On Fuzzy Cluster Validity Indexes for High Dimensional Feature Space

Fernanda Eustáquio[1], Heloisa Camargo[2], Solange Rezende[3], and Tatiane Nogueira[1(✉)]

[1] Department of Computer Science, Federal University of Bahia, Salvador, Brazil
fernandase@dcc.ufba.br , tatiane.nogueira@ufba.br

[2] Department of Computer Science, Federal University of São Carlos, São Carlos, Brazil
heloisa@dc.ufscar.br

[3] Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil
solange@icmc.usp.br

Abstract. Fuzzy document clustering aims at automatically organizing related documents into clusters in a flexible way. At this context, the topics identification addressed by documents in every cluster is performed by automatically discovering cluster descriptors, which are relevant terms present in these documents. Since documents are represented by a high-dimensional feature space, the extraction of good descriptors is a big problem to be solved. This problem is even bigger using fuzzy clustering, since the same descriptor can be representative for more than one cluster. Moreover, it is well-known that the Fuzzy C-Means clustering algorithm is also affected by documents dimensionality and the choice of correct partition of a given document collection into clusters is still a challenging problem. In order to overcome this drawback, we have investigated the most common fuzzy clustering validity indexes to validate the organization of data with high dimensional feature space, since they are commonly used to evaluate fuzzy clusters from low dimensional data sets.

Keywords: Fuzzy clustering · Validity indexes · Flexible organization · Documents · Text mining

1 Introduction

In general, data stored in a textual form (documents) cover a wide set of topics that are updated constantly and the organization of them in categories can not be predefined. Additionally, the automatic organization of documents in categories face the subjectivity problem, since it is related to the concept of relevance itself. It is a well known fact that the same document may be more relevant to a

J. Kacprzyk et al. (eds.), *Advances in Fuzzy Logic and Technology 2017*,
Advances in Intelligent Systems and Computing 642,
DOI 10.1007/978-3-319-66824-6_2

topic than to another topic, although both topics pose similar or complementary information.

According to [9], it has been shown by many researches that humans can only represent their information need in vague and imprecise terms to characterize documents about an specific topic. In order to provide abilities to a system to manage imprecise and/or uncertain information inherent to the data, Flexible Document Organization (FDO) has been proposed in [10].

To illustrate the usefulness of such a flexibility in the management of documents, consider a context in which news are organized in categories according to their main topic. Consider a news (textual document) with the title "*Experts affirm the adventure sport strengthens heart health*", which addresses complementary topics: *Sports* and *Health*. This news can be assigned to categories related to the *Sports* topic or the categories related to the *Health* topic.

Nevertheless, the cited news deals with both topics simultaneously, which suggests that the assignment of this news to categories that represent both topics would be more appropriate than choosing predefined categories that represents just one of them. Therefore, supposing that a human user is requiring documents of the *Sports* topic, if the cited document is assigned only to the *Health* topic, this document would not be recovered for the user, despite being useful for his/her requirements.

In this context, FDO is handled by a document clustering process, where there are no predefined topics and no examples that would show what kind of desirable relations should be valid among the data. Document clustering is an unsupervised process used in a variety of applications because if there is a document in a cluster that is relevant to a user, then it is likely that other documents from the same cluster are also relevant [6].

Furthermore, to overcome the drawback concerning multi-topic documents, FDO also consider the use of overlapping clustering [12]. In document clustering, representatives such as the cluster prototype are not very useful to identify the topic addressed by documents in each cluster. By FDO, overlapping cluster descriptors are automatically discovered, which are terms present in the documents and significant to the topic addressed in the documents [11].

From [12] and [5] investigations, it was concluded that FDO with Fuzzy C-Means (FCM) is a promising approach. However, selecting the correct partition of a given collection into clusters that reveal meaningful information for advanced data analysis, such as data mining, is a challenging problem.

Several comparative studies of validity indexes have been proposed to evaluate clustering partitions but most of them was performed by analyzing data sets with low dimensionality. For example, [13] have published a review analyzing the most commonly used fuzzy clustering indexes which the data sets used had 4 dimensions, in [4] the biggest dataset had also 4 dimensions and the rest of data sets had only 2. In [18] the biggest one had 30 dimensions. In this work, the document collections used have the number of dimensions ranging from 2804 to 22926. Since documents are represented in FDO by a high-dimensional feature space and a document needs to be compatible with more than one cluster with

different compatibility degrees, the selection of the correct partition of a given collection is a problem in evidence.

