Journal of Theoretical and Applied Information Technology
15th December 2017. Vol.95. No 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A HYBRID APPROACH FOR UNSUPERVISED PATTERN CLASSIFICATION

[1]FADOUA GHANIMI, [1]ABDLWAHED NAMIR, [1]EL HOUSSIN LABRIJI
[1]University Hassan II, Department of Mathematics and Computer Sciences, Morocco
E-mail: [1]ghanimi_fadoua@yahoo.fr,

ABSTRACT

In this paper, we present a new data classification approach in an unsupervised context, which is based on both numeric discretization and mathematical pretopology. The pretopologicals tool, specially the adherency application are used in the modes extraction process. The first part of the proposed algorithm consists to a presentation of the set of the multidimensional observations as a mathematical numeric discrete set; the second part of the algorithm consists in detecting clusters as separated subsets by means of pretopological transformations.

Keywords: *Pretopology, Cluster Analysis, Adherence, Pretopological Closure, Unsupervised Classification*

1. INTRODUCTION

Clustering or unsupervised classification has long been an active area of research. The aim was always to divide a set of objects into subsets according to their similarities and dissimilarities. The objects are generally by N-dimensional vectors of observed features.

Several studies attacked these problems [1], [6]. Among the various types of procedures for clustering such sets, morphological operators are the most common used [1]. In the years 80, new tools based on pretopology and well adapted to cluster analysis has been developed [9], [2].

The pretopology is a mathematical tool for the analysis modeling and construction in various fields: social sciences, game theories, networks... It provides us a clustering process based on adherence and pretopological closures [5], [17] and establishes a powerful tool for structure analysis and automatic classification [4].

The adherence a(.) defined on the subsets of a finite set E, has the advantage to express the extension phenomena. Contrarily to what occurs in topology, a(.) is not always a closure, but its successive aggregations lead to produce closed subsets which characterize homogenous or interdependent parts of E.

The intent of the present paper is to develop a new pretopological approach for classification of multidimensional data.

The choice of the pretopology is motivated by the fact that it has less axioms then the topology which facilitates its adaptation to discrete spaces [3].

Thus, under the commonly accepted assumption that each class of the distribution corresponds to a region with a high concentration of observations and based on the mathematical adherence to express the proximity relation existing between these observations, we have developed a new mode detection algorithm. At the end of each of this algorithm, a partition of the discrete space E, formed by the closures of the elements is obtained. The different modes of the sample correspond, in fact, to these closures.

The detected modes do not include all observations submitted for analysis. To classify the observations not yet assigned to these modes, we adopt the classification procedure assigning each observation to the center of the nearest mode or to the class considering each prototype classified as a new prototype.

The results of this classification procedure is compared to those obtained by the classical algorithms Isodata and K-means.

In section 2, it's first shown how any finite set of multivariate observations can be represented as a mathematical discrete set in Euclidian space.

In section 3, we expose the notion of pretopology and the basic pretopological tools and concepts.

Journal of Theoretical and Applied Information Technology
15th December 2017. Vol.95. No 23

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A new iterative approach, based on the detection of cluster cores by means of a pretopological formalism by region growing is proposed in section 4.

The performance of this approach is demonstrated in section V using artificially generated data sets.

2. REPRESENTATION OF A SET OF MULTIDIMENSIONAL OBSERVATIONS AS A DISCRETE NUMERICAL SET

We consider a set of Q, N-dimensional observations.

$$\{X_1, X_2, \dots, X_q \dots . X_Q\} \quad (1)$$

Such that

$$X_q = \{X_{q,1}, X_{q,2}, \dots, X_{q,n}, \dots, X_{q,N}\} \quad (2)$$

In order to extend the theory of mathematical pretopology to cluster analysis, the set of available observations must be represented as a mathematical discrete set in a Euclidean space.

For this purpose, the origin O of the data space is first translated to the point O' defined as:

$$O' = \{\min_q X_{q,1}, \dots, \min_q X_{q,n}, \dots, \min_q X_{q,N}\} \quad (3)$$

The normalization of the range of variations of observations coordinates is then performed using a diagonal transformation such that:

$$X'_{q,n} = \frac{X_{q,n} - \min_{q'} X_{q',n}}{\max_{q'} X_{q',n} - \min_{q'} X_{q',n}} \times L \quad (4)$$

Where $X'_{q,n}$ is the nth coordinate of the observation X_q in the new data space.

