A Mixed Pattern Classification Method

Moussa Diaf and K.K.Chopra

Institute of Electronics
University of Tizi-Ouzou
B.P. 95 - Chikhi
15008 Tizi-Ouzou
ALGERIA

Abstract: This paper presents a new approach to pattern classification. Two methods are combined, namely a graphical interactive method involving the motion of an observer in a multidimensional space and the Iterative Adjustments About Mobile Centers. The former method transforms the multidimensional space into a bidimensional space allowing the visualisation of clusters. In case of overlapping data, when a user is not capable of decision- making, the Iterative Adjustments About Mobile Centers for an automatic problem- solving can be introduced.
Keywords: Pattern classification,pattern recognition, clustering,automatic classification,data analysis

Moussa Diaf was born in Algeria in 1953. He received the Engineer degree in Electrical Engineering from the Algerian Polytechnical School, Algiers, in 1978, and the D.E.A. (Diplome d'Etudes Approfondies) in 1981 and Docteur-Ingenieur in Control Engineering in 1983 both from the University of Lille,France.
Moussa Diaf is currently working with and directing a group of researchers in the Pattern Classification Field. He is also a lecturer at postgraduate as well as at undergraduate departments.
He is in charge of the Postgraduate Department. He was a member of the Organizing Committee of ICEA '92.

K.K.Chopra was born in India. He received the M.Sc. in Physics from the University of Agra, and the Ph.D from the University of Meerut, India, in 1970. K.K.Chopra worked as Reader and Director of the Physics Institute S.D. Postgraduate College MUZAFFARNAGAR (Meerut,India) till 1980. Currently , he is working as a professor at the University of Tizi-Ouzou,Algeria. He is author of several books and papers.

Introduction

In the statistical formulation of the pattern classification, several approaches have been proposed [1,2,3]. Most of them are based on the main assumption stating that patterns are drawn from a multidimensional probability density function where each mode corresponds to a class [4,5]. Unfortunately, the number of the required samples should be large enough and the problem of sensitivity to noise,to data distribution irregularities and to the discretisation period has not been solved yet [6,7]. The other tool of classification, using clustering algorithms, is valuable since it helps someone "look" at data and ascertain their structure by organizing data into subgroups or clusters. Various algorithms have been proposed [8,9,10]. They are based on 1) using centers of clusters as modes of the pattern distribution [11,12], 2) minimizing the appropriate criteria [13,14], 3) using graph-theoretical methods [15,16], 4) hierarchically ordering data [17], 5) using non-linear mapping [18] and 6) the appropriate similarity [19].

In case of distinct clusters, using any of the above algorithms will certainly do. However, detecting overlapping clusters independently of their number and size is still to be solved. In order to solve the former problem, the present paper includes a novel graphical representation of multidimensional data. Until now, visual multidimensional representations have used projections or mathematical transformations into 2D or 3D space [21,22]. However, the local data structures are seldom maintained and the distortions(resulting from either transformations or projections) bring about ambiguous interpretations.

The graphical representation proposed in this paper consists in moving an artificial observer in a multidimensional space. According to the position of the artificial observer and to the direction its look takes, clusters may be viewed different ways. On plotting data points in a 2D space, with one co-ordinate being the distances of data points to the observer , and the other one being the angular positions of data points with respect to the observer's look, clusters' display will suffer no distortion. Obtaining several graphical representations depends on the observer's position and/or on the direction the observer's look takes.
If the well-known algorithms of cluster analysis are used,the number of clusters is to be specified or determined. The present method let a user

choose one mapping configuration from among various existing mapping configurations, even though he/she has no prior knowledge of the structure of data concerned.Once the clusters are formed, different pattern classification methods can be introduced. Clusters are being visualised, no matter their overlapping, and a human can see them and that is why the Iterative Adjustments About Mobile Centers method, which is an automatic pattern classification process, is used.
Usually, in order to maximize the performances of the family of the iterative processes [22], adequate initial guesses of the cluster centers are needed. In case of unknown data under nonsupervised hypothesis, obtaining good results is uncertain. In the proposed method, the initial guess on the centers of the displayed clusters is made by the user. Provided that distinct clusters exist, a user validates a classification by bounding the clusters. Otherwise, a combination of the two methods- the graphical representation and the Iterative Adjustment About Mobile Centers-will be required.

The Bidimensional Representation

The idea underlying the proposed graphical representation is the motion of an observer in a multidimensional space. It considers the way how a human sees the objects in space. One's position and the direction one's look takes make the objects be viewed in their contextual clustering. In this way, each multidimensional data point is labelled by letters d (Euclidean distance between the point and the observer) and α (angular shift of the look) as shown in Figure l.