To present such analysis, this paper is organized as follows. In Sect. 2, basic concepts concerning flexible organization of documents are reviewed. In Sect. 3, the experimental results concerning the performance of FCM for FDO under different cluster validity indexes are presented, followed by discussions about the achieved results. Finally, in Sect. 4, the conclusion and the future directions of this research are also presented.

2 Background

In this section, we review basic concepts and methods used in the approach proposed in [10] and improved in [11] to organize documents in a flexible way.

2.1 Document Preprocessing

The preprocessing of documents is necessary to structure documents in order to make them processable by algorithms of pattern extraction. The most common output of a document preprocessing is the representation of a document collection in a vector space in the form of a document-term matrix. Each matrix row corresponds to one document in the collection and each matrix column (attribute) corresponds to one term in the entire collection of documents.

The terms in the document-term matrix are first examined in an initial effort to disregard terms that do not represent useful knowledge. In this step of examination, three tasks are very common: (1) Elimination of stopwords, which are words that are not relevant in the analysis of documents and usually consist of prepositions, pronouns, articles, interjections, among others; (2) Stemming, a technique that reduce words to their root form in order to reduce the number of terms needed to represent a document collection; (3) n-gram extraction, which is the extraction of terms represented by n consecutive words, since words that occur in sequence in a document may contain more information than isolated words.

After selecting the terms that represent the document collection, for the proposed approach, the document-term matrix contains in its cells $tf\text{-}idf$ (Term Frequency-Inverse Document Frequency) as defined in [12]. By this measure, the importance of the terms in a document is weighted, so that terms which are present in a lot of documents have a smaller weight than the terms that occur more rarely in a collection.

The preprocessed document-term matrix as defined in [12] is composed by a document represented by a vector that comprises the frequency of each term $\mathbf{t}$ in a document $\mathbf{d}$, weighted by how often this term occurs in a collection. This document-term matrix is inherently high-dimensional and sparse, which sometimes can make the document organization computationally very expensive or even impossible. This negatively affects the outcome of some knowledge extraction algorithms.

2.2 Fuzzy Document Clustering

To make the flexible organization of documents possible, the preprocessed documents are clustered by means of overlapping clustering algorithm Fuzzy C-Means (FCM) [2], since by it a document can belong to more than one cluster with different membership degrees. The interpretation of these membership degrees can be used to quantify the compatibility of a document with a topic, which is identified by cluster representatives.

FCM was proposed to be used to cluster low dimensional data and, because of that, uses the Euclidean distance in its process. However, this metric distance is not appropriate for high-dimensional and sparse data as documents. According to [17], the similarity measures plays an important role in documents clustering and the use of cosine coefficient similarity is more appropriate.

Therefore, the FCM clustering algorithm was slightly modified to handle the flexible organization of documents facing the high-dimensionality problem. The modification was done in the similarity measure norm function, and this measure is defined as follows.

Definition 1. *Let* $\boldsymbol{n}$ *be the number of documents in a collection and* $\boldsymbol{c}$ *be the number of clusters. Consider a document* $\boldsymbol{d}_k$, $k = 1, ..., n$, *and a cluster prototype* $\boldsymbol{v}_i$, $i = 1, ..., c$. *The dissimilarity between a document and a prototype* $\|\boldsymbol{d}_k - \boldsymbol{v}_i\|$ *is measured using the cosine coefficient similarity according to (1).*

$$\|\boldsymbol{d}_k - \boldsymbol{v}_i\| = 1 - \frac{\boldsymbol{d}_k \cdot \boldsymbol{v}_i}{|\boldsymbol{d}_k|\,|\mathbf{v}_i|} \in [0, 1] \tag{1}$$

The FCM algorithm used for document clustering is an iterative process that updates the prototypes of the clusters defined initially from a fuzzy pseudo-partition and the partition matrix giving the membership degree of each document to each cluster. This update tries to minimize the dissimilarity between a document and a cluster prototype. The pseudo-partition is defined as follows.

