Fractional Metrics for Fuzzy c-Means

Amina Dik
Laboratoire Conception et Systèmes (LCS)
Mohammed V-Agdal University (UM5A)
Rabat Morocco
Email : a.dik70 {at} yahoo.fr

Khalid Jebari
Laboratoire Conception et Systèmes (LCS)
Mohammed V-Agdal University (UM5A)
Rabat Morocco

Abdelaziz Bouroumi
Information Processing Laboratory
Hassan II Mohammedia-Casablanca (UH2MC)
Casablanca, Morocco

Aziz Ettouhami
Laboratoire Conception et Systèmes (LCS)
Mohammed V-Agdal University (UM5A)
Rabat Morocco

Abstract— **The fuzzy c-means algorithm (FCM) is a widely used for fuzzy clustering. Usually, FCM uses the Euclidean distance as similarity measure among data points. However, this distance is strongly influenced by the larger units of measure and promotes the circular forms of data. A wide variety of distance measures have been suggested to detect different forms of cluster in data sets. A typical example of these distances is the L_p distance. In this paper, we show that values of the parameter p less than 1 can improve significantly the performance of FCM, especially when the data set contains outliers. This measure is called fractional metric. For this, we realise a comparative study of FCM with different values of p on six data sets. The results show clearly that fractional metric allows FCM to produce good results in a wide variety of real world applications.**

Keywords- similarity; fractional metric; fuzzy c-means; fuzzy clustering; distance Measures.

I. Introduction

Clustering is an unsupervised learning process of exploring unlabeled input data of the form X={x_1, x_2,…, x_n} $\subset \Re^N$ where $x_i \in R^N$ represent a vector object and x_{ij} its jth feature. Clustering may be found under different appellations in different contexts, such as unsupervised learning in pattern recognition, numerical taxonomy in biology and typology in social science [1]. It has been widely applied in several different fields and various disciplines [2-4].

Several clustering algorithms are proposed in the literature. The most widely used clustering algorithm is FCM originally proposed by Bezdek [5]. Based on Fuzzy set theory which models uncertainty of belonging, this algorithm partitions the considered objects such as similar objects are in the same cluster and dissimilar objects belong to different clusters. Hence the crucial need for FCM of an appropriate way of measuring similarities between pairs of objects.

One popular way to cluster a data set is to define a distance measuring similarity between pairs of objects. Gustafson and Kessel [6] have generalized the FCM algorithm and used an adaptive distance measure to detect ellipsoidal structures of the clusters. However, this algorithm needs added constraint and can only be used for a specific data [7]. Gath and Geva [8] have defined an «exponential» distance measure. This algorithm performs when clusters are spherical or ellipsoidal. Other distance measures were conceived in the literature such Sorensen [9] or Bray-Curtis[10], Canberra, Gower, Spearman or Squared Euclidean, Mahalanobis [11] and Minkowski ($p \geq 1$) distances. Hathaway [12] had proposed to generalize the use of the Lp Norm Distances and given examples with $0.5 \leq p$.

In this paper, we show that using fractional metric with $0 < p < 0.5$ in clustering context improves results when compared to $0.5 \leq p$ or the usual distances, especially when there are outliers. Six data sets including three data sets with outliers are tested. Our approach consists to repeat FCM using different values of p between two chosen values: 0,01 and 30. We also search if there is a correlation between characteristic and value of p that provide best results.

The remainder of the paper is organized as follows. The next section presents p-metric and a succinct presentation of FCM. Test results are detailed and discussed in section 3. Conclusions are given in section 4.

II. Related Work

The fuzzy c-means proposed by Bezdec [5] is based on Euclidean distance. This clustering algorithm presented a drawback when the data contain outliers.

This section introduces the fuzzy c-mean and presents some research that overcome the inconvenience of Euclidean distance.

A. *Fuzzy c-means*

FCM is a generalization of K-means algorithm which is a hard clustering algorithm. K-means assigns each vector object

x_i to a unique cluster with a degree of membership equal to one. As a consequence, clusters are disjointed and have well-defined boundaries. In contrast, FCM assigns each data point to every cluster with different degrees of membership, and boundaries between generated clusters may be vague [5]. FCM is more successful compared to the crisp clustering with overlapping and not well separated clusters.

