

Kernel-Based Fuzzy Clustering of Interval Data

Bruno A. Pimentel
Centro de Informatica/UFPE
Recife, Brazil
Email: bap@cin.ufpe.br

Anderson F. B. F. da Costa
Centro de Informatica/UFPE
Recife, Brazil
Email: afbfc@cin.ufpe.br

Renata M. C. R. de Souza
Centro de Informatica/UFPE
Recife, Brazil
Email: rmcrs@cin.ufpe.br

Abstract—Kernel clustering methods have been very important in application of non-supervised machine learning to real problems. Kernel methods possess many advantages other than non-linearity such as modularity, ability to work with heterogeneous descriptions of data, incorporation of prior knowledge etc. In this paper, we present a clustering method based on kernel functions for partitioning a set of interval-valued data. In addition, this method is compared to a fuzzy partitioning approach for interval data introduced previously. Experiments with real and syntectic symbolic interval-valued data sets are presented. The evaluation of the clustering results furnished by the methods is performed regarding the computation of an external cluster validity index and the global error rate of classification.

Index Terms—Kernel, clustering, interval-valued data, fuzzy.

I. INTRODUCTION

With the growing interest in automatically to understand, process and summarize data, several application domains like pattern recognition, machine learning, data mining, computer vision, computational biology and classification of statistical findings in social studies have used clustering algorithms [1] [2]. The objective of the cluster analysis is to group a set of elements into clusters such that elements within a cluster have a high degree of similarity, while elements belonging to different clusters have a high degree of dissimilarity. Clustering techniques may be divided into hierarchical clustering and partitioning clustering: hierarchical clustering groups data objects with a sequence of partitions dividing one group with all individuals in singleton clusters or vice versa, while partitioning clustering directly divides data objects into some fixed number of clusters with no hierarchical structure [3] and makes this division based on the optimization of a suitable objective function.

The partitioning methods may be classified in hard partitioning and fuzzy partitioning. The hard partitioning associates each element of the data to a single class, while in the fuzzy partitioning each element of the data is associated with all classes of the partition, the elements not belong a single cluster only and so that the memberships are calculated. When the memberships have values between 0 and 1, the sum of the all possibilities of an element belong to some classes is equal to 1 and the method is the type probabilistic [4]. But when the memberships have values 0 or 1, the method is the type crisp and the fuzzy method becomes equivalent to a hard method [5].

Kernel clustering methods have been very important in the

application of non-supervised machine learning to real problems. These methods are able to produce nonlinear separating hyper surfaces among clusters [6]. The use of kernels allows to map implicitly data into a high dimensional space, called feature space; computing a linear partitioning in this feature space results in a nonlinear separation between clusters in the input space. Several clustering methods has been modified including the use of the kernel functions such as SOM [7], k-means [8], fuzzy c-means (FCM)[9] and so on. The kernel k-means was proposed [10] as an extension of the standard hard partitioning method k-means and the kernel fuzzy c-means was proposed [11] aiming to replace a kernel-induced distance metric for the original Euclidean distance of the fuzzy partitioning method.

In classical cluster analysis, data are often represented as arrays of quantitative or qualitative values, where each column represents a variable. In particular, each individual takes a single value for each variable. However, research study has shown that this presentation is not as rich in information to represent the complexity of data such as intervals or distributions found in real problems [12]. The main goal of *Symbolic Data Analysis* (SDA) concerns to model more complex information through a set-valued (interval or multi-valued) or modal (weight or probability distribution) variables also termed symbolic and extend classical data analysis techniques, such as, clustering, factorial techniques, decision trees and other to symbolic data. These data allow to take account variability and/or uncertainty, making the symbolic data analysis more comprehensive than the classical data analysis.

In this paper, we present a new clustering method based in kernel functions for partitioning a set of interval-valued data. This method is an extension of the fuzzy c-means cluster algorithm based on kernelization of metric [11] to interval-valued data. In addition, this method is compared with a fuzzy c-means method for interval data introduced in [13]. The paper is organized as follows. In Section 2, we describe the fuzzy c-means clustering method for symbolic interval data. In Section 3, we introduce the new kernel-based fuzzy clustering method for symbolic interval data. In order to validate this new method, Section 4 presents experiments with real and syntectic symbolic interval data sets. The evaluation of the clustering results furnished by the methods is based on the computation of an external cluster validity index and the global error rate of classification. In Section 5, the concluding remarks are given.

