The IEMA Fuzzy *c*-Means Algorithm for Text Clustering

Domenica Fioredistella Iezzi[1], Mario Mastrangelo[2]

[1] Tor Vergata University – stella.iezzi@uniroma2.it
[2] Tor Vergata University – mario.mastrangelo@ uniroma2.it

Abstract

The fuzzy *c*-means algorithm is a soft version of the popular *k*-means clustering. As is well known, the *k*-means method begins with an initial set of randomly selected exemplars and iteratively refines this set so as to decrease the sum of squared errors. The *k*-centers clustering is moderately sensitive to the initial selection of centers, so it is usually rerun many times with different initializations in an attempt to find a good solution. We propose a new version of the fuzzy *c*-means algorithm for unstructured data to detect the best centroid of clusters, and we choose the final partition according to the validation of the Xie-Beni index. We apply our method to three different corpora (literature, forum, and ads) to verify the quality of the procedures.

Riassunto

L'algoritmo fuzzy *c*-means è la versione soft del popolare algoritmo di classificazione *k*-medie. Il metodo *k*-medie seleziona casualmente i primi centroidi, dopo avere fissato un numero *k* di gruppi in cui partizionare il dataset, poi, iterativamente, migliora la partizione ottenuta calcolando la distanza euclidea minima tra i centroidi e gli oggetti da classificare. Il centroide è sensibile alla selezione iniziale dei punti, infatti, usualmente si cercano differenti inizializzazioni dei punti per trovare una buona soluzione finale. L'obiettivo di questo lavoro è proporre una nuova versione dell'algoritmo fuzzy *c*-means (IEMA fuzzy *c*-means) per dati non strutturati. Per verificare la qualità dei risultati ottenuti, sono stati utilizzati tre differenti corpora (Alice, pubblicità natalizie e analisi di un forum di turismo) e validati mediante l'indice Xie-Beni.

Keywords: centroids, correspondence analysis of lexical table, *k*-means, fuzzy *c*-means

1. Introduction

Text clustering refers to a broad variety of methods that subdivide a corpus C into *c* subsets (clusters), which are pairwise disjoint and all nonempty and reproduce C through union. The clusters then are termed by a hard (i.e., nonfuzzy) *c*-partition of C. Accordingly, text clustering is an unsupervised learning task that aims at decomposing a given set of words/documents into subgroups or clusters based on the grade of similarity. The goal is to divide the corpus in such a way that words/documents belonging to the same cluster are as similar as possible, whereas objects belonging to different clusters are as dissimilar as possible. A possible drawback of these algorithms is that each element in C (words/documents) is unequivocally grouped with other members of its cluster and thus bears no apparent similarity to other members of C. This risk is higher in a corpus as regards language ambiguity. A manner to overcome this obstacle is to represent the similarity between words/documents using a fuzzy approach, where each cluster assumes a value between zero and one (Zadeh, 1965). The fuzzy approach allows us to deal with lexical ambiguity because a single word mostly belongs to multiple semantic categories, and a hard approach does not allow us to suitably treat this issue (e.g., in the tourism sector, the expression *holiday* is very close to *trip*, but generally, *holiday* is synonymous with *vacation* or *days off* (Iezzi & Mastrangelo, 2013). Fuzzy *c*-means (FCM) clustering is one of the well-known unsupervised clustering techniques, which can also be used for document clustering.

In this paper, we will introduce a new FCM algorithm (IEMA) that improves the final partition, choosing the centroids according to the Xie-Beni index (Xie-Beni, 2001), and we compare the results with FCM (Dunn, 1973; Bezdek, 1973). We apply our method to three different corpora (literature, forum, and ads) to verify the quality of the procedures. The paper is structured as follows: in section 2, we describe the state of the art about *c*-means algorithm; in section 3, we explain the IEMA FCM; in section 4, we illustrate the applications carried out on three corpora: *Alice in Wonderland* (literature), posts from tourism forums (forum), and articles on big data collected in newspapers; in section 5, we explain the conclusions and final remarks.

2. Fuzzy *c*-Means

FCM algorithm is a soft version of the popular *k*-means clustering. The *k*-means method, after choosing the number of clusters *k,* starts from a partition of points that may be random or given by an ad hoc rule. Given the initial partition, the following two steps are repeated until convergence—that is, no change of cluster memberships: (1) calculate the center of each cluster as the center of gravity, which is also called the *centroid*; (2) reallocate every point to the nearest cluster center (Iezzi, 2012a ; Miyamoto *et al*., 2008).

Fuzzy cluster analysis allows gradual memberships of data points to clusters measured as degrees in [0,1]. This gives the flexibility to express that data points can belong to more than one cluster. The basic algorithm for the *c*-means method is as follows:

1. Specify the number of clusters *k* and then randomly select *k* observations to initially represent the *k* cluster centers. Each observation is assigned to the cluster corresponding to the closest of these randomly selected objects to form *k* clusters.
2. Centroids of the clusters are calculated, and each observation is reassigned (based on the new means) to the cluster whose mean is closest to it to form new *k* clusters.
3. Repeat step 2 until the algorithm stops when the means of the clusters are constant from one iteration to the next.

