Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 20, No. 2, June 2020, pp. 156-167
http://doi.org/10.5391/IJFIS.2020.20.2.156

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

Automatic Determination of the Number of Clusters for Semi-Supervised Relational Fuzzy Clustering

Norah Ibrahim Fantoukh , Mohamed Maher Ben Ismail , and Ouiem Bchir
Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

ljfis

Received: Feb. 16, 2020
Revised : May 10, 2020
Accepted: May 26, 2020

Correspondence to:
Mohamed Maher Ben Ismail and Ouiem Bchir
(maher.benismail@gmail.com, ouiem.bchir@gmail.com)

Abstract

Semi-supervised clustering relies on both labeled and unlabeled data to steer the clustering process towards optimal categorization and escape from local minima. In this paper, we propose a novel fuzzy relational semi-supervised clustering algorithm based on an adaptive local distance measure (SSRF-CA). The proposed clustering algorithm utilizes side-information and formulates it as a set of constraints to supervise the learning task. These constraints are expressed using reward and penalty terms, which are integrated into a novel objective function. In particular, we formulate the clustering task as an optimization problem through the minimization of the proposed objective function. Solving this optimization problem provides the optimal values of different objective function parameters and yields the proposed semi-supervised clustering algorithm. Along with its ability to perform data clustering and learn the underlying dissimilarity measure between the data instances, our algorithm determines the optimal number of clusters in an unsupervised manner. Moreover, the proposed SSRF-CA is designed to handle relational data. This makes it appropriate for applications where only pairwise similarity (or dissimilarity) information between data instances is available. In this paper, we proved the ability of the proposed algorithm to learn the appropriate local distance measures and the optimal number of clusters while partitioning the data using various synthetic and real-world benchmark datasets that contain varying numbers of clusters with diverse shapes. The experimental results revealed that the proposed SSRF-CA accomplished the best performance among other state-of-the-art algorithms and confirmed the outperformance of our clustering approach.

Keywords: Semi-supervised clustering, Relational data, Fuzzy clustering, Local distance measure learning, Optimal number of clusters

1. Introduction

Clustering is one of the most popular unsupervised learning techniques that are commonly used in data mining and pattern recognition fields [1, 2]. The resulting categories include sets of homogeneous patterns [1]. Accordingly, the distances between the data instances that belong to the same cluster exhibit high similarity to each other compared to those from other clusters. Clustering can be perceived as a data modeling technique that yields concise data summarization. Recently, clustering approaches have gained attention because they play a key

role in a broad range of applications. In fact, it has become relevant to various contexts and disciplines including artificial intelligence, pattern recognition, information retrieval, image analysis, and bioinformatics [3].

Clustering is typically an NP-hard problem where the unsupervised learning task uses no prior information on the clusters/categories or on the relationship between the data points [4]. Furthermore, the same dataset may require different categorizations based on its application purposes. The prior information can be integrated to steer the clustering process and yield significant performance improvement. Clustering algorithms that employ both unlabeled and labeled data as prior knowledge are called semi-supervised clustering algorithms. Recently, researchers from data mining and pattern recognition fields have shown considerable interest in semi-supervised clustering [5, 6].

Another limitation of typical clustering algorithm is their inability to handle complex data using cluster prototypes. In fact, conventional clustering approaches that rely on cluster prototypes cannot be used with all datasets. In particular, relational data, which is commonly encountered in fields where the representation of individual data instances using low-level features is not possible, cannot be handled by such prototype-based clustering algorithms. A relational clustering technique [7] should be employed when the data is purely relational and only pairwise relations (similarities or dissimilarities) between pairs of points are known. However, relational clustering has received considerably lesser attention than prototype-based clustering approaches. This can be attributed, in part, to the fact that most engineers and mathematicians often deal with object data, and infrequently deal with purely relational data.

Moreover, hard/crisp clustering yields distinct clusters where each data instance belongs exclusively to one cluster. However, real datasets usually include overlapping clusters. Consequently, it is more convenient to use fuzzy clustering [8, 9] that allows data points to simultaneously belong to multiple clusters with different degrees of membership.

Another major issue is the difficulty in specifying the optimal number of clusters. Various clustering techniques determine this parameter by repeating the clustering process using different number of clusters and then selecting the value that yields the best cluster validity measure [10]. Such an approach is impractical for large datasets owing to its expensive computational cost. Despite the considerable effort made by the research community to introduce more advanced and accurate clustering algorithms [3], determining the optimal number of clusters in an automatic manner remains a challenging problem that affects the performance of most state-of-the-art techniques.

The main contribution of this work is the introduction of a novel semi-supervised fuzzy relational clustering algorithm, based on local distance learning that automatically determines the optimal number of clusters. In addition, the supervision information used in this approach steers the clustering process towards the optimal partition and escape from the local minima. Additionally, the proposed method is formulated to handle relational data. This makes it useful for real applications where only pairwise similarities (or dissimilarities) between data instances are available.

2. Related Works

We dedicate this section to survey the most related works to our proposed approach. We partition this survey into two main categories based on the technical foundations of the published works. The first part of this section focuses on the different studies relevant to semi-supervised fuzzy clustering with measure learning. The approaches that form the second part enclose several methods that can be adopted to obtain the optimal number of clusters for a given dataset.

2.1 Semi-Supervised Clustering with Measure Learning

The research in [11] introduced the metric learning and pairwise constraints K-means (MPCK-means) algorithm, which is a semi-supervised clustering algorithm that unifies constraint-based and measure-based techniques. It is an extension of the well-known K-means algorithm [1] that employs both labeled and unlabeled data to guide the clustering and learn the distance measure simultaneously. However, this approach allows only linear mapping for the input data. Therefore, this clustering method cannot separate clusters that are nonlinearly separable.

In [12], an adaptive semi-supervised clustering kernel method (SCKMM) has been proposed. This study has proposed a kernel semi-supervised clustering algorithm, which uses an adaptive version of K-means that incorporates the pairwise constraints along with measure learning into a nonlinear framework. Although the experimental results on various datasets proved the effectiveness of the SCKMM, it remains practical with small datasets only owing to the computational complexity of solving the measure matrix.