II. INTRODUCTION TO A FUZZY C-MEANS METHOD FOR INTERVAL DATA

This section initially shows a brief description of the fuzzy c-means method for symbolic interval data [13] (here called FCM-IV). This method is a suitable extension of the standard fuzzy c-means clustering algorithm that furnishes a fuzzy partition and a prototype for each cluster by optimizing an adequacy criterion based on a suitable squared Euclidean distance between vectors of intervals.

Consider Ω a set of n objects indexed by i and formed by p symbolic interval-valued variables indexed by j . Each object i is represented by a vector of intervals $\mathbf{x}_i = ([a_i^1, b_i^1], \dots, [a_i^p, b_i^p])^T$. Similar to the object, the prototype of a cluster C_k ($k = 1, \dots, K$) is also represented as a vector of intervals $\mathbf{y}_k = ([\alpha_k^1, \beta_k^1], \dots, [\alpha_k^p, \beta_k^p])^T$. Thus the partition $P = (C_1, \dots, C_K)$ from Ω in K clusters where each cluster has a prototype L_k described by the vector $\mathbf{y}_k$.

The problem of clustering concerns to find a partition of Ω into K disjoint clusters $C_1, \dots, C_K$ such that minimizes locally a clustering criterion defined as

$$J(P, L) = \sum_{k=1}^K \sum_{i \in C_k} (u_{ik})^m d^2(x_i, y_k). \quad (1)$$

where the squared Euclidean distance d^2 measuring the dissimilarity between a pair of vectors of intervals is given as

$$d^2(x_i, y_k) = \sum_{j=1}^p \left[(a_i^j - \alpha_k^j)^2 + (b_i^j - \beta_k^j)^2 \right]. \quad (2)$$

The term u_{ik} is the membership degree of pattern i in cluster C_k and $m \in]1, +\infty[$ is a parameter that controls the fuzziness of membership for each pattern i . The membership degree u_{ik} ($i = 1, \dots, n$) of each pattern i in each cluster C_k , minimizing the clustering criterion J under $u_{ik} \geq 0$ and $\sum_{k=1}^K u_{ik} = 1$, is updated according to the following expression:

$$u_{ik} = \left[\sum_{h=1}^K \left(\frac{\sum_{j=1}^p \left[(a_i^j - \alpha_h^j)^2 + (b_i^j - \beta_h^j)^2 \right]}{\sum_{j=1}^p \left[(a_i^j - \alpha_k^j)^2 + (b_i^j - \beta_k^j)^2 \right]} \right)^{\frac{1}{m-1}} \right]^{-1}. \quad (3)$$

The prototype $\mathbf{y}_k = ([\alpha_k^1, \beta_k^1], \dots, [\alpha_k^p, \beta_k^p])^T$ of C_k which minimizes the clustering criterion J has the bounds of each interval $[\alpha_k^j, \beta_k^j]$ ($j = 1, \dots, p$) updated according to the following expressions:

$$\alpha_k^j = \frac{\sum_{i=1}^n (u_{ik})^m a_i^j}{\sum_{i=1}^n (u_{ik})^m} \quad (4)$$

$$\beta_k^j = \frac{\sum_{i=1}^n (u_{ik})^m b_i^j}{\sum_{i=1}^n (u_{ik})^m} \quad (5)$$

The algorithm schema of the FCM-IV algorithm for interval data is as follows:

SCHEMA OF THE FCM-IV ALGORITHM

1) Initialization step

The parameters c ($2 \leq c \leq n$), m ($1 < m < \infty$) and $\varepsilon > 0$ are fixed.

1.1 Initialize u_{ik} ($i = 1, \dots, n$) and ($k = 1, \dots, K$) such that $u_{ik} \geq 0$ and $\sum_{k=1}^K u_{ik} = 1$.

1.2 Compute J^0 using Eq. 1 and do $t = 1$.

2) Representation step

Compute the prototype $\mathbf{y}_k$ of the cluster C_k ($k = 1, \dots, K$) using the Eqs. 4 and 5.

3) Fuzzy membership step

Update the fuzzy membership degree u_{ik} of the pattern i belonging to cluster C_k ($k = 1, \dots, K$) using the Eq. 3.

