Comparative study on textual data set using fuzzy clustering algorithms

Rjiba Sadika
University of Economics and Management, Sousse, Tunisia
Moez Soltani
Department of Electrical Engineering,
High School of Engineers of Tunis (ENSIT), Tunis, Tunisia, and
Saloua Benammou
Faculté de Droit et des Sciences Economiques et Politiques de Sousse,
Sousse, Tunisia

Abstract

Purpose – The purpose of this paper is to apply the Takagi-Sugeno (T-S) fuzzy model techniques in order to treat and classify textual data sets with and without noise. A comparative study is done in order to select the most accurate T-S algorithm in the textual data sets.

Design/methodology/approach – From a survey about what has been termed the "Tunisian Revolution," the authors collect a textual data set from a questionnaire targeted at students. Five clustering algorithms are mainly applied: the Gath-Geva (G-G) algorithm, the modified G-G algorithm, the fuzzy c-means algorithm and the kernel fuzzy c-means algorithm. The authors examine the performances of the four clustering algorithms and select the most reliable one to cluster textual data.

Findings – The proposed methodology was to cluster textual data based on the T-S fuzzy model. On one hand, the results obtained using the T-S models are in the form of numerical relationships between selected keywords and the rest of words constituting a text. Consequently, it allows the authors to interpret these results not only qualitatively but also quantitatively. On the other hand, the proposed method is applied for clustering text taking into account the noise.

Originality/value – The originality comes from the fact that the authors validate some economical results based on textual data, even if they have not been written by experts in the linguistic fields. In addition, the results obtained in this study are easy and simple to interpret by the analysts.

Keywords Analysis data, Fuzzy c-means algorithm, Gath-Geva algorithm,
Kernel fuzzy c-means algorithm, Modified Gath-Geva algorithm, Textual data

Paper type Research paper

1. Introduction

Data mining is the process of extracting appropriate and useful information from a big data set. The overall goal is to transform big data into an understandable structure for further use. Moreover, it identifies original structures and significant correlations from a database (Azzalini *et al.*, 2012; Hastie *et al.*, 2001). It is applied in several fields such as medicine (Buk *et al.*, 2012), finance (Ngaia *et al.*, 2011) and management (Ngaia *et al.*, 2009).

Text mining (TM) was developed nearly a half century ago. The purpose of TM is to explore textual data sets (Lebart and Salem, 1988) extracted from natural language text. Hence, TM is applied to extract numerical information in order to highlight meaningful patterns between texts or words constituting a text (Manning and Schutze, 1999). Several TM methods are developed such as information retrieval (Do prado and Ferneda, 2007; Weiss *et al.*, 2010), information extraction (Yong and Mooney, 2002), web mining

Kybernetes
Vol. 45 No. 8, 2016
pp. 1232-1242

0368-492X
DOI 10.1108/K-11-2015-0301

(Kosala and Blockeel, 2000), K-means clustering text (Rakhlin and Caponnetto, 2007), hierarchical clustering text (Cai *et al.*, 2004) and spectral clustering text (Song *et al.*, 2011). Yao *et al.* (2012) successfully applied improved K-means (IKM) algorithm to cluster text in two steps of learning algorithm. The first one is to select documents more similar to the original cluster center. The second step is to calculate average value which is effectively considered as the new cluster center. The main finding to note in their study is that the IKM method ameliorates the clustering purity compared with the classical K-means method. Later, Yinglong *et al.* (2014) proposed a three-phase approach to document clustering based on topic significance degree. First, they determine the most significant topics by LDA technique. Second, the K-means++ algorithm was applied to choose the initial clustering centers. In the third phase, K-means method was used for document clustering. Recently, Wei *et al.* (2015) proposed a semantic approach based on lexical chains and using WordNet to exploit ontology hierarchical structure and relations in order to provide a more accurate assessment of the similarity between terms for word sense disambiguation.

However, the aforementioned algorithms consider only the semantic relationships among words. Moreover, in real applications, the textual data sets may be also subject to noise so that it cannot accurately represent the meaning of text. In order to overcome these problems, we propose a novel methodology to cluster textual data based on Takagi-Sugeno (T-S) fuzzy model (Takagi and Sugeno, 1985). On one hand, the results obtained using the T-S models are in the form of numerical relationships between selected keywords and the rest of the words constituting a text. Consequently, it allows us to interpret these results not only qualitatively but also quantitatively. On the other hand, the proposed method is applied for clustering text, taking into account the noise. Finally, a comparative study is done in order to select the most accurate algorithm for clustering text.

