Possibilistic clustering with seeds
Violaine Antoine, Jose Guerrero, Tanya Boone, Gerardo Romero

To cite this version:


HAL Id: hal-03939504
https://hal.science/hal-03939504
Submitted on 26 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Possibilistic clustering with seeds

Violaine Antoine*, Jose A. Guerrero†‡, Tanya Boone§, Gerardo Romero§

* Clermont Auvergne University, UMR 6158, LIMOS, F-63000, Clermont-Ferrand, France
Email: violaine.antoine@uca.fr
† Irstea, 9 avenue Blaise Pascale, 63170 Aubire, France
‡ Litis, EA 4108, INSA de Rouen, 76800 Saint Etienne du Rouvray, France
Email: jguerrer@ieee.org
§ Electronic Department at U.A.M. Reynosa-Rodhe, UAT, México
Email: gromero@docentes.uat.edu.mx

Abstract—Clustering methods assign objects to clusters using only as prior information the characteristics of the objects. However, clustering algorithms performance can be improved when background knowledge is available. Such background knowledge can be incorporated in a clustering method as label constraints which results in a semi-supervised clustering algorithm. We propose to extend two possibilistic clustering algorithms to make use of available a priori information. The goal is twofold: to improve the accuracy of the clustering result by leading the method towards a desired solution and to detect outliers by taking advantage of the generated possibilistic partition. The proposed methods are called semi-supervised repulsive possibilistic c-means (SRPCM) and semi-supervised possibilistic fuzzy c-means (SPFCM). They correspond to possibilistic clustering algorithms that introduce label constraints. Experimental results show that the proposed algorithms using label constraints improve (1) the clustering result and (2) the outliers detection.

I. INTRODUCTION

Clustering methods are part of exploratory data analysis techniques that aim to group unlabeled objects into clusters thanks to a similarity notion. Different approaches have been proposed in the literature such as hierarchical clustering and partition based clustering [1]. On one hand, hierarchical methods generate dendrograms based on a proximity matrix. On the other hand, partition based methods organize data into groups or clusters using either crisp (hard) partitions or soft partitions. Crisp methods divide data into groups by assigning each object of the dataset to a single cluster with total certainty. In the domain of partition based clustering, the most popular algorithms generating crisp partitions are k-means and density based algorithms such as DBSCAN. The k-means algorithm, which is an optimization based method, corresponds to a classical data analysis tool used in many topics [2], [3], [4]. On the contrary, soft clustering methods allow to express a degree of uncertainty for the membership of each object to each cluster. Fuzzy clustering methods, based on the fuzzy sets theory [5], are the most commonly used soft methods. Amongst fuzzy clustering methods, Fuzzy C-Means (FCM) is a well known variant of k-means which assigns for each object a probability value to belong to each cluster. Its main drawback is a poor performance on noisy data. To overcome this weakness, [12] proposed a possibilistic clustering algorithm called Possibilistic C-Means (PCM) which is based on FCM with a relaxed membership degree constraint. However, as discussed in [13], PCM highly depends on initial conditions to produce good results. Indeed, based on initial conditions, PCM often generates coincident clusters. This problem has been addressed in [13], [14], [15]. In [14], the authors propose to add a repulsion based penalty term to the PCM objective function to avoid the coincident cluster problem resulting in an algorithm named Repulsive PCM (RPCM). In [15], the authors propose a new algorithm which is a linear combination of FCM and PCM named PFCM.

Simultaneously, it has been shown that the performance of hard and soft clustering algorithms can be improved by using a limited amount of background knowledge expressed as contraints [16], [17], [18]. These methods, called semi-supervised clustering methods, can be divided following the type of constraints employed: pairwise contraints [17], [19], label constraints [16], [18], [20], or others [21], [22].