In [13], a relational fuzzy semi-supervised clustering method with local distance measure learning (SURF-LDML) has been introduced. The supervisory information is integrated to guide the clustering process towards target categorization and to avoid

the local minima. The side-information is also exploited to learn the underlying distance measure while categorizing the data. Moreover, SURF-LDML is formulated to work with relational data. However, SURF-LDML may not be suitable when the distribution of the different clusters in the input space exhibits large variations.

More recently, a novel clustering algorithm named fuzzy clustering with learnable cluster-dependent kernels (FLeCK) has been proposed [14]. The proposed algorithm learns the local Gaussian kernel parameters while categorizing the data. FLeCK is also designed to work with relational data. However, all the above state-of-the-art approaches assume that the number of clusters is given before starting the clustering process based on some experience or domain knowledge, which is considered a significant limitation that affects them in real-world clustering applications.

2.2 Approaches to Determine the Number of Clusters

The competitive agglomeration (CA) [15] is an approach that can learn the number of clusters in an unsupervised way. The CA algorithm is an iterative process where the number of clusters decreases gradually with the number of iterations. It starts the clustering process with the maximum number of clusters and the final categorization represents the optimal number of clusters. The core challenge related to the CA algorithm is that it is prototype-based, and thus, not suitable for applications where the data points are not expressed using feature vectors. The possible alternative to overcome this drawback is to use relational data, which consists of the pairwise relations between each pair of points.

The CA for relational data (CARD) clustering algorithm [16], is an extended version of the fuzzy CA [15] that can deal effectively with complex, non-Euclidean relational data. The CA and CARD [16] algorithms can learn the optimal number of clusters in an unsupervised manner. However, unsupervised clustering is an optimization problem that is prone to many local minima [17]. A possible alternative to overcome this drawback can be accomplished by including pairwise constraints along with the unlabeled data to steer the learning task towards the optimal categorization and escape from local minima.

The semi-supervised fuzzy clustering algorithm with pairwise constraints (SCAPC) [18] was introduced to improve the typical semi-supervised fuzzy clustering with pairwise constraints. A novel penalty term was incorporated to the CA [15] objective function to discover accurately the optimal partition of the data by moderately changing the disagreement on the magnitude order between the penalty term and original objective function. Although the side information in the SCAPC approach is used to bias the clustering algorithm towards an optimal partitioning, it does not learn and adapt the underlying distance measure. Moreover, the performance of the clustering is significantly sensitive to the choice of the distance metric [19].

To overcome the above limitations, we introduce a novel semi-supervised relational fuzzy clustering algorithm with measure learning based on CA (SSRF-CA). The proposed algorithm is an extension of the SURF-LDML introduced in [13]. More precisely, we intend to exploit the CA [15] approach to learn the number of clusters while categorizing the data. The proposed SSRF-CA uses side information to penalize or reward the objective function to learn a better partition. Moreover, it learns the dependent dissimilarity measure with respect to each cluster.

3. Proposed Approach

Given the dataset $X = \{x_i \mid x_i \in \mathbb{R}^d, i = 1, 2, \ldots, N\}$, where N represents the number of data points. Let ML be the "Must-Link" matrix, where pair of points such as $ML(x_j, x_k) = 1$ implies that x_j and x_k must belong to the same cluster and 0 otherwise. Likewise, let CL be the "Cannot-Link" matrix, where pair of points such as $CL(x_j, x_k) = 1$ implies that x_j and x_k must not belong to the same cluster and 0 otherwise. The SSRF-CA minimizes the following objective function:

$$\begin{aligned} \mathrm{J} = & \sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N} u_{ij}^m u_{ik}^m d_{jki} \\ & + \alpha_1 \sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N} (M - u_{ij}^m u_{ik}^m) ML(j,k) d_{jki} \\ & - \alpha_2 \sum_{i=1}^{C}\sum_{j=1}^{N}\sum_{k=1}^{N} (M - u_{ij}^m u_{ik}^m) CL(j,k) d_{jki} \\ & - \sum_{i=1}^{N} \log(det(A_i)) \\ & - \alpha_3 \sum_{i=1}^{C} \left[\sum_{j=1}^{N} u_{ij} \right]^2, \end{aligned} \tag{1}$$

subject to

$$0 \leq u_{ij} \leq 1 \ \text{ and } \ \sum_{j=1}^{N} u_{ij} = 1, \ \ i, j \in \{1, \ldots N\}. \tag{2}$$

In Eq. (1), C denotes the number of clusters, in which ($2 \leq C \leq N$), m$\in (1, \infty)$ is a parameter that determines the level of cluster fuzziness, u_{ij} represents the degree of membership of x_j to the ith cluster, and M is a constant $\in (0, 1)$. α_1, α_2, and α_3 are three constants.

The distance d_{jki} to be learned between two points x_j and x_k, in terms of the ith cluster is defined as

$$d_{jki} = (x_j - x_k)A_i(x_j - x_k)^t, \tag{3}$$

where A_i is a $d \times d$ matrix that is learned for each cluster. It permits clusters to lie in different subspaces and have different shapes as it adapts itself to the shape of each cluster.

SSRF-CA depends on optimizing a combined objective function with five terms (as illustrated in Eq.(1)). The first term finds compact clusters using fuzzy relational clustering [9]. The second term represents the penalty term for violating the ML constraints. It is formulated in such a way that the penalty among distant ML points is higher than the nearby points. This penalty term is weighted by the fuzzy memberships assigned to the data points. If the distance among two ML points is high, then the learned distance measure for this cluster is grossly not suitable, and the penalty should be increased accordingly.

Analogously, the third term is the reward of satisfying a CL constraint. It is formulated in such a way that the reward among distant CL points is higher than the nearby points. This reward term is also weighted by the fuzzy memberships. If the distance among two CL points is high, then the learned distance measure for this cluster is suitable, and the reward should be increased to permit the modification.

The fourth term of the objective function in Eq. (1) is incorporated to adapt the underlying distribution of the data points and learn the proper size of each cluster as the determinant of the distance matrix, A_i, is proportional to the volume of the cluster.

The last term in Eq. (1) is the bias term that can be defined as the squared sum of the cardinalities of the clusters. The neighboring clusters compete for data points, and the clusters that lose the competition are progressively discarded. In fact, the initial iteration has a number of clusters that is greater than the expected value. As the clustering algorithm progresses, the algorithm learns the optimal number of clusters by removing and combining the clusters.