The major contributions of the current study, with respect to the related literature, can be summarized as follows:

- We use some clustering algorithms based on T-S fuzzy models such as Gath-Geva (G-G) (Gath and Geva, 1989) algorithm, the modified Gath-Geva (MGG) algorithm (Abonyi *et al.*, 2002), the fuzzy c-means (FCM) algorithm (Bezdek *et al.*, 1984) and the kernel fuzzy c-means (KFCM) algorithm (Wu *et al.*, 2003) to cluster textual data.
- The FCM, KFCM, G-G and MGG algorithms are not used in the cited above literature, precisely in the textual field.
- The proposed study treated the case of data without and with noise.

The remainder of this paper is organized as follows. In Section 2, brief reviews of the fuzzy clustering algorithms formulation are given. The sample selection is presented in Section 3. Simulation results are shown in Section 4, and Section 5 summarizes the important features of our approach.

2. Materials and methods

2.1 FCM clustering algorithm

The FCM clustering algorithm was proposed by Bezdek *et al.* (1984). It is the most popular fuzzy clustering algorithm. Given an unlabeled data set $S=[x_1, x_2, \ldots, x_N]$ composed of N observations, the FCM algorithm minimizes the following objective function (Wu *et al.*, 2003):

$$J(S; U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m d_{ik}^2; \;\; i = 1, \ldots, c \tag{1}$$

with:

$$d_{ik} = \left\| \boldsymbol{x}_k - \boldsymbol{V} \right\| \tag{2}$$

where d_{ik} is the Euclidean distance, $\boldsymbol{x}_k = [x_{k1}, \ldots, x_{kM}]^T \in \Re^M$ the input vector, M the dimension of input vector, c the number of clusters, m the weighting exponent and μ_{ik} the membership degree of each object belonging to the ith cluster. The membership values μ_{ik} have to satisfy the following conditions:

$$\mu_{ik} \in [0\ 1];\ i = 1, \ldots, c;\ \ k = 1, 2, \ldots, N \tag{3}$$

$$\sum_{i=1}^{c} \mu_{ik} = 1;\ \ k = 1, 2, \ldots, N \tag{4}$$

$$0 < \sum_{k=1}^{N} \mu_{ik} < N \tag{5}$$

Let the partial derivative of J(S; U, V) with respect to μ_{ik} and V_i equal to 0. Then, we obtain the following:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}}},\ i = 1, \ldots, c \tag{6}$$

$$\boldsymbol{V}_i = \frac{\sum_{k=1}^{N} \left(\mu_{ik} \right)^m \boldsymbol{x}_k}{\sum_{k=1}^{N} \left(\mu_{ik} \right)^m} \tag{7}$$

The FCM algorithm is summarized as follows (Nidhi, 2014).

Given data S, set $m > 1$, initialize $U^{(0)}$ and the initials centers $\boldsymbol{V}^{(0)}$ (e.g. random). Choose the number of clusters c and pick a termination threshold $\varepsilon > 0$.

Repeat for $l = 1, 2, \ldots$

Step 1. Calculate each clusters centers $\boldsymbol{V}_i^{(l)}$ via Equation (7).

Step 2. Compute the distance measures via Equation (2).

Step 3. Update the fuzzy partition matrix $U^{(l)}$ via Equation (6).

Until $\left\| U^{(l)} - U^{(l-1)} \right\| \leqslant \varepsilon$, then stop. Otherwise, set $l = l + 1$ and return to Step 1.