A partition of $X=\{x_1, x_2,\ldots, x_n\} \subset \Re^N$ into c fuzzy clusters can be defined by a fuzzy membership matrix $U=[u_{ik}]$ satisfying these three conditions[13]:

$$0 \le u_{ik} \le 1;\quad 1 \le i \le c;\quad 1 \le k \le n \tag{1}$$

$$0 < \sum_{k=1}^{n} u_{ik} < n;\quad 1 \le i \le c \tag{2}$$

$$\sum_{i=1}^{c} u_{ik} = 1;\quad 1 \le k \le n \tag{3}$$

where u_{ik} is the degree to which the pattern x_k belongs to the i^{th} cluster ($1 \le i \le c$ and $1 \le k \le n$).

The first constraint reflects the generalization of the characteristic function which assumes values in {0, 1}. For a given vector object, a value close to 1 indicates a high grade of belonging to the cluster. Inversely, value close to 0 indicates a low grade of belonging to the cluster. The second constraint guaranties that no cluster is empty or totally equal to X. The last constraint assures that the membership of each object is distributed over all the c clusters.

FCM is an iterative procedure that optimizes an objective function J_m. This objective function depends on the distances of the data to the cluster centres weighted by the membership degrees. By varying the distance function, different forms of cluster in data sets can be detected. The objective function J_m is defined by:

$$J_m(U,V;X) = \sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik})^m d^2(x_k, v_i) \tag{4}$$

where:

- m ($1 < m < \infty$) is a weighting exponent used to control the relative contribution of each object vector xi and the fuzziness degree of the final partition.
- $V= (v_1, v_2, \ldots , v_c)$ represents a c-tuple of prototypes, each prototype characterizes one of the c clusters.
- $d(x_k,v_i)$ is the distance between the i^{th} prototype and the k^{th} data point.

Bezdek proved that FCM converges to an approximate solution under two conditions [13]:

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d(x_k,v_i)}{d(x_k,v_j)} \right)^{2/m-1} \right]^{-1};\ 1 \le i \le c; 1 \le k \le n \tag{5}$$

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m};\ 1 \le i \le c \tag{6}$$

The pseudo-code of FCM algorithm is given in Table I [**13].**

TABLE I. FCM ALGORITHM

```
Store unlabeled Dataset $X=\{x_1, x_2, \ldots, x_n\} \subset \Re^p$;
Choose
    • $1<c<n$;
    • $m>1$;
    • $t_{max}$ (iteration limit);
    • the $\varepsilon$ (tolerance bound);
    • norm for clustering criterion $J_m$;
    • norm for termination error $E_t=\|V_t-V_{t-1}\|_{err}$;
Initialize
    • prototypes $V_0= (v_{1,0}, v_{2,0}, \ldots, v_{c,0}) \in \Re^{cxp}$
    • $t=0$; (iteration index)
do { t++;
    • Calculate $U_t$ using $V_{t-1}$ and (Eq.5);
    • Calculate $V_t$ using $U_t$ and (Eq.6);
} while ($\|V_t-V_{t-1}\|_{err} > \varepsilon$) and ($t < t_{max}$) );
$U^* = U_t$; $V^* = V_t$;
Use $U^*$ and/or $V^*$;
```

B. Distance Measures

Mathematically, a distance measure "d" on a set of points E is a function d: E x E→R+ such as d(x, y) between two points x and y should satisfy the following conditions:

$$d(x,y) = 0 \Leftrightarrow x = y \quad \forall x, y \tag{7}$$

$$d(x, y) = d(y, x) \quad \forall\ x, y \tag{8}$$

$$d(x, z) \le d(x, y) + d(y, z) \quad \forall\ x, y, z \tag{9}$$

The most commonly used distance measure is the p-metric defined for two points x, y in $\Re^k$ by:

$$d(x,y)=\left(\sum_{j=1}^{k}\left|x_j-y_j\right|^{p}\right)^{\frac{1}{p}} \qquad (10)$$

The L_p distance measure is a metric for $p \geq 1$ but not for p<1 because the triangular inequality property $d(x,z)\leq d(x,y)+d(y,z)\ \forall\ x,y,z$ is not satisfied for p<1 [14]. However, this property is not necessarily required for clustering tasks.