In this paper, we are interested in the semi-supervised possibilistic clustering problem since the possibilistic framework enables to handle noisy data. Indeed, as explained in [15], a probabilistic partition forces an outlier to belong to one or more clusters with a high membership degree. Thus, this make impossible to mark it as an outlier. In contrast, a possibilistic partition allows for an object to be assigned with low membership degree to every clusters. When it happens, the object can be interpreted as an outlier. We propose to extend the RPCM and PFCM algorithms by incorporating a penalty term in their initial objective functions such that label based constraints are taken in account.

This work is organized as follows: in section II, possibilistic clustering preliminaries are introduced. Section III presents...
a semi-supervised possibilistic clustering algorithm based on the repulsive PCM (RPCM) [14]. Section IV presents a semi-supervised possibilistic clustering algorithm based on the PFCM method [15]. In section V, the experimental results are discussed. Finally, the conclusions and future work are presented in section VI.

II. POSSIBILISTIC CLUSTERING PRELIMINARIES

Let us consider \( n \) objects represented by a set of feature vectors \( \mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \) in \( \mathbb{R}^p \) and \( c \) centroids defined by a matrix \( \mathbf{V} = (\mathbf{v}_k) \) such that each centroid \( \mathbf{v}_k \in \mathbb{R}^p \) corresponds to a cluster. The PCM algorithm [12] iteratively updates the centroids \( \mathbf{V} \) and a possibilistic partition \( \mathbf{T} = (t_{ik}) \) in order to minimize the \( J_{PCM} \) objective function:

\[
J_{PCM}(\mathbf{T}, \mathbf{V}) = \sum_{i=1}^{n} \sum_{k=1}^{c} t_{ik}^m d_{ik}^2 + \sum_{k=1}^{c} \gamma_k \sum_{i=1}^{n} (1 - t_{ik})^m,
\]

subject to constraint (1). The parameters \( \eta_k \geq 0 \), \( \forall k \in \{1 \ldots c\} \) define specific degrees of repulsion for each cluster to the other ones.

The authors propose to minimize the objective function as PCM, i.e. by carrying out an iterative optimization of the possibilistic partition \( \mathbf{T} \) and the centroids \( \mathbf{V} \). Since the new term does not depend on \( \mathbf{T} \), its update is identical to PCM:

\[
t_{ik} = \frac{1}{1 - \frac{d^2_{ik}}{\gamma_k}}, \forall i \in \{1 \ldots n\}, k \in \{1 \ldots c\}.
\]

The update of the centroids \( \mathbf{V} \) composed of \( \mathbf{v}_1 \ldots \mathbf{v}_k \ldots \mathbf{v}_c \) is more intricate. Indeed, in \( J_{RPCM} \) there exists a dependency between centroids. Thus, the optimization with the respect to \( \mathbf{V} \) cannot be performed by updating each centroid \( \mathbf{v}_k \) separately. However, in order to facilitate the optimization and to update prototypes one by one, [14] proposed to set as a constant \( r_{kl} = (\mathbf{v}_k - \mathbf{v}_l)^T (\mathbf{v}_k - \mathbf{v}_l) \), \( \forall k, l \in \{1 \ldots c\} \), making the hypothesis that such values are not significantly changing during the optimization process. The update formula is then obtained by setting the gradient \( \frac{\partial J_{RPCM}}{\partial \mathbf{v}_k} = 0 \):

\[
\mathbf{v}_k = \frac{\sum_{i=1}^{n} t_{ik}^m \mathbf{x}_i - \eta_k \sum_{l \neq k} r_{kl} \mathbf{v}_l}{\sum_{i=1}^{n} t_{ik}^m - \eta_k \sum_{l \neq k} r_{kl}}, \forall k \in \{1 \ldots c\}.
\]

Notice that the centroids update equation (4) does not always allow to minimize the objective function (3). Moreover, in [14], the authors state that if \( \sum_{i=1}^{n} t_{ik}^m < \eta_k \sum_{l \neq k} r_{kl} \), concerned centroids have to be relocated at random positions. In [24], the authors propose to use a second order approximation method such as Newton, instead of a gradient based algorithm. Their choice is based on the fact that the gradient with respect to the centroids is non linear.