Definition 2. *Let* $A_i(\boldsymbol{d}_k)$ *be the membership degree of a document* $\boldsymbol{d}_k$ *in a cluster* $\boldsymbol{i}$, $i = 1, ..., c$. *A fuzzy pseudo-partition* $U = [A_i(\boldsymbol{d}_k)]$ *is a family of fuzzy sets of a collection* $D = \{\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_n\}$ *denoted by* $\mathrm{P} = \{A_1, A_2, ..., A_c\}$. *In P,* $\sum_{i=1}^{c} A_i(\boldsymbol{d}_k) = 1$ *and* $0 < \sum_{k=1}^{n} A_i(\boldsymbol{d}_k) < n$.

During the clustering procedure, the prototypes and the partition matrix are updated until a stopping criterion be satisfied. Since the FCM algorithm is well known, details of its steps are not described here. The interested reader is refered to [2]. The best number of clusters for the document organization can be selected using a suitable fuzzy cluster validity index, as described next.

2.3 Fuzzy Validity Indexes

To organize a document collection in clusters that represents topics, finding the appropriate number of clusters is not a trivial task, specially when these topics

represent overlap knowledge. To select the most appropriate number of clusters refers to the selection of the correct partitions of a given collection. Usually, it is performed clusters validity indexes to select such partition. Therefore, we have investigated the behavior of well known cluster validity indexes in the context of document clustering.

Some indexes use in their evaluation only the membership degrees obtained by the clustering algorithm. This kind of index produces good results for some document collection domains, but they have the disadvantage that are not connected directly to the document collection. Among the most common indexes, Partition Coefficient (PC) [1], Partition Entropy (PE) [3] and Modified Partition Coefficient (MPC) [7] were used in the experiments.

The PC index measures the amount of fuzzy intersection among the documents clusters in a partition. PC is a maximization index and its values range in $[1/c, 1]$. The index is defined as:

$$PC = \frac{1}{n} \sum_{i=1}^{c} \sum_{k=1}^{n} (A_i(\mathbf{d}_k))^2 \tag{2}$$

The PE index measures the amount of fuzziness in a partition and it is characterized as a minimization index. PE values range in $[0, \log_a c]$ where $a \in (1, \infty)$ and in this paper was performed using $a = 10$. The index is defined as:

$$PE = -\frac{1}{n} \sum_{i=1}^{c} \sum_{k=1}^{n} A_i(\mathbf{d}_k) \log_a(A_i(\mathbf{d}_k)) \tag{3}$$

PC and PE have the disadvantage of monotonicity, where as c grows the PC value decreases and the opposite occurs with PE.

MPC was proposed as an effort to try to correct the monotonic trend that PC has. MPC is a maximization index and its values range in $[0, 1]$. The index is defined as:

$$MPC = 1 - \frac{c}{c-1}(1 - PC) \tag{4}$$

Besides that, there are some indexes that use, in addition to the membership degree of documents in each cluster, the dissimilarity between a document d_k and a cluster prototype $\mathbf{v}_i$ or the dissimilarity between $\mathbf{v}_i$ and the prototypes mean $\bar{\mathbf{v}} = \sum_{i=1}^{c} \frac{\mathbf{v}_i}{c}$. The dissimilarities are measured using the cosine coefficient similarity according to (1). Among the most common indexes, Fukuyama Sugeno (FS) [8], Xie and Beni (XB) [19] and Fuzzy Simplified Silhouette (SF) [4] were carried out in the experiments.

FS is a minimization index given by the difference between the clusters compactness $\|\mathbf{d}_k - \mathbf{v}_i\|^2$ and the separation between them $\|\mathbf{v}_i - \bar{\mathbf{v}}\|^2$. A fuzzification factor $m > 1$ is a real number that controls the influence of the membership degrees in the fuzzy clustering. The index is defined as:

$$FS = \sum_{i=1}^{c} \sum_{k=1}^{n} (A_i(\mathbf{d}_k))^m (\|\mathbf{d}_k - \mathbf{v}_i\|^2 - \|\mathbf{v}_i - \bar{\mathbf{v}}\|^2) \tag{5}$$