2.2 KFCM algorithm

The KFCM algorithm (Wu *et al.*, 2003) uses a kernel function for calculating the distance. To construct a kernel version of the FCM algorithm, we define a non-linear mapping function ϕ from the original input space to high-dimensional space in which the data are more clearly separable. The kernel distance with the mapping ϕ is defined as follows:

$$d_{ik} = \left\| \phi(\boldsymbol{x}_k) - \phi(\boldsymbol{V}_i) \right\| \tag{8}$$

with:

$$\|\phi(\boldsymbol{x}_k)-\phi(\boldsymbol{V}_i)\| = \sqrt{K(\boldsymbol{x}_k, \boldsymbol{x}_k)+K(\boldsymbol{V}_i, \boldsymbol{V}_i)-2K(\boldsymbol{x}_k, \boldsymbol{V}_i)} \tag{9}$$

where the kernel function K is chosen as Gaussian function as follows:

$$K(\boldsymbol{x}_k, \boldsymbol{V}_i) = \exp\left(\frac{\|\boldsymbol{x}_k, \boldsymbol{V}_i\|^2}{\sigma^2}\right) \tag{10}$$

leads to $K(\boldsymbol{x}_k, \boldsymbol{x}_k) = 1$ and $K(V_i, V_i) = 1$, then Equation (8) is rewritten as follows:

$$d_{ik} = \sqrt{2-2K(\boldsymbol{x}_k, \boldsymbol{V}_i)} \tag{11}$$

and then, the objective function Equation (1) is modified as follows:

$$J(S; U, \boldsymbol{V}) = 2\sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m(1-K(\boldsymbol{x}_k, \boldsymbol{V}_i)) \tag{12}$$

thus Equations (6) and (7) are rewritten as follows:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{1-K(\boldsymbol{x}_k, \boldsymbol{V}_i)}{1-K(\boldsymbol{x}_k, \boldsymbol{V}_i)}\right)^{\frac{1}{m-1}}} \tag{13}$$

$$\boldsymbol{V}_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^m K(\boldsymbol{x}_k, \boldsymbol{V}_i)\boldsymbol{x}_k}{\sum_{k=1}^{N}(\mu_{ik})^m K(\boldsymbol{x}_k, \boldsymbol{V}_i)} \tag{14}$$

The KFCM clustering algorithm is summarized as follows (Daniel and Pedrycz, 2010).

Given data S, set $m > 1$, initialize $\mathrm{U}^{(0)}$ and the initials centers $\mathbf{V}^{(0)}$ (e.g. random). Choose the number of clusters c and pick a termination threshold $\varepsilon > 0$.

Repeat for $l = 1, 2, \ldots$

Step 1. Calculate each clusters centers $\boldsymbol{V}_i^{(l)}$ via Equation (14).

Step 2. Compute the distance measures via Equation (11).

Step 3. Update the fuzzy partition matrix $U^{(l)}$ via Equation (13).

Until $\|U^{(l)}-U^{(l-1)}\| \leqslant \varepsilon$, then stop. Otherwise, set $l = l+1$ and return to Step 1.

2.3 G-G algorithm

The G-G algorithm is an extension of Gustafson-Kessel algorithm (Gustafson and Kessel, 1979) that takes the size and density of the clusters into account. It employs a distance norm based on the fuzzy maximum likelihood estimate, which is defined as follows:

$$d_{ik} = \frac{(2\pi)^{M+1/2}\sqrt{\det(A_i)}}{\alpha_i}\exp\left[\frac{1}{2}(\boldsymbol{x}_k-\boldsymbol{V}_i)^T A^{-1}(\boldsymbol{x}_k-\boldsymbol{V}_i)\right] \tag{15}$$

The objective function of G-G algorithm is as follows:

$$J_{GG}(S, U, \boldsymbol{V}) = \sum_{i=1}^{c}\sum_{k=1}^{N}(U_{ik})^m d_{ik}^2 \tag{16}$$

Minimizing the objective function respect to all parameters in Equation (16), the G-G fuzzy clustering algorithm can briefly be described as follows.

Given data S, set $m > 1$, initialize $U^{(0)}$ and the initials centers $\boldsymbol{V}^{(0)}$ (e.g. random). Choose the number of clusters c and pick a termination threshold $\varepsilon > 0$.

Repeat for $l = 1, 2, \ldots$

Step 1. Calculate each clusters centers $V_i^{(l)}$ via Equation (7).

Step 2. Compute the fuzzy covariance matrices:

$$A_i^{(l)} = \frac{\sum_{k=1}^{N}\left(\mu_{ik}^{(l-1)}\right)^m (\boldsymbol{x}_k - \boldsymbol{V}_i)(\boldsymbol{x}_k - \boldsymbol{V}_i)^T}{\sum_{k=1}^{N}\left(\mu_{ik}^{(l-1)}\right)^m} \tag{17}$$

Step 3. Compute the distance measures via Equation (15) with the prior probability $\alpha_i^{(l)}$:

$$\alpha_i^{(l)} = \frac{1}{N}\sum_{i=1}^{c}\mu_{ik} \tag{18}$$

Step 4. Update the fuzzy partition matrix using Equation (6).