4) Stopping criterion

If $|J_t - J_{t-1}| \leq \varepsilon$, STOP; else update $t = t + 1$ and go step 2.

III. A KERNEL FUZZY C-MEANS METHOD FOR INTERVAL DATA

The kernel fuzzy c-means (KFCM) method is an extension of the fuzzy c-means (FCM) cluster algorithm proposed in [11]. KFCM adopts a new kernel-induced metric in the data space to replace the original Euclidean norm metric in FCM, thus KFCM can implicitly perform a nonlinear mapping to a high dimensional feature space.

This section presents an extension of the kernel fuzzy c-means method to interval-valued data. The main idea of this method (here called KFCM-IV) consist in manipulating a kernel function for interval data and to use the approach fuzzy for generate a fuzzy partitions based on the idea of partial membership expressed by the degree of membership of each pattern in a given cluster.

Regarding interval data, each object i ($i = 1, \dots, n$) of Ω is represented as a vector of intervals $\mathbf{x}_i = ([a_i^1, b_i^1], \dots, [a_i^p, b_i^p])^T$ and a nonlinear function $\xi : \mathbf{x}_i \rightarrow \xi(\mathbf{x}_i)$ performs the mapping from the input space to a high dimensional feature space H . Consider k centers $\mathbf{g}_1^\xi, \dots, \mathbf{g}_K^\xi$ as prototypes of the clusters where $\mathbf{g}_k^\xi \subset H$ ($k = 1, \dots, K$).

Using the classic kernel fuzzy c-means methodology, the problem of clustering concerns to find a partition of Ω into K disjoint clusters $C_1, \dots, C_K$ such that minimizes a clustering error defined as:

$$E = \sum_{k=1}^K \sum_{i \in C_k} u_{ik}^m \|\xi(\mathbf{x}_i) - \xi(\mathbf{v}_k)\|^2 \quad (6)$$

The squared Euclidean distance between $\xi(\mathbf{x}_i)$ and $\xi(\mathbf{v}_k)$ is given by:

$$\|\xi(\mathbf{x}_i) - \xi(\mathbf{g}_k)\|^2 = K(\mathbf{x}_i, \mathbf{x}_i) - 2K(\mathbf{x}_i, \mathbf{v}_k) + K(\mathbf{v}_k, \mathbf{v}_k) \quad (7)$$

The Gaussian kernel function $K_{il} = K(\mathbf{x}_i, \mathbf{x}_l)$ for interval data is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_l) = \exp(-\|\mathbf{x}_i - \mathbf{x}_l\|^2 / 2\delta^2) \quad (8)$$

with

$$\|\mathbf{x}_i - \mathbf{x}_l\|^2 = \sum_{j=1}^p \left[(a_i^j - a_l^j)^2 + (b_i^j - b_l^j)^2 \right]$$

The membership u_{ik} of the object x_i in class C_k assuming values in the interval $[0, 1]$ and satisfying $\sum_{k=1}^K u_{ik} = 1$ is defined as:

$$u_{ik} = \frac{(1/(1 - K(x_i, v_k)))^{1/(m-1)}}{\sum_{l=1}^K (1/(1 - K(x_i, v_l)))^{1/(m-1)}} \quad (9)$$

The prototypes are given by:

$$v_k = \frac{\sum_{i=1}^n u_{ik}^m K(x_i, v_k) x_i}{\sum_{i=1}^n u_{ik}^m K(x_i, v_k)} \quad (10)$$

The weighting exponent m is called the fuzzifier and it can have an influence on the clustering performance of the method. The algorithm schema of the KFCM-IV algorithm for interval data is as follows:

SCHEMA OF THE KFCM-IV ALGORITHM

- 1) **Pre-processing step: compute a kernel matrix**
Compute the matrix $\mathbf{K}$ using the Eq. 8 for $i = 1, \dots, n$ and $l = 1, \dots, n$.
- 2) **Initialization step**
The parameters c ($2 \leq c \leq n$), m ($1 < m < \infty$) and $\varepsilon > 0$ are fixed.
 - 1.1 Initialize u_{ik} ($i = 1, \dots, n$) and ($k = 1, \dots, K$) such that $u_{ik} \geq 0$ and $\sum_{k=1}^K u_{ik} = 1$.
 - 1.2 Compute the criterion error E^0 using Eq. 6 and do $t = 1$.
- 3) **Representation step**
Update the prototypes v_k ($k = 1, \dots, K$) with Eq. 7.
- 4) **Membership step**
Update the memberships u_{ik} ($k = 1, \dots, K$) and ($i = 1, \dots, n$) with Eq. 6.
- 5) **Stopping criterion**
If $|E_{t+1} - E_t| \leq \varepsilon$ then STOP calculate error criterion E using the Eq. 5, otherwise do $t = t + 1$ and go to step 3.

