A Possibilistic Multivariate Fuzzy c-Means Clustering Algorithm

Ludmila Himmelspach(✉) and Stefan Conrad

Institute of Computer Science, Heinrich-Heine-Universität Düsseldorf,
40225 Düsseldorf, Germany
{himmelspach,conrad}@cs.uni-duesseldorf.de

Abstract. In this paper, we present a new *possibilistic multivariate fuzzy c-means* (PMFCM) clustering algorithm. PMFCM is a combination of multivariate fuzzy c-means (MFCM) and possibilistic fuzzy c-means (PFCM) that produces membership degrees of data objects to each cluster according to each feature and typicality values of data objects to each cluster. In this way, PMFCM produces a multivariate partitioning of a data set detecting clusters with unevenly distributed data over different features. It also reduces the influence of noise and outliers to computation of cluster centers.

Keywords: Fuzzy clustering · c-Means models · Possibilistic clustering · Multivariate memberships

1 Introduction

Clustering is an unsupervised learning technique for identifying groups of similar data objects within a data set. It is used in many fields, including image processing, bioinformatics, text mining where high dimensional data objects have to be grouped. Clustering high dimensional data bears several challenges that can be explained on the example of text clustering where a document is represented by a vector of *tf-idf*s of terms in the collection [1]. Due to the documents related to several topics, there are usually overlapping clusters in the data set. Depending on the range of topics, only few feature values in data vectors are significantly greater than zero. This implies that only few dimensions determine clusters. The information about the belonging of data objects to clusters in each dimension might be of great use. Finally, documents in the collection that do not belong to any cluster have to be recognized as noise and outliers. In this paper, we propose a new objective function based possibilistic multivariate fuzzy clustering algorithm that aims at satisfying these requirements.

The rest of the paper is organized as follows: in Sect. 2 we give a short overview over the different fuzzy clustering algorithms that we used as a basis for our approach. The possibilistic multivariate fuzzy c-means algorithm is presented in Sect. 3. The evaluation results of our method and the comparison with the basic approach are presented in Sect. 4. Section 5 closes the paper with a short summary and the discussion of future research.

S. Schockaert and P. Senellart (Eds.): SUM 2016, LNAI 9858, pp. 338–344, 2016.
DOI: 10.1007/978-3-319-45856-4_24

2 Related Works

The first problem described in introduction can be solved by using the *fuzzy c-means* (FCM) [2] clustering algorithm that assigns each data object to each cluster with a membership degree. The objective function of the fuzzy c-means algorithm is defined as follows:

$$J_m(U, V; X) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m d^2(v_i, x_k), \tag{1}$$

where c is the number of clusters, $u_{ik} \in [0, 1]$ is the membership degree of data object x_k to cluster i, $m > 1$ is the fuzzification parameter, $d(v_i, x_k)$ is the distance between cluster center v_i and data object x_k. The objective function of FCM has to be minimized under constraint (2) to obtain a good partitioning of the data set.

$$\sum_{i=1}^{c} u_{ik} = 1 \quad \forall k \in \{1, ..., n\} \text{ and } \sum_{k=1}^{n} u_{ik} > 0 \quad \forall i \in \{1, ..., c\}. \tag{2}$$

FCM is able to model the soft transitions between clusters. The information about the clustering structure, especially about the overlaps between clusters can be derived from the partitioning results.

The problem about FCM is that due to constraint (2) it assigns outliers and noise points to clusters in the same way as data objects within clusters. On the one hand, the information about whether a data object is a typical representative of the data structure or whether it is an outlier or noise point cannot be derived from the membership degrees. On the other hand, the outliers affect the computation of cluster centers. This problem can be solved by using the *possibilistic fuzzy c-means* (PFCM) [3] clustering algorithm that additionally produces the typicality values of data objects to clusters which express a relative degree of typicality of a data object to the overall structure of data. The objective function of PFCM is defined as follows:

$$J_{m,\eta}(U, T, V; X) = \sum_{k=1}^{n} \sum_{i=1}^{c} (a u_{ik}^m + b t_{ik}^\eta) d^2(v_i, x_k) + \sum_{i=1}^{c} \gamma_i \sum_{k=1}^{n} (1 - t_{ik})^\eta, \tag{3}$$