After the normalization procedure, all observations are then situated within an hypercube of side length L. Each axis of this new data space is then partitioned into L intervals of unit width. This discretization, defines a set of a L_n hypercubes of side length unity which cover all the normalized space (see Figure.1).

The centers of these hypercubes constitute a lattice of sampling points. Each hypercube can be defined by N integers H_1, H_2... H_n... H_N such that:

$$H_q = \{\text{int}\, X'_{q,1}, \dots, \text{int}\, X'_{q,n}, \dots, \text{int}\, X'_{q,N}\} \quad (5)$$

Where $\text{int}\, X'_{q,n}$ denotes the integer part of $X'_{q,n}$.

Several hypercubes may contain more than one data point. The hypercubes not in this list are known to be empty.

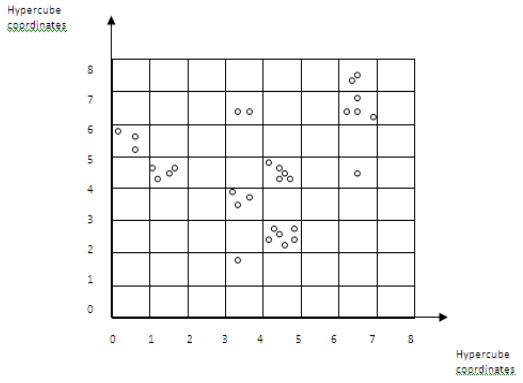


Figure 1: Non Empty Hypercubes In The Sampling Lattice (L = 10)

A transformation of the initial data to a discrete set of numerical elements is then performed [15]. It consists on giving to every non empty hypercube the value corresponding to the number of data points it contains, and to every empty hypercube the value 0. This procedure is equivalent to an encoding operation of the N component of the Q data points of the data set which reduce the running time in implementation.

The set of the centers of nonempty hypercubes will be referred as the discrete numerical set associated to the input data and we will denote it E. (see Figure.2)

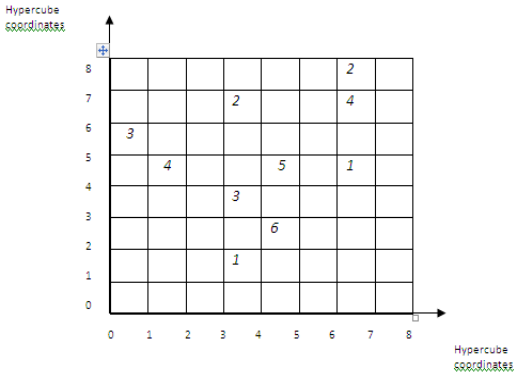


Figure. 2. The Discrete Set Of Numerical Elements

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. PRETOPOLOGY: BASIC CONCEPTS

Pretopology is an extension of topology that differs by the non idempotence of the adherency function which is a fundamental property for our purpose of following up the process of structuring a data set.
The fact that pretopology integrates fewer axioms makes it more adapted to discrete spaces [8], [9].
We suppose in all that paper that E is a non empty finite space.

3.1 Pretopological Adherence

Given a non empty set E, a function a(.) from P(E) into P(E) is called an adherence if and only if

$$\forall A \in P(E), a(\emptyset) = 0 \quad (6)$$

And

$$\forall A \in P(E), a(A) \subset A \quad (7)$$

Then (E,a) is said a pretopological space.
According to properties of a(.), we obtain more or less complex pretopological spaces from the most general spaces to topological spaces. Pretopological spaces of V type are the most interesting case. In that case, a(.) fulfills the following property:

$$\forall A \in P(E), \forall B \in P(E), A \subset B \Rightarrow a(A) \subset A \quad (8)$$

3.2 Closed Subsets, Pretopological Closure: Definitions

Closed subset: Lets (E,a) be a pretopological space. A subset A of E is said a closed subset if a(A) = A.