XB is a minimization index also intended to measure the compaction and separation of clusters. The smaller value of XB implies that clusters are more compact and separate. The index is defined as:

$$XB = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}(A_i(\mathbf{d}_k))^m \|\mathbf{d}_k - \mathbf{v}_i\|^2}{n \times min_{k\neq i}\|\mathbf{v}_i - \mathbf{v}_k\|^2} \tag{6}$$

SF is a maximization index that considers the two clusters in which d_k has the greatest membership degrees. The index is defined as:

$$S(d_k) = \frac{\beta(\mathbf{d}_k, g_i) - \delta(\mathbf{d}_k, g_i)}{max\left\{\delta(\mathbf{d}_k, g_i), \beta(\mathbf{d}_k, g_i)\right\}} \tag{7}$$

$$SF = \frac{\sum_{k=1}^{n}(A_1(\mathbf{d}_k) - A_2(\mathbf{d}_k))S(\mathbf{d}_k)}{\sum_{k=1}^{n}(A_1(\mathbf{d}_k) - A_2(\mathbf{d}_k))} \tag{8}$$

In (7), a document $\mathbf{d}_k$ belongs to the cluster g_i, where $g_i \in (g_1, g_2, ..., g_c)$. $\delta(\mathbf{d}_k, g_i)$ is the average distance between $\mathbf{d}_k$ and all other documents belonging to g_i, i.e. the intra-cluster distance. $\beta(\mathbf{d}_k, g_i)$ is the distance between $\mathbf{d}_k$ and the neighbor cluster closest to g_i, i.e. the inter-cluster distance.

Finally, once documents are clustered, the overlapping cluster descriptors can be extracted by the method described next.

2.4 Fuzzy Cluster Descriptor Extraction

The method proposed in [10] carries out a procedure that uses an adaptation of the classical measures of information retrieval namely precision, recall, and $f1$-measure, which is the weighted harmonic mean of precision and recall.

In the fuzzy cluster descriptor extraction, all the terms found in a document preprocessing step are initially considered as descriptor candidates. Additionally, a document $\mathbf{d}_k$ is considered to belong to a cluster i if it has a membership degree $A_i(\mathbf{d}_k) \geq s$, where $s = \frac{1}{c}$. The threshold s is considered for two reasons. Firstly, its use allows the selection of descriptor candidates from documents that belong to more than one cluster with different compatibility degrees, instead of considering only the cluster with the highest compatibility degree. Secondly, using this threshold it is possible to penalize the descriptor candidates that occur in documents with low compatibility degree in a cluster.

A rank of terms weighted by their $f1$-measure is obtained for each cluster, considering the contingency matrix presented in [10]. Since the ranking of descriptor candidates is obtained, the descriptors are selected. The number of descriptors to be selected depends on the application.

Next, some experimental results are presented concerning the improvement that can be obtained in FDO, and consequently the meaningful cluster descriptors extraction, by selecting a correct partition to the documents collection.

3 Experimental Results

To evaluate the improvement obtained by FDO using different cluster validity indexes, eight different real document collections widely used about different topics were clustered by FCM and each index analysed. The fuzzy cluster validity indexes sanctioned the partition resulted from the FCM clustering.

All collections were preprocessed using the Pretext[1] tool [16]. Any term that occurs in fewer than two documents was eliminated and 1-gram terms were selected.

The FCM algorithm was performed using the following values of parameter: error $E = 10^-2$ as stop criterion and the fuzzification factor $m = 2.5$. The indexes were tested using the number of clusters c ranging from $c(min) = 2$ to $c(max) = \#class + c(min)$, since FCM is an iteration process in which the number of clusters is predefined. Once an index obtain its best value using a number of clusters between $c(min)$ and $c(max)$, this is its best partition.

The rule of thumb $c(max) \leq \sqrt{n}$ suggested by [13] is commonly used. However, in [13] the number of examples of the used data sets is very short compared with our data sets.

Table 1. Document collections with their number of documents (#docs), topics (#class), maximum number of clusters ($c(max)$) and dimensions (#terms)

Dataset	NSF	Hitech	WAP	Irish-Sentiment	Opinosis	20 Newsgroups	La1s	Reviews
#docs	1600	600	1560	1660	51	2000	3204	4069
#terms	2804	6593	8068	8658	10784	11026	13195	22926
#class	16	6	20	3	3	4	6	5
$c(max)$	18	8	22	5	5	6	8	7

After clustering, the number of clusters obtained by each index was faced with the number of topics of each dataset. From the results of validation, PC and PE have recognized their best partition as $c(min)$ for all data sets. On the other hand, FS has recognized its best partition as $c(max)$ for all data sets.