where $t_{ik} \leq 1$ is the typicality value of data object x_k to cluster i, $m > 0$ and $\eta > 0$ are user defined constants. The first term in the objective function of PFCM has the same meaning as in FCM, where constants $a > 0$ and $b > 0$ control the relative influence of fuzzy memberships and typicality values. The second term ensures that the typicality values are determined as large as possible. The second summand is weighted by the parameter $\gamma_i > 0$ that the authors in [4] recommend to choose by computing:

$$\gamma_i = K \frac{\sum_{k=1}^{n} u_{ik}^m d^2(v_i, x_k)}{\sum_{k=1}^{n} u_{ik}^m} \quad 1 \leq i \leq c, \tag{4}$$

where the $\{u_{ik}\}$ are the terminal membership degrees computed by FCM and $K > 0$ (usually $K = 1$). The objective function of PFCM has to be minimized under constraint (2) and $\sum_{k=1}^{n} t_{ik} > 0, \forall i \in \{1, ..., c\}$.

The possibilistic fuzzy c-means algorithm solves the first and the third problems described above but it assumes that all features are equally important for all clusters. Since few features usually determine particular clusters in high dimensional data sets, using either the attribute weighting fuzzy clustering algorithm [5] or the *multivariate fuzzy c-means* (MFCM) [6] method might be a better choice in such domains. We abstain from using the subspace clustering algorithms because they determine clusters in subspaces disregarding values of data objects in other features. In our case we aim for finding clusters where data objects have similar values in all features. Since MFCM produces the membership degrees of data objects to each cluster according to each feature which is beneficial for subsequent use of clustering results, we use it as a basis for our approach. The objective function of MFCM is defined as follows:

$$J_m(U, V; X) = \sum_{k=1}^{n} \sum_{i=1}^{c} \sum_{j=1}^{p} u_{ikj}^m (v_{ij} - x_{kj})^2, \tag{5}$$

where p is the number of features and $u_{ikj} \in [0, 1]$ is the membership degree of data object x_k to cluster i on feature j. Similarly to FCM, the objective function of MFCM has to be minimized under constraint (6).

$$\sum_{i=1}^{c} \sum_{j=1}^{p} u_{ikj} = 1 \quad \forall k \in \{1, ..., n\} \text{ and } \sum_{j=1}^{p} \sum_{k=1}^{n} u_{ikj} > 0 \quad \forall i \in \{1, ..., c\}. \tag{6}$$

In order to obtain the membership degrees of data objects to clusters, the authors propose to sum up the multivariate membership degrees over the dimensions [6]. Like FCM, its multivariate version does not recognize outliers and noise points as such assigning them to clusters in the same way as data objects within clusters.

3 A Possibilistic Multivariate Fuzzy c-Means Clustering Algorithm

The possibilistic FCM algorithm simultaneously produces membership degrees and the typicality values of data objects to clusters which makes it possible to derive the information about the overlaps between clusters and the noise points from the partitioning results. Unfortunately, it does not provide the information about the dimensions in which clusters overlap. This information might be valuable for subsequent use. Therefore, in our new approach called *possibilistic multivariate fuzzy c-means* (PMFCM) we combine the ideas of PFCM and MFCM algorithms. We define the objective function of PMFCM as follows:

$$J_{m,\eta}(U, T, V; X) = \sum_{k=1}^{n} \sum_{i=1}^{c} \sum_{j=1}^{p} (a u_{ikj}^m + b t_{ik}^\eta)(v_{ij} - x_{kj})^2 + p \sum_{i=1}^{c} \gamma_i \sum_{k=1}^{n} (1 - t_{ik})^\eta. \tag{7}$$

In (7) the typicality values of data objects to clusters are included in the weighting of the dimension-wise distances between data objects and cluster centers. We do not compute the typicality values of data objects to clusters at each feature because we consider the noise points as data objects that have a large overall distance (here, the Euclidean distance) to cluster centers. Unlike MFCM, we do not constrain the sum over all clusters and variables to a particular data object to be 1. In order to keep the equal weighting of distances by the membership degrees and the typicality values, we only constrain the sum over all clusters to a particular data object in each feature to be 1. Thus, the objective function of PMFCM has to be minimized under constraint (8).

