Tecniche di validazione per il clustering di documenti*. Tesi di Dottorato. Università Federico II di Napoli: http://www.fedoa.unina.it/10417

Wang W., Wang C., Cui X. and Wang A. (2008). Fuzzy C-Means Text Clustering with Supervised Feature Selection. In *Fuzzy Systems and Knowledge Discovery, FSKD'08. Fifth International Conference*, 1: 57-6.