Pretoplouological closure: Given a pretopological space (E,a), for any subset a of E, we can consider the whole family of closed subsets of E which contain A. if exists, we determine the smallest element of that family for the including relationship. That element is called the closure of A and denoted F(A) .
In any pretopological space of type V, given a subset A of E, the closure of A always exists. In particular for each singleton {x} of P(E) [10].
If a(.) fulfils on the top of the properties (6), (7), (8) the following property:

$$\forall A \in P(E), \quad a(A) = \cup_{x \in A}\{a(x)\} \quad (9)$$

The pretopological space (E, a) is called a Vs - type space.
The Vs pretopological space (E,a) are very interesting since the pretopological closures of the set E will be made by the pretopological closures of its elements.

4. THE NEW HYBRID APPROACH

The clustering approach developed here is based on the fact that a region of high local concentration of observations is associated with each cluster. Such region is called the core of the cluster.The key problem is to detect the cluster cores in the discrete numerical set associated to the input data by defining a pretopological structure of Vs type on E. The closures of E will be the best candidates to be the cluster cores.
Let S(x) be the value taken by the hypercube x.
We define an adherence on E by:

$$a(x) = \begin{cases} \{y \in V(x) & si\ |S(x) - S(y)| \leq T \\ \{x\} & si\ |S(x) - S(y)| > T \end{cases} \quad (10)$$

Where V(x) designates the set formed by the hypercube x and its 3^N -1 hypercubes neighbors, and T is an integer number defined by the tuning algorithm[16].

We put:

$$a(A) = \cup_{x \in A}\{a(x)\} \quad \forall A \in P(E) \quad (11)$$

(E,a) is a V_s type pretopological space by definition.

As a result, the pretopological closures of a set E will be made by the pretopological closures of its elements [8].
This property will allows us to engage the process of the detection of the cluster cores by taking some chosen points of the discrete numerical set as germs.

The proposed algorithm

The purpose of our approach is to detect the cluster cores, by obtaining all the pretopological closures Ei. The different steps of the algorithm are:

1. Choose an initial germ X in E such that S(V(X) is the maximum of $S(V(x_i))$
2. Search the pretopological closure of X , C(X) following this process:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

C(X) = a(X);
While (a(C (X) ≠ C (X));
C(X) = a(C(X))

3. Put $X_1 = X$
4. Choose another germ X2 in E − C (X_1) such that $|S(X_2) - S(X_1)| \leq T$ and search for his pretopological closure following the same process as step 2.
5. If there is any x in $E - \bigcup_i C(X_i)$ fulfilling $|S(X) - S(X_1)| \leq T$, we take it as a new germ and we go to step 2 else we go to step 5
6. Finally, we present the different cluster cores as $E_i = C(X_i)$.

5. EVALUATION

Several grouping process can be used to assign the input data points to the clusters that remain after the pretopological filtering. One approach is to use the data points falling into the pretopological cores as prototypes. The data points that do not fall in one of the detected cores are finally assigned to the clusters attached to their nearest neighbour (NN) among these prototypes [6].
Cover and Hart have shown that the use of the nearest neighbor method ensures a misclassification rate that cannot exceed twice the error rate defined by the Bayes Rule. Moreover, when using the classification method by the k nearest neighbors, the choice of the value of k is often arbitrary [10]. That is why we adopted the nearest neighbor method, the generalization to the nearest neighbors being immediate.
When the clustering process is applied to artificially generated data sets, the results are evaluated by means of the classification error rate, which is estimated as the ratio of the number of misclassified data points to the total number of available samples [9].
To assess the efficiency of the proposed pretopological clustering approach, it is compared with the ISODATA [7] and the K-means [8] procedures on several sets of data.

Example 1

The data of this example is presented in Figure. 3(a). It is composed of 600 bidimensional observations drawn from three equiprobable normal distributions specified. The cores detection is made by applying our algorithm to the discrete numerical set E associated to the input data shown in Figure3(b).

The Parameter L was fixed by applying the procedure to the data for different values of the parameter L. We choose the value of L which is in the middle of the largest range where the number of detected clusters remains constant [16], [17]. The three cores detected using our approach is displayed in Figure.3(c) and the result of the classification is shown in Figure.3(d).

Table 1. Statistical Parameters Of The Two Nongaussian Distributions Of Example 1.