$$\sum_{i=1}^{c} u_{ikj} = 1 \quad \forall k, j \quad \wedge \quad \sum_{k=1}^{n} u_{ikj} > 0 \quad \forall i, j \quad \wedge \quad \sum_{k=1}^{n} t_{ik} > 0 \quad \forall i. \tag{8}$$

In PMFCM, the membership degrees, the typicality values, and the cluster centers are updated according to formulae (9), (10), and (11).

$$u_{ikj} = \frac{1}{\sum\limits_{l=1}^{c} \left(\frac{(x_{kj} - v_{ij})^2}{(x_{kj} - v_{lj})^2} \right)^{\frac{1}{m-1}}} \qquad 1 \le i \le c,\ 1 \le k \le n,\ 1 \le j \le p. \tag{9}$$

$$t_{ik} = \frac{1}{1 + \left(\frac{b \sum_{j=1}^{p} (x_{kj} - v_{ij})^2}{\gamma_i p} \right)^{\frac{1}{\eta-1}}} \qquad 1 \le i \le c,\ 1 \le k \le n. \tag{10}$$

$$v_{ij} = \frac{\sum\limits_{k=1}^{n} (a u_{ikj}^m + b t_{ik}^\eta) x_{kj}}{\sum\limits_{k=1}^{n} (a u_{ikj}^m + b t_{ik}^\eta)} \qquad 1 \le i \le c,\ 1 \le j \le p. \tag{11}$$

The membership degrees of data objects to clusters can be computed in our model as the average of the multivariate membership degrees over all variables, $u_{ik} = \frac{1}{p} \sum_{j=1}^{p} u_{ikj}$.

The working principle of PMFCM is basically the same as of PFCM. So, due to the lack of space we omit the details and refer to [3]. As in [4] we also recommend using terminal outputs of FCM for the initialization of our algorithm.

4 Data Experiments

The proposed algorithm PMFCM is tested on artificial data in order to examine its ability to correctly determine the centers of clusters that have different extends in different dimensions in presence of noise and outliers. Unfortunately, we could not test its ability to distinguish between the data objects belonging to clusters and the noise points because the transitions between the data objects on the border of clusters and noise points are rather soft. Therefore, we could not find any meaningful threshold for typicality values to differentiate between cluster objects and noise.

Figure 1(a) shows the data set *4-clusters* with 1245 data objects unequally distributed on one spherical cluster and three clusters that have a low variance in one dimension. The sum of the distances between the means of clusters in this data set is 31.5785. We generated the data set *4-clusters-noise* by adding 150 noise points to the data set *4-clusters*. The data set *4-clusters-noise* is depicted in Figure 1(b).

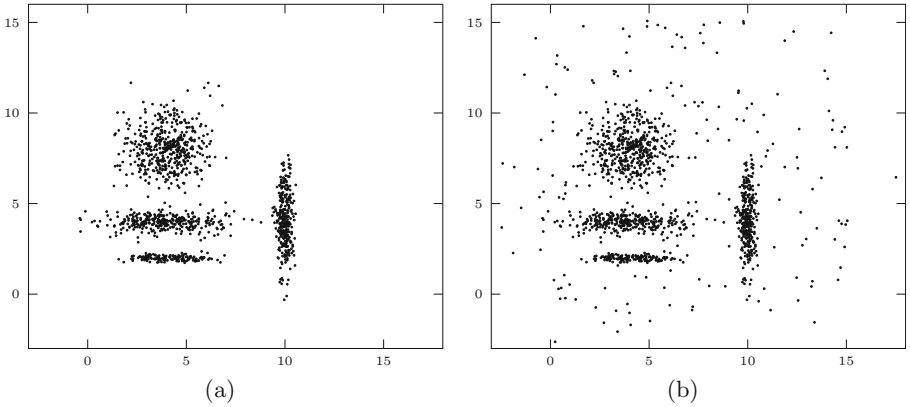

Fig. 1. Test data: (a) 4-clusters, (b) 4-clusters-noise.