The weights α_1 and α_2 provide a way of identifying the relative importance of the ML and CL constraints compared to the sum of inter-cluster distances, and the weight α_3 provides a way of identifying the relative importance of the CA term compared to the sum of inter-cluster distances.

The proposed approach can handle both data matrices of the pairwise distance or pairwise dissimilarity. The method that we apply in order to achieve this goal is inspired by Hathaway et al. [20] where the relational version of the popular fuzzy C-mean (FCM) was introduced [9]. It has been proved in [20] that the squared Euclidean distance $dist^2_{ik} = \|x_k - c_i\|^2$, between feature vector x_k and the center of the ith cluster, c_i, can be expressed with respect to the relational matrix R as

$$dist^2_{ik} = (Rv_i)_k - \frac{v_i^t R v_i}{2}, \tag{4}$$

where v_i is the normalized membership vector of all N samples in cluster i defined as

$$v_i = \frac{(u^m_{i1}, \ldots, u^m_{iN})^t}{\sum_{j=1}^{N} u^m_{ij}}. \tag{5}$$

In Eq. (4), let R_{A_i} be the learning measure dissimilarity matrix. It can be written as

$$R_{A_i}(j,k) = (1 - \alpha_1 ML(j,k) + \alpha_2 CL(j,k))d_{jki}. \tag{6}$$

The aim of the proposed SSRF-CA algorithm is to learn the fuzzy membership values u_{ij} of each data point x_i in cluster i, the measure matrix A_i, and the number of clusters. To optimize the objective function in Eq. (1) with respect to u_{ij}subject to Eq. (2), we use the Lagrange multiplier technique [21] and obtain the following update equation:

$$u_{ij} = u_{ij}{}^{\mathrm{SURF}} + u_{ij}{}^{\mathrm{Bias}}, \tag{7}$$

where

$$\mathrm{u}_{ij}^{\mathrm{SURF}} = \frac{1}{\sum_{t=1}^{C} \left(\mathrm{dist}^2_{\mathrm{ij}}/\mathrm{dist}^2_{\mathrm{tj}}\right)^{\frac{1}{(m-1)}}}, \tag{8}$$

and

$$u_{ij}^{\mathrm{Bias}} = \left(\frac{2\alpha_3}{m\ \mathrm{dist}^2_{\mathrm{ij}}}\right)^{\frac{1}{(m-1)}} (N_i - \overline{N}_j)^{\frac{1}{(m-1)}}. \tag{9}$$

The first term in Eq. (7), u_{ij}^{SURF}, is the membership term in the SURF-LDML [13]. The second term in Eq. (7), u_{ij}^{Bias}, is a bias term which mainly relies on the difference among the cardinality of the cluster and the weighted average of cardinalities of data

point x_j. If the cardinality of the cluster is greater than the weighted average, the bias term will be positive; hence, the overall membership value u_{ij} will increase. In contrast, if the cardinality of the cluster is less than the weighted average, the bias term will be negative, thus the overall membership value u_{ij} will decrease. In Eq. (9), N_i is the cardinality of cluster i and can be calculated as follows:

$$N_i = \sum_{j=1}^{N} u_{ij}, \tag{10}$$

and $\overline{N}_j$ is basically a weighted average of the cluster cardinalities. $\overline{N}_j$ is expressed as

$$\overline{N}_j = \frac{\sum_{k=1}^{C} \left(1/\mathrm{dist}_{\mathrm{jk}}^2\right)^{\frac{1}{(m-1)}} N_k}{\sum_{k=1}^{C} \left(1/\mathrm{dist}_{\mathrm{jk}}^2\right)^{\frac{1}{(m-1)}}}. \tag{11}$$

Furthermore, to optimize the objective functions in Eq. (1) with respect to A_i, we apply the Lagrange multiplier technique [21] and obtain the following optimal update equation of the matrix A_i:

$$A_i = \left(\sum_{j=1}^{N} \sum_{k=1}^{N} \mathrm{H}_{jki} (x_i - x_j)(x_i - x_j)^t \right)^{-1}, \tag{12}$$

where

$$\begin{aligned} \mathrm{H}_{jki} = &\left(u_{ij}^m u_{ik}^m\right) + \alpha_1 \left(\mathrm{M} - u_{ij}^m u_{ik}^m\right) ML(j,k) \\ &- \alpha_2 (\mathrm{M} - u_{ij}^m u_{ik}^m) CL(j,k). \end{aligned} \tag{13}$$

3.1 Updating the Number of Clusters

SSRF-CA is initialized with a large number of clusters C_{max} that is considerably higher than the optimal value. As the algorithm progresses, the neighboring clusters compete for data points, and the spurious clusters (low cardinality) that lose the competition are progressively discarded. After eliminating and merging the clusters, the algorithm converges towards the optimal number of clusters.

In Eq. (7), the second term, u_{ij}^{Bias}, is a bias term which mainly relies on the difference among the cardinality of the cluster and the weighted average of cardinalities of data point x_j. If the cardinality of the cluster is greater than the weighted average, the bias term will be positive; hence, the overall membership value u_{ij} will increase. In contrast, if the cardinality of the cluster is less than the weighted average, the bias term will be negative, thus the overall membership value u_{ij} will decrease. In fact, the membership value u_{ij} in spurious (low cardinality) clusters are reduced when their distances to such clusters are low, which provides a gradual reduction of the cardinality of the spurious clusters. When the cardinality of a cluster declines below a weighted average, we remove the cluster, and thus update the number of clusters.

It is important to note that when a data point x_i is near to only a cluster i and far away from the other clusters, we obtain

$$N_i \approx \overline{N}_j, \quad \text{or} \quad {u_{ij}}^{\mathrm{Bias}} \approx 0.$$

To put it simply, if a data point x_i is close to only one cluster, it will not be involved in any competition. In contrast, if a data point x_i is close to several clusters, these clusters will compete for this data point according to their cardinality.