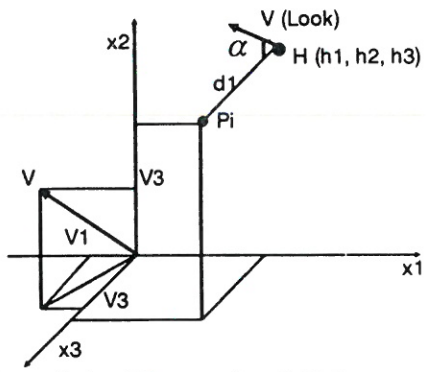

Figure l. An Observer in a 3-D Space

In the case of N data points, each data point is described by M parameters, and the Euclidean distance d_i of a data point P_i (having the co-ordinates $x_{i1}, x_{i2},..., x_{iM}$) from the observer located at a point H ($h_1, h_2, ... h_M$) is given by:

$$d_1=\sqrt{\sum_{j=1}^{M}\left(x_j^i-h_i\right)}\ ,\ i=1,2,...,N \qquad (1)$$

The angular shift α_i between the point p_i (i=1, 2,...N) and the direction of the look defined by the vector V ($v_1, v_2, ..., v_M$) is given by Eq. (2) calculated from the scalar products of the vectors V and HP_i ($x_1 - h_1, x_2 - h_2, ..., x_M - h_M$).

$$\alpha_i=\cos^{-1}\frac{\sum_{j=1}^{M}\left(x_j^i-h_j\right)v_j}{d_i\sqrt{\sum_{j=1}^{M}v_j^2}},\quad i=1,2,...,N \qquad (2)$$

It follows that this 2D representation visualises data points under the same clusterings as they are configured in the multidimensional space. If any changes occur in the observer's position and in the direction the observer's look takes, a different graphical representation, that is a new image, does result.

The Adjustments About Mobile Centers

There is a principle of this process based on which the centers of the classes initially fixed by a user are calculated. Objects are then iteratively re-assigned to the nearest class up to the moment the classification crystallizes. The calculations involved are described by the following equations.

$$g_j^k=\frac{1}{N}\sum_{i=1}^{L_k}x_j^i \qquad \begin{matrix} j=1,2,...,M \\ k=1,2,...,M \end{matrix} \qquad (3)$$

$$M2\left(C_k/G_k\right)=\sum_{i=1}^{L_k}\sum_{j=1}^{M}\left(x_j^i-g_j^k\right)^2 \qquad (4)$$

$$M2\left(C_k/G\right)=\sum_{k=1}^{K}\sum_{i=1}^{L_n}\sum_{j=1}^{M}\left(x_j^i-g_j^k\right)-\sum_{k=1}^{K}\sum_{j=1}^{M}\left(g_j-g_j^k\right)^2 \qquad (5)$$

$$R = \frac{\sum_{k=1}^{K}\sum_{i=1}^{L_n}\sum_{j=1}^{M}\left(x_j^i - g_j^k\right)^2}{\sum_{k=1}^{K}\sum_{j=1}^{M}\left(g_j - g_j^k\right)^2} \qquad (6)$$

Eq.(3) calculates the gravity center of the cluster C_k containing L_k objects (k = 1,2,...K). Eq.(4) gives the second central moment of the cluster with respect to its center of gravity. Note that the lower the value of this intraclass moment, the higher the concentration of the objects about their centers of gravity. Using Huyghens' theorem, the second central moment of the clusters C_k, with respect to the center of gravity G of all data points, is given in Eq.(5). In this equation, the second term designates the interclass moment. A higher value of this moment will be a net separation of the classes. On the other hand, a low value of the ratio of the intraclass and the interclass moments, given by Eq. (6), is characteristic of a good classification.

The Mixed Method

Data are plotted in the graphical part of the present method according to their Euclidean distances d_i from an observer placed at point H $(h_1,h_2,...h_M)$, and according to the angular shifts α_i between the points p_i and the directions of the look vector V $(v_1, v_2,..., v_M)$ as described above. The co- ordinates of the point H $(h_1, h_2,...h_M)$ and the components of the look vector are chosen by the user as explained below. With user's every option, the values of d_i and α_i given by Eqs. (1) and (2) change, yielding different graphical clusterings. Several experiments are carried out before the user opts for the most suitable representation and validates the classification by bounding the clusters as shown in Figure 2.

Eliminating and isolating a cluster or providing an unknown sample are among other possibilities offered.

This interactive graphical part is, as one can see, only effective in case of disjointed clusters. In this case, the Iterative Adjustments About Mobile Centers method is introduced in order to put into effect an automatic separation of classes. Number and locations of the initial centers required for this automatic process will be appropriately chosen if graphically displayed.

The software associated with the present method shows performant man-machine interaction using a structured menu,a keyboard and a mouse. The observer's position co-ordinates in the multidimensional space and the components of the "look" vector are represented by rectangles . Thus,the user will easily modify their values. Variable segments within the corresponding rectangles (see Figure 2) will account for this modification.