IV. EXPERIMENTAL RESULTS

In order to validate the proposed method for interval data, this section presents the experimental results regarding synthetic and real symbolic data sets. Our aim is to achieve a comparison between this method and the fuzzy c-means method introduced in the Section II. The evaluation was performed in the framework of a Monte Carlo experience with 100 replications for each synthetic interval data set. Concerning real symbolic data, an application with two data sets is also considered.

To assess the clustering results provided by FCM-IV and KFCM-IV methods, we used the corrected Rand index (CR) and the overall error rate of classification (OERC). The index CR [14] takes its values in the interval $[-1, 1]$ where the value 1 indicates a perfect agreement between the partitions, values close to 0 correspond to more or less random cluster agreement, and negative values to insufficient recovery properties.

The OERC takes its values in the interval $[0, 1]$ where the value 0 indicates a perfect agreement between the partitions, unlike the value 1.

In all experiments for KFCM-IV method, it was used the Gaussian kernel defined in Eq. 8 with parameter $\sigma = 1$. For FCM-IV and KFCM-IV methods, it used the parameter m equal to 2.

A. The Monte Carlo experiment results and comparison

First, we simulated two classical quantitative data sets in $\mathbb{R}^2$ with nonlinear clusters. Thus, interval data sets is built from these quantitative data sets, each point (y_1, y_2) of these data sets is considered as the 'seed' of a rectangle. Each rectangle is therefore a vector of two intervals defined by:

$$([y_1 - \gamma_1/2, y_1 + \gamma_1/2], [y_2 - \gamma_2/2, y_2 + \gamma_2/2])$$

The parameters γ_1 and γ_2 are the width and the height of the rectangle. They are drawn randomly within a given range of values.

Data set 1 is formed by two rings, each representing a class. The first ring has radius equals to 12 and 100 individuals. The second ring has radius equals to 5 and 50 individuals. From this data, we can observe that the classes are nonlinearly separable and similar points sets can be found in [10]. Figure 1 represents an interval data set with two rings using width and the height of all the rectangles drawn randomly within the interval $[1, 2]$.

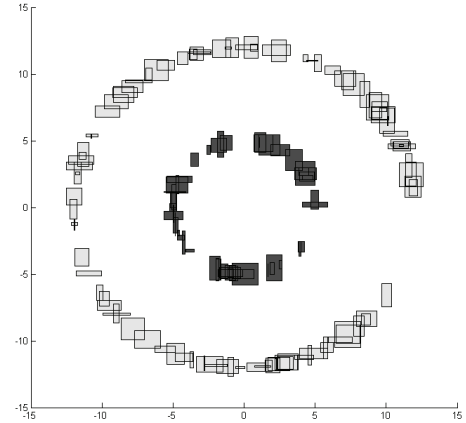


Fig. 1. Symbolic interval data set 1: two rings

Data set 2 is formed by data follow a similar training to Archimedean spiral, where each arm of the spiral is defined in polar coordinates (r, θ) by the equation:

$$r = a + b\theta \quad (11)$$

where a and b are real numbers and θ in angular measurements. The parameter a controls the initial radius of the each arm of the spiral, while b is related with distance between successive turns. This data set has 90 individuals distributed among the classes contained 15, 20, 25 and 30. The total angle $\theta = 120$ degrees with $a = 10$ and $b = 5$. Figure 2 represents the interval data set 2 width and the height of all the rectangles drawn randomly within the interval $[1, 5]$.