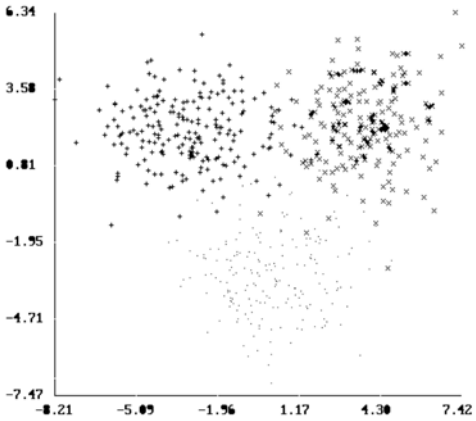
	Generated data	
Distributions	Mean vector	Covariance matrix
1	0,1695 -2,671	2,8274 0,0687 0,0687 2,0576
2	-2,563 1,5673	3,3969 -0,105 -0,105 1,3701
3	3,1907 1,8051	2,1336 0,2865 0,2865 2,1583

Table 2: Comparison Of The Results For The Distribution Of Example 1.

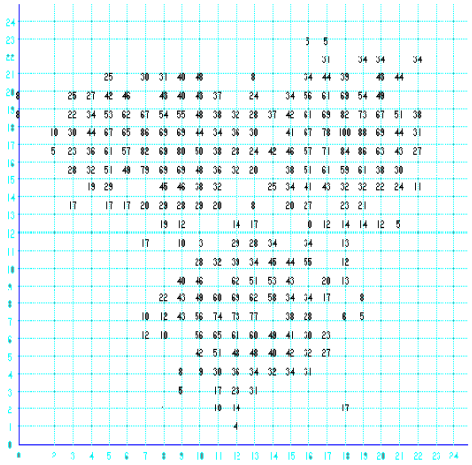
Our approach results (%)	Isodata results(%)	Kmeans results
10.1	9.5	10.16

The Parameter L was fixed by applying the procedure to the data for different values of the parameter L. We choose the value of L which is in the middle of the largest range where the number of detected clusters remains constant [16], [17]. The three cores detected using our approach is displayed in Figure.3(c) and the result of the classification is shown in Figure.3(d).

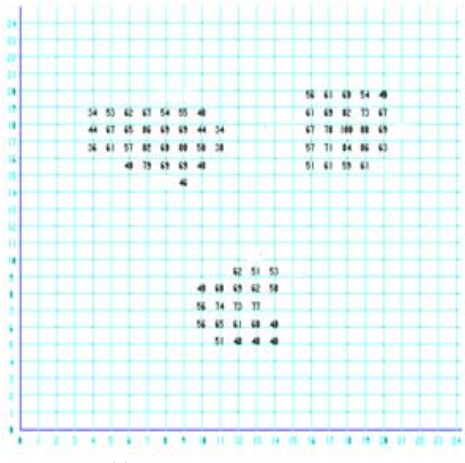
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



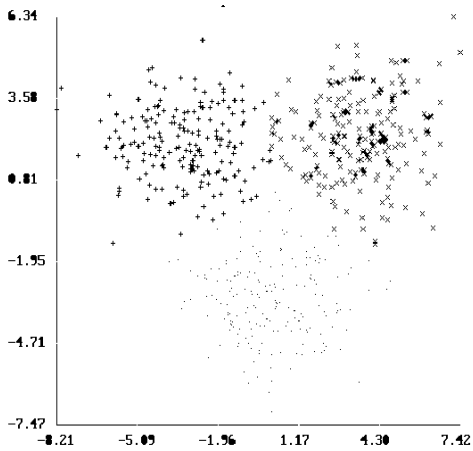
(a)



(b)



(c)



(d)

Figure 3: Cluster Detection Of The Data Of Example 1:(A) Raw Data Set ;(B) The Discrete Numeric Set;(C) The Detected Cores;(D) The Result After Classification

The error rate obtained with our hybrid approach is equal to 10.1 which is near of the error rates obtained using the ISODATA and the K-MEANS (see Table 2).

Example 2

As a second example, we choose a data composed of two populations (see Figure 4). The first is composed of 550 points and the second of 600 drawn from non spherical distributions.
The error rate associated to the hybrid procedure is 8.6% whereas it reaches 11.3% and 10.6% the errors rates of ISODATA and K-means algorithms (see Table 3).