The objective of fuzzy clustering is to partition a data set into *c* homogeneous fuzzy clusters. This algorithm is based on the minimization of the following objective function:

$$Fcm_m = \sum_{i=1}^{c}\sum_{j=1}^{n} {u_{ij}}^m \left\|x_j - v_i\right\|^2, 1 \le m \le \infty$$

where *m* is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i^{th} of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center (Iezzi et al., 2013).

3. IEMA Fuzzy *c*-Means

There are some disadvantages in using FCM: if the membership of an object is not strong enough or significantly high for a particular cluster, it will mean that the equation of calculating membership is not effective, and sometimes the equation for updating prototypes is incapable of working with data that are greatly affected by noise. Thus, the equation for updating centroids leads to the result of clustering that might be uncorrected. The main reason for the underlying drawbacks of the above is that FCM is employed based on existing Euclidean distance measures (Ganesh & Palanisamy, 2012). The centroids, however, are the

arithmetic average of the points that belong to the cluster, and this procedure is good only for spherical clusters, but more frequently, we have clusters with elliptical shape.
In our strategy, we apply a preprocessing to reduce the sparseness of term-document matrix using Correspondence Analysis of lexical tables (LCA). The input matrix for IEMA FCM is the first component obtained by LCA. The aim of this procedure was to improve the results of FCM, choosing the appropriate centroids or prototypes. The steps of the procedure are as follows:

1. Let **C** be a corpus; after preprocessing, we obtain a term-by-document matrix $\mathbf{X} = [w_{ij}]$, building with a bag-of-words weighting scheme.

2. We reduce the dimensionality of the **X** matrix by applying LCA, obtaining a new matrix **M** of size (words × first factorial components). The **X** matrix decomposition is expressed by a rectangular matrix as a product of three matrices of a simple structure $\mathbf{X} = \mathbf{U}\mathbf{D}\alpha\mathbf{V}^{T}$, where the columns of the matrices **U** and **V** are the left and the right singular vectors, respectively, and the positive values αk down the diagonal of **D**α, in descending order, are the singular values (Lebart, Salem, & Berry, 1988; Greenacre, 2007). We decide, time to time, the number of components so that the explained variability is at least 70% of the total.

3. We apply FCM algorithm to the matrix **M**; we apply the Xie-Beni index to decide the number of clusters *c* and to measure the clustering quality [F1]:

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} {u_{ij}}^{2}\left\|x_j - v_i\right\|^2}{n\, min_{i,j}\left\|v_i - v_j\right\|^2} \quad \text{[F1]}$$

4. The XB index is a fuzzy clustering validity function that evaluates the goodness of a fuzzy *c*-partition depending on two parameters, the overall average compactness and the separation of the clusters. The first one refers to the mean of compactness of data of each cluster, measured by the ratio of the variation of each cluster—that is, the summation of the fuzzy deviation's squares for each data point—to the (fuzzy) number of data points belonging to the cluster. The second one is expressed by the minimum distance between cluster centers.

In particular, by decomposing the XB index, we can identify, if it exists, the less compact cluster and its centroid, and we can repeat the *c*-means algorithm replacing the latter from time to time with the original data points, leaving unchanged the remaining *c*-1 centroids. In this way, we can obtain a better fuzzy *c*-partition, evaluated again by the XB index.

5. This procedure can be repeated iteratively, as long as we obtain a better partition, which means more compact clusters and/or a greater separation, expressed by a greater minimum distance between cluster centers.

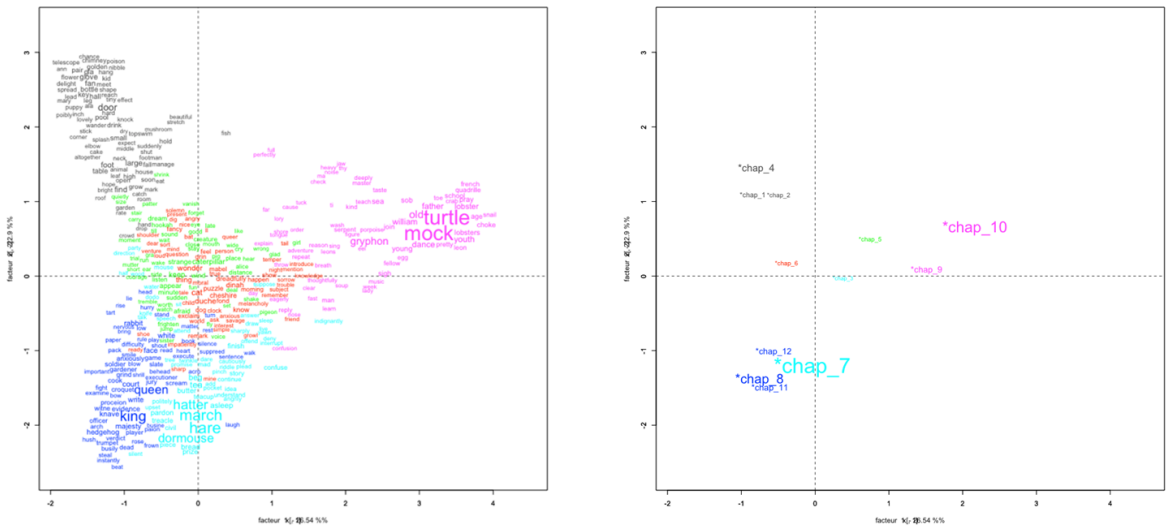
Data were processed with an R software; in particular, an R program for performing steps 4 through 9 has been developed by the authors.