3.2 Updating Trade-Off Parameter α_3

The value of α_3 can be reduced gradually in each iteration to help the proposed algorithm to search for compact partitions with the learned number of clusters that are close to the "optimal" in the first few iterations. The value of α_3 can be calculated using the following equation:

$$\alpha_3(k) = \eta(k) \cdot \frac{\sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij})^2 d^2{}_{jki}}{\sum_{i=1}^{C} \left[\sum_{j=1}^{N} u_{ij} \right]^2}. \tag{14}$$

In Eq. (14), α and η are functions of the iteration number k. An appropriate choice for η is the exponential decay defined by

$$\eta(k) = \eta_0 \cdot e^{-k/\tau}, \tag{15}$$

where η_0 is an initial value, k is the current number of iterations, τ is the time constant.

3.3 The SSRF-CA Algorithm

The SSRF-CA algorithm is summarized as Algorithm 1.

3.4 Time Complexity

The computational complexity of the proposed SSRF-CA is $O((N + d^3) * C)$, where N, d, C denote the number of data points, dimensions, and clusters. respectively. The cubic computational complexity with the number of dimensions is due to

Algorithm 1. SSRF-CA algorithm.

Fix the maximum number of clusters $C = C_{max}$ and m $\in [1, \infty)$;
Initialize the fuzzy partition matrix $U^{(CL)}$ randomly;
Set α_1, α_2, and α_3;
Initialize the matrices A_i to the identity matrix;
Initialize the threshold ϵ_1;
Compute the initial cardinalities N_i for $1 \leq i \leq C$ by using Eq. (10);
REPEAT
Compute the dissimilarity R_{A_i} for all clusters using Eq. (6);
Compute the membership vectors v_i using Eq. (5);
For each point x_k, $k \in 1, 2, \ldots, N$
Compute the distance $dist^2_{ik}$ using Eq. (4);
Update $\alpha_3(k)$ using Eq. (14) and (15);
Update the fuzzy membership using Eq. (7);
Update the measure distance matrices $A_i$using Eq. (12);
Compute the cardinalities of all clusters N_i by using equation Eq. (10);
If $(N_i < \epsilon_1)$ remove cluster i and update the number of clusters C;
UNTIL (The max number of iterations is reached or the fuzzy memberships stabilize)

the computation of the inverse of each matrix A_i as it requires computing the determinants and Eigen decomposition of the matrix A_i with respect to each cluster i in each iteration.

4. Experimental Results

In the following, we first describe the considered datasets and experimental settings. Then, we introduce the performance metrics used to assess and analyze the obtained results. Finally, we prove the success of the proposed SSRF-CA and compare its performance to the relevant clustering algorithms below:

- The CA [15]: The prototype-based clustering approach, outlined in Section 2.2, can learn the optimal number of clusters in an unsupervised way. The pairwise constraints, ML and CL, are ignored in this algorithm.
- The SCAPC [18]: A semi-supervised version of the CA algorithm, depicted in Section 2.2, uses the pairwise constraints, ML and CL, to provide partial supervision.

4.1 Datasets

We compare the proposed SSRF-CA with relevant clustering approaches using different synthetic and real-world benchmark datasets provided by the UCI machine-learning repository.

As the proposed algorithm is designed as a semi-supervised clustering technique, we randomly keep 10% of the labeled data points (called seed points) for guiding the learning process. In particular, the pairs of seed points residing in the same cluster compose the ML subset. Likewise, the pairs of seed points belonging to different clusters compose the CL subset. Tables 1 and 2 summarize the synthetic and real datasets, respectively.

4.2 Experimental Settings

The performance of the clustering algorithm is significantly sensitive to the choice of the parameter m, which determines the level of cluster fuzziness. If $m = 1$, the memberships u_{ij} converge to 0 or 1, which implies a hard/crisp clustering, where each data point belongs to exactly one cluster. Alternatively, higher values of m will blur the partitions. Consequently, the data points tend to belong to all clusters with the same degree of membership, which is not preferable. It is generally suggested that the parameters m are between 1.5 and 2 [22]. In our experiments, the fuzziness index m is set to 2, and the initial fuzzy partition matrix U is generated randomly. The maximum number of clusters C_{max} is set to 30, and the time constant τ is chosen to be 3. The termination condition of the clustering process is reached when the maximum number of iterations is equal to 100, or when the difference between the fuzzy memberships of two successive iterations is less than 0.002.

For the proposed SSRF-CA, the two regularization parameters, α_1 and α_2, denote the importance of the ML and CL constraints, respectively, compared to the sum of inter-cluster distances. Similarly, the parameter α_3 reflects the importance of the bias term. We tune the parameters α_1, and α_2, and then we select the values that yields the partition with the highest performance. For the third regularization parameter, α_3, the value is initialized to be 1 and it updates gradually in each iteration by using the Eq. (14) mentioned in Section 3.2.

It is important to note that because of the bias term, the membership values, u_{ij}, may not be within the range between 0 and 1. In this situation, it is possible to change the negative values of u_{ij} to zero to indicate that the data points are not assigned to the cluster i. Similarly, it is feasible to change the values of u_{ij} that are larger than 1 to 1 to show that the data points are certainly assigned to the cluster i.

Table 1. Four synthetic datasets used in our experiments

	# of data points	# of clusters	Cluster sizes	# seeds per cluster	Balanced/unbalanced
Dataset 1	87	2	43,44	5,5	Unbalanced
Dataset 2	238	3	100.56,82	10,6,9	Unbalanced
Dataset 3	185	4	51,42,53,39	6,5,6,4	Unbalanced
Dataset 4	250	5	50,50,50,50,50	5,5,5,5,5	Balanced

Table 2. Four real datasets used in our experiments

	# of data points	# of data attributes	# of clusters	Cluster sizes	# seeds per cluster	Balanced/unbalanced
Bankruptcy	250	7	2	142,107	15,11	Unbalanced
Seeds	140	7	3	70,70	7,7	Balanced
Iris	150	4	3	50,50,50	5,5,5	Balanced
Wi-Fi localization	470	7	4	101,108,116,145	11,11,12,15	Unbalanced

4.3 Performance Measures

To evaluate the performance of the proposed SSRF-CA and compare it with the considered relevant clustering algorithms, we assume that the true labels are given. Then, we calculated the clustering accuracy, Rand index, Jaccard coefficient, Fowlkes-Mallows [23], and the learned number of clusters to reflect the overall performance of each algorithm.