On introducing the Iterative Adjustments About Mobile Centers method, by making the option "Automatic", a user fixes up the centers of the initial partitions with the help of the mouse. Clusters are then formed and displayed iteratively as described above until the convergence is reached. Graphically displayed results, in different colours, are stored as clusters with their statistical parameters.

Experimental Results

In order to demonstrate the performance of the method, some examples are given.

Example 1

The first example demonstrates the ability of the procedure to separate three clusters of randomly produced data following the Gaussian distribution [23]. This random set contains three classes of 750 samples with the following mean vectors:

$$X_1 = \begin{vmatrix} 0.26 \\ -0.24 \\ 0.25 \\ 0.20 \end{vmatrix} \quad X_2 = \begin{vmatrix} 12.11 \\ 05.67 \\ 04.21 \\ 08.17 \end{vmatrix} \quad X_3 = \begin{vmatrix} 00.03 \\ 12.94 \\ 08.13 \\ 03.88 \end{vmatrix},$$

co-variance matrix:

$$M = \begin{vmatrix} 5.67 & 5.64 & 6.36 \\ 5.65 & 6.03 & 6.11 \\ 6.74 & 5.78 & 5.41 \\ 6.46 & 6.25 & 6.27 \end{vmatrix}$$

and *a priori* probabilities:

P1 = 0.20; P2 = 0.30; P3 = 0.50

The resulting classes shown in Figure 3 present almost the same mean vectors, co-variance matrix