4. Experimental Results

We analyze three different data sets: *Alice's Adventures in Wonderland* (literature); *Christmas ads* from *Il Corriere della Sera*, since 1876 (foundation year of the newspaper) to 2001; and *tourism forum posts on travel packages*. To classify those corpora, we use the same strategy: (1) preprocess (normalization and lexicalization), (2) build the term-document matrix, (3) reduce dimensionality using correspondence analysis of lexical table (LCA), (4) explore the corpus applying descending clustering (GNAPA), (5) use fuzzy *c*-means, (6) use IEMA (**IE**zzi and **MA**strangelo) fuzzy *c*-means, and (7) validate and compare the results of clustering using the Xie-Beni index (Iezzi and Mastrangelo, *in press*).

4.1. Corpus 1: Alice's Adventures in Wonderland

Alice's Adventures in Wonderland is a novel written in 1865 by English author Charles Lutwidge Dodgson under the pseudonym Lewis Carroll. The book is composed of 12 chapters, 27,379 words, 2,086 types, and 766 hapax legomena (2.80% of words and 36.72% of types). To reduce the dimensionality of the words, we lexicalized the corpus. In this way, we obtained a lexical table **A** of size 925 × 12, where 925 is the number of types after lexicalization and 12 is the number of chapters. We apply LCA on **A** to find the lower-dimensional subspaces, which most accurately approximates the original distributions of points (Lebart & Salem, 1994). We consider the first four factors that explain 86.828% of the total variance. As is well known, percentages of variance measure the relative importance of each eigenvalue in the trace. In this case, the plane spanned by the first two principal axes "explains" 49.440% of the total variance. Figure 3a shows that several clusters are overlapped, especially those that are located near the origin of the Cartesian axes. This configuration suggests that we cannot adopt a partition mutually exclusive of the contents, but it is preferable to use a fuzzy algorithm.



(a) *(b)*

Figure 1. Plane of the first two axes from the LMC of the words (a) and chapters (b)

Moreover, chapters 3, 5, 6, and 9 have overlapped contents (figure 1b). In an explorative phase, we use a descending classification, proposed by Reinard (1987, 1990), to detect the number of clusters in a corpus. We applied this algorithm because it subdivides the corpus into units of elementary contexts (CEU) and combines these CEUs so that we analyze dimension variables more widely, and then it performs a classification based on vocabulary distribution. This step allows to classify documents, characterized by their dominant vocabulary, in such a way as to obtain a classification based on an interpretative approach. Figure 2 shows that there are six clusters but four of them are nested (classes 1, 3, 4, and 5). The dendrogram displays that the clusters tend toward a minimally homogeneous distribution, in effect, the Gini index (hgi)[1] = 0.21.

1. $Hgi = \frac{k}{k-1}\sum_{i=1}^{k} f_i$, where f_i is relative frequency of class i^{th} and k is the number of cluster.