4.4 Evaluation of the Learned Distance Measure

In this experiment, we illustrate how the new mapping enhances the clustering performance even when we include the new bias term in the objective function to learn the optimal number of clusters. We have chosen dataset 2 from the synthetic data to illustrate the learned distance measure because it includes clusters of diverse number of instances, shapes, and orientations.

In Figure 1, the small circles are the data points and each cluster is illustrated using a different color. The black color represents the 10% seed points used for supervision. For the visualization and interpretation of the learned measures, we used the fact that learning the underlying distance measure is equal to finding a rescaling that substitutes each input data assigned to cluster with $x \rightarrow \boldsymbol{x}\boldsymbol{A}_{\boldsymbol{i}}^{\frac{1}{2}}$, and then using the Euclidean distance measure to the rescaled data. If the rescaling works successfully, similar points move closer to each other while keeping different ones apart.

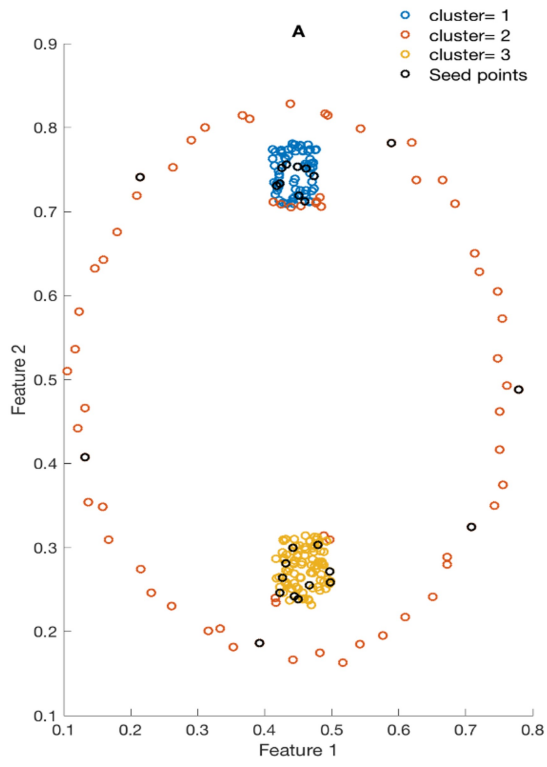
In Figure 1(a), the data is plotted in the original feature space, which means that the A_i is fixed to be equal to the identity matrix ($A_i = I$). We can see from Figure 1(a) that the three clusters are merged together. On the contrary, we can observe from Figure 1(b) that the new mapping separates the three clusters well such that they are not merged anymore.

4.5 Learning the Optimal Number of Clusters

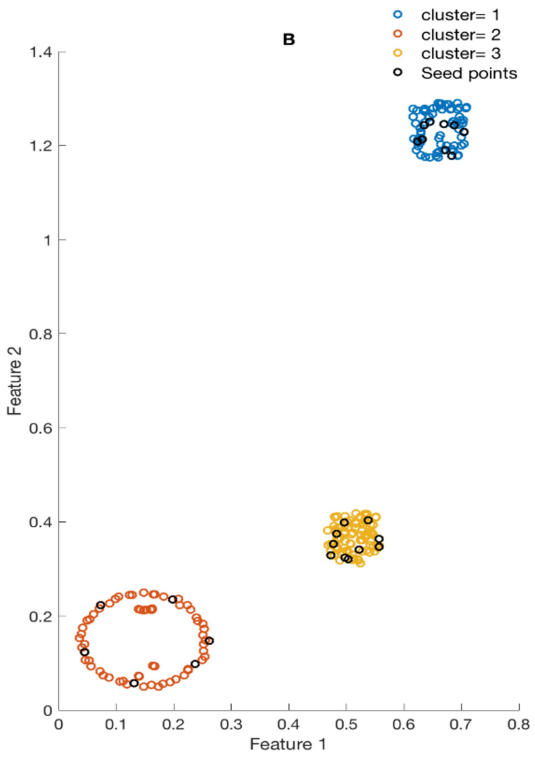
This experiment focuses on how the proposed SSRF-CA is able to learn the optimal number of clusters that reflects the correct partition of the data. Specifically, we compare the results obtained using our proposed approach with those obtained using the state-of-the-art clustering approaches for the same synthetic and real datasets introduced in Section 4.1.

Figures 2-5 show the clustering results for the four synthetic datasets using three different algorithms: CA [15], SCAPC [18], and the proposed SSRF-CA. We observe from Figure 2(a) that dataset 1 is easy to partition as the clusters are well separated. It contains two Gaussian clusters where each cluster has one low variance and one high variance feature. Figure 2(a) indicates that all the clustering algorithms perform well and also they learn the optimal clustering number correctly.

The geometric characteristics of dataset 2 renders it slightly difficult to categorize and learn the optimum number of clusters (as indicated in Figure 3(a)). The CA algorithm [15] was not able to partition this data and has brought the algorithm toward a local minimum. The CA algorithm has merged all the dataset in one cluster, as illustrated in Figure 3(b) because it cannot identify the circular-shaped clusters. Furthermore, integrating partial supervision in the SCAPC algorithm [18] did not provide



(a)



(b)

Figure 1. Representation of the learned distance measure for dataset 2. (a) Data is plotted on the original feature space. (b) rescaled data is plotted using $x \rightarrow xA_i^{\frac{1}{2}}$.

