FEDERAL UNIVERSITY OF BAHIA - BRAZIL
DEPARTMENT OF COMPUTER SCIENCE

On Fuzzy Cluster Validity Indexes for High Dimensional Feature Space

Fernanda Eustáquio, Heloisa Camargo, Solange Rezende, and Tatiane Nogueira

Flexible Document Organization

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.

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.

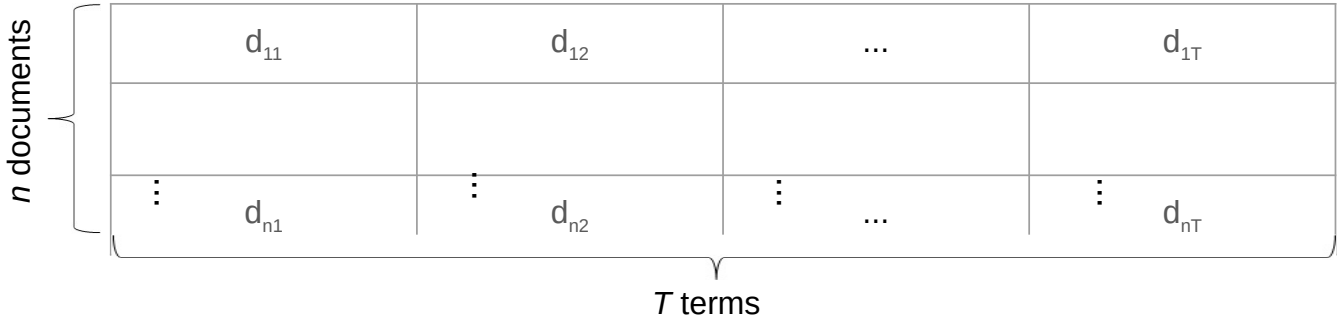
Since documents are represented in Flexible Document Organization (FDO) (1) by a high dimensional feature space and a document needs to be compatible with more than one cluster with different compatibility degrees, the selection of the correct partition of a given collection is a problem in evidence.

Document preprocessing

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;
2. Stemming, a technique that reduce words to their root form;
3. *n-gram* extraction.

After selecting the terms that represent the document collection, for the proposed approach, the document-term matrix contains in its cells *tf* - *idf* (Term Frequency-Inverse Document Frequency). Document x term matrix (2):



Fuzzy validity indexes

The best number of clusters for the document organization can be selected using a suitable fuzzy cluster validity index. These indexes are organized as:

- Indexes that use only the matrix of membership degrees
- Indexes that use the matrix of membership degrees and the dataset
  - These indexes use, in addition to the membership degree of documents in each cluster, the dissimilarity between a document d_k and a cluster prototype v_i or the dissimilarity between v_i and the prototypes mean defined as:

$$\bar{v} = \sum_{i=1}^{c} v_i / c$$

Indexes that use only the matrix of membership degrees

Partition Coefficient - PC (3): the PC index measures the amount of fuzzy intersection among the documents clusters in a partition. The index is defined as:

$$PC = \frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n} A_i^2(d_k)$$

PC values range in [1/c; 1]
Optimal number of clusters: *c* = max {PC}

Modified Partition Coefficient - MPC (4): this index was proposed as an effort to try to correct the monotonic trend that PC has. The index is defined as:

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

MPC values range in [0; 1]
Optimal number of clusters: *c* = max {MPC}

Indexes that use only the matrix of membership degrees

Partition Entropy - PE (5): the PE index measures the amount of fuzziness in a partition and it is characterized as a minimization index. In this paper it was performed using *a = 10*. The index is defined as:

$$PE = -\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n} A_i(d_k)\log_a(A_i(d_k))$$