B. Possibilistic Fuzzy c-Means

A different approach to solve the problem raising with PCM is to combine possibility and probability membership values [15]. The objective function is then defined as follows:

\[
J_{PFCM}(\mathbf{U}, \mathbf{T}, \mathbf{V}) = \sum_{i=1}^{n} \sum_{k=1}^{c} (a u_{ik}^m + b t_{ik}^m) d_{ik}^2 + \sum_{k=1}^{c} \gamma_k \sum_{i=1}^{n} (1 - t_{ik})^m,
\]

where \( a, b \) and \( m, \eta \) are positive coefficients. The objective function, called PFCM for Possibilistic Fuzzy c-means, is subject to the constraints (1), (6) and (7).

\[
\sum_{k=1}^{c} u_{ik} = 1, \forall k \in \{1 \ldots c\}, \quad \sum_{k=1}^{c} u_{ik} \geq 0, \forall i \in \{1 \ldots n\}, \forall k \in \{1 \ldots c\},
\]

subject to constraint (1). The parameters \( \eta_k \geq 0 \), \( \forall k \in \{1 \ldots c\} \) define specific degrees of repulsion for each cluster to the other ones.

The authors propose to minimize the objective function as PCM, i.e. by carrying out an iterative optimization of the possibilistic partition \( \mathbf{T} \) and the centroids \( \mathbf{V} \). Since the new term does not depend on \( \mathbf{T} \), its update is identical to PCM:

\[
t_{ik} = \frac{1}{1 + \frac{1}{\gamma_k} d_{ik}^2}, \forall i \in \{1 \ldots n\}, k \in \{1 \ldots c\}.
\]
In [15], the authors have demonstrated that \( J_{PFCM} \) is minimized by iteratively updating the fuzzy partition \( U = (u_{ik}) \), the possibilistic partition \( T \) and the centroids \( V \) using equations \((8)-(10)\) until convergence. Update formulas were obtained using the Lagrange multiplier method. Remark that the update formula of \( U \) is identical to FCM and the update formula of \( T \) is similar to PCM.

\[
u_{ik} = \left( \sum_{l=1}^{c} \left( \frac{d_{ik}}{d_{il}} \right)^{\frac{2}{m-1}} \right)^{-1}. \tag{8}
\]

\[
t_{ik} = \left( 1 + \left( \frac{b}{\gamma_k} \right)^{\frac{2}{m+1}} \right)^{-1}. \tag{9}
\]

\[
v_k = \frac{\sum_{i=1}^{n} (au_{ik}^m + bt_{ik}^n)x_i}{\sum_{i=1}^{n} (au_{ik}^m + bt_{ik}^n)}. \tag{10}
\]

### III. SEMI-SUPERVISED REPULSIVE PCM

In real applications, a priori information is available with various degrees of certainty. We propose to exploit soft label knowledge, i.e. objects labeled with a degree of membership with various degrees of certainty. We propose to exploit soft label knowledge, i.e. objects labeled with a degree of membership. We propose to exploit soft label knowledge, i.e. objects labeled with a degree of membership. We propose to exploit soft label knowledge, i.e. objects labeled with a degree of membership.

Let \( f_{ik} \in [0, 1] \) be the possibility known a priori that \( x_i \) belongs to the cluster \( k \). This value, is equal to 0 when it is sure that the object \( i \) does not belong to the cluster \( k \). Conversely, \( f_{ik} = 1 \) indicates that \( x_i \) has a strong possibility to belong to the cluster \( k \), even if it also let the possibility to have degrees of belief for the other clusters.

A natural requirement for the Semi-supervised Repulsive PCM algorithm (SRPCM) is to obtain a possibilistic value \( t_{ik} \) the closest to \( f_{ik} \). In particular cases and as in [16], constraints are softened to avoid sudden disturbance of the structure. Indeed, it can lead to inconsistent solutions.