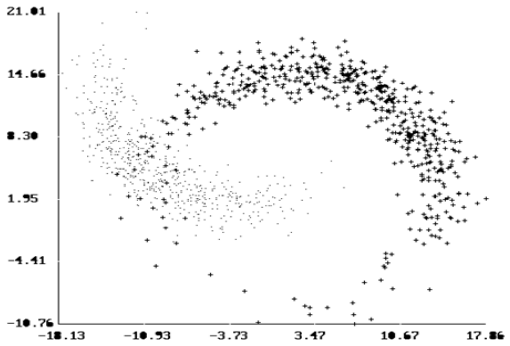
Table 3: Comparison Of The Results For The Distribution Of Example 2.

Our approach results (%)	Isodata results(%)	Kmeans results
8.6	11.3	10.6

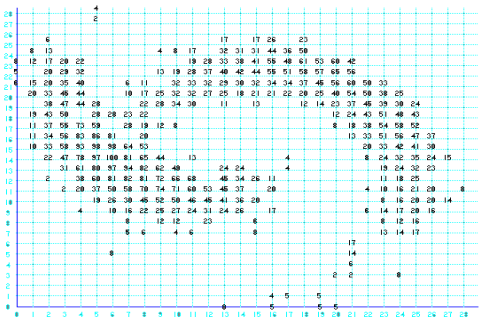
This example shows that, the hybrid technique is more efficient then the two classical algorithms for non spherical clusters.

Journal of Theoretical and Applied Information Technology
15th December 2017. Vol.95. No 23

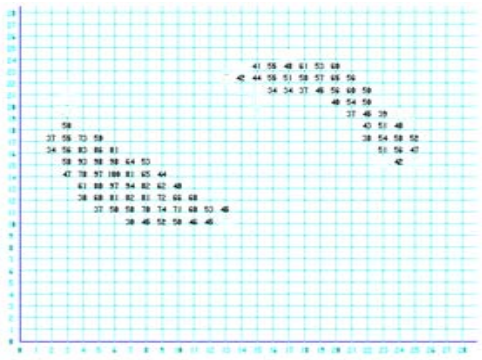
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



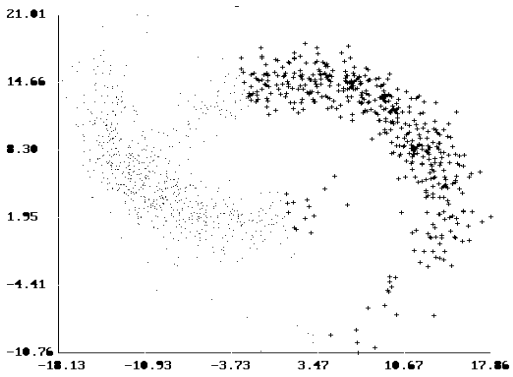
(a)



(b)



(c)



(d)

Figure 4: Cluster Detection Of The Data Of Example 2:(A) Raw Data Set ;(B) The Discrete Numeric Set;(C) The Detected Cores;(D) The Result After Classification

Example 3

As fourth example, we use a multidimensional data composed of the two Gaussian clusters presented in Figure. 5 .
The error rate of the proposed pretopological algorithm which is equal to 2.1, remains very close of those obtained using the basic ISODATA and K-means procedures.

Table 4. Statistical Parameters Of The Two Nongaussian Distributions Of Example 3.

	Generated data			
Distributions	Mean vector	Covariance matrix		
1	1,0138	2,9465	-0,129	-0,1987
	0,9545	-0,1299	3,0983	0,1024
	1,9488	-0,1987	0,1224	2,9211
2	6,0754	2,9723	-0,039	-0,1523
	4,0872	-0,039	1,9843	-0,009
	5,039	-0,0153	-0,009	3,0858