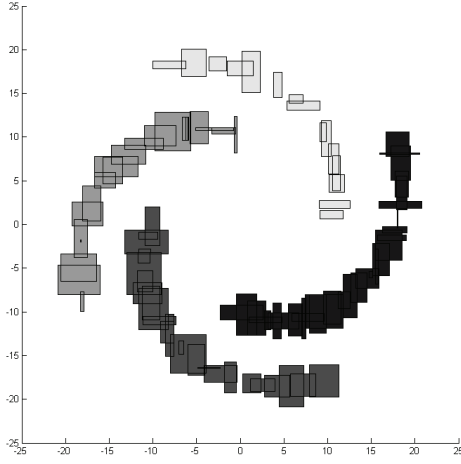


Fig. 2. Symbolic interval data set 2: spiral of four classes

In the framework of a Monte Carlo experiment, 100 replications of the previous process were carried out for parameters γ_1 γ_2 drawn randomly 100 times from each of the following intervals: $[1, 2]$, $[1, 5]$, $[1, 10]$, $[1, 15]$ and $[1, 20]$. This process has also been repeated for interval data sets 1 and 2. The KFCM-IV and FCM-IV methods have been performed on these data sets. The cluster partition obtained with these clustering methods were compared with the class partition known a priori based on CR index and the OERC metric. For each 100 replications, the average CR index and the average OERC are calculated.

Tables I and II give the values of the average and standard-deviation (in parenthesis) of the CR index and OERC for the FCM-IV and KFCM-IV methods and the data sets 1 and 2, respectively. As expected, the KFCM-IV method outperforms the FCM-IV one considering CR index and OERC metric.

TABLE I
COMPARISON OF THE CLUSTERING METHODS FOR DATA SET 1

Predefined Interval	FCM-IV		KFCM-IV	
	CR index	OERC	CR index	OERC
[1,2]	0.0000 (0.0114)	0.4771 (0.0345)	0.4391 (0.0665)	0.1352 (0.0155)
[1,5]	-0.0001 (0.0099)	0.4767 (0.0354)	0.4401 (0.0756)	0.1359 (0.0177)
[1,10]	0.0001 (0.0113)	0.4776 (0.0354)	0.3725 (0.0573)	0.1516 (0.0132)
[1,15]	0.0012 (0.0134)	0.4769 (0.0389)	0.2695 (0.0615)	0.1742 (0.0140)
[1,20]	-0.0005 (0.0091)	0.4782 (0.0328)	0.1609 (0.0787)	0.1996 (0.0176)

B. Results and comparison with a city-temperature interval data set

This interval data set [15] concerns 37 cities, each city is described by 12 interval variables which are minimum and the maximum temperatures of 12 months in degree centigrade. Table III shows part of this data set.

The 4-cluster partitions obtained with these clustering methods were compared with the 4-cluster partition known a priori.

TABLE II
COMPARISON OF THE CLUSTERING METHODS FOR DATA SET 2

Predefined Interval	FCM-IV		KFCM-IV	
	CR index	OERC	CR index	OERC
[1,2]	0.3207 (0.0030)	0.5482 (0.0285)	0.3814 (0.0412)	0.2809 (0.0274)
[1,5]	0.3204 (0.0028)	0.5481 (0.0318)	0.3826 (0.0425)	0.2837 (0.0295)
[1,10]	0.3204 (0.0026)	0.5471 (0.0306)	0.3843 (0.0423)	0.2799 (0.0264)
[1,15]	0.3204 (0.0023)	0.5523 (0.0267)	0.3815 (0.0382)	0.2827 (0.0292)
[1,20]	0.3213 (0.0028)	0.5492 (0.0291)	0.3752 (0.0381)	0.2871 (0.0303)

TABLE III
MINIMUM AND THE MAXIMUM TEMPERATURES OF CITIES IN DEGREE CENTIGRADE

Cities	January	February	...	November	December
Amsterdam	[-4, 4]	[-5, 3]	...	[1, 10]	[-1, 4]
Athens	[6, 12]	[6, 12]	...	[11, 18]	[8, 14]
Bahrain	[13, 19]	[14, 19]	...	[20, 26]	[15, 21]
Bombay	[19, 28]	[19, 28]	...	[14, 26]	[10, 20]
Calcutta	[13, 27]	[16, 29]	...	[18, 29]	[13, 26]
Colombo	[22, 30]	[22, 30]	...	[23, 29]	[22, 30]
...	...	...	...	...	...
Toronto	[-8, -1]	[-8, -1]	...	[-1, 17]	[-5, 1]
Vienna	[-2, 1]	[-1, 3]	...	[2, 7]	[1, 3]
Zurich	[-11, 9]	[-8, 15]	...	[0, 19]	[-11, 8]