Thus, we suggest to introduce a penalty term in the objective function of \( J_{RPCM} \) so that the label constraints are respected. The distance \( d_{ik} \) for a constrained object \( x_i \) on cluster \( k \) is employed to relax the constraint:

\[
J_{SRPCM}(T, V) = J_{RPCM} + \alpha \sum_{i=1}^{n} \sum_{k=1}^{c} b_{ik}(t_{ik} - f_{ik})^m d_{ik}^2,
\]

subject to constraint \((1)\) and where \( m > 1 \) is even and \( \alpha \geq 0 \) is a tradeoff coefficient between the inherent structure unsupervisedly retrieved and the consideration of the constraints. The variable \( b_{ik} \) enables to select only constrained values in the penalty term:

\[
b_{ik} = \begin{cases} 
1 & \text{if } x_i \text{ and class } k \text{ are constrained}, \\
0 & \text{otherwise}.
\end{cases}
\]

The penalty term can be decomposed to obtain the value for a single element of \( v_k \), e.g., \( v_{k_j} \):

\[
\frac{\partial J_{SRPCM}}{\partial v_{k_j}} = -2 \sum_{i=1}^{n} t_{ik}^2(x_{ij} - v_{kj}) - 2\eta_k \sum_{l \neq k} v_{kl} - v_{lj} \frac{v_k - v_l}{\|v_k - v_l\|^4} - 2\alpha \sum_{i=1}^{n} b_{ik}(t_{ik} - f_{ik})^2(x_{ij} - v_{kj}).
\]

A. Optimization with the respect to the possibilistic partition

The first step of the algorithm consists in fixing \( V \) to find an update formula of \( T \) that minimize \( J_{SRPCM}(T) \). Since each element \( t_{ik} \) of \( T \) are independent, we compute the derivative of the objective function \((11)\) with the respect to \( t_{ik} \) and \( m = 2 \):

\[
\frac{\partial J_{SRPCM}}{\partial t_{ik}} = 2t_{ik}d_{ik}^2 - 2\gamma_k(1 - t_{ik}) + 2\alpha b_{ik}d_{ik}^2(t_{ik} - f_{ik}).
\]

Each value \( t_{ik} \) of the possibilistic partition minimizing \( J_{SRPCM} \) is obtained by setting \( \frac{\partial J_{SRPCM}}{\partial t_{ik}} = 0 \):

\[
t_{ik} = \frac{\gamma_k + \alpha b_{ik}d_{ik}^2f_{ik}}{\gamma_k + (\alpha b_{ik} + 1)d_{ik}^2}.
\]

B. Optimization with the respect to the centroids

In a second step, the possibilistic partition \( T \) is fixed and the objective function \( J_{SRPCM} \) is minimized with the respect to the centroids. In order to choose a good optimization method, a study concerning the convexity of the objective function is first performed. The gradient is then calculated:

\[
\frac{\partial J_{SRPCM}}{\partial v_{k_j}} = -2 \sum_{i=1}^{n} t_{ik}^2(x_{ij} - v_{kj}) - 2\eta_k \sum_{l \neq k} v_{kl} - v_{lj} \frac{v_k - v_l}{\|v_k - v_l\|^4} - 2\alpha \sum_{i=1}^{n} b_{ik}(t_{ik} - f_{ik})^2(x_{ij} - v_{kj}).
\]

Then, elements composing the Hessian matrix \( H_k \in R^{p \times p} \) deduced from the second derivatives of \( J_{SRPCM} \) with the respect to the centroids are the following:

\[
\frac{\partial^2 J_{SRPCM}}{\partial v_{k_j}^2} = 2 \sum_{i=1}^{n} t_{ik}^2 - 2\eta_k \sum_{l \neq k} \frac{1}{d_{ik}^2} \frac{(v_{kj} - v_{lj})^2}{\|v_k - v_l\|^6} + 2\alpha \sum_{i=1}^{n} b_{ik}(t_{ik} - f_{ik})^2,
\]