PE values range in [0; $\log_a c$]
Optimal number of clusters: c = min {PE}

Indexes that use the matrix of membership degrees and the dataset

Fukuyama-Sugeno index - FS (6): FS is a minimization index given by the difference between the clusters compactness and the separation between them. The index is defined as:

$$FS = \underbrace{\sum_{i=1}^{c}\sum_{k=1}^{n} A_i^m(d_k)\,|d_k - v_i|^2}_{\text{Compactness}} - \underbrace{\sum_{i=1}^{c}\sum_{k=1}^{n} A_i^m(d_k)\,|v_i - \bar{v}|^2}_{\text{Separation}}$$

Optimal number of clusters: c = min {FS}

Xie-Beni index - XB (7): XB 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:

$$XB = \frac{\overbrace{\sum_{i=1}^{c}\sum_{k=1}^{n} (A_i^m(d_k)\,|d_k - v_i|^2}^{\text{Compactness}}}{\underbrace{n \times min_{k\neq i}\,|v_i - v_k|^2}_{\text{Separation}}}$$

Optimal number of clusters: c = min {XB}

Indexes that use the matrix of membership degrees and the dataset

Fuzzy Simplified Silhouette - SF (8): the SF index considers the two clusters in which d_k has the greatest membership degrees. The index is defined as:

$$SF = \frac{\sum_{k=1}^{n} (A_1(d_k) - A_2(d_k))S(d_k)}{\sum_{k=1}^{n} (A_1(d_k) - A_2(d_k))},$$

$$S(d_k) = \frac{\beta(d_k, g_i) - \delta(d_k, g_i)}{max\{\delta(d_k, g_i), \beta(d_k, g_i)\}}$$

Compactness

Separation

Silhouette of document d_k

where d_k belongs to the cluster g_i;

$\beta(d_k, g_i)$ is the average distance between d_k and all other documents belonging to g_i;

$\delta(d_k, g_i)$ is the distance between d_k and the neighbor cluster closest to g_i;

Optimal number of clusters: c = max {SF}

Fuzzy cluster descriptor extraction

All the terms found in a document preprocessing step are initially considered as descriptor candidates. Additionally, a document d_k is considered to belong to a cluster i if it has a membership degree $A_i(d_k) \geq s$, where $s = 1/c$. The threshold s is considered for two reasons:

1. 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;
2. 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 and then the descriptors are selected.

Document collections

Dataset	#terms	#docs	#class
NSF	2804	1600	16
Hitech	6593	600	6
WAP	8068	1560	20
Irish-Sentiment	8658	1660	3
Opinosis	10784	51	3
20Newsgroups	11026	2000	4
La1s	13196	3204	6
Reviews	22926	4069	5

Experimental results

The Fuzzy C-Means (FCM) (9) algorithm was performed using the following values of parameter:

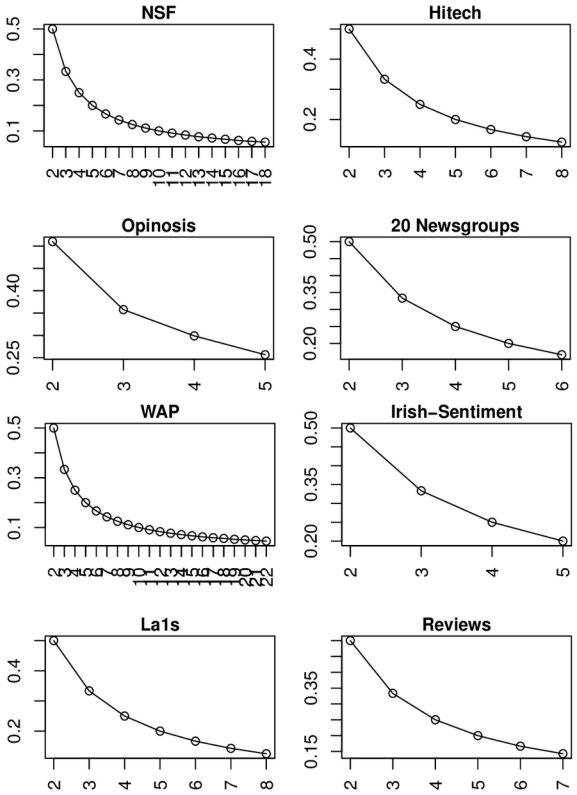
- Error $\mathcal{E} = 0.01$ as stop criterion;
- The fuzzification factor $m = 2.5$;
- The number of clusters c ranging from $c(min) = 2$ to $c(max) = \#class + c(min)$