Until $\|U^{(l)} - U^{(l-1)}\| \leqslant \varepsilon$, then stop. Otherwise, set $l = l + 1$ and return to Step 1.

2.4 MGG algorithm

The MGG clustering is an extension of G-G algorithm. This algorithm is proposed based on the expectation-maximization identification of Gaussian mixture of models in order to form an easily interpretable model that does not rely on transformed input variables. The MGG clustering algorithm is summarized in five steps.

Given data S, set $m > 1$, initialize $U^{(0)}$ and the initials centers $\boldsymbol{V}^{(0)}$ (e.g. random). Choose the number of clusters c and pick a termination threshold $\varepsilon > 0$.

Repeat for $l = 1, 2, \ldots$

Step 1. Calculate each clusters centers $V_i^{(l)}$ via Equation (7).

Step 2. Compute standard deviations of the Gaussian membership functions:

$$\sigma_{i,j}^{2(l)} = \frac{\sum_{k=1}^{N}\mu_{i,k}^{(l-1)}\left(\boldsymbol{x}_{j,k} - \boldsymbol{V}_{j,k}\right)^2}{\sum_{k=1}^{N}\mu_{i,k}^{(l-1)}} \tag{19}$$

Step 3. Compute the prior probability α_i via Equation (18).

Step 4. Compute the distance measure $D_{i,k}^2$:

$$\frac{1}{D_{i,k}^2} = \prod_{j=1}^{M} \frac{\alpha_i}{\sqrt{2\pi\sigma_{i,j}^2}} \exp\left(-\frac{1}{2}\frac{(\boldsymbol{x}_{j,k} - \boldsymbol{V}_{i,j})^2}{\sigma_{i,j}^2}\right) \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(\boldsymbol{x}_{j,k} - \boldsymbol{V}_{i,j})^T (\boldsymbol{x}_{j,k} - \boldsymbol{V}_{i,j})}{2\sigma_i^2}\right) \tag{20}$$

Step 5. Update the fuzzy partition matrix using Equation (6).
Until $\| U^{(l)} - U^{(l-1)} \| \leqslant \varepsilon$, then stop. Otherwise, set $l = l + 1$ and return to Step 1.

3. Sample selection

The textual data set comes from a survey about what has been termed the Tunisian Revolution in 2011. Data were collected via a questionnaire targeted to students or young professionals, men and women of different origins, all enrolled in schools, institutes and universities of Sousse and Monastir. These institutes include about 60 thousands young Tunisians. We focused on students because they were at the very beginning of the revolution process (social networks) and went on to being the major actors in it. This survey was conducted during the academic year 2012/2013, by direct face-to-face interviews. The duration of an interview is approximately about one hour. Sampling is up to now a very difficult challenge in the Tunisian context both for official or private companies. Particularly in our questionnaire, which includes open-ended questions, major difficulties are first of all the languages in use (Arabic or French, answers had to be written in French), along with the fact that people are globally not positive about participating in surveys, and most of all that at that time particularly there was defiance and suspicion in the context of the Tunisian post-revolution period. We tried to be as much as possible in a probability sampling frame: first a geographical cluster sample including heterogeneous respondents and within each region a stratified random sample including homogeneous respondents in the different grades of each university. Moreover, a few unplanned respondents were included when submitting the questionnaire, ones passing by the university campus, for example. We eventually collected about 600 students, and the size of the final subset is 541 people because open-ended answers in Arabic were not used. Note that the classical sampling procedure may not be relevant when collecting phrases and words instead of measuring a numerical variable in order to make inference on the target population.

Two open-ended questions are asked. The first question is in the form of comments related to the closed question "How proud are you about the Tunisian revolution?" and the second one is "What opinion do you have about the Tunisian economical situation" after the Tunisian revolution. To be noted that all questions and all answers have to be in French.

4. Results and discussion

In this section, we are going to examine the performances of the clustering algorithms mentioned above (FCM, G-G, MGG and KFCM) in order to select the most reliable one otherwise which realized the most accurate clustering.