\[
\frac{\partial J_{SRPCM}}{\partial v_{k_j} \partial v_{k_j'}} = 2\eta_k \sum_{i=1}^{n} \frac{(v_{kj} - v_{lj})(2v_{k_j'} - 2v_{lj})}{\|v_k - v_l\|^6}.
\]

where \( v_{k_j'} \) corresponds the \( j' \)th element of the centroid \( v_k \) such that \( j' \neq j \). Notice that a similar result, without the label constraints, is available in [24].
Finally, the Hessian matrix $H_k$ can be rewritten as follows:

$$
H_k = 2 \left( \sum_{i=1}^{n} t_{ik}^2 \right) I + 8\eta_k \sum_{l \neq k} (v_k - v_l)(v_k - v_l)^T \frac{(v_k - v_l)^6}{\|v_k - v_l\|^6} - 2\eta_k \left( \sum_{l \neq k} \frac{1}{d_{lk}^2} \right) I + 2\alpha \sum_{i=1}^{n} b_{ik}(t_{ik} - f_{ik})^2 I,
$$

where $I$ is the identity matrix of proper dimension.

Let us remind that the sum of positive (semi)definite matrices results in a positive (semi)definite matrix. Since $I$ is positive definite, $t_{ik}^2 \geq 0$ and $b_{ik}(t_{ik} - f_{ik})^2 \geq 0$ then the first and the last term of $H_k$ are positive semidefinite. Similarly, the matrix $(v_k - v_l)(v_k - v_l)^T$ is positive semidefinite, $8\eta_k \geq 0$ and $\|v_k - v_l\|^6 \geq 0$ so the second term of $H_k$ is also positive semidefinite. Finally, the third term is negative semidefinite. Indeed, the scalar coefficient applied for $I$ is negative or equal to 0. Consequently, $H_k$ is not guaranteed to be positive semidefinite.

Thus, in order to update the centroids, a standard trust-region method for non linear minimization is employed [25]. Although such method may just reach a local minimum, it assures the convergence of the clustering algorithm. Indeed, the scalar coefficient applied for $I$ is negative or equal to 0. Consequently, $H_k$ is not guaranteed to be positive semidefinite.

B. Update of the possibilistic partition

By fixing $J$ objective function to be minimized is the following: called SPFCM. It enables to handle soft label constraints. The integrated in PFCM in order to produce a new algorithm are independent, letting us the possibility to update each value

$$
J_{SPFCM}(U, T, V) = J_{PFCM} + \alpha \sum_{i=1}^{n} \sum_{k=1}^{c} b_{ik}(t_{ik} - f_{ik})\eta d_{ik}^2,
$$

subject to (1), (6) and (7).

Fixed parameters have the same limits as PFCM, except for $\eta$ that should be positive and even. The $\alpha$ coefficient follows the same rule as SRPCM, i.e. $\alpha \geq 0$.

The objective function (12) is optimized using a heuristic method which consists in iteratively minimizing $J_{SPFCM}$ with respect to $U$, then $T$, then $V$ until convergence.

A. Update of the probabilistic partition

The optimization of $J_{SPFCM}$ with respect to $U$ is achieved by fixing $T$ and $V$ as constants. Since the penalty term incorporated for SPFCM does not contain any probabilistic partition values, the update of the membership degrees $U$ are identical to PFCM and corresponds to the equation (8).

B. Update of the possibilistic partition

In order to minimize $J_{SPFCM}$ with respect to $T$, the variables $U$ and $V$ are fixed. The columns and rows of $T$ are independent, letting us the possibility to update each value $t_{ik}$ separately. By setting $\eta = 2$ to facilitate the optimization process, the problem becomes quadratic. The derivative is then calculated:

$$
\frac{\partial J_{SPFCM}}{\partial t_{ik}} = 2bt_{ik}d_{ik}^2 - 2\gamma_k(1 - t_{ik}) + 2\alpha b_{ik}d_{ik}^2(t_{ik} - f_{ik}).
$$

Setting the derivative to 0 enables to obtain the following update formula:

$$
t_{ik} = \gamma_k + \alpha b_{ik}d_{ik}^2f_{ik} \frac{bt_{ik}d_{ik}^2 + \gamma_k + \alpha b_{ik}d_{ik}^2}{bd_{ik}^2 + \gamma_k + \alpha b_{ik}d_{ik}^2}.
$$

Notice that when $b = 1$, the update of the possibilistic partition for SPFCM is identical to SRPCM.

