CERN - DATA HANDLING DIVISION
E.S. Gelsema
DD/74/16
May 1974

DESCRIPTION OF AN INTERACTIVE CLUSTERING TECHNIQUE AND ITS APPLICATIONS

DD-pp

1. Introduction

Clustering analysis is a tool much used in pattern recognition. It has a wide range of applications in diverse fields such as biology, medicine, high-energy physics, space research and many others.

The problem that cluster analysis addresses itself to may be stated as follows: Suppose there is a sample of objects which may be flowers, biological cell images, interactions in a bubble chamber. Let each object be characterized by a number (n) of observations or parameters. Then each object may be represented by a point in an n-dimensional space. How should one partition this n-dimensional space into regions, such that objects in any one such region have more resemblance to each other than to objects in other regions. The aim, therefore, is to identify, in the n-dimensional space, subspaces with a high object density. Such a subspace will then be the representation of a whole class of objects. Any subsequent object with a point representation within such a subspace would then be assigned to that same class.

Such a partition of the n-dimensional space may be arrived at in one of two different ways: supervised or unsupervised. Clustering analysis falls in the category of unsupervised learning.

In supervised classification a representative sample of labelled events is normally used prior to the actual classification procedure, the label denoting membership of a class. Class properties in terms of the observations are established using such a "training set". Supervision consists of assigning the labels to the events in the training set. Alternatively, a training set may stem from an entirely different source, such as a Monte-Carlo procedure which may generate labelled, artificial events.

In unsupervised learning no labelled training set is used; categories are identified in the observation space as regions of high event density, surrounded by low density regions. Many methods to achieve this have been described.[1,2]. They differ in the assumptions made about the distribution of the parameters within classes (no assumptions lead to the so-called parameter-free techniques) and in the amount of other a-priori knowledge that one admits to start the procedure. For example, one may or may not specify the number of classes to be identified.

Once an unsupervised procedure has been found applicable to a given problem, its advantages over supervised classification techniques are:

i) One does not need to provide a separate training set, which may involve lengthy Monte-Carlo calculations.

ii) Appropriately designed clustering techniques may be able to find new, unexpected categories or subdivisions within known classes.

iii) Clustering techniques are insensitive to a slight shift in one or more of the parameters, which may occur in one batch of data with respect to another, as long as the batches are treated separately. For example, in white blood cell recognition, the staining of the cells may differ considerably between different preparations. This will result in a shift of the colour parameters in the observation space. Both preparations, as long as they are not mixed, will yield clusters, but they will occupy slightly different regions in the observation space. In supervised classification, unless appropriate calibration is done, such a shift presents a problem.

In all classification problems, the result, i.e. the partitioning of the observation space is difficult to visualize when the dimensionality is greater than two. It is therefore useful if a procedure uses a built-in criterion to judge its own performance and/or if it somehow presents the achieved separation of classes in visual form.

The clustering procedure described here has the following features:

i) It assumes a Gaussian distribution of the values of each parameter within classes.

ii) It uses a criterion function which evaluates the quality of the clustering achieved.

iii) It is interactive and iterative.

iv) Between iterations it displays in some way the separation of classes achieved so far and provides information on the state of convergence.

In Section 2 the mathematical model will be outlined briefly. Section 3 gives details about the procedure and in Section 4 the results for a number of applications will be given.

2. Mathematical Model

Consider a random vector x from a probability density distribution $h(x|b^*)$. The distribution is characterized by a distribution parameter vector b^*, which we shall assume to be fixed but unknown. The vector b^* may specify a mixture of M distributions:

$$h(x|b^*) = \sum_{k=1}^{M} f(x|b_k^*)\ P_k \quad ,$$

where the parameters b_k^* characterize the distribution k, occurring in the mixture with mixing parameter or weight P_k.

The problem of cluster analysis may now be stated as follows: Given a set of observations from a density $h(x|b^*)$, find the distribution parameter vector b which is the best estimate of b^*, consistent with this set of observations. It should be noted that b contains information about the individual b_k as well as about the mixing parameters P_k.

For a given choice b of the distribution parameter vector, define the information function $\eta(b,b^*)$ as the expected value of the natural logarithm of the mixture density:

$$\eta(b,b^*) \equiv E\{\ln h(x|b)\} = \int \ln\{h(x|b)\}\ h(x|b^*)\ dx$$

It can be shown [1] that the vector b that maximizes $\eta(b)$ (from now on we will omit the fixed vector b^* from the notation) corresponds to the asymptotic minimum-risk solution.