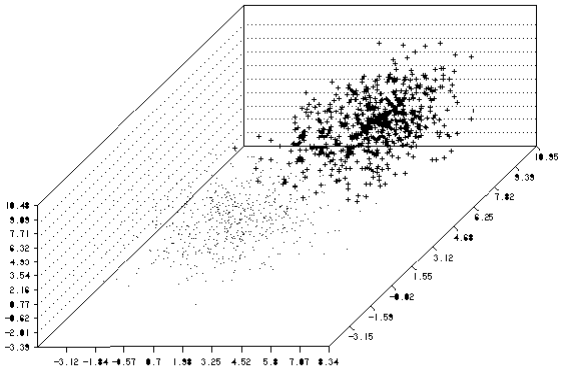
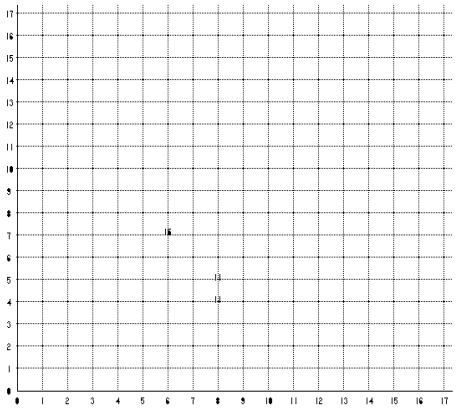
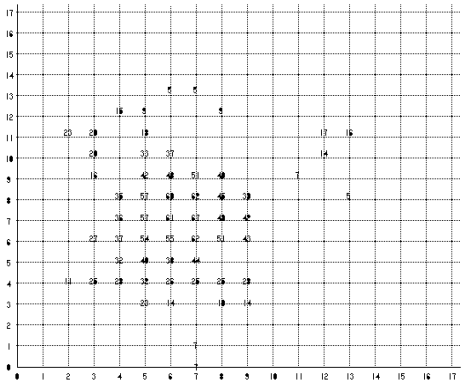


Figure 5:Raw Data Set Of The Data Of Example 3

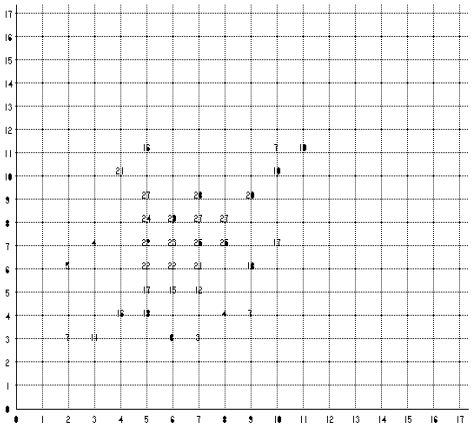
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