For each method, the following formula is used to compute the accuracy rate:

$$accurate\ rate = \frac{TW}{N} \times 100 \tag{21}$$

where TW is the number of the observations correctly classified.

A pre-treatment of the collected data was carried out before the application of different algorithms. We simulated two experimental cases. For each case, we compare the clustering results based T-S fuzzy models with and without noise.

4.1 Case 1

The first input data set contains 492 observations (all current words in the comments repeated more than four times). The parameter settings of all algorithms are $c = 5$ and $m = 2.5$. Table I shows the various performance results obtained by different algorithms with and without noise. As can be seen in Table I, the MGG gives the best accurate rate of 95.7317 percent compared with that of other algorithms. It has also a slight sensibility against noise (92.0696 percent).

Otherwise, each word used in our text more than four times can be represented by the following form:

$$T_i = \{\alpha t_1, \beta t_2, \theta t_3, \delta t_4, \lambda t_5\} \tag{22}$$

where α, β, θ, δ and λ are the membership values for each word. We note that $\{T_1, \ldots, T_N\}$ are the words explained in terms of membership degrees of the keywords $\{x_1, \ldots, x_n\}$. According to the results obtained by MGG algorithm, the 492 words can be explained by the selected keywords. In Table II, we illustrate membership degrees between some words (T_1: *poverty* T_2: *dictatorship* T_3: *economy*) with the five selected keywords (x_1: *revolution*, x_2: *unemployment*, x_3: *freedom*, x_4: *high* and x_5: *bad*). Table II gives an example of membership degrees between three words and the five selected keywords. The same procedure is applied for the rest of words using these selected keywords.

Using Equation (22), the three words can be represented as follows:

$$\{T_1 = \{0.94x_2, 0.01x_3, 0.05x_5\} \tag{23}$$

$$\{T_2 = \{0.02x_1, 0.02x_2, 0.95x_4, 0.01x_5\} \tag{24}$$

$$\{T_3 = \{0.74x_2, 0.02x_3, 0.23x_5\} \tag{25}$$

From these equations, three important and evident economic relationships are validated. First, *Poverty* and *unemployment* are highly dependent. For a long time, the relationship between these two indicators has been proved by several techniques applied to several data. Indeed, high rate of unemployment leads to financial crisis. Also it influences the overall purchasing capacity of a nation. So, unemployment leads automatically to poverty. The second proved result is that the use of the adjective dictatorship is high and up to 95 percent compared to other membership functions.

	Without noise		Noisy data	
Algorithms	Accuracy rate (%)	Misclustering observations	Accuracy rate (%)	Misclustering observations
MGG	95.7317	21	92.0696	41
G-G	95.5285	22	90.9091	47
KFCM	95.5285	22	90.4472	47
FCM	92.8862	35	89.8374	50

Table I.
Performance results

In fact, our data are collected just after the Tunisian Revolution, so it is evident to conclude that before the revolution the term dictatorship is applied in an abusive manner. Finally, we remark on an extensive use of the term *economy* jointly with the term *unemployment* (74 percent). Certainly this relationship is proved with many ways for many years.

4.2 Case 2

In the second case, the input data set contains 668 words. The parameter settings of all algorithms are $c = 5$ and $m = 2.5$. Table III provides the performance of different T-S clustering algorithms. It gives an example of membership degrees between three words and the five selected keywords. The simulation results obtained show that the MGG algorithm outperforms that of other algorithms with and without noise.

Equally to the first case, we can validate some obvious relationships between economic factors according to the views of young Tunisian students which are ultimately realistic opinions. We also quote five keywords: x_1: *rate*, x_2: *increase*, x_3: *country*, x_4: *price* and x_5: *poverty* (Table IV).