For PC and PE this invariant optimal number of clusters can be explained by their monotonic tendency as described in Sect. 2.3. Moreover, the fuzzification factor $m = 2.5$ represents a large value for PC and PE in all data sets, resulting in both selecting $c = 2$ as shown in Figs. 1 and 2. This occurs because of their limits when $m \to \infty$ are defined as $\lim_{m\to\infty} PC = 1/c$ and $\lim_{m\to\infty} PE = \log_a c$. Besides that, FS is very unstable for low and high values of m, as observed in the experiments presented in [13].

Figures 1, 2, 3, 4, 5 and 6 show the behavior of all indexes for all data sets when c varies from $c = 2$ until $c(max)$. In Fig. 1 it is possible to check that for

[1] http://sites.labic.icmc.usp.br/pretext2/.

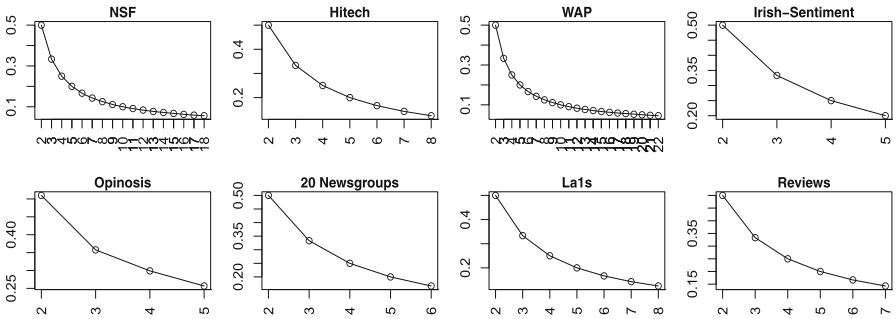

Fig. 1. PC results by number of clusters for each dataset

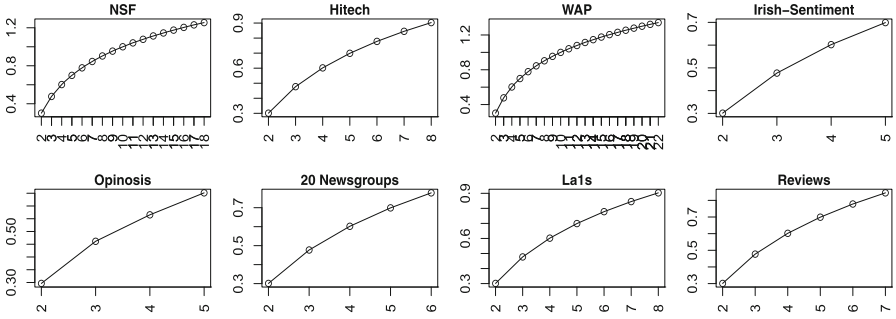

Fig. 2. PE results by number of clusters for each dataset

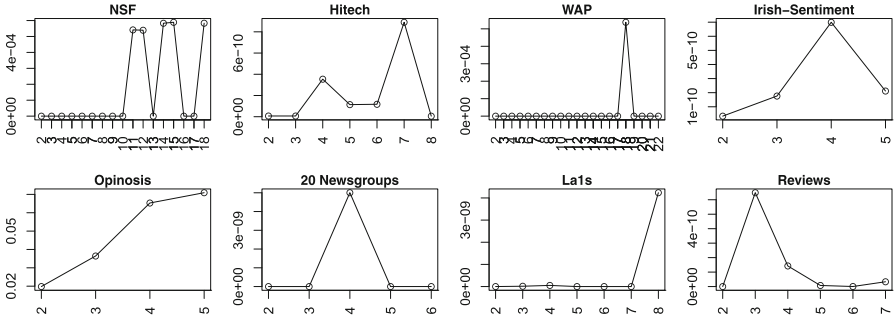

Fig. 3. MPC results by number of clusters for each dataset

all data sets PC had the same behavior as a decreasing function with values little above $1/c = 1/2$. In Fig. 2 is shown that for all data sets PE had the same behavior of a increasing logarithmic function with values very close to $\log 2 = 0.3$ that corresponds to the maximum value when $c = 2$ (3). Opinosis that has the lowes value of n (Table 1) had the best results of all data sets for PC and PE.