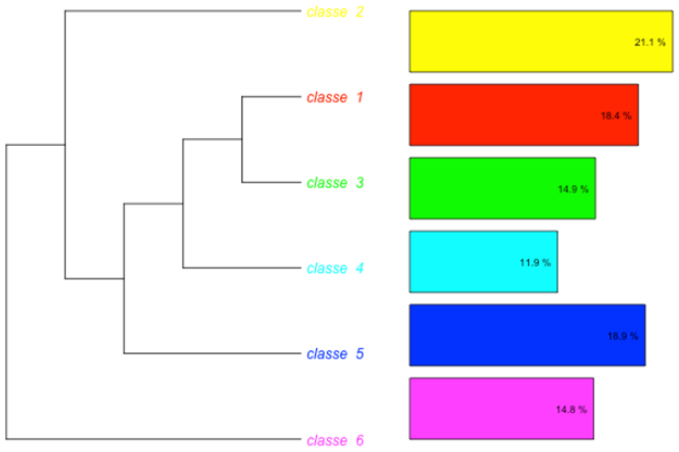
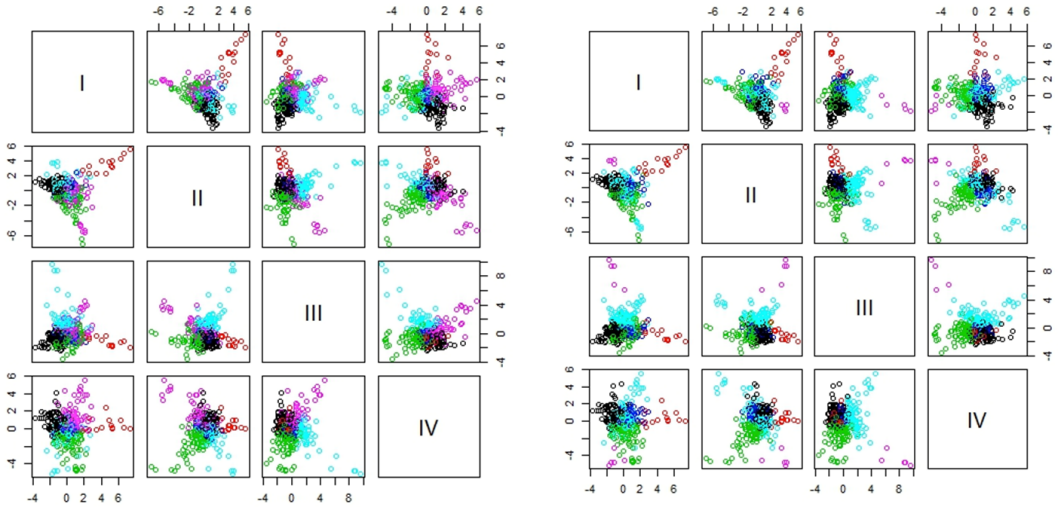


Figure 2. Hierarchical descending classification method of the corpus Alice

We detect six classes : (1) the pink cluster that describes mythological animals (e.g., *gryphon*) and marine creatures (e.g., *turtle*), (2) the light blue cluster that illustrates the movements of the characters (e.g., *to hare*, *to march*), (3) the dark blue cluster that illustrates the most important characters of the book (e.g., *king*, *queen*, *rabbit*), (4) the black cluster that focuses on the details of the environments (e.g., *door*, *table*, *bottle*, *small*, *large*), (5) the green cluster that represents the time-space dimensions (e.g., *clock*, *minute*, *distance*, *close*), and (6) the red cluster that depicts the imaginary world (e.g., *fantasy*, *remember*). The Xie-Beni index suggests five classes for the FCM and six groups for the IEMA FCM, but as can be seen from Table 1, the IEMA FCM always produces better results than the FCM. Figure 3 shows the final partitions of the FCM *(a)* and the IEMA FCM *(b)* for the six groups. In the IEMA FCM, the quality of the groups, in terms of internal cohesion and better separation, is very evident.



(a) *(b)*

Figure 3. Final partition of the FCM (a) and the IEMA FCM (b) - corpus Alice

4.2. Corpus: Christmas Advertisements

We analyzed 993 advertisements published in the newspaper *Corriere della Sera* on December 24 from 1876 (year of foundation) to 2000. The tokens are 25,108, and the types

are 6,422, and the number of hapax is 2,833 (Iezzi, 2012). If we apply LCA to the bag-of-words weighting scheme, the clusters will largely overlap (figure 4), and hierarchical classification will produce two mutually exclusive raw classes (figure 6). We grouped the ads into 25 chunks, where each chunk is composed of five observed years. The Christmas advertisements form non-overlapping groups for about 20 adjacent years, with the exception of the periods before and after the two world wars (figure 5). The descending hierarchical clustering produces two big clusters, and it cannot find the macro contents within the corpus (figure 6). The Xie-Beni index suggests fixing the number of classes into five groups, although when applying the IEMA FCM clustering, it can be seen that the ideal number of groups is four (table 1). This last result allows an optimal interpretation of the groups and corresponds to the results obtained using the centrality measures in place of the factorial coordinates (Iezzi, 2012b).