The p metric depends on the parameter "p" ($p\in\Re^{+}$) called exponent of the metric, and covers the Minkowski distance ($p >= 1$) and fractional metrics ($p<1$).

Moreover, from equation (10) it is easy to see that the Euclidean distance, L_2, the Manhattan or city block distance, L_1, and the Chebychev distance, L_∞, are particular cases of L_p distance.

The Euclidean distance ($p=2$) is often used in spaces with two or three dimensions, but it creates a problem in case of large dimensions. Besides, it's strongly influenced by the larger units of measure and varies with the scale of each feature [15]. To deal with this problem, some authors proposed to calculate the Euclidean distance after centering, reduction or normalization of variables [2-4].

Manhattan distance ($p=1$) between two vectors is computed by summing the absolute value of the difference on each dimension. Schematically, it consists to determine the distance that would be travelled to get from one point to the other if a grid-like path is followed. It is significantly less costly to calculate than Euclidean distance that requires taking a square root.

Chebychev distance, or L_∞ metric, is also known as "chessboard" distance. It returns the maximum distance among coordinates of a pair of objects.

Spearman distance is the square of the Euclidean distance. It's easier to calculate than the Euclidean distance.

Canberra distance is used where elements in the vector are non-negative. As defined, individual elements in distance could have zero for the numerator or denominator.

Bray-Curtis or Sorensen distance is also called ecological distance. However, this measure does not satisfy the triangle inequality axiom, and then is not a true distance.

III. THE PROPOSED TECHNIQUE

A clustering algorithm can lead to different clusters. The selection of a distance measure may affect the final results, especially when data contain outliers. Our technique consist to repeat FCM algorithm using the measure Lp with different values of p as a measure of similarity among objects. The aim of these experiments is to search the impact of p on the obtained clusters and to predict the optimal set for p that can improve significantly the computational performance of fuzzy clustering. For this, we change the values of the coefficient p between two random chosen limits: 0,01 and 30.

Generally, the use of a fractional norm in clustering context reduces the impact of extreme individual attribute differences [16]. But with p > 1, L_p distances emphasise the larger attribute dissimilarities between two vectors. An example for this is illustrated by Figure 1. This figure represents the first quadrant plot of unit length loci from the origin with different values of p.

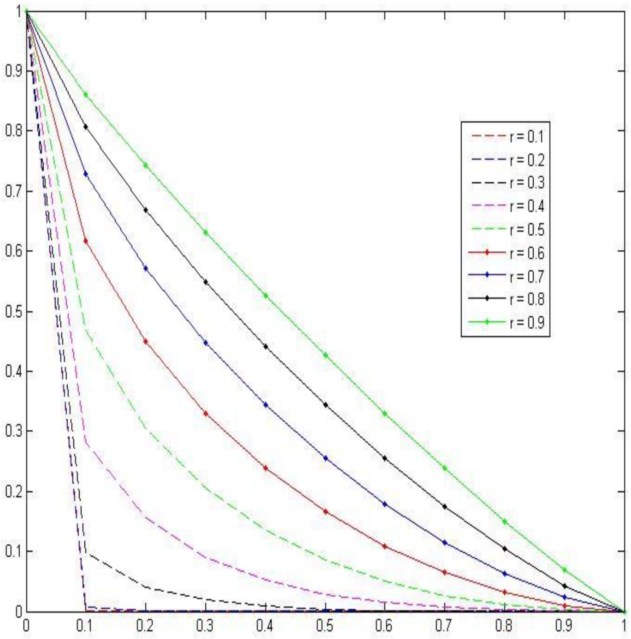

Figure 1 Representation of the Lp according to value of p $(0.1 \leq p \leq 0.9)$