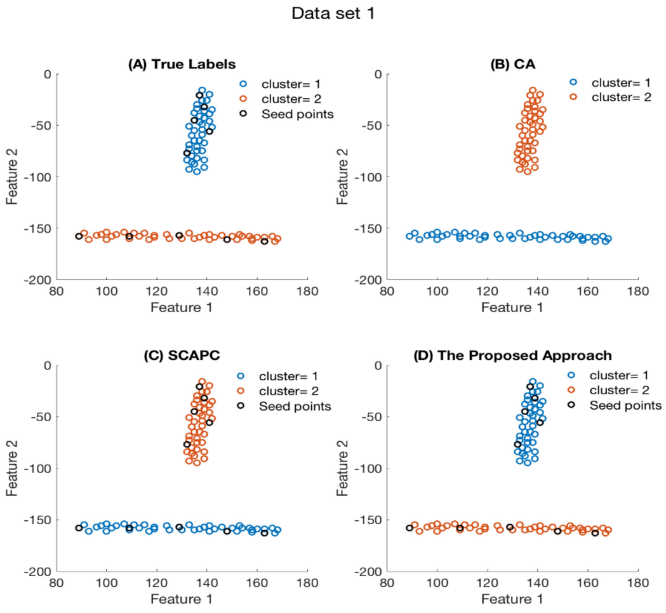


Figure 2. Results of clustering dataset 1 using three different algorithms: (a) true labels, (b) CA, (c) SCAPC, and (d) the proposed approach.

significant help. The proposed SSRF-CA, on the other hand, overtakes all the other algorithms in partitioning the data and learning the optimal number of clusters despite the complexity of the geometry of this data (see Figure 3(d)). The proposed SSRF-CA uses the pairwise constraints information effectively to learn a cluster dependent distance measure and learn the optimal number of clusters.

For Dataset 3, defining the optimal number of clusters in this dataset is challenging as it contains four clusters with different degrees of overlapping, and various sizes and shapes that are adjacent to each other (as indicated in Figure 4(a)). We observe from Figure 4 that neither CA nor SCAPC has learned the number of clusters correctly, which is reflected from the categorization of the data. Figure 4 illustrates that SCAPC is close to CA because there is no distinct function for semi-supervised learning in SCAPC. This is mainly owing to the fact that SCAPC is using a global parameter for both ML and CL terms that makes the algorithm less effective in manipulating this data geometry. Furthermore, the proposed algorithm has failed in categorizing some data points that are at the boundaries of the clusters; however, it has learned the exact number of clusters and classified most of this dataset successfully. In fact, despite the complexity of the structure of this dataset, only the proposed SSRF-CA has relatively achieved a reasonable partition.

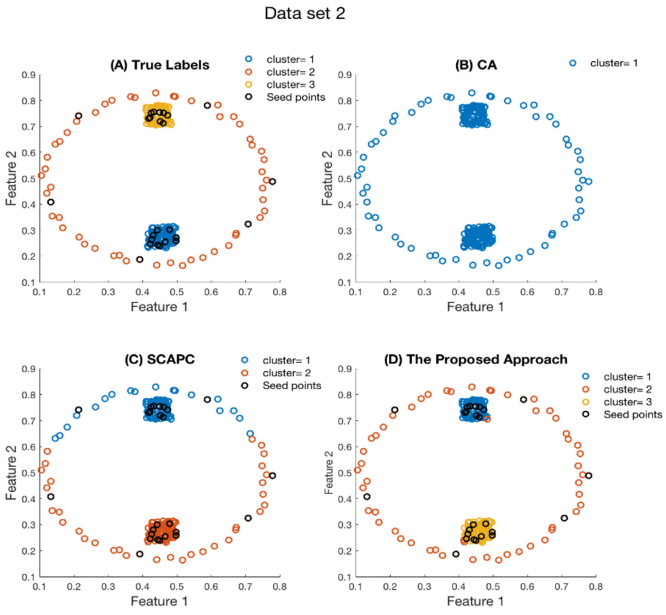


Figure 3. . Results of clustering dataset 2 using three different algorithms: (a) true labels, (b) CA, (c) SCAPC, and (d) the proposed approach.

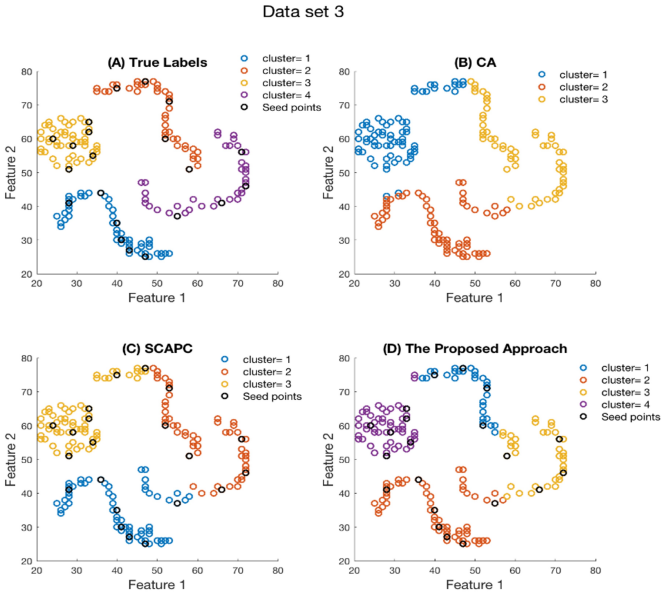


Figure 4. Results of clustering dataset 3 using three different algorithms: (a) true labels, (b) CA, (c) SCAPC, and (d) the proposed approach.

The fourth synthetic dataset contains five Gaussians distributed clusters that have similar shapes, densities, and balanced sizes that can be categorized by our proposed approach (see Figure 5(d)). The proposed algorithm can easily learn the

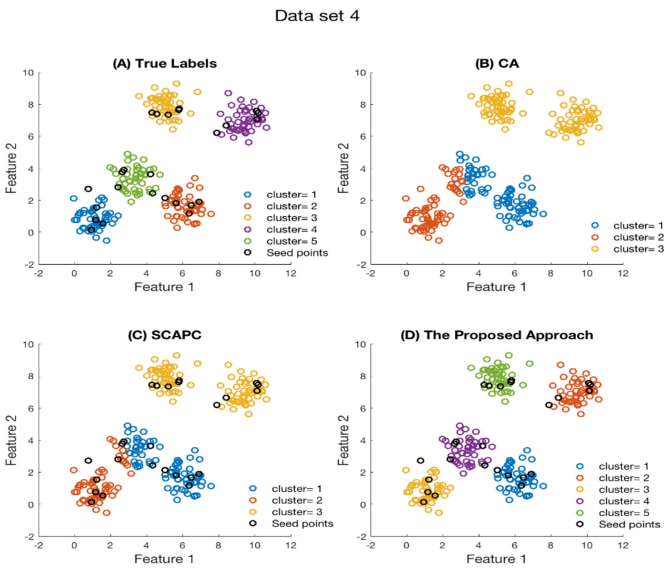


Figure 5. Results of clustering dataset 4 using three different algorithms: (a) true labels, (b) CA, (c) SCAPC, and (d) the proposed approach.

optimal number of clusters while the other algorithms, CA and SCAPC, were not able to learn the exact clustering numbers and thus cannot separate the five clusters successfully. These two algorithms were prone to several local minima owing to the fact that the CA and SCAPC algorithms defined the distance measure as a priori. In fact, the performance of the clustering algorithms relies critically on the choice of the distance measure. However, the proposed SSRF-CA learned the measure by the supervision to escape from local minima and reflect the target categorization correctly.