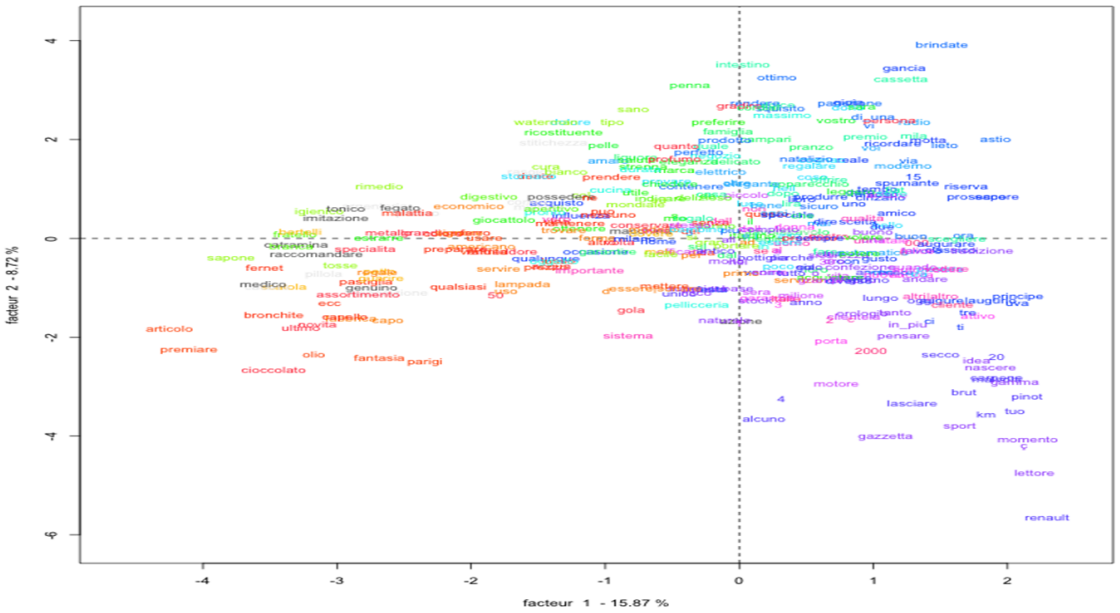


Figure 4. Plane of the first two axes from the LMC of the ads

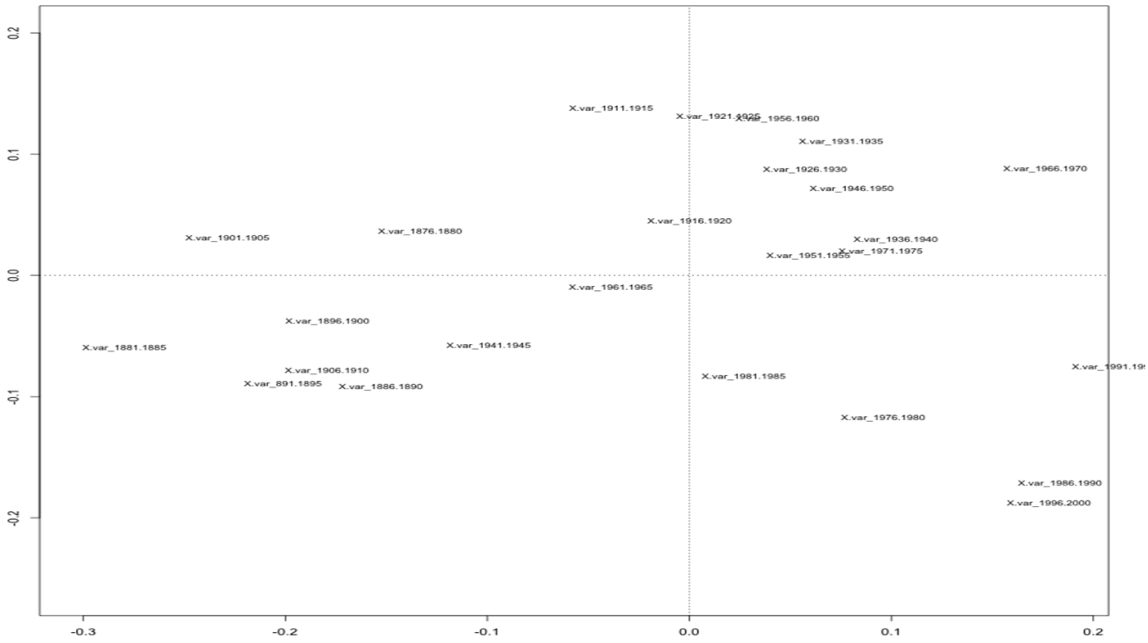


Figure 5. Plane of the first two axes from the LMC of the years

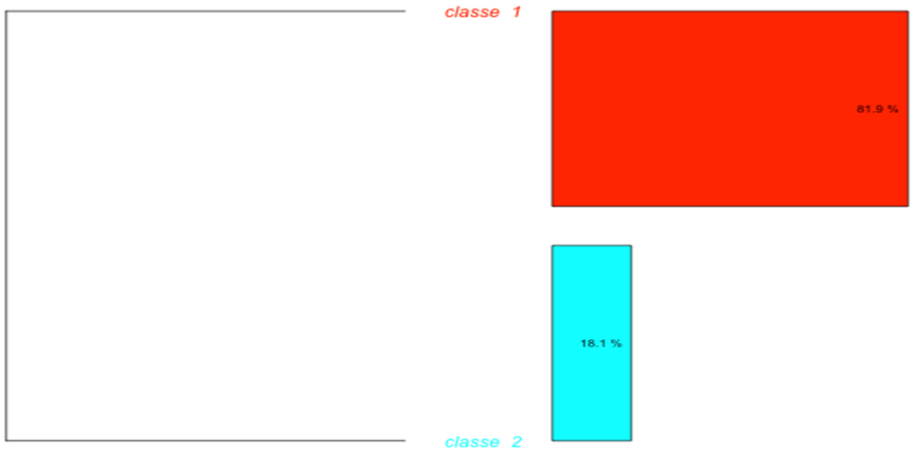
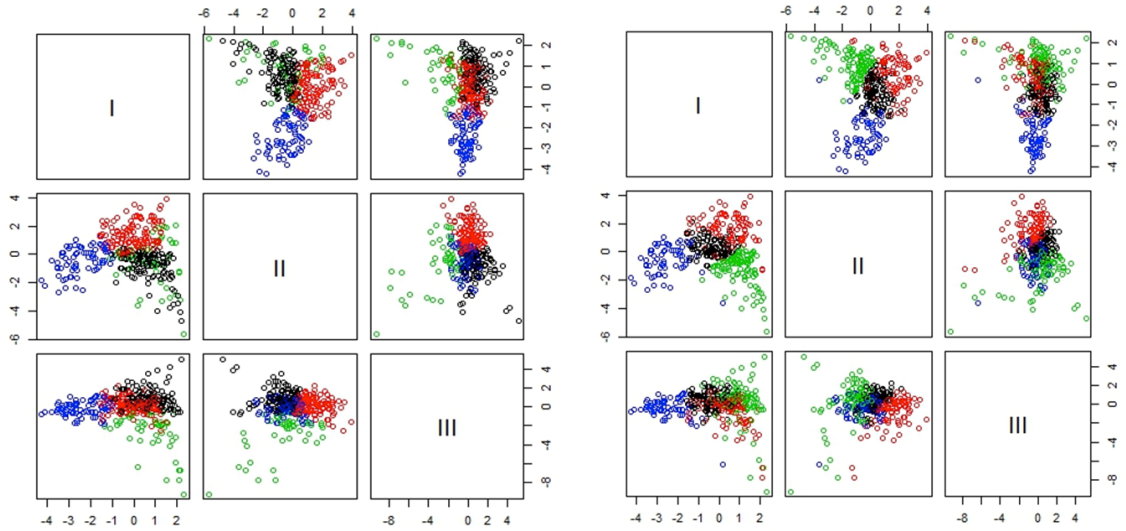


Figure 6. Hierarchical descending classification method of the corpus Christmas advertisements

Figure 7 represents the final partition of the FCM algorithm (*a*) and the IEMA FCM (*b*); also, in this case, the quality of the representation of the IEMA FCM is higher than the FCM.



(a) *(b)*

Figure 7. Final partition of the FCM algorithm (a) and the IEMA FCM (b) corpus Christmas ads

4.3. CORPUS Tourism Forum on Travel Packages

The corpus is composed of 817 posts, published on the website forum www.tripadvisor.it from 2008 to 2012, regarding the subject of travel packages. It is constituted of 73,168 words, 9,225 types, and 5,035 hapaxies. The term-document matrix was based not only on words but also on some repeated segments, for example, *web site*, *e-mail*, *travel agency*, *last minute*, *credit card*, *all inclusive*, *on line*, *best price*, and *legal fees*.