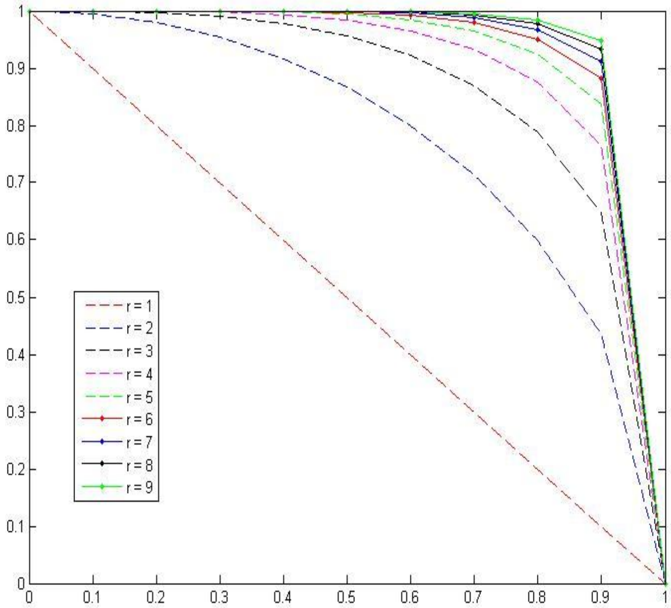

Figure 2 Representation of the Lp according to value of p $(1 \leq p \leq 0.9)$

IV. Numerical Results and Discussion

In order to illustrate the impact of the parameter p on clustering, experiments are conducted on six datasets available from the UCI Machine Learning Repository [17]: Iris, Wine, BCW, Spect Heart, BreastTissu and Indian.

These bases are supervised, but any information about classes is given to the algorithm. Thus, it is possible to determine the number of misclassified objects, and then the recognition rate.

Table II describes the type of data and gives information about attributes, size and number of classes.

To implement FCM, the values of the parameters should set up in advance. They consist of the following items:

- The parameter m=2.
- The maximum of iteration's number is 500.
- ε = 0.00001

TABLE II. DESCRIPTION OF 10 DATASETS

Dataset	Instances	Attributes	Classes
Iris	150	4	3
BCW	699	9	2
Wine	178	13	3
Heart	267	22	2
BreastTissu	106	9	6
Indian	583	10	2

The representation on the plan of those datasets (figure 3, figure 4 and figure 5) shows that Iris, BCW and Heart do not contain outliers. Whereas the others datasets (figure 6, figure 7 and figure 8) contain outliers.