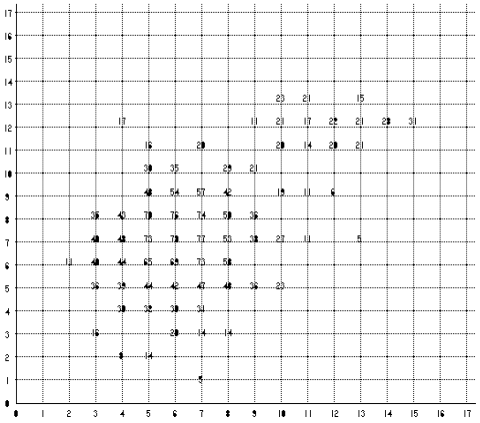
X =0



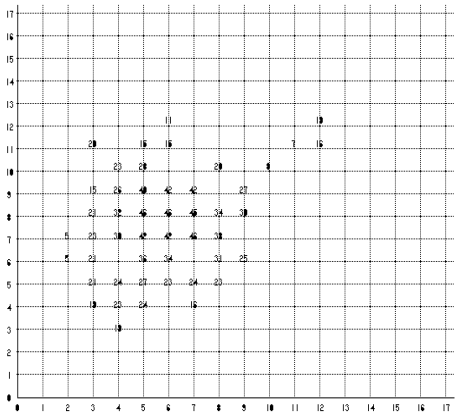
X = 3



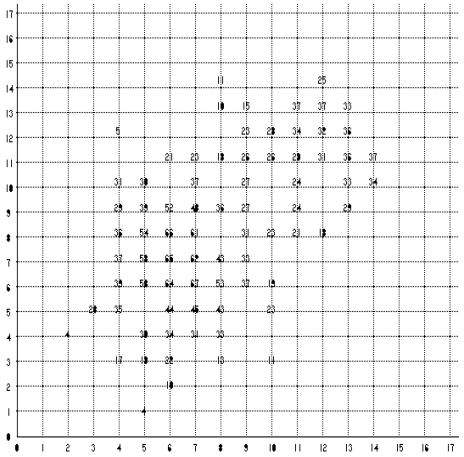
X = 1



X = 4



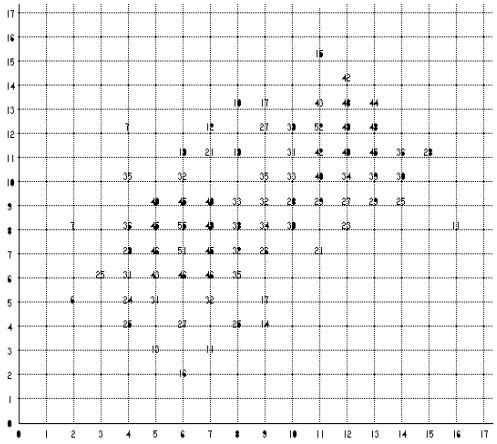
X = 2



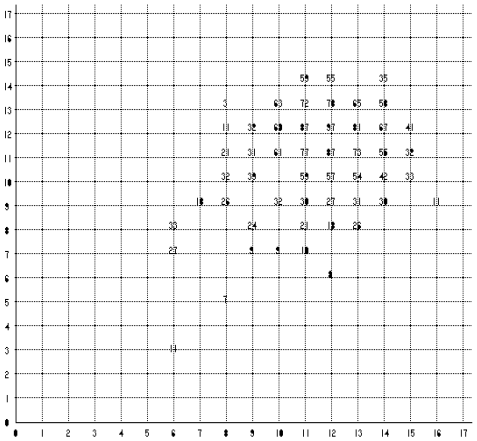
X = 5

Journal of Theoretical and Applied Information Technology
15th December 2017. Vol.95. No 23

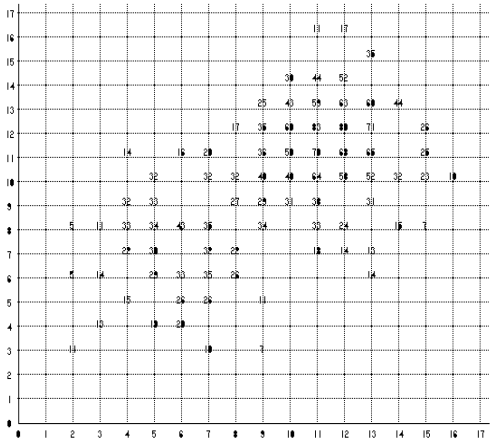
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



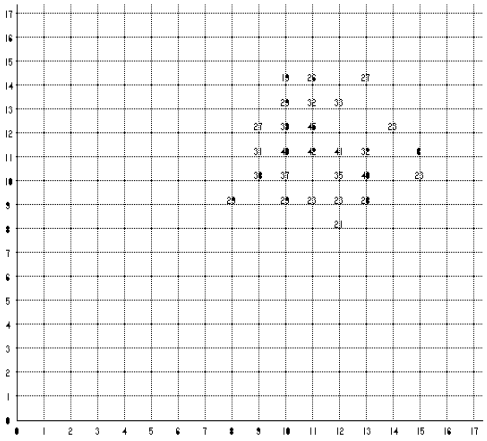
X = 6



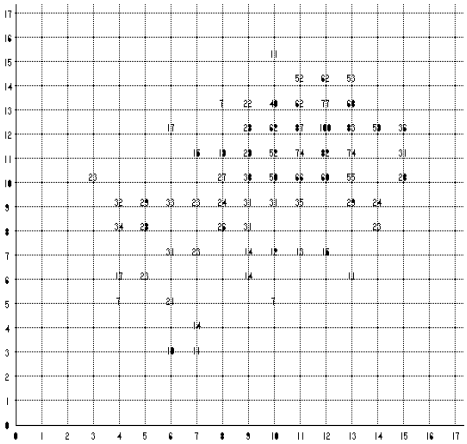
X = 9



X = 7



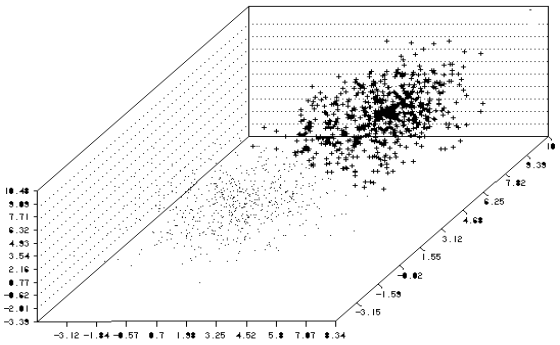
X = 10



X =8

Figure 6: The Discrete Numeric Set Presented Plan By Plan

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(b)

Figure 7:The Result After Classification.