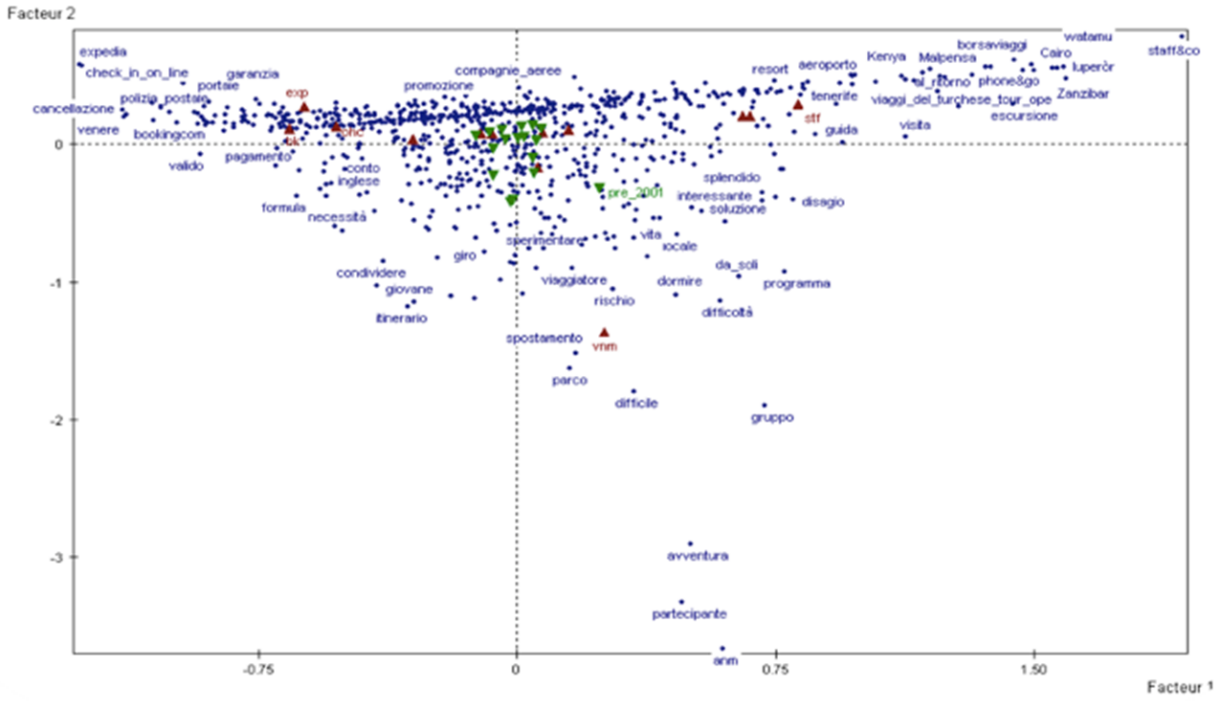


Figure 8. Plane of the first two axes from the LMC of the tourism forum Tripadvisor

The factorial plane shows the three main topics on which users tend to focus: (1) **reservations and payment**, located in the negative semiaxis of the first factor, the specific expressions of this issue are *expedia*, *delete*, *booking.com*, *outlethotels.com*, *amount*, *credit card*, *rate vouchers*, and *payment*; (2) **travel experience**, with specific reference to exotic destinations, this area is placed in the positive semiaxis of the first factor, and the most representative words are *excursion*, *airport*, *cruise*, *resort*, and *village*; (3) **travel kind**, *alternative* or *differentiated*, this group are collocated in the negative semiaxis of the second factor, and the most relevant words are *adventure*, *awkward*, *risk*, *group*, *parent*, and *share* (figure 8).

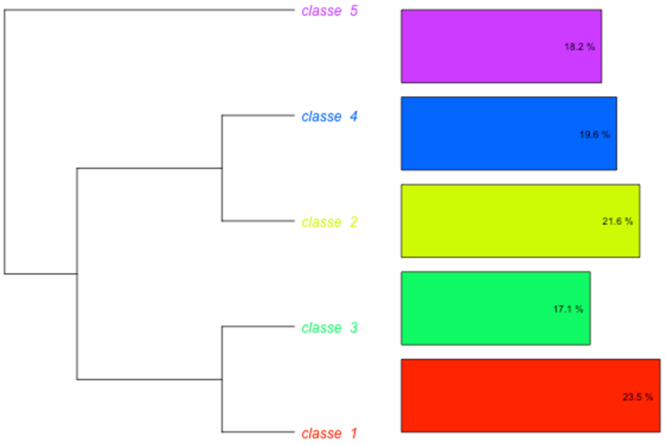


Figure 9. Hierarchical descending classification method of the corpus tourism forum

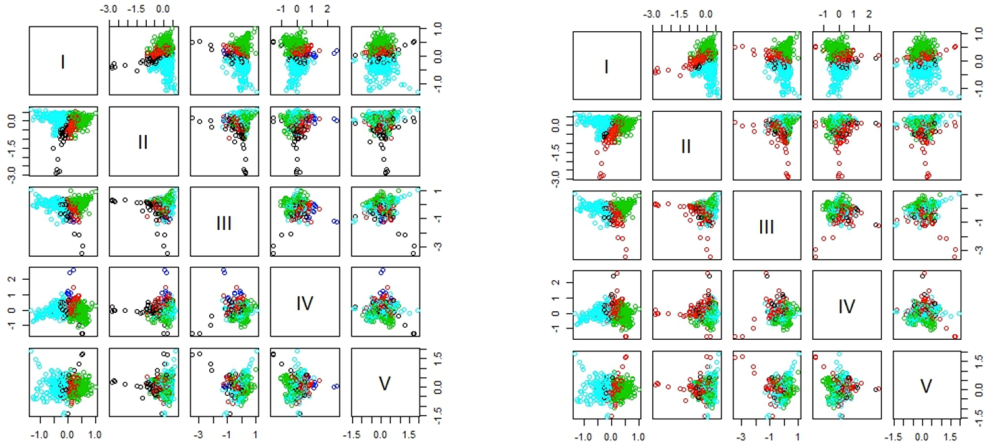
The descending hierarchical clustering detects five classes (figure 9), but the Xie-Beni index suggests six clusters (table 1).

Cluster no.	Alice		Christmas ads		Tourism forum	
	XB	***XB****	***XB***	***XB****	***XB***	***XB****
4	2,153	2,068	4,982	1,915	1,770	1,726
5	2,142	1,905	2,053	2,029	1,450	1,393
6	2,231	1,870	2,114	2,112	1,167	1,149

Legend: XB = Xie-Beni index for FCM; XB= Xie-Beni index for IEMA FCM.*

Table 1. Xie-Beni index for Alice, Christmas ads, and forum tourism corpora using FCM and IEMA FCM

The best number of groups is five, both for the FCM and for the IEMA FCM (table 1). The procedure IEMA improves in all cases, especially for the five groups (figure 10). Figure 10 shows that in this last case study, the quality of the clustering is also better than the IEMA FCM.