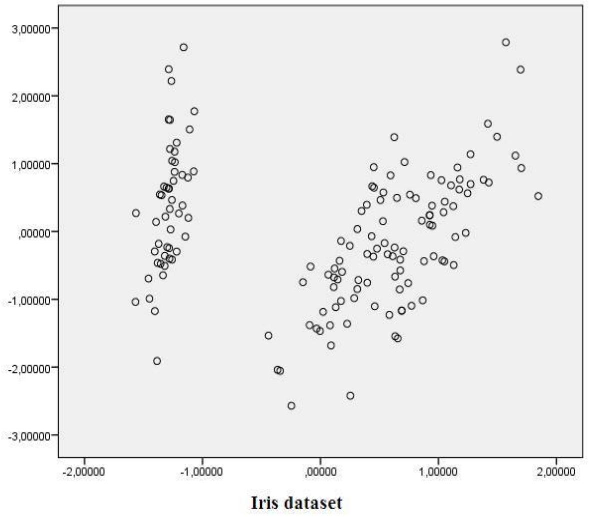

Figure 3 Representation of the IRIS without outliers

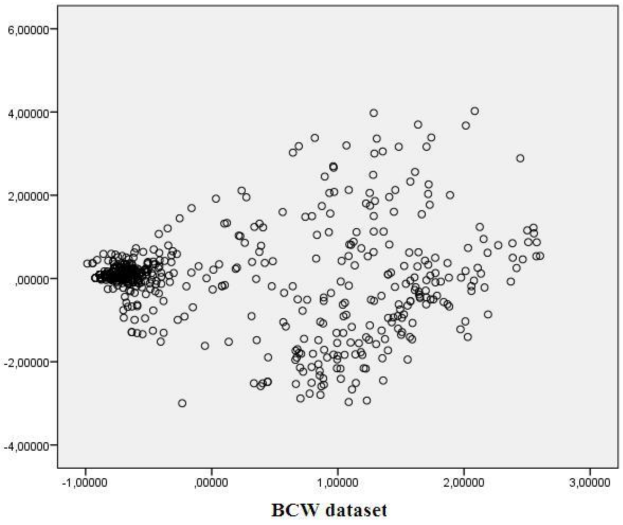

Figure 4 Representation of the BCW without outliers

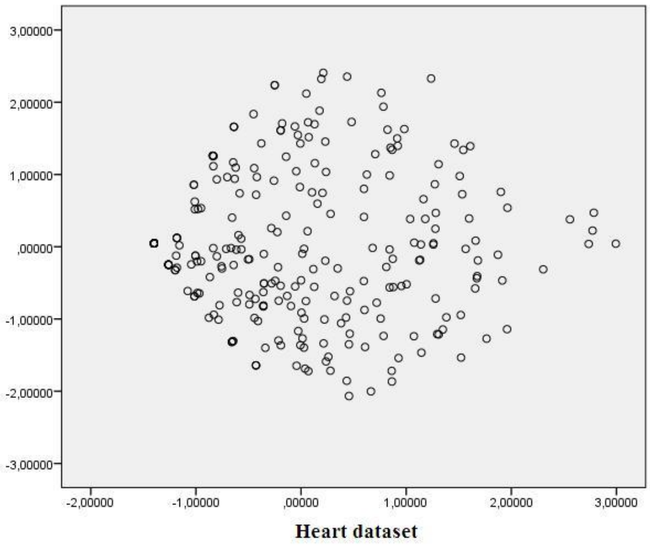

Figure 5 Representation of the Heart without outliers

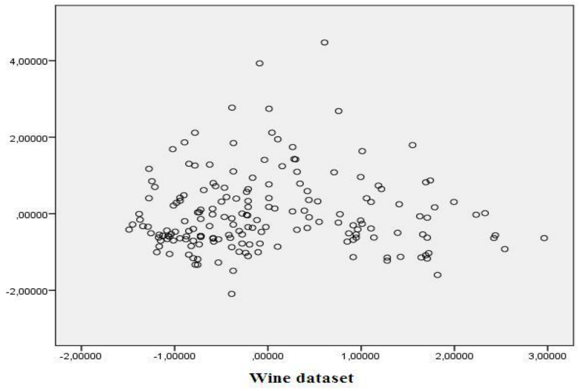

Figure 6 Representation of the Wine dataset with outliers

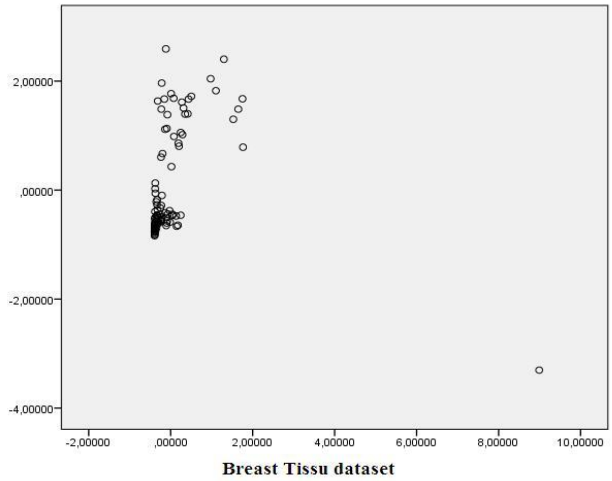

Figure 7 Representation of the Wine dataset with outliers

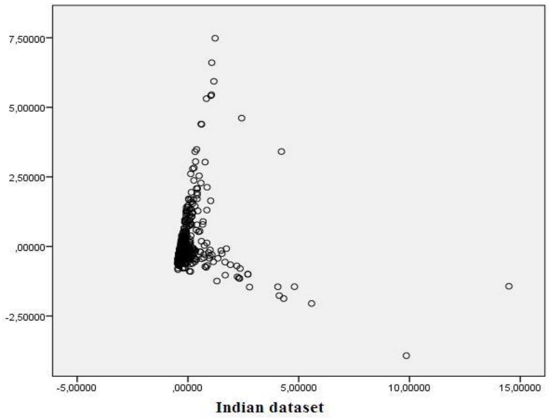

Figure 8 Representation of the Wine dataset with outliers

We repeat this algorithm by using different values of p between two arbitrarily chosen limits: 0,01 and 30. However, best results had always obtained for p between 0,01 and 11.

In table 3, our results confirm those of some work on the limit of the Euclidean distance to solve the problems of classification [3,15]. Euclidian distance gave best recognition rate 89.34% for Iris dataset, but the same rate is also obtained for the values 5, 6, 7 and 8 of p.

Moreover, some authors suggest using Euclidian distance in low dimensional spaces and Manhattan or fractional metrics in high dimensional spaces [3]. But our results contradict this. For example, dimension is 9 in BCW data, and the recognition rate given by Euclidian distance is 95.43%, that outperforms 94.14% obtained with Manhattan distance. Also, Heart Data has 22 features, and recognition rate obtained by Euclidian distance is 58.06% whereas Manhattan has given 56.18%.

The results show that there isn't any relation between the value of p that provide best result, and the dimension of data. For example, Breast Tissu and BCW datasets have both 9 attributes, but best result for the first is obtained with p= 0.1 and p =0.4, whereas Chebychev distance (L_∞) allows better result for the second dataset.