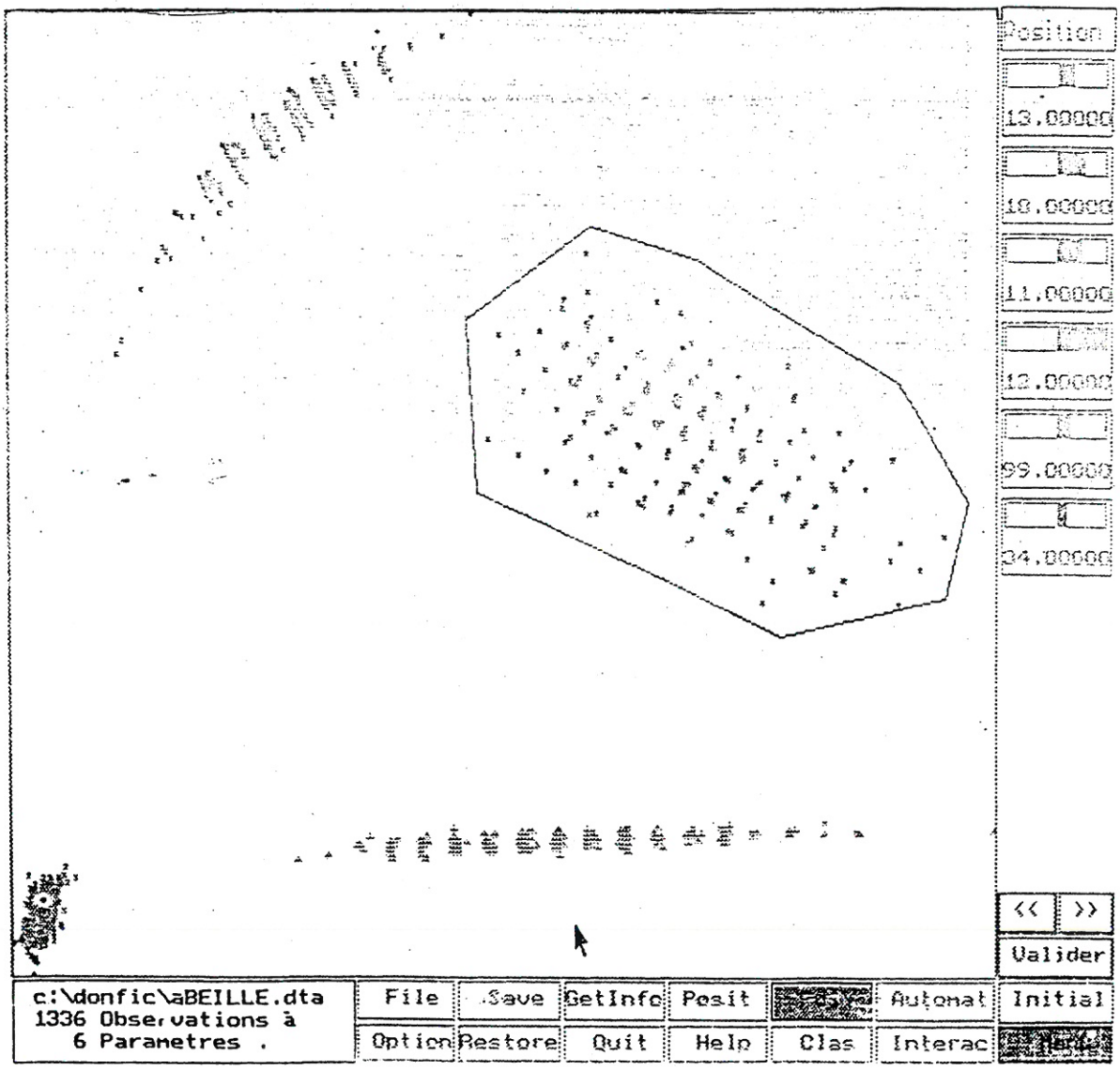

Figure 2. Classification by Bounding Clusters

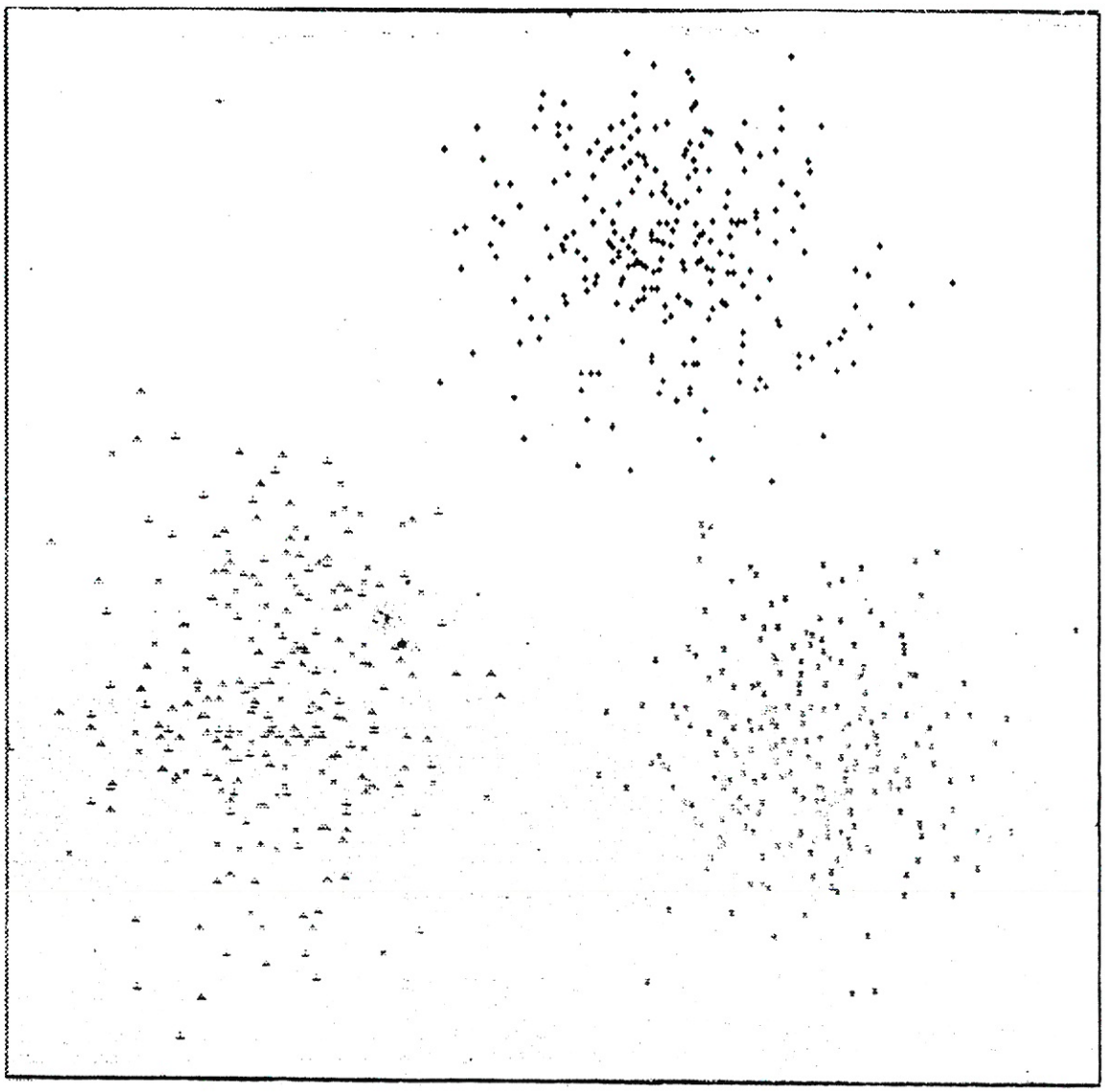

Figure 3. Illustration of Example – 1

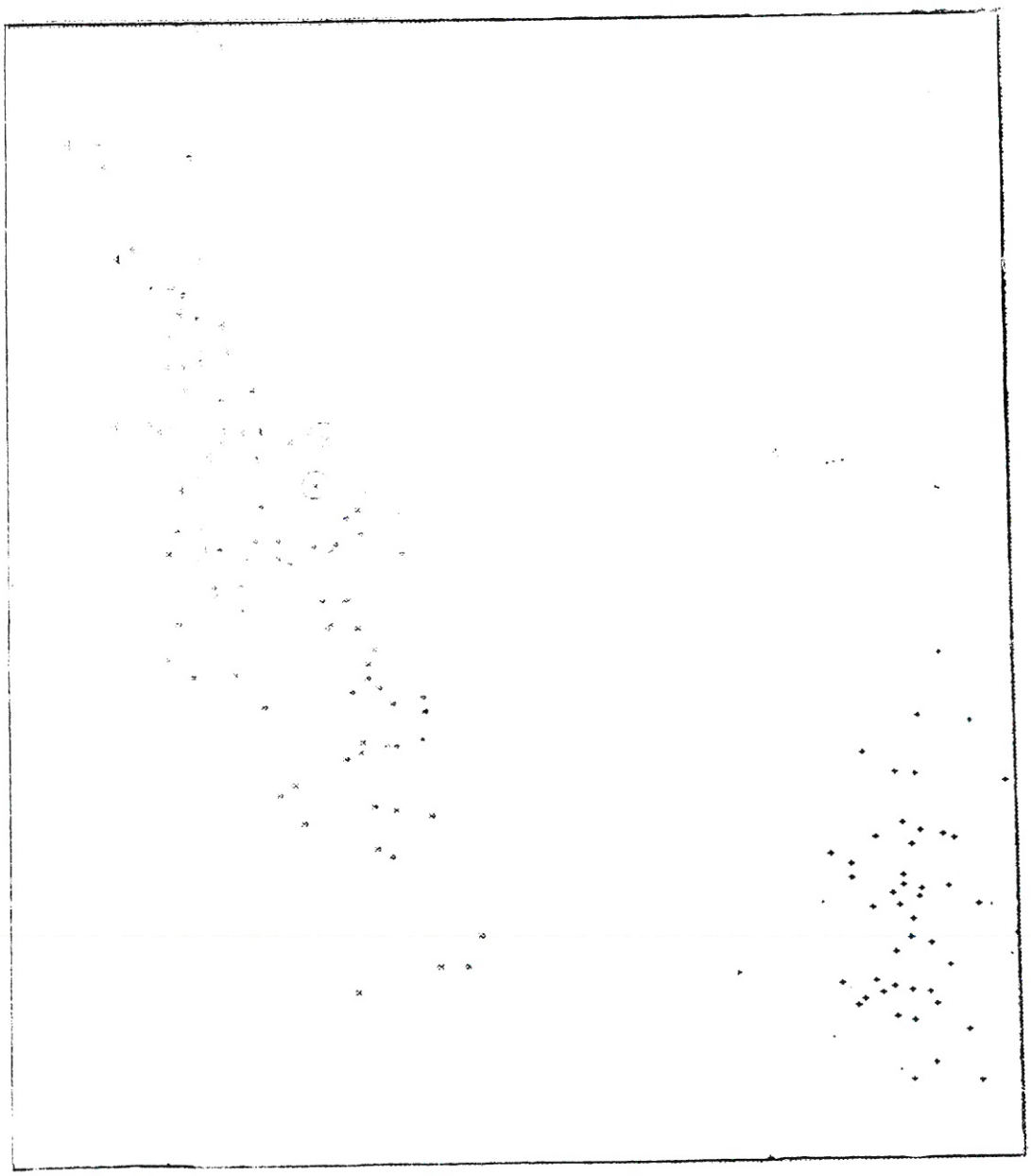

Figure 4. Classification of Fisher's *Iris* Data

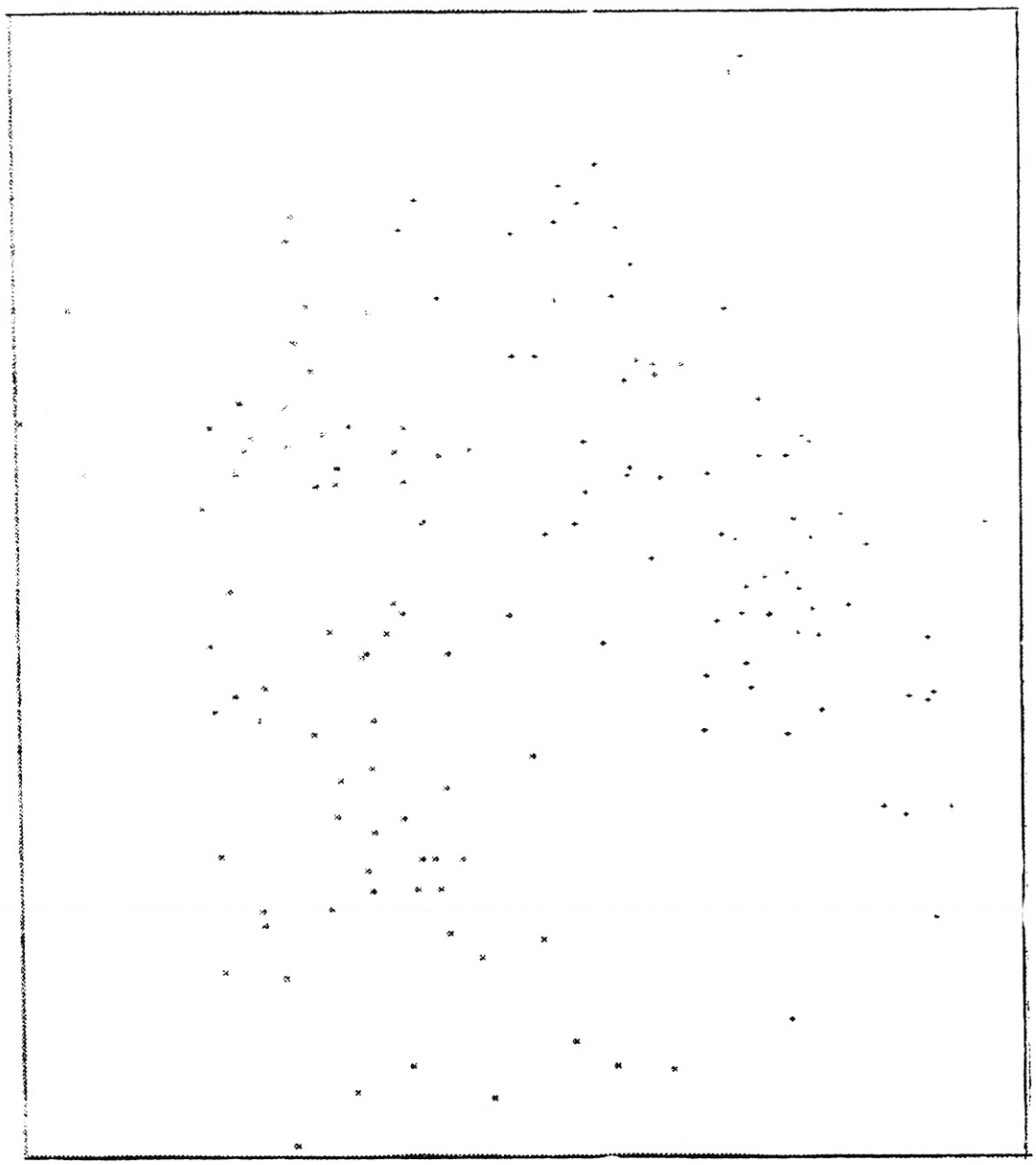

Figure 5. Illustration of Example – 3

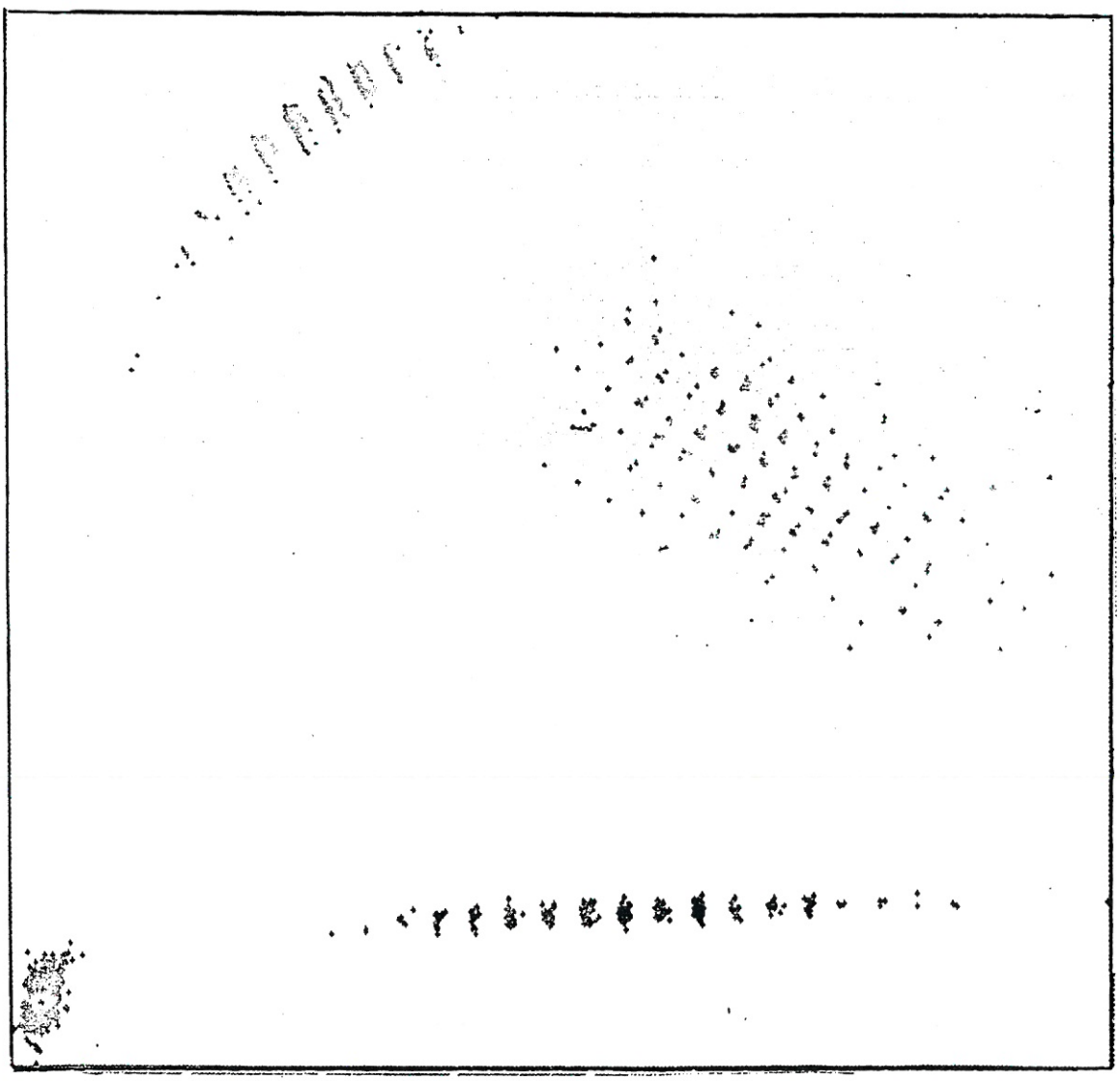

Figure 6. Illustration of Example – 4

and *a priori* probabilities.

Example 2

This example comes to demonstrate the capacity of the procedure to identify Fisher's iris species [24]. The *iris setosa* species gets separated but two others overlap. A user is not able to distinguish their boundaries. By introducing the automatic process, the fifty *iris vesicolor* and the fifty *iris virginica* are separated as shown in Figure 4.