(a) *(b)*

Figure 10. Final partition of the FCM algorithm (a) and the IEMA FCM (b) for the corpus tourism forum

5. Conclusions

In the clustering of unstructured data, it is very frequent to use algorithms tested for structured data. In many cases, this approach leads to a lack of adaptation to the peculiarities of textual data. In the literature, especially in the presence of corpora with millions of occurrences, it is common to use the *k*-means algorithm because it has a very quick elaboration time. Unfortunately, this algorithm, using the Euclidean distance for optimization function, misinterprets outliers and noises. This distance only makes it possible to identify spherical clusters. Several variants have been proposed—for example, the fuzzy Gustafson-Kessel algorithm (Gustafson & Kessel, 1979) replaces the Euclidean distance with a Mahalanobis distance, so as to adapt to various sizes and forms of the clusters or the fuzzy *c*-medoids and fuzzy *c* trimmed medoids (Krishnapuram et al., 1999), where the objective functions are based on selecting *c*-representative objects (medoids) from the data set in such a way that the total dissimilarity within each cluster is minimized. In the last case, the medoid is also a more robust estimator of noise and outliers. We used a fuzzy approach because it is better suited to the ambiguity of the words, but algorithms used for text mining, proposed in the literature for

structured data, then need to be adapted for unstructured data. The IEMA FCM algorithm is a new suggestion that goes in this direction, and in all corpora examined by us, our algorithm always gets the best results compared with the FCM. To validate the outcomes, we used the Xie-Beni index, an important index for the validation of fuzzy clustering. In the future, we will also work with several indices—such as partition coefficient, partition entropy, modified partition coefficient, silhouette, and fuzzy silhouettes—to obtain more stable results. We have developed an R program for performing the IEMA FCM, then validation using the Xie-Beni index, and we will prepare a new routine for meeting the needs of different corpora and for different aims.

References

Bezdek J. C. (1973). *Fuzzy Mathematics in Pattern Classification*, PhD thesis, Cornell University Ithaca.

Dunn J. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters. *Journal of Cybernetics*, vol. 3(3): 32–57.

Krishnapuram, Raghu, Joshi, Anupam, Yi and Liyu (1999). Fuzzy relative of the *k*-medoids algorithm with application to web document and snippet clustering. *IEEE International Conference on Fuzzy Systems*, 3, pp. 1281–1286. Cited 44 times.

Ganesh M. and Palanisamy V. (2012). A modified adaptive fuzzy c-means clustering algorithm for brain MR image segmentation. *International Journal of Engineering Research & Technology* (IJERT), 1: ePub 1: 125–135.

Greenacre M. (2007). *Correspondence Analysis in Practice*, Chapman & Hall, New York.

Gustafson E. E. and Kessel, W. C. (1979). Fuzzy clustering with a fuzzy covariance matrix. *Proc. of the IEEE Conference on Decision and Control*, San Diego, California, pp. 761–766. IEEE Press, Piscataway, NJ.

Iezzi D. F. (2012a). A new method for adapting the *k*-means algorithm to text mining. *Statistica Applicata – The Italian Journal of Applied Statistics*, vol. 22(1): 69–80.

Iezzi D. F. (2012b). Centrality measures for text clustering. *Communications in Statistics. Theory and Methods*, vol. 41: 3179–3197.

Iezzi D. F. and Mastrangelo M. (in press). Fuzzy *c*-means for web mining: The Italian tourist forum case. In (Eds) Vicari D., Okada A., Ragozini G., Weihs C., *Analysis and Modeling of Complex Data in Behavioral and Social Sciences*, Berlin: Springer.

Iezzi D. F., Mastrangelo M. and Sarlo S. (2013). A new fuzzy method to classify professional profiles from job announcements. In (Eds) Giudici P., Ingrassia S., Vichi M., *Statistical Models for Data Analysis*, 151–159, Berlin: Springer.

Lebart L. and Salem A. (1994). *Statistique Textuelle*, Dunod, Paris.

Miyamoto S., Ichihash H. and Honda K., (2008). *Algorithms for Fuzzy Clustering. Methods in c-Means Clustering with Applications*, Berlin: Springer.

Reinert M. (1987). Classification descendante hiérarchique et analyse lexicale par contexte: application au corpus des poésies d'Arthur Rimbaud (Descending hierarchical classification and context-based lexical analysis: Application to the corpus of poems by A. Rimbaud). *Bulletin de Méthodologie Sociologique*, vol. 13: 53–90.

Reinert M. (1990). ALCESTE. Une methodologie d'analyse des donnees textuelles et une application: Aurelia de Gerard de Nerval, *Bulletin de Méthodologie Sociologique*, vol. 26: 25–54.

Xie X. L. and Beni G. (1991). A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 13: 841–847.

Zadeh L. A. (1965). Fuzzy sets. *Information Control*, vol. 8: 338–353.