We can conclude from Figures 2-5 that the proposed SSRF-CA outperforms all the other algorithms in categorizing the data points and learning the optimal number of clusters regardless of the complexity of the geometry of this data. Especially in the second and the third synthetic datasets that could not be handled satisfactorily by the other algorithms. This is because the proposed algorithm uses the pairwise constraints information effectively to learn a cluster dependent distance measure and learn the optimal number of clusters.

4.6 Comparison of the Clustering Performance

This experiment demonstrates how our method improves the clustering performance. A summarization of the clustering results for both the synthetic and real datasets is represented in the following subsections. Figures 6 and 7 display the clustering

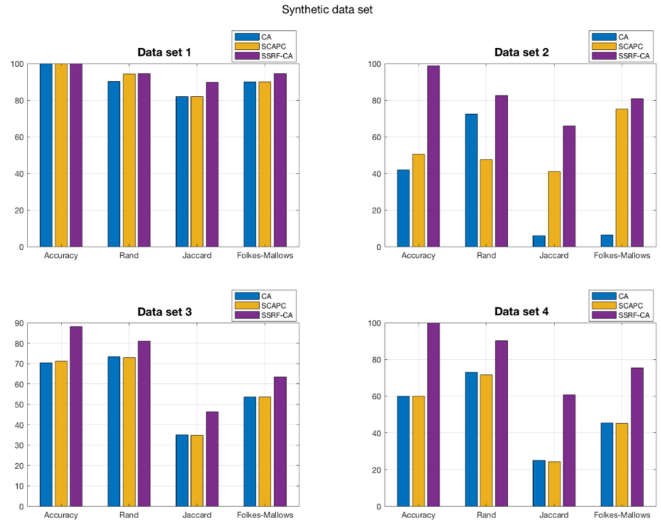


Figure 6. Performance measures obtained on categorizing synthetic datasets: (a) dataset 1, (b) dataset 2, (c) dataset 3, and (d) dataset 4.

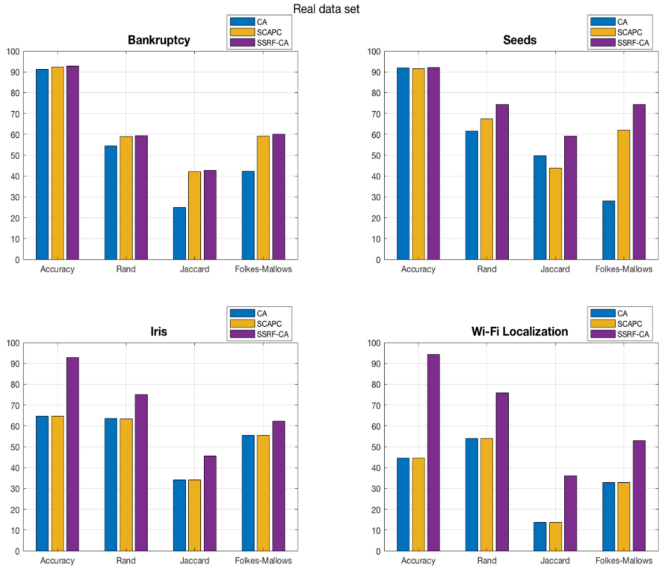


Figure 7. Performance measures obtained on categorizing real datasets: (a) bankruptcy, (b) seeds, (c) iris, and (d) Wi-Fi localization.

results for the synthetic and real datasets, respectively. The statistical results prove the success of the proposed SSRF-CA regardless of the variability in cluster shape, compactness, and size. Not all the datasets can be partitioned well using the other considered clustering algorithms. Conversely, the proposed approach can perform relatively well on all the synthetic and real datasets. We have observed that the categorization performance is improved when we incorporate the bias term in spite of the variability in shape and density of each cluster. Moreover, integrating the pairwise constraints in the proposed SSRF-CA enables more effective grouping of the data and avoid the local minima. However, the pairwise constraints might include noisy constraints that negatively affect the structure of the clusters and thus mislead the categorization of the data.

5. Conclusions

In this paper, we propose a novel fuzzy relational semi-supervised clustering algorithm based on an adaptive local distance measure i.e., the SSRF-CA. The proposed clustering algorithm exploits side-information and uses it under the form of constraints to steer the learning task. Along with its ability to perform data clustering, our algorithm is designed to handle relational data. Moreover, it learns the dependent dissimilarity measure between the data instances and also learns the optimal number of clusters automatically.

In our experiments, we prove the ability of our proposed algorithm to learn the local distance measures and the optimal number of clusters while finding a compact cluster. We use various synthetic and real-world benchmark datasets that contain different number of clusters with diverse shapes. Based on the experimental results, we concluded that the proposed SSRF-CA outperforms the other state-of-the-art algorithms. This performance can be attributed to the effective use of the pairwise constraints in the proposed SSRF-CA to guide the algorithm towards the optimal partition and also to learn the underlying cluster distance measure.

Currently, we tune the parameters α_1 and α_2 manually, then select the values that yield the partition with the highest performance. In ongoing experiments, we intend to automatically tune these parameters to obtain viable performance.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

The authors are grateful for the support they received from the Research Center of the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia.