C. Update of the centroids

Since $au_{ik}^m + bt_{ik}^m > 0$ and $(t_{ik} - f_{ik})^m > 0$, $J_{SPFCM}$ is positive semidefinite with respect to $V$. As a consequence, the minimum of the objective function corresponds to the value of $V$ vanishing the derivative. Notice that each centroid $v_k$ is independent to each other and can then be managed separately:

$$
\frac{\partial J_{SPFCM}}{\partial v_k} = -2\sum_{i=1}^{n} (au_{ik}^m + bt_{ik}^m)(x_i - v_k) - 2\alpha \sum_{i=1}^{n} b_{ik}(t_{ik} - f_{ik})^2(x_i - v_k).
$$

Let $z_{ik}$ be a scalar such that $z_{ik} = (au_{ik}^m + bt_{ik}^m) + \alpha b_{ik}(t_{ik} - f_{ik})^2$. Setting the derivative to 0 leads to the following result:

$$
v_k = \frac{\sum_{i=1}^{n} z_{ik}x_i}{\sum_{i=1}^{n} z_{ik}}.
$$

V. EXPERIMENTAL RESULTS

A. Experimental protocol

We have run extensive experimental tests on three well known datasets from the UCI repository, i.e. Ecoli, Iris and Wine, and on a toy dataset called GaussK2. The characteristics of each dataset are presented in Table I.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># classes used</th>
<th># classes</th>
<th># attributes</th>
<th># instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ecoli</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>336</td>
</tr>
<tr>
<td>Iris</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>150</td>
</tr>
<tr>
<td>Wine</td>
<td>3</td>
<td>3</td>
<td>13</td>
<td>178</td>
</tr>
<tr>
<td>GaussK2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>400</td>
</tr>
</tbody>
</table>

Ecoli dataset contains eights classes from which three classes contain very few instances (2, 2, and 5 instances). We consider the instances from these classes as outliers, leading us to take into account five classes only.

The GaussK2 dataset is a two-dimensional space dataset generated by two gaussians with different covariance matrices.
The result is that the clusters have different densities, as shown in Figure 1.

The parameters for SRPCM and SPFCM algorithms were defined as follows: $\alpha = 1$ and $\gamma_k$ is retrieved as explained in [12] with equation (2) and $K = 1$. The parameters $\eta_k \forall k \in \{1 \ldots c\}$ for SRPCM are identical and fixed manually for each dataset. Similarly, for SPFCM, we set $a = 1$ and $b$ is manually chosen following each dataset. Details are presented Table II.

### Table II

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$\eta_k$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ecoli</td>
<td>0.8</td>
<td>2.2</td>
</tr>
<tr>
<td>Iris</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Wine</td>
<td>0.2</td>
<td>1.6</td>
</tr>
<tr>
<td>GaussK2</td>
<td>0.1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

In order to evaluate the proposed methods, a comparison with SKMEANS [26] and SFCM [16] has been performed. The SKMEANS algorithm corresponds the k-means method taking label constraints as background knowledge while the SFCM represents the fuzzy c-means version with label constraints.

Final acquired partitions are hard for SKMEANS, fuzzy for SFCM and possibilistic for SRPCM and SPFCM. To perform a comparison between these methods, fuzzy and possibilistic partitions are transformed into hard partition by selecting the cluster with the maximum of probability or possibility. Then, since the true classes of the datasets are known, we compute the ARI [27] to measure the performance of the clustering algorithms.