Example 3

A two-dimensional data set is used in checking on the ability of the method to separate two small-size clusters. These Gaussian clusters are generated following the mean vectors:

$$\Upsilon_1 = \begin{vmatrix} 0.47 \\ 0.25 \end{vmatrix} \quad \Upsilon_2 = \begin{vmatrix} 0.45 \\ 0.71 \end{vmatrix}$$

co-variance matrix:

$$M' = \begin{vmatrix} 0.053 & 0.048 \\ 0.012 & 0.010 \end{vmatrix}$$

Each cluster contains 75 samples. An automatic process separates the two clusters as shown in Figure 5. The ratio R is 0.46.

Example 4

The method also applies to a large-size data file: a 6-dimensional set of 1336. Figure 6 shows four clusters that can be interactively separated by either bounding them or using the automatic process as shown in Figure 2. In addition, when isolating the thick cloud on the left of the Figures, the graphical visualisation will show it as consisting of two sub-clusters.

Conclusions

This paper proposes a new approach to automatic classification by combining an iterative graphical method with the Iterative Adjustments About Mobile Centers method.The graphical part will consist in transforming the multidimensional space into a bidimensional one, making it possible to display clusters. Just using this graphical display is enough for the cases of compact and not overlapping clouds. Otherwise, the Iterative Adjustments About Mobile Centers method is going to be used in order to automatically separate the classes.

Data sets have been simulated in order to demonstrate how satisfactory the graphically displayed results of the method actually are and how its statistical parameters indicate.

An interactive method like this supports , thanks to a bidimensional representation of the multidimensional data, decision- making and helps a user estimate correctly. The user acts for bounding distinct clusters or for choosing number and location of the initial centers required by the automatic process. He/she also handles the procedures about the results.

The programs have been tested on several data sets. The results of the studies showed that the method matched the initial data. If compared to the ascending hierarchical method and the k-nearest neighbours,the Iterative Adjustments About Mobile Centers is a quick and accurate method.

The method seems to be a viable tool for classificiaton and analysis of new data.