The a priori classification, indicated by the suffix attached to the city temperature, is as follows:

- Class 1: Bahrain: 3, Bombay: 4, Cairo: 5, Calcutta: 6, Colombo: 7, Dubai: 9, Hong Kong: 12, Kula Lumpur: 13, Madras: 16, Manila: 18, Mexico: 20, Nairobi: 23, New Delhi: 24, Sydney: 30 and Singapore: 32
- Class 2: Amsterdam: 1, Athens: 2, Copenhagen: 8, Frankfurt: 10, Geneva: 11, Lisbon: 14, London: 15, Madrid: 17, Moscow: 21, Munich: 22, New York: 25, Paris: 26, Rome: 27, San Francisco: 28, Seoul: 29, Stockholm: 31, Tokyo: 34, Toronto: 35, Vienna: 36 and Zurich: 37
- Class 3: Mauritius: 19
- Class 4: Tehran: 33

Table IV shows the clusters (individual labels) given by the a priori partition according to the categorical variable, as well as the clusters obtained by the kernel fuzzy c-means and fuzzy c-means methods concerning the temperature interval data set.

The CR index obtained from the comparison between the a priori partition and the partitions given by the FCM-IV and KFCM-IV methods are, respectively, 0.466 and 0.714. The OERC for FCM-IV and KFCM-IV methods are, respectively, 0.405 and 0.216. Therefore, the KFCM-IV method is very superior to FCM-IV for this data set.

C. Results and comparison with the agaricus interval data set

This data set was extracted from [16] and contains information on species of fungi. Agaricus is a genus of mostly medium to large gilled mushrooms. Well known members of the genus

TABLE IV
CLUSTERING RESULTS FOR TEMPERATURE OF CITIES DATA SET

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
A priori Partition	3 4 5 6 7 9 12 13 16 18 20 23 24 30 32	1 2 8 10 11 14 15 17 21 22 25 26 27 28 29 31 34 35 36 37	19	33
FCM-IV	3 5 9 12 20 23 24 32	1 8 10 11 15 21 22 26 31 35 36 37	2 14 17 25 27 28 29 33 34	4 6 7 13 14 18 19 30
KFCM-IV	3 4 6 7 9 13 16 18 24 30	1 8 10 11 14 15 17 21 22 25 26 27 28 29 31 34 35 36 37	2 5 12 20 33	19 23 32

Agaricus are the bisporus, the “button mushroom” common in supermarkets and the campestris, the “meadow mushroom”, a popular edible mushroom which is common in pastures and grasslands in many temperate areas of the world. This data set consists of 24 mushrooms, each mushroom described by 5 interval variables: width of the pileus, width of the stipe, thickness of the stipe, height and width of spores. Table V shows part of this data set. It is composed of two classes: edible (E) and inedible (I).

TABLE V
INFORMATION ABOUT AGARICUS SPECIES KNOWN TO OCCUR IN CALIFORNIA

Species	pileus's width	stipe's width	stipe's thickness	spores's height	spores's width
Albolutescens	[6, 12]	[2, 7]	[1.5, 3]	[6, 7.5]	[4, 5]
Arvensis	[6, 21]	[4, 14]	[1, 3.5]	[6.5, 9]	[4.5, 6]
Benesii	[4, 8]	[5, 11]	[1, 2]	[5, 6]	[3, 4]
Bernardii	[7, 16]	[4, 7]	[3, 4.5]	[5.5, 7]	[5.5, 6.5]
Bisporus	[5, 12]	[2, 5]	[1.5, 2.5]	[6, 8.5]	[4.5, 6]
⋮	⋮	⋮	⋮	⋮	⋮
Subrutilescens	[6, 14]	[6, 12]	[1, 2]	[4, 6]	[3.5, 4.5]
Subrufescens	[6, 13]	[6, 12]	[1.5, 2.5]	[5.5, 6.5]	[4, 4.5]
Xanthodermus	[5, 17]	[4, 14]	[1, 3.5]	[5, 6]	[4, 5.5]

Table VI shows the elements grouped by clusters given by the a priori partition according to the categorical variable edible or inedible, as well as the clusters obtained by the kernel fuzzy c-means and fuzzy c-means methods concerning the agaricus interval data set.