We considered a possible influence of the number of classes. But results were not encouraging. Indeed, Iris and Wine datasets had both 3 classes, but best result for the Iris is obtained for values 5,6,7 and 8 of p. Wine had best result for p=0.2.

We considered a possible relation between p and variable's correlation. But the result was not encouraging. For example, both the Indian and BreastTissu datasets have best results for p<1, but the correlation is medium for the first dataset and high for the second dataset.

This mentioned, representing data in the plan shows that $p \geq 1$ gives good results when the data sets do not contain outliers (Table III). Whereas the values of p<1 gives good results when there are outliers in data sets. This was confirmed by the obtained results (Table III), and extends the previous results [16] to fuzzy clustering framework.

TABLE III. RECOGNITION RATE FOR USUAL DISTANCES AND LP METRIC ON DATASETS WITHOUT OUTLIERS.

	Iris	**BCW**	**Heart**
Best recognition rate (with p correspondent)	89.34% (p = 2 and p for 5 to 8)	96.71% (p= 11)	**84.27%** (p = 5 – 6 - 7- 8)
Low recognition rate (with p correspondent)	84.67% (p = 0.02 - 0.03 - 0.04 - 0.05 - 0.06)	88.7% (p = 0.03 - 0.04)	50.19% (p = 12)
Manhattan	88.67%	94.14%	56.18%
Euclidian	89.34%	95.43%	58.06%
Chybechev	88.67%	**97%**	52.06%
Canberra	**94.67%**	95.28%	*
BrayCurtis	88%	**97%**	64.05%

TABLE IV. RECOGNITION RATE FOR USUAL DISTANCES AND LP METRIC ON DATASETS WITH OUTLIERS.

	Wine	BreastTissu	Indian
Best recognition rate (p correspondent)	93.83% (p = 0.2)	46.23% (p = 0.09- 0.1 - 0.4)	**55.58%** (p = for 0.01 to 0.09)
Low recognition rate (p correspondent)	69.11% (p = 0.9)	22.65% (p = 14)	28.99% (p = for 3 to 15)
Manhattan	73.04%	27.36%	30.37%

Euclidian	69.67%	30.19%	30.37%
Chybechev	69.67%	27.36%	28.99%
Canberra	**94.95%**	**52.84%**	*
BrayCurtis	71.92%	46.23%	*

In the table IV, the Canberra distance give the best rate for a specific data like Iris, Wine and BreasTissu. However, this distance is not used for Indian and Heart data sets.