As showed in Table IV, the three words can be represented as following:

$$\{T_1 = \{0.05x_1, 0.03x_3, 0.92x_4\} \tag{26}$$

$$\{T_2 = \{0.02x_1, 0.27x_2, 0.10x_3, 0.60x_5\} \tag{27}$$

$$\{T_3 = \{0.02x_1, 0.17x_2, 0.07x_3, 0.41x_4, 0.33x_5\} \tag{28}$$

Table II. Membership degrees between three words and the selected keywords

	x_1: *revolution*	x_2: *unemployment*	x_3: *freedom*	x_4: *high*	x_5: *bad*
T_1: *poverty*	0.00	0.94	0.01	0.00	0.05
T_2: *dictatorship*	0.02	0.02	0.00	0.95	0.01
T_3: *economy*	0.00	0.74	0.02	0.00	0.23

Table III. Performance results

	Without noise		Noisy data	
Algorithms	Accuracy rate (%)	Misclustering observations	Accuracy rate (%)	Misclustering observations
MGG	97.9042	14	94.3723	39
G-G	97.9042	14	94.2280	40
KFCM	88.0240	80	83.8323	108
FCM	84.8802	101	80.5389	130

Table IV. Membership degrees between three words and the selected keywords

	x_1: *rate*	x_2: *increase*	x_3: *country*	x_4: *price*	x_5: *poverty*
T_1: *economy*	0.05	0.00	0.03	0.92	0.00
T_2: *unemployment*	0.02	0.27	0.10	0.00	0.60
T_3: inflation	0.02	0.17	0.07	0.41	0.33

From Equation (26), the word *economy* is influenced by the word *prices* up to 92 percent. In fact, it is effectively evident that economy is indirectly influenced by the rise or the fall of prices. But what seems strange is that *economy* in not at all influenced by *poverty*. These results may be justified by the restricted vision of the respondents and also the simple phrases used in reply to the survey. Another result to mention from Equation (27) is the word *unemployment* depends mainly on the word *poverty* (60 percent) and up to 27 percent to the word *increase*. After all the revolutions that have taken place in the world, unemployment always increased. Finally, the word "inflation" is connected to the word *price* with 41 percent, 33 percent to the word *poverty* and 0.17 percent to the word *increase*.

5. Conclusion

In this paper, several clustering algorithms based on the T-S fuzzy model are used on textual clustering. These algorithms like FCM, G-G, MGG and KFCM are applied for textual data with and without noise. Moreover, two real data sets are used to test the performance of these algorithms. It has been found that the MGG significantly outperforms the other existing methods. In addition, we validate some obvious economical relationships characterized by their membership degrees. The originality comes from the fact that we validate these economical results based on textual data, even if they have not been written by experts in the linguistic fields. In addition, the results obtained in this study are easy and simple to interpret by analysts. Consequently, the proposed methodology can be extended to include more keywords and other databases and to develop an identification procedure using least squares methods and optimization techniques for future work.

References

Abonyi, J., Babuška, R. and Szeifert, F. (2002), "Modified Gath-Geva fuzzy clustering for identification of Takagi-Sugeno fuzzy models", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 32 No. 5, pp. 612-621.

Azzalini, A., Scarpa, B. and Gabriel, W. (2012), *Data Analysis and Data Mining: An Introduction*, Oxford University Press, New York, NY.

Bezdek, J.C., Ehrlich, R. and Full, W. (1984), "FCM: the fuzzy c-means clustering algorithm", *Computers and Geosciences*, Vol. 10 Nos 2-3, pp. 191-203.

Buk, Z., Kordik, P., Bruzek, J., Schmitt, A. and Snorek, M. (2012), "The age at death assessment in a multi-ethnic sample of pelvic bones using nature-inspired data mining methods", *Forensic Science International*, Vol. 220 Nos 1-3, pp. 294.e1-294.e9.

Cai, D., Xiaofei, H., Zhiwei, L. and Wei, Y.M. (2004), "Hierarchical clustering of WWW image search results using visual, textual and link information", *Proceedings of the 12th Annual ACM International Conference on Multimedia*, pp. 952-959.

Daniel, G. and Pedrycz, W. (2010), "Kernel-based fuzzy clustering and fuzzy clustering: a comparative experimental study", *Fuzzy Sets and Systems*, Vol. 161 No. 4, pp. 522-543.

Do Prado, H.A. and Ferneda, E. (2007), *Emerging Technologies of Text Mining: Techniques and Applications*, Information Science Reference.

Gath, I. and Geva, A.B. (1989), "Unsupervised optimal fuzzy clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 7 No. 7, pp. 773-781.

Gustafson, D.E. and Kessel, W.C. (1979), "Fuzzy clustering with a fuzzy covariance matrix", IEEE Conference on Decision and Control, San Diego, CA.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning – Data Mining, Inference and Prediction*, Springer, New York, NY.