TABLE VI
CLUSTERING RESULTS FOR AGARICUS DATA SET

	Cluster 1	Cluster 2
A priori Partition	1 2 3 4 5 6 8 9 10 11 12 14 15 16 17 19 20 21 22 23	7 13 18 24
FCM-IV	2 4 6 11 13 14 16 17 18 20 21 22 23 24	1 3 5 7 8 9 10 12 15 19
KFCM-IV	1 2 3 4 5 6 7 8 9 10 11 12 15 16 17 19 20 21 22 23	13 14 18 24

The CR index calculated from the comparison between the a priori partition and the partitions given by the FCM-IV and KFCM-IV methods are, respectively, -0.037 and 0.613 . The OERC for the FCM-IV and KFCM-IV methods are, respectively, 0.500 and 0.082 . Again, the KFCM-IV is very superior to the FCM-IV for the agaricus data set.

V. CONCLUSIONS

In this paper, we have presented a new clustering method for symbolic interval data based on kernel. This method is an extension of the kernel fuzzy c-means to symbolic interval data. The evaluation of this method in comparison with a fuzzy c-means clustering method for interval data was carried out. The accuracy of the results provided by these clustering methods regarding synthetic interval data sets were assessed by the corrected Rand index and overall error rate in the framework of a Monte Carlo experience. An application with two real interval data sets was also performed. Concerning the CR index and overall error rate, the proposed method in this paper surpassed the fuzzy c-means method in all interval data sets.

ACKNOWLEDGMENT

The authors would like to thank CNPq and FACEPE (Brazilian Agencies) for their financial support.

REFERENCES

- [1] A. K. Jain, M. N. Murty and P. J. Flynn Data clustering: a review. ACM Computing Surveys, 31(3):264-323, 1999.
- [2] R. O. Duda, P. E. Hart, and D. G. Stork Pattern Classification, John Wiley and Sons, 2001.
- [3] Xu, R.; Wunsch, D. I. I. Survey of clustering algorithms. IEEE Transactions on Neural Networks, p. 645-678, 2005.
- [4] Nikhil, R. Pal; Pal, K.; Keller, James M.; Bezdek, James C. A Possibilistic Fuzzy c-Means Clustering Algorithm. IEEE Transactions on fuzzy Sysyems, 2005.
- [5] Yang, M.-S., Wu, K.-L. Unsupervised possibilistic clustering. Pattern Recog. p. 521, 2006.
- [6] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta. A survey of kernel and spectral methods for clustering. Pattern recognition, p. 176-190, 2008.
- [7] R. Inokuchi and S. Miyamoto. LVQ clustering and som using a kernel function. volume 3, p. 1497-1500, 2004.
- [8] M. Girolami. Mercer kernel based clustering in feature space. I.E.E.E. Transactions on Neural Networks, p. 780-784, 2002.
- [9] Dao Q. Zhang and Song C. Chen. Kernel based fuzzy and possibilistic cmeans clustering. In Proceedings of the International Conference Artificial Neural Network, p. 122-125. Turkey, 2003.
- [10] I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means, spectral clustering and normalized cuts. in Proc. ACM SIGKDD Intl Conf Knowledge Discovery and Data Mining, Seattle, WA, 2004.
- [11] D.-Q. Zhang and S.-C. Chen. Kernel based fuzzy and possibilistic c-means clustering. in Proc. ICANN, Istanbul, Turkey, p. 122125, 2003.
- [12] E. Diday and M. Noirhomme-Fraiture. Symbolic Data Analysis And The Sodal Software. Wiley-Interscience Publishers, 2008.
- [13] de Carvalho, F.d.A.T., Fuzzy c-means clustering methods for symbolic interval data, Pattern Recog. Lett. doi:10.1016/j.patrec.2006.08.014, 2006.
- [14] Hubert, L. and Arabie. P. Comparing Partitions. Journal of Classification. p. 193-218, 1985.
- [15] D.S. Guru, B.B. Kiranagi, and P. Nagabhushan. Multivalued type dissimilarity measure and concept of mutual dissimilarity value for clustering symbolic patterns. Pattern Recognition, p. 1203-1213, 2004.
- [16] Wood, M.; Stevens, F. Agaricus. <http://www.mykoweb.com/CAF/genera/Agaricus.html>. Accessed July, 2010.