Table 1 shows the comparison results between the algorithms MFCM and PMFCM for $a = 0.5$ and different values of b on the data set with four clusters without noise. We computed the Frobenius distance d_{orig} between the original means of clusters and the cluster centers produced by the clustering algorithms. We also computed the sum of the distances between the cluster centers d_{means} produced by the clustering algorithms. For a small values of b our approach produced less accurate cluster prototypes than MFCM. It determined cluster centers too close to each other, while MFCM produced cluster centers that were farther from each other than the original means of clusters. However, our approach produced much more accurate cluster centers for $b = 12$ than MFCM. With the increasing weight of the typicality values, our algorithm produced cluster centers that were slightly farther from each other than the original means of clusters. Unfortunately, we did not manage to find the golden mean where the sum of

Table 1. Comparison between MFCM and PMFCM on data set *4-clusters*.

	MFCM: $m = 2$	PMFCM: $m = 2$, $\eta = 2$ $a = 0.5$, $b = 4$	PMFCM: $m = 2$, $\eta = 2$ $a = 0.5$, $b = 12$	PMFCM: $m = 2$, $\eta = 2$ $a = 0.5$, $b = 20$
d_{orig}	2.49	2.92	1.12	1.17
d_{means}	33.73	27.87	31.52	31.65

the distances between the cluster centers produced by PMFCM corresponded to the sum of the distances between the original means of clusters in order to test whether or not our algorithm could produce the cluster centers which met the original means of clusters.

Table 2 shows the comparison results between MFCM and our algorithm on the data set *4-clusters-noise*. Unsurprisingly, the MFCM algorithm produced less accurate cluster centers than on the data set *4-clusters*. This is due to the fact that MFCM does not deemphasize the noise points while clustering. Consequently, it adjusted the cluster centers according to the distribution of all data objects in the data set. In contrast, PMFCM did not sustain a loss of performance in comparison to the data set without noise points. The fact that our approach produced more accurate cluster centers than on the data set *4-clusters* is due to the presence of noise points located close to the cluster borders. Apparently, such noise points advantageously completed the clusters so that PMFCM was able to produce more accurate cluster centers. As on the data set *4-clusters*, PMFCM produced cluster centers farther from each other with the increasing b and achieved the best results for $a = 0.5$ and $b = 12$.

Table 2. Comparison between MFCM and PMFCM on data set *4-clusters-noise*.

	MFCM: $m = 2$	PMFCM: $m = 2$, $\eta = 2$ $a = 0.5$, $b = 4$	PMFCM: $m = 2$, $\eta = 2$ $a = 0.5$, $b = 12$	PMFCM: $m = 2$, $\eta = 2$ $a = 0.5$, $b = 20$
d_{orig}	8.82	2.43	0.93	0.94
d_{means}	41.14	27.38	31.27	31.89

5 Conclusion and Future Works

In this paper, we proposed a possibilistic multivariate fuzzy c-means (PMFCM) algorithm that produces a multivariate partitioning of a data set detecting clusters with unevenly distributed data over different features in presence of noise points and outliers. In experiments, we showed that our algorithm is able to produce more accurate cluster centers than the MFCM algorithm on data sets with and without noise. Like the PFCM algorithm, the performance of the proposed method depends on the choice of the parameters that control the influence of the membership degrees and the typicality values. Therefore, in the future we plan to adapt MFCM to other possibilistic clustering models to test if the role of the right choice of user defined parameters can be minimized. Furthermore, we aim to apply our method for the text clustering to find out whether the subsequent text retrieval can be improved by the multivariate membership degrees. In this context it would be very helpful to find a heuristic for a distinction between the data objects belonging to clusters and noise points.

References

1. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Document. **28**(1), 11–21 (1972)
2. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell (1981)
3. Pal, N.R., Pal, K., Keller, J.M., Bezdek, J.C.: A possibilistic fuzzy c-means clustering algorithm. IEEE Trans. Fuzzy Syst. **13**(4), 517–530 (2005)
4. Krishnapuram, R., Keller, J.M.: A possibilistic approach to clustering. IEEE Trans. Fuzzy Syst. **1**(2), 98–110 (1993)
5. Keller, A., Klawonn, F.: Fuzzy clustering with weighting of data variables. Int. J. Uncertainty Fuzziness Knowl.-Based Syst. **8**(6), 735–746 (2000)
6. Pimentel, B.A., de Souza, R.M.C.R.: A multivariate fuzzy c-means method. Appl. Soft Comput. **13**(4), 1592–1607 (2013)