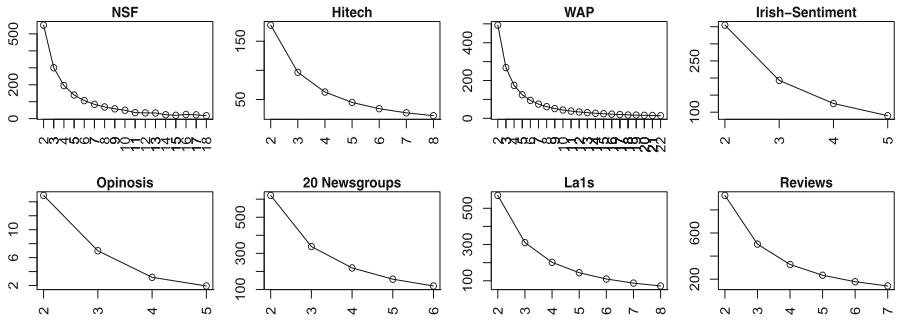

Fig. 4. FS results by number of clusters for each dataset

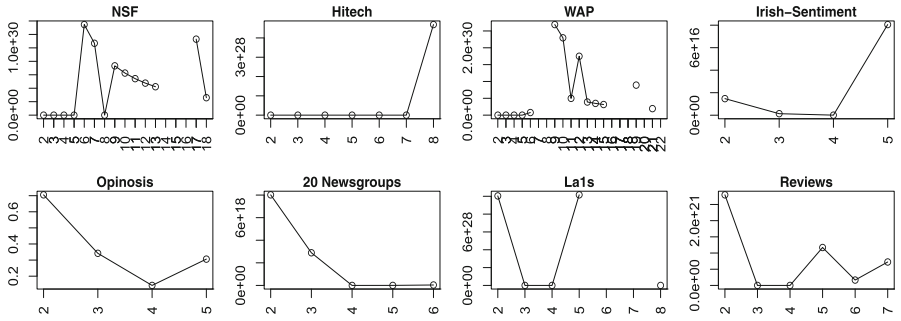

Fig. 5. XB results by number of clusters for each dataset

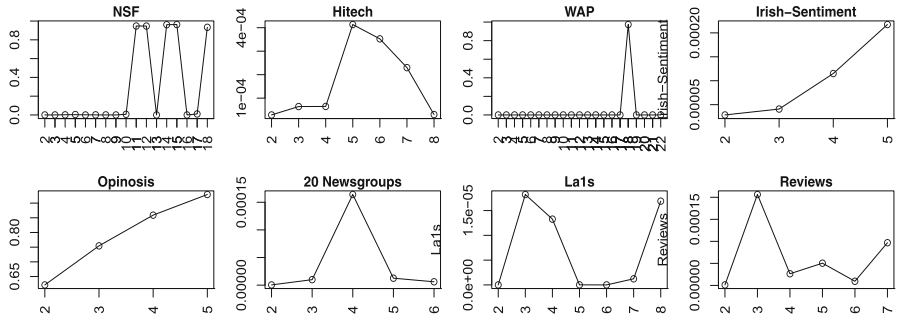

Fig. 6. SF results by number of clusters for each dataset

FS (Fig. 4) had a similar behavior of PC (Fig. 1) with the difference that FS can oscillate as can be verified in Fig. 4 for NSF dataset. From $c = 15$ to $c = 16$ FS increased its value and from $c = 16$ FS comes back to decrease. This oscillation of FS is explained for the difference between compactness and separation of the clusters (5) that does not allows that it has the same behavior of PC with $m = 2.5$.

In Fig. 3, MPC did not have a standard behavior like PC, PE or FS and the data sets WAP, Irish-Sentiment, 20 NewsGroup, La1s and Reviews have only one oscillation that results in the maximum value of MPC. The Opinosis dataset had a similar behavior as in PE (Fig. 2) and again had the best value of all data sets for MPC.