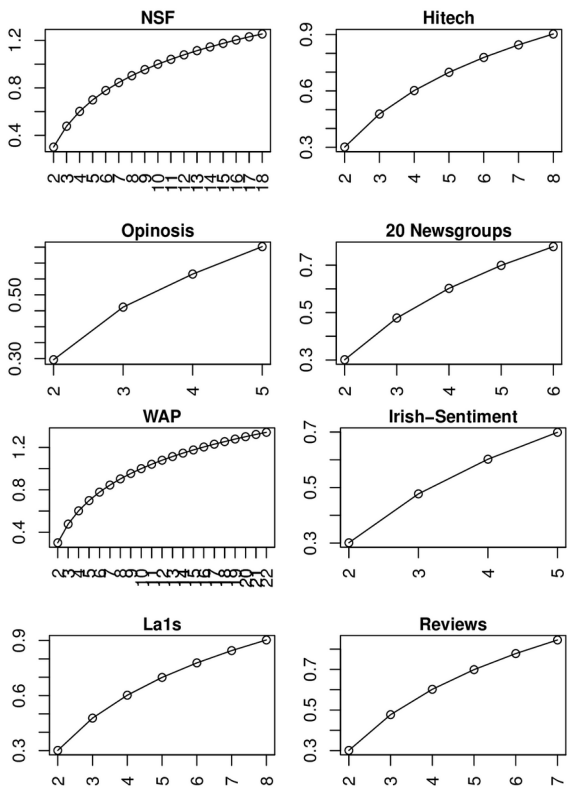
The dissimilarities between a document d_k and a cluster prototype v_i or between v_i and the prototypes mean v was calculated using the cosine coefficient similarity defined as:

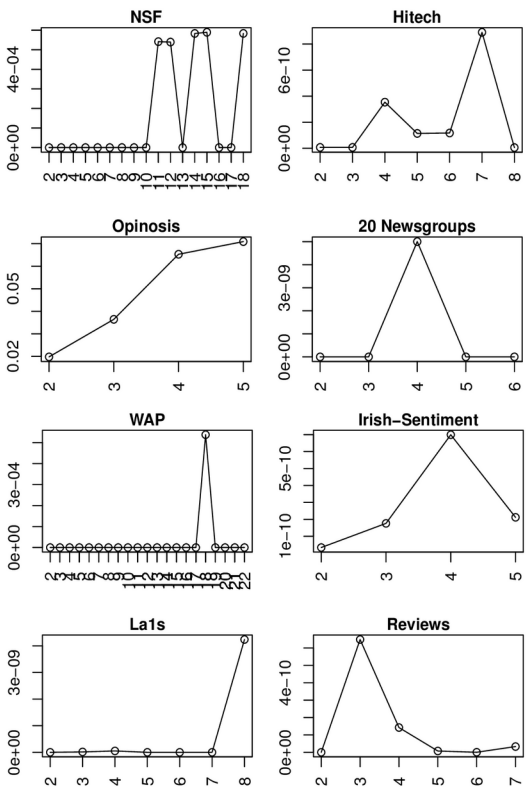
$$sim(d_k, v_i) = cos\theta = \frac{d_k . v_i}{|d_k||v_i|} \in [0;1]$$