### B. Clustering results

All clustering algorithms use the same initial conditions for centroids and constraints. The constraints correspond to totally certain labels. Experiments consist in 100 trials with a given percentage of constraints. Each trial corresponds to 5 executions of an algorithm with different centroid initializations. The partition with the minimum objective function value is then selected.

Figures 2, 3, 4 and 5 present the average ARI and its confidence interval against the proportion of labeled constraints for SKMEANS, SFCM and the proposed clustering algorithms, i.e. SRPCM and SPFCM.

The results show that the addition of constraints on SRPCM clearly improves the partition compared to the initial partition found without constraints, i.e. with RPCM. For GaussK2 and Iris, the best performances are achieved by SRPCM. However, the same clustering method have low results on Wine and Ecoli. We can also observe that RPCM has always a low ARI. The reason is that the algorithm is more sensitive to local minima than the other clustering methods.

Conversely, the SPFCM algorithm gives more stable results than SRPCM. Then, it can be seen as a more secure choice than SRPCM when no expert assessment can be performed a posteriori. Indeed, it still obtains better results than SKMEANS and SFCM for Iris and GaussK2.

The results on the Ecoli dataset can be surprising on the first approach. Indeed, without constraints, PFCM outperforms the other algorithms. The performance of this algorithm is however inferior when constraints are taken into account. Such
behavior can be explained by the inherent structure of Ecoli: over the 5 classes of the dataset, a couple of pairs of classes are quasi-overlapped. Thus, none of the clustering algorithms are able to find the real structure of Ecoli. Experimental results show that PFCM has the best ARI by finding coincident clusters. The addition of constraints in SPFCM tends to separate the clusters, thus decreasing its ARI value.

Inversely, few number of constraints enables SKMEANS to obtain the best accuracy for Ecoli. The reason is that SKMEANS forces the total respect of the constraints whereas the other algorithms, with the use of a penalty term, can let some constraints not respected to obtain a more coherent final structure. The SKMEANS algorithm has then the possibility to converge faster to the desired solution.

C. Outliers detection

An interesting feature of a possibilistic clustering algorithm is its ability to identify outliers. Consequently, an experiment on the Ecoli dataset has been performed to show how SRPCM and SPFCM handle outliers.

The same experimental protocol described above is used, i.e. 100 trials for each specific set of constraints. When a possibilistic partition is retrieved, a simple rule to detect outliers is applied: an object $x_i$ is an outlier if $t_{ik} \leq 0.1$, $\forall k \in \{1...c\}$. The average rates of good detection as well as their confidence interval are illustrated Figure 6. As it can be observed, labels constraints helps to the detection of outliers. The SRPCM algorithm detects better outliers than SPFCM. This can be explained by the fact that the optimization of the fuzzy partition for SPFCM has an impact on the possibilistic partition.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, two novel semi-supervised clustering algorithms have been presented: SRPCM and SPFCM. Both of them incorporate label constraints and handle a possibilistic partition. The use of a possibilistic framework enables to express various type of uncertainty. The advantage of such framework is twofold in SRPCM and SPFCM: first, constraints are represented in the form of possibility to belong to clusters. Such representation enables an expert to include partial information into the clustering algorithms more than any other semi-supervised variants of k-means and FCM. Second, the generation of a possibilistic partition by SRPCM and SPFCM enables to obtain rich information about the dataset and makes easy for instance the detection of outliers.

Encouraging results have emerged from experiments and comparisons with other constraint-based methods. It leads to consider several possible future works: first, the impact of labels having a degree of uncertainty can be study and an active learning scheme can be developed. Second, an investigation about the parameters that are currently manually fixed can be performed in order to acquire them automatically. Future works includes extensions to use other distance measures such as Mahalanobis distance as well as extending the proposed method to consider more complex uncertainty models.

ACKNOWLEDGMENT

This work has been partially supported by the Autonomous University of Tamaulipas through the grants No: PEI-2014-209955, PEI-2014-213516, P/PFCE-2016-28MSU0010B-22 and by Conacyt through the Master’s Scholarships of Tanya Boone R.N 588607.
REFERENCES