References

[1] R. Greenlaw and S. Kantabutra, "Survey of clustering: algorithms and applications," *International Journal of Information Retrieval Research*, vol. 3, no. 2, pp. 1-29, 2013. https://doi.org/10.4018/ijirr.2013040101

[2] S. Wazarkar and B. N. Keshavamurthy, "A survey on image data analysis through clustering techniques for real world applications," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 596-626, 2018. https://doi.org/10.1016/j.jvcir.2018.07.009

[3] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165-193, 2015. https://doi.org/10.1007/s40745-015-0040-1

[4] Z. Yu, Z. Kuang, J. Liu, H. Chen, J. Zhang, J. You, H. S. Wong, and G. Han, "Adaptive ensembling of semi-supervised clustering solutions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1577-1590, 2017. https://doi.org/10.1109/TKDE.2017.2695615

[5] S. Saha, A. K. Alok, and A. Ekbal, "Brain image segmentation using semi-supervised clustering," *Expert Systems with Applications*, vol. 52, pp. 50-63, 2016. https://doi.org/10.1016/j.eswa.2016.01.005

[6] S. Xiong, J. Azimi, and X. Z. Fern, "Active Learning of Constraints for Semi-Supervised Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 43-54, 2014. https://doi.org/10.1109/TKDE.2013.22

[7] A. Skabar and K. Abdalgader, "Clustering Sentence-level text using a novel fuzzy relational clustering algorithm," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 62-75, 2013. https://doi.org/10.1109/TKDE.2011.205

[8] G. Raju, B. Thomas, S. Tobgay, and S. Kumar, "Fuzzy clustering methods in data mining: a comparative case analysis," in *Proceedings of 2008 International Conference on Advanced Computer Theory and Engineering*, Phuket, Thailand, 2008, pp. 489-493. https://doi.org/10.1109/ICACTE.2008.199

[9] S. Zeng, X. Tong, N. Sang, and R. Huang, "A study on semi-supervised FCM algorithm," *Knowledge and Information Systems*, vol. 35, no. 3, pp. 585-612, 2013. https://doi.org/10.1007/s10115-012-0521-x

[10] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Perez, and I. Perona. "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243-256, 2013. https://doi.org/10.1016/j.patcog.2012.07.021

[11] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proceedings of the Twenty-First International Conference on Machine Learning*, New York, NY, 2004. https://doi.org/10.1145/1015330.1015360

[12] X. Yin, S. Chen, E. Hu, and D. Zhang, "Semi-supervised clustering with metric learning: an adaptive kernel method," *Pattern Recognition*, vol. 43, no. 4, pp. 1320-1333, 2010. https://doi.org/10.1016/j.patcog.2009.11.005

[13] O. Bchir, H. Frigui, and M. M. Ben Ismail, "Semi-supervised relational fuzzy clustering with local distance measure learning," in *Proceedings of 2013 World Congress on Computer and Information Technology (WCCIT)*, Sousse, Tunisia, 2013, pp. 1-4. https://doi.org/10.1109/WCCIT.2013.6618764

[14] O. Bchir and H. Frigui, "Fuzzy clustering with Learnable Cluster dependent Kernels," in *Proceedings of 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011),* Taipei, Taiwan, 2011, pp. 2521-2527. https://doi.org/10.1109/FUZZY.2011.6007411

[15] H. Frigui and R. Krishnapuram, "Clustering by competitive agglomeration," *Pattern Recognition*, vol. 30, no. 7, pp. 1109-1119, 1997. https://doi.org/10.1016/S0031-3203(96)00140-9

[16] O. Nasraoui, H. Frigui, R. Krishnapuram, and A. Joshi, "Extracting web user profiles using relational competitive fuzzy clustering," *International Journal on Artificial Intelligence Tools*, vol. 9, no. 4, pp. 509-526, 2000. https://doi.org/10.1142/S021821300000032X

[17] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: a brief survey," 2005, Available http://cedric.cnam.fr/~crucianm/src/BriefSurveyClustering.pdf

[18] C. F. Gao and X. J. Wu, "A new semi-supervised clustering algorithm with pairwise constraints by competitive agglomeration," *Applied Soft Computing*, vol. 11, no. 8,

pp. 5281-5291, 2011. https://doi.org/10.1016/j.asoc.2011.05.032

[19] J. M. Yih, Y. H. Lin, H. C. Liu, and C. F. Yih, "FCM & FPCM algorithm based on unsupervised Mahalanobis distances with better initial values and separable criterion," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 3, no. 1,pp. 9-18, 2009.

[20] R. J. Hathaway, J. W. Davenport, and J. C. Bezdek, "Relational duals of the *c*-means clustering algorithms," *Pattern Recognition*, vol. 22, no. 2, pp. 205-212, 1989. https://doi.org/10.1016/0031-3203(89)90066-6

[21] D. M. Duc, N. D. Hoang, and L. H. Nguyen, "Lagrange multipliers theorem and saddle point optimality criteria in mathematical programming," *Journal of Mathematical Analysis and Applications*, vol. 323, no. 1, pp. 441-455, 2006. https://doi.org/10.1016/j.jmaa.2005.10.038

[22] E. Frias-Martinez, S. Y. Chen, and X. Liu, "Survey of data mining approaches to user modeling for adaptive hypermedia," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 36, no. 6, pp. 734-749, 2006. https://doi.org/10.1109/TSMCC.2006.879391

[23] C. Borgelt and R. Kruse, "Finding the number of fuzzy clusters by resampling," in *Proceedings of 2006 IEEE International Conference on Fuzzy Systems*, Vancouver, BC, 2006, pp. 48-54. https://doi.org/10.1109/FUZZY.2006.1681693

Norah Ibrahim Fantoukh is a teaching assistant in the Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. She received her master's degree in Computer Science from King Saud University in 2020. Her research interests include pattern recognition, machine learning, and data mining.
E-mail: n.fantoukh@gmail.com

*The photo is not included according to the author's request.

Mohamed Maher Ben Ismail is an associate professor in the Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He received his Ph.D. in Computer Science from the University of Louisville in 2011. His research interests include pattern recognition, machine learning, data mining, and image processing.
E-mail: maher.benismail@gmail.com

Ouiem Bchir is an associate professor in the Computer Science Department, College of Computer and Information Systems, King Saud University, Riyadh, Saudi Arabia. She got her Ph.D. from the University of Louisville, KY, USA. Her research interests are pattern recognition, machine learning, and hyperspectral image analysis. She received the University of Louisville Dean's Citation, the University of Louisville CSE Doctoral Award, and the Tunisian presidential award for her electrical engineering diploma.
E-mail ouiem.bchir@gmail.com