$$|d_k - v_i| = 1 - sim(d_k, v_i) \in [0;1]$$

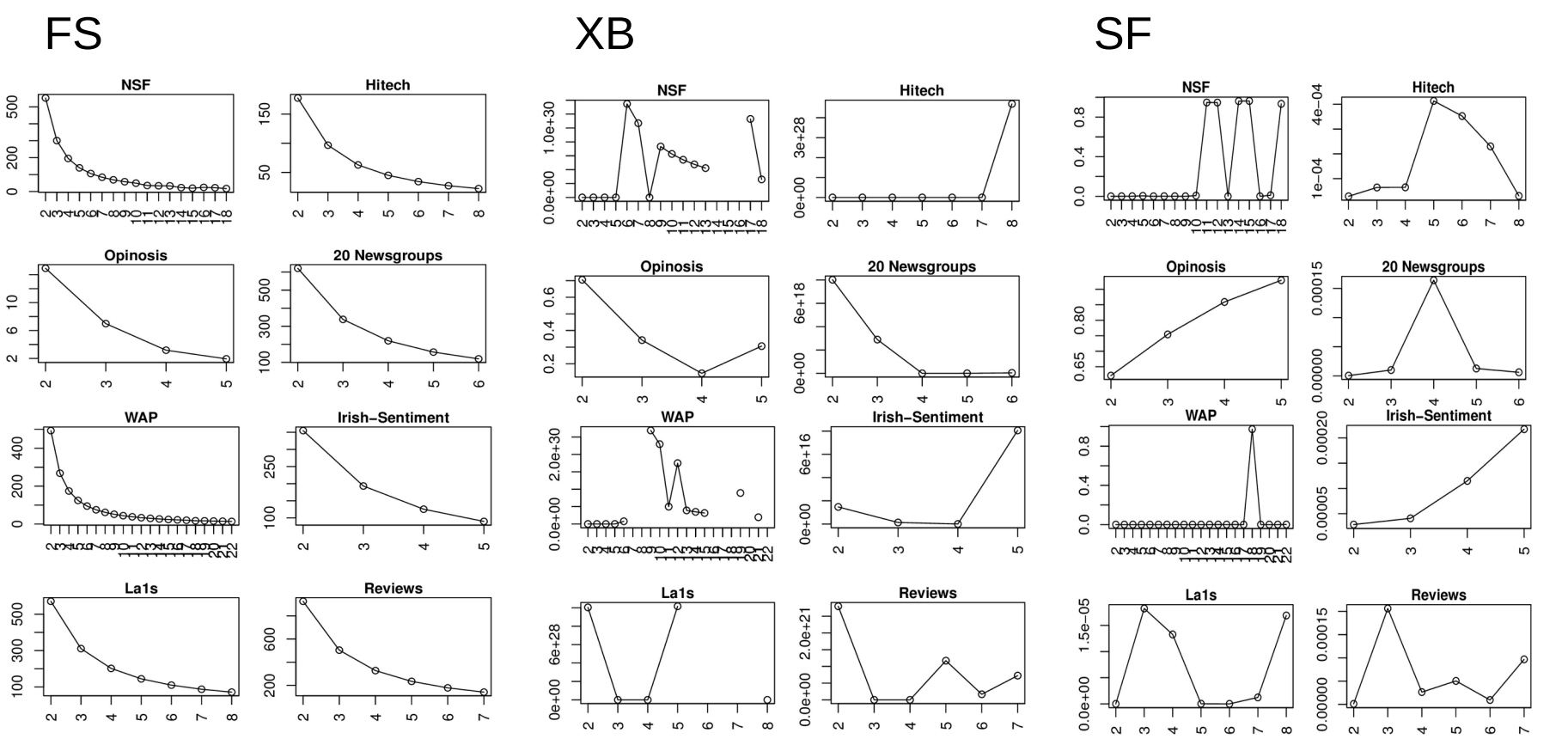
Experimental results







Experimental results



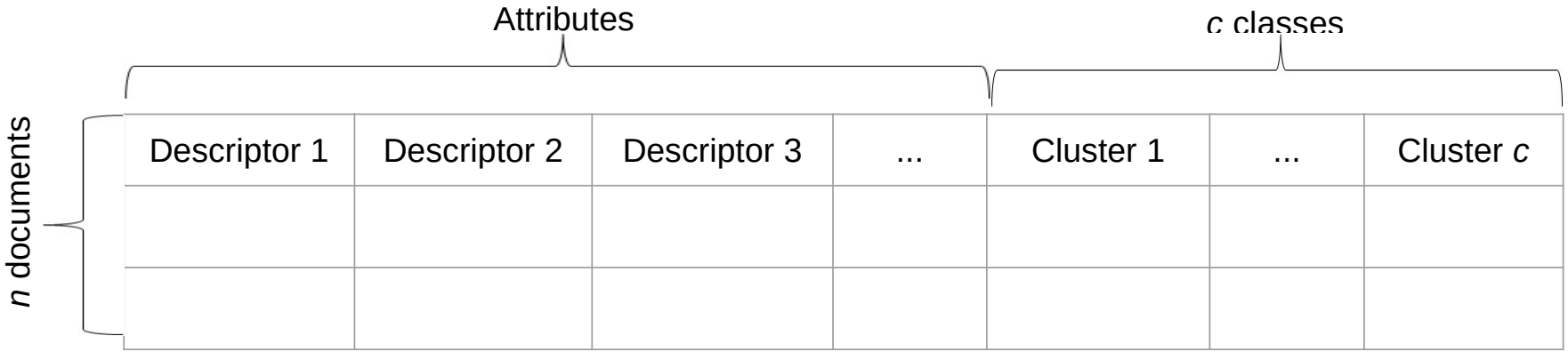
Experimental results

From the figures it was 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. The fuzzification factor *m = 2.5* represents a large value for PC and PE in all data sets, resulting in both selecting *c = c(min) = 2;*
3. FS value trends to get down when *c* grows showing that as more clusters are formed, the value of $(A_i(d_k))^m$ is smaller than less clusters and/or the difference between intra and inter clusters distances ($||d_k - v_i||^2 - ||v_i - v||^2$) is small;
4. FS has recognized its best partition as *c(max)* for all data sets;
5. In many data sets, the biggest difference of all the indexes values was from *c = 2* to *c = 3*.

Experimental results

FDO was also evaluated checking its performance face to the power prediction of the descriptors obtained after document clustering. After checking the indexes results, each cluster obtained by FCM was considered as a class. The matrix entries are the frequency of the descriptors in each document. The attribute-value matrix is defined as:



Using such matrix, we have performed the machine learning algorithm Support Vector Machine (SVM) often used for text classification.

Experimental results

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	c	MPC	c	XB	c	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	64.0 (10.8)	5	**69.0** (5.5)
Opinosis	5	**68.1** (18.9)	4	58.5 (15.4)	5	66.7 (14.6)
20Newsgroups	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)

Conclusion

- 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.
- It was possible to recognize that MPC can be considered as a good validation index for flexible high dimensional document organization for having had the highest classification rates. Moreover, MPC has been 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.
- 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.

Thank you!

Questions?

solange@icmc.usp.br, fernandase@dcc.ufba.br,
heloisa@dc.ufscar.br, tatiane.nogueira@ufba.br

References

(1) 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).

(2) 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).

(3) Bezdek, J.C.: Numerical taxonomy with fuzzy sets. Journal of Mathematical Biology 1(1), 57-71 (1974), http://dx.doi.org/10.1007/BF02339490.

(4) Dave, R.N.: Validating fuzzy partitions obtained through c-shells clustering. Pattern Recognition Letter 17(6), 613{623 (May 1996).

(5) Bezdek, J.C.: Cluster validity with fuzzy sets. Journal of Cybernetics 3(3), 58-73 (1974).

(6) 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).

(7) Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(8), 841-847 (Aug 1991).

(8) Campello, R., Hruschka, E.: A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Sets and Systems 157(21), 2858 - 2875 (2006).

(9) Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell, MA, USA (1981).