SF (Fig. 6) did not have a standard behavior on the data sets but showed a similar behavior of MPC for WAP, 20 NewsGroups and NSF. XB also did not show a standard behavior but was the only index that obtained infinite value for WAP, NSF and La1s in some values of c in Fig. 5. The Opinosis dataset again had the lowest values of XB that represent that it had clusters more compact and separate.

From Figs. 1, 2, 3, 4, 5 and 6 it is possible to check that using $m = 2.5$ for high dimensional data sets: (1) The monotonic tendency of PC and PE beyond their invariant behavior; (2) FS value trends to get down when c grows (as verified in [13] when for well-known Iris dataset, FS points to the maximum value of c, in this case $c = 10$) showing that as more clusters are formed, the value of $(A_i(\mathbf{d}_k))^m$ is smaller than less clusters and/or the difference between intra and inter clusters distances ($\|\mathbf{d}_k - \mathbf{v}_i\|^2 - \|\mathbf{v}_i - \bar{\mathbf{v}}\|^2$) is small; (3) In many data sets, the biggest difference of all the indexes values was from $c = 2$ to $c = 3$.

Such results were analysed because it is not enough an index to find the number of clusters equal to the number of classes. Since the organization of documents in a flexible way is a completely unsupervised approach, in which document labels are not predefined, the definition of a good partition to be obtained from an appropriate index is very important.

Therefore, FDO was also evaluated checking its performance face to the power prediction of the descriptors obtained after document clustering. To do that, the best value of MPC, XB and SF indexes to select the optimal number of clusters were used to check the performance rate of FDO obtained by a text classification algorithm. Only such indexes were used because PC, PE, and FS are more sensitive to the parameters, as explained previously.

After checking the indexes results, each cluster obtained by FCM was considered as a class. After labeling each document in the collection with the corresponding clusters, an attribute-value matrix was created with each descriptor being an attribute. The matrix entries are the frequency of the descriptors in each document. Using such attribute-value matrix, we have performed the machine learning algorithm Support Vector Machine (SVM) often used for text classification [15]. The performance of SVM was tuned up using Normalized Polynomial Kernel and the complexity parameter p=2.0. The 10-fold cross validation method was used in all experiments. Since in fuzzy clustering documents can belong to more than one cluster, the multi-label classification was used by means of the MEKA tool [14]. The number of clusters by which each cluster obtained its best value and their classification rates are presented in Table 2.

From Table 2, we recognize that MPC can be considered as a good validation indexes for flexible high-dimensional document organization for having had the highest classification rates (showed in bold in Table 2). Moreover, MPC has been

Table 2. Flexible document organization performance measured by means of the classification rate (%) and standard deviation obtained using MPC, XB and SF fuzzy clustering validity indexes.

Dataset	#clusters	MPC	#clusters	XB	#clusters	SF
NSF	15	**99.9** (0.6)	4	94.6 (2.1)	15	98.9 (0.6)
Hitech	7	52.5 (3.7)	7	**52.8** (3.4)	5	52.1 (5.3)
WAP	18	**96.7** (2.2)	4	58.2 (7.2)	18	**96.7** (2.3)
Irish-Sentiment	4	63.1 (6.1)	4	64 (10.8)	5	**69.0** (5.5)
Opinosis	5	**68.1** (18.9)	4	58.5 (15.4)	5	66.7 (14.6)
20 Newsgroups	**4**	**98.2** (1.3)	**4**	86.9 (2.9)	**4**	74.1 (3.2)
La1s	8	**75.7** (2.3)	4	65.7 (3.3)	3	58.5 (4.0)
Reviews	3	**76.7** (14.0)	4	57.7 (20.0)	3	**76.7** (14.0)

the index that most correctly selected the optimal number of clusters. Thus MPC was the best index to identify the distribution of topics in the used collections.

4 Conclusion

We have presented a comparative study concerning the performance of FDO using cluster descriptors obtained after fuzzy clustering and different validity indexes. Although there exists plenty of work analyzing indexes used to assess clustering methods, such analysis on high-dimensional collections is still an open problem.

The results of this study suggest that the descriptors extracted after FCM and an appropriate clustering validity index can achieve good attributes for text categorization of high-dimensional collections. Moreover, MPC cluster validity index may guide the selection of the appropriate partition of a collection.