REFERENCES

1. COVER ,T.M. and WAGNER,T.J., **Topics in Statistical Pattern Recognition**, in K.S.Fu (Ed.) Digital Pattern Recognition, SPRINGER-VERLAG ,New York,1976.
2. DEVIJVER, P.A., **Decision-theoretical and Related Approaches to Pattern Classification**, Proc.NATO Advanced Study Institute on Pattern Recognition-Theory and Applications, 1975.
3. FUKUNAGA, K., **Introduction to Statistical Pattern Recognition**, SPRINGER-VERLAG, New York, 1976.
4. 4. POSTAIRE, J.G., **Une approche statistique unique pour l'analyse des mélanges et la détéction des modes en classification automatique**, REVUE DE STATISTIQUE APPLIQUEE, Vol.4,1983.
5. VASSEUR, C. and POSTAIRE, J.G., **Convexité des fonctions de densité. Application à la détéction des modes en reconnaissance des formes**, RAIRO AUTOMATIQUE, Vol.13, No.2, 1979, pp.171-188.
6. GLENN, W.M., SOON,S.C. and SOKOL,

L.M., **The Effect of Cluster Size, Dimensionality and the Number of Cluster on Recovery of True Structure**, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,Vol. pami-5, No.l, 1983, pp.40-47.

7. HO, Y.C. and AGRAWALA, A.K., **On Pattern Classification Algorithms -Introduction and Survey**, IEEE TRANSACTIONS ON AUTOMATIC CONTROL, Vol. ac-13, No.6, 1968.
8. POSTAIRE, J.G. and TOUZANI,A., **Mode Boundary Detection by Relaxation for Cluster Analysis**, PATTERN RECOGNITION, Vol. 22, No.5, 1989, pp.447-489.
9. MIZOGUCHI, R. and SHIMURA,M., **A Non-parametric Algorithm for Detecting Clusters Using Hierarchical Structure**, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol. pami-2, No.4, 1980, pp.273-3oo.
10. FROMM, F.R. and NORTHOUSE,R.A., **CLASS: A Non-parametric Clustering Algorithm**, PATTERN RECOGNITION ,Vol. 8, 1976, pp.107- 114.
11. GITMAN,I. and LEVINE, M.D., **An Algorithm for Detecting Unimodal Fuzzy Sets and Its Application as a Cluster Technique**, IEEE TRANS. COMP. Vol. C19, 1973, pp.583-593.
12. JOHNSON,S.C., **Hierarchical Clustering Schemes**, PSYCHOMETRIKA,Vol. 32, 1976, pp. 241-254.
13. FUKUNAGA, K. and KOONZ, W.L.G., **A Criterion and an Algorithm for Grouping Data**, IEEE TRANS. COMPT., Vol.C19, 1973, pp.917-923.
14. KOONZ, W.L.G. and FUKUNAGA,K., **A Non-parametric Valley- Seeking Technique for Cluster Analysis**, IEEE TRANS.COMPT., Vol.C21, 1976, pp.171-178.
15. AUGUSTSON, J.G. and MINKER, J., **An Analysis of Some Graph- Theoretical Cluster Techniques**, J.ASS.COMPT., Vol. 17, 1970, pp.571-588.
16. ZAHN, C.T., **Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters**, IEEE TRANS. COMPT. ,Vol.C20, 1971, pp.68-86.
17. HUBERT, L., **Monotonous Invariant Clustering Procedures**, PSYCHOMETRIKA, Vol. 38, 1973, pp.47-62.
18. PATRICK,E.A. and FISHER, F.P., **Cluster Mapping with Experimental Computer Graphics**, IEEE TRANS.COMPT., Vol. C18, 1969, pp.987-991.
19. JARVIS, R.A. and PATRICK, E.A., **Clustering Using a Similarity Measure Based on Shared Near Neighbours**, IEEE TRANS. COMPT., Vol.C22,1973, pp.1025-1034.
20. **Graphical Devices**, IEEE Proceedings, Special Issue on Computer Graphics, Vol. 62,No.4,April 1974.
21. WILLIAMS, R., **A Survey of Data Structures for Graphics Systems**, COMPUTING SURVEYS, Vol. 3, No.l, March 1971, pp.2-21.
22. DIDAY, E., **La méthode des nuées dynamiques**, REV.STAT.APPLIQUEE, Vol. XIX, No.2, 1971, pp.19-97.
23. BAHADUR,T.W and BAHADUR, R.R., **Classification into Two Multivariate Normal Distributions with Different Co-variance Matrices**,ANN.MATHS.STAT., 33,June 1962, pp.420-431.
24. FISHER, R.A., **The Use of Multiple Measurements in Taxonomic Problems**, HUMAN GENETICS, 6, 1986, pp.179-188.