Table 5: Comparison Of The Results For The Distribution Of Example 3.

Our approach results (%)	Isodata results(%)	Kmeans results
2.1	2.25	2.25

6. CONCLUSION

This paper presented a new pretopological method for clustering based on the concepts of adherency function and pretopological closure. The pretopological adherence possesses the property of non idempotence which allows the development of iterative algorithms.

A step by step cluster detection process is then built using a well chosen germ as start.

Compared to various classical classification schemes, it appears of that this pretopological approach performs well for detecting clusters and constitutes then an interesting application of pretopology to cluster analysis field.

However, this approach requires a discretization of the space which requires too many observations when the size of the space is high. Moreover it is less adapted to Gaussian classes than to other nonlinearly separable classes. We therefore think of trying to define adherency applications that take into account the local form of the distributions, which suggests an improvement in the classification.

REFERENCES:

[1] R.O. Duda, & P.E. Hart, Pattern Classification and Scene Analysis, New York: Wiley, 1973.

[2] G. Duru, Contribution à l'étude des structures des systèmes complexes dans les sciences humaines, Thèse de Doctorat d'Etat, UCB Lyon, France, pp.1980, 229.

[3] M. Lamure, Contribution à l'analyse des espaces abstraits - application aux images digitales, Thèse de Doctorat d'Etat, UCB Lyon, France, 1987.

[4] H. Emptoz, Modèle Prétopologique pour la reconnaissance des Formes : Application en Neurophysiologie, Thèse de Doctorat d'Etat, UCB Lyon, France, 1983.

[5] S. Bonnevay, & C. Largeron, Data analysis based on minimal closed subsets, Classification and related Methods, Kiers et al. editors, Springer, 2000, pp303-308,.

[6] J. Serra, Image Analysis and mathematical Morphology, New York : Academic Press, 1982.

[7] Z. Belmandt, Manuel de Prétopologie et ses applications, Edition Hermès, 1993.

[8] M. Lamure, & J.J. Milan, A system of image analysis based on a pretopological approach, in Intelligent Autonomous System Procceding, 1987 pp.340-345.

[9] J.P. Aurey, G. Duru, M. Lamure, & M. Terrenoire, Outils prétopologiques pour le traitement des images. Actes du Coloque, Analyse des Problèmes Décisionnels dans un Environnement Incertain Et Imprécis, Reims, France, 1985 pp.135-145.

[10] T.M. Cover, & P.E. Hart, Nearest neighbor pattern classification, IEEE Transactions On Information Theory, Vol IT-13,1967, pp.340-345.

[11] G.H. Ball, & D.J. Hall, ISODATA, a novel method of data analysis and pattern classification, AD-699616, Stanford Research Institute,1965.

[12] J.B.MacQueen, Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1960, pp. 281-297.

[13] J.T. Tou Dynoc, A dynamic optimal cluster seeking technique, International Journal of Computing and Information Sciences, Vol. 8, n°6, 1979, pp. 541-547.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[14] D. L. Davies, & D. W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 1, n°2, 1979.

[15] J.-G. Postaire, R.D. Zhang & C. Lecoq-botte, Cluster analysis by Binary Morphology, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, n°2, 1993.

[16] J.-G. Postaire, & C. P. A. Vasseur, An approximate solution to normal mixture identification with application to unsupervised pattern classification, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-3, n°2,1981, pp. 163-179.

[17] A. Touzani & J.-G. Postaire, Mode detection by relaxation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 10, n°6, 1988, pp. 970-978.

[18] D. Mammass, S. Djeziri, & F. Nouboud, A Pretopological Approach for Image Segmentation and Edge Detection, Journal of Mathematical Imaging and Vision, Vol 15, 2001, pp. 169–179,.