Furthermore, the values of m and $c(max)$ used in validity indexes and the FCM algorithm are also determinant to the indexes calculation that consider the high-dimensionality of a data set.

As future work, we intend to perform more experiments using different parameters of the fuzzification factor m and $c(max)$, in order to identify drawbacks present in current cluster validity indexes, motivating the design of new ones.

References

1. Bezdek, J.C.: Numerical taxonomy with fuzzy sets. J. Math. Biol. **1**(1), 57–71 (1974). doi:10.1007/BF02339490
2. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell, MA (1981)
3. Bezdek, J.C.: Cluster validity with fuzzy sets. J. Cybern. **3**(3), 58–73 (1974)

4. Campello, R., Hruschka, E.: A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Sets Syst. **157**(21), 2858–2875 (2006)
5. Carvalho, N.V., Rezende, S.O., Camargo, H.A., Nogueira, T.M.: Flexible document organization by mixing fuzzy and possibilistic clustering algorithms. In: IEEE International Conference on Fuzzy Systems, pp. 790–797 (2016)
6. Chiang, I.J., Liu, C.H., Tsai, Y.H., Kumar, A.: Discovering latent semantics in web documents using fuzzy clustering. IEEE Trans. Fuzzy Syst. **23**(6), 2122–2134 (2015)
7. Dave, R.N.: Validating fuzzy partitions obtained through c-shells clustering. Pattern Recogn. Lett. **17**(6), 613–623 (1996)
8. Fukuyama, Y., Sugeno, M.: A new method of choosing the number of clusters for fuzzy c-means method. In: Fuzzy Systems Symposium, pp. 247–250 (1989)
9. Ingwersen, P.: Information Retrieval Interaction. Taylor Graham, London (1992)
10. Nogueira, T.M., Rezende, S.O., Camargo, H.A.: Fuzzy cluster descriptor extraction for flexible organization of documents. In: International Conference on Hybrid Intelligent Systems, pp. 528–533 (2011)
11. Nogueira, T.M., Rezende, S.O., Camargo, H.A.: Fuzzy cluster descriptors improve flexible organization of documents. In: International Conference on Intelligent Systems Design and Applications, pp. 616–621 (2012)
12. Nogueira, T.M., Rezende, S.O., Camargo, H.A.: Flexible document organization: comparing fuzzy and possibilistic approaches. In: IEEE International Conference on Fuzzy Systems, pp. 1–8 (2015)
13. Pal, N.R., Bezdek, J.C.: On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Syst. **3**(3), 370–379 (1995)
14. Read, J., Reutemann, P., Pfahringer, B., Holmes, G.: MEKA: A multi-label/multi-target extension to Weka. J. Mach. Learn. Res. **17**(21), 1–5 (2016)
15. Shanahan, J., Roma, N.: Improving SVM text classification performance through threshold adjustment. Machine Learning, Lecture Notes in Computer Science, vol. 2837, pp. 361–372 (2003)
16. Soares, M.V.B., Prati, R.C., Monard, M.C.: PreTexT II: Description of restructuring tool preprocessing of texts. Technical report 333, ICMC-USP (2008). (in Portuguese)
17. Subhashini, R., Kumar, V.: Evaluating the performance of similarity measures used in document clustering and information retrieval. In: International Conference on Integrated Intelligent Computing, pp. 27–31 (2010)
18. Wang, W., Zhang, Y.: On fuzzy cluster validity indices. Fuzzy Sets Syst. **158**(19), 2095–2117 (2007)
19. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. **13**(8), 841–847 (1991)

http://www.springer.com/978-3-319-66823-9

Advances in Fuzzy Logic and Technology 2017
Proceedings of: EUSFLAT- 2017 – The 10th Conference of the European Society for Fuzzy Logic and Technology, September 11-15, 2017, Warsaw, Poland IWIFSGN'2017 – The Sixteenth International Workshop on Intuitionistic Fuzzy Sets and Generalized Nets, September 13-15, 2017, Warsaw, Poland, Volume 2
Kacprzyk, J.; Szmidt, E.; Zadrożny, S.; Atanassov, K.T.; Krawczak, M. (Eds.)
2018, XI, 628 p. 198 illus., Softcover
ISBN: 978-3-319-66823-9