V. Conclusion

Clustering analysis technique has an important role in data analysis. However it depends on the concept of dissimilarity (or distance). The choice of this is highly dependent on the data itself; and generally, there is no prior information in the unsupervised context. Several distances were proposed in the literature. However, Fractional metric ($p < 1$) has rarely been used in the clustering tasks.

In this paper, we show that values of the parameter p less than 1can improve significantly the performance of FCM, especially when the data set contains outliers. This study gives encouraging results. Future work could be done, with the relation between the value of p and the entropy measure.

References

[1] M. Halkidi, Y.Batistakis, and M.Vazirgiannis, “On clustering validation techniques”. Journal of Intelligent Information Systems, 17, pp.107-145, 2001.

[2] S-H.Cha, C.Tappert, and S.Yoon, “Enhancing Binary Feature Vector Similarity Measures”, Journal of Pattern Recognition Research 1, pp.63-77, 2006.

[3] W. M Rand, “Objective criteria for the evaluation of clustering methods”, Journal of the American Statistical Association, 66 (336, pp. 846–850), 1971.

[4] K. West, S.Cox, and P.Lamere, “Incorporating machine learning into music similarity estimation”. In Proceedings of the 1st ACM workshop on Audio and music computing multimedia, ACM. pp. 89–96, 2006.

[5] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, 1981.

[6] D.E. Gustafson and W.C. Kessel, “Fuzzy clustering with a fuzzy covariance matrix”. Proc. IEEE CDC, San Diego, CA, pp.761-766, 1979.

[7] J-C Tseng, “Clustering Accuracies on Concepts of Nursing”, In Journal of Public Health Frontier, Vol. 2 Iss. 3, pp. 133-140, Sept. 2013.

[8] I.Gath and A.B. Geva, “Unsupervised optimal fuzzy clustering”. IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. 11, pp. 73-781, 1989.

[9] T. Sorensen, “A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons”. Biologiske Skrifter /Kongelige Danske Videnskabernes Selskab, 5 (4). pp. 1-34, 1948.

[10] J. R. Bray, J. T.Curtis, “An ordination of the upland forest of the southern Winsconsin”. Ecological Monographies, 27. pp. 325-349, 1957.

[11] H-C Liu, B-C Jeng, J-M Yih and Y-K Yu, “Fuzzy C-Means Algorithm Based on Standard Mahalanobis Distances”, In Proceedings of the 2009 International Symposium on Information Processing (ISIP’09), Huangshan, P. R. China, August 21-23, pp. 422-427, 2009.

[12] R. J. Hathaway, J.C. Bezdek, and Y. Hu, “Generalized Fuzzy c-Means Clustering Strategies Using Lp Norm Distances”, IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 8, NO. 5, OCTOBER 2000.

[13] A.Bouroumi, M.Limouri, and A.Essaïd, “Unsupervised Fuzzy Learning and Cluster Seeking”, Intelligent Data Analysis, vol. 4, no. 3-4, pp. 241-253. 2000.

[14] D. François, V.Wertz, and M.Verleysen. “The concentration of fractional distances”. IEEE Transactions on Knowledge and Data Engineering, Vol 19, N° 7, July 2007.

[15] E. Guaus, “Audio content processing for automatic music genre classification: descriptors, databases, and classifiers”. PhD thesis, Universitat Pompeu Fabra. 2009.

[16] K.A.J.Doherty, R.G.Adams and N.Davey, “Non-Euclidean Norms and Data Normalisation”, ESANN'2004 proceedings - European Symposium on Artificial Neural Networks, Bruges (Belgium), pp. 181-186. 28-30 April 2004.

[17] A.Asuncion and D. J. Newman, “UCI Machine Learning Repository,” Irvine, CA: University of California, School of Information and Computer Science, 2007. Available: http://archive.ics.uci.edu/ml/datasets.html.

[18] D. H. T. Clifford and W. Stephenson, “An Introduction to Numerical Classification”. New York: Academic, 1975.