Kosala, R. and Blockeel, H. (2000), "Web mining research: a survey", *ACM SIGKDD Explorations Newsletter*, Vol. 2 No. 1, pp. 1-15.

Lebart, L. and Salem, A. (1988), "Analyse Statistique des Données Textuelles, Questions ouvertes et lexicométrie", Dunod, Paris.

Manning, C. and Schutze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.

Nidhi, G. (2014), "A study of various fuzzy clustering algorithms", *International Journal of Engineering Research*, Vol. 3 No. 3, pp. 177-181.

Ngaia, E.W.T., Xiub, L. and Chaua, D.C.K. (2009), "Application of data mining techniques in customer relationship management: a literature review and classification", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 2592-2602.

Ngaia, E.W.T., Yong, H., Wong, Y.H., Chen, Y. and Sun, X. (2011), "The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature", *On Quantitative Methods for Detection of Financial Fraud*, Vol. 50 No. 3, pp. 559-569.

Rakhlin, A. and Caponnetto, A. (2007), *Stability of k-Means Clustering Advances, Neural Information Processing Systems*, Vol. 12, MIT Press, Cambridge, MA, pp. 216-222.

Song, Y., Chen, Y., Hongjie, B., Chih, J.L. and Chang, E.Y. (2011), "Parallel spectral clustering in distributed systems", *IEEE Transactions Pattern Analysis and Machine Intelligence*, Vol. 33 No. 3, pp. 568-586.

Takagi, T. and Sugeno, M. (1985), "Fuzzy identification of systems and its application to modeling and control", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 15 No. 1, pp. 116-132.

Wei, T., Yonghe, L., Huiyou, C., Qiang, Z. and Xianyu, B. (2015), "A semantic approach for text clustering using WordNet and lexical chains", *Expert Systems with Applications*, Vol. 42 No. 4, pp. 2264-2275.

Weiss, S.M., Indurkhya, N. and Zhang, T. (2010), *Fundamentals of Predictive Text Mining*, Chapter 4, Information Retrieval and Text Mining, Springer.

Wu, Z.D., Xie, W.X. and Yu, J.P. (2003), "Fuzzy c-means clustering algorithm based on kernel method", *Proceedings of the 5th International Conference on Computational Intelligence and Multimedia Application, IEEE Computer Society*, *Washington, DC*, pp. 49-54.

Yao, M., Dechang, P. and Xiangxiang, C. (2012), "Chinese text clustering algorithm based K-means", *Physics Procedia*, Vol. 33 No. 8, pp. 301-307.

Yinglong, M., Yao, W. and Beihong, J. (2014), "A three phase approach to document clustering based on topic significance degree", *Expert Systems with Applications*, Vol. 41 No. 18, pp. 8203-8210.

Yong, U.N. and Mooney, R.J. (2002), "Text mining with information extraction", *AAAI Technical Report*, Austin, pp. 63-67.

Further reading

Ashok, N.S. and Sahami, M. (2009), *Text Mining: Classification, Clustering, and Applications*, Chapman and Hall/CRC, New York, NY.

Lebart, L., Salem, A. and Berry, L. (1998), *Exploring Textual Data*, Springer Science & Business Media, Dordrecht.

About the authors

Rjiba Sadika was born in Tunisia. She obtained her Master's Degree in 2010 from the high Institute of Management of Sousse (ISGS). Her main research interest is treatment of textual data analysis with modern methods. Rjiba Sadika is the corresponding author and can be contacted at: rjibasadika@yahoo.fr

Moez Soltani was born in 1980, Tunisia. He is an Assistant Professor at the Institut Supérieur des sciences appliquées et de technologie de Mateur (ISSATM). He obtained his PhD Degree in 2013 from the National High School of Engineers of Tunis (ENSIT), Tunisia. He obtained his BSc and MSc Degrees in electrical engineering in 2004 and 2006, respectively, from the same school. His main research interests are fuzzy logic and its application in the identification and control of non-linear systems.

Saloua Benammou was born in Tunisia. She is a Mature Student of the ENSAE. She obtained her PhD Degree from the Dauphine University of Paris. She is a Professor and also a Doyen at the University of Economics and Management of Sousse (FSEGS). Her main interest is data analysis and its application in economical and financial contexts.