In ref.1 it is shown that for a given dimensionality of the vector x and under the assumptions that the category densities are Gaussian (i.e. b_k is given by a vector of means m_k, and a covariance matrix Σ_k) and separated (i.e. the densities are truncated at the category boundaries), maximizing $\eta(b)$ corresponds to maximizing

$$\eta_G(b) = \sum_{k=1}^{M} P_k \ln\left\{ \frac{P_k}{|\Sigma_k|^{\frac{1}{2}}} \right\}$$

Here $|\Sigma_k|$ denotes the determinant of the covariance matrix of category k. Indeed, as $|\Sigma_k|^{\frac{1}{2}}$ is related to the volume in the observation space occupied by category k, maximizing $\eta_G(b)$ corresponds to finding that partition of the observation space into regions such that the average probability density is maximum. A procedure that maximizes $\eta_G(b)$ will therefore tend to locate clusters in the observation space.

3. Description of the procedure

This procedure is a modified version of a clustering technique described by Patrick [3]. This book also contains the mathematical justification, of which the previous section is a condensed account.

The essential features of the modified version are as follows:

a) The number of classes to be detected is fixed.

b) It is assumed that initial cluster-nuclei can be defined on the basis of the observations, or otherwise.

c) Any event may be assigned to one of the classes, or it may be left unclassified as an isolated point. This permits the treatment of the case of clusters superimposed on a background.

d) The procedure uses a criterion function $\eta(b)$ to guide the partitioning of the observation space. It is assumed that the multi-variate mixing distributions are Gaussian and separated.

e) The procedure works interactively, in successive iterations during which any event may change its class-membership, globally under control of the user. Updating is performed at the end of each iteration step.

f) Termination of the iteration process is decided interactively by the user.

At the start, cluster nuclei are formed by events fulfilling selection criteria to be supplied for each class. These criteria may be specified as regions in one or some of the parameters, or they may be defined by values other than those used in the clustering procedure. The example concerning the prism-plot variables in bubble chamber events (Section 4) uses the latter alternative.

For each cluster nucleus i, the vector of averages, m_i, the covariance matrix Σ_i and the fraction of events contained in this cluster nucleus, P_i are evaluated, thus giving rise to an initial value of $\eta(b)$, expressing the quality of the initial clustering solution.

The iteration procedure is then started. In each step, the parameter values of an event are used and, regardless of whether it is an event already assigned to a cluster or not, M new values of $\eta_i(b)$ are evaluated, assuming that the event be assigned to any one of the M existing clusters in turn. The best assignment corresponds to the larger value $\Delta\eta_{max}$ in the sequence

$$\Delta\eta_i(b) = \eta_i(b) - \eta_{old}(b) \quad , \quad i = 1,2,\ldots,M$$

The value $\Delta\eta_{max}$ and the corresponding value of i are recorded together with the event. No updating of the classes is done at this stage.

The calculation of the sequence $\Delta\eta_i$ is repeated for all events and at the end, for each class i in turn a histogram of all the values $\Delta\eta_i$ thus obtained is displayed on the screen.*) Events which really belong to class i will have larger values of $\Delta\eta_i$ than other events. Examples of such histograms are given in figure 2. The user is then asked to provide interactively a

*) The procedure is implemented on a CDC-6600 to which is attached an Argus 500 computer driving a Ferranti refresh display.

threshold value τ_i of $\Delta\eta_i$, above which events should be accepted in category i. This is repeated for all categories i.

At this stage the categories are updated. An event not yet contained in class i is entered into that class if for that event both conditions:

$$\Delta\eta_{max} = \Delta\eta_i \quad \text{and} \quad \Delta\eta_i \geq \tau_i$$

are satisfied. Inversely, an event already contained in class i is removed if for that event one or both of the conditions:

$$\Delta\eta_{max} \neq \Delta\eta_i \quad \text{or} \quad \Delta\eta_i < \tau_i$$

are satisfied.

A table is then displayed, giving the numbers of events added to and removed from each of the clusters in the course of the last iteration step. When the solution stabilizes, these numbers become small. After each iteration step the user has the choice of terminating the procedure or going into a next step.

4. Applications

A. The first test to which this program has been submitted is in the field of white blood cells. Mature white blood cells may be classified in one of the following five categories: neutrophils, lymphocytes, monocytes, eosinophils and basophils. These cell types may be distinguished from each other on the basis of criteria, measured on microscopic images, such as cell size, colour, shape of the nucleus and many others. A file containing over 40 of such criteria for each of a total of over 1000 cells was available from previous work [4].

A sample of 100 of these cells was chosen as a sample to test the program performance. The program was asked to grow clusters corresponding to neutrophils (NEU), lymphocytes (LYM) and eosinophils (EOS) on a background of monocytes and basophils. Clustering was to be performed on the basis of 5 criteria, known from previous work to be among the best in separating power. Fig.1 is a two dimensional projection of the test sample of 100 cells.

After application of the initial selection criteria, the program started from a situation as described by the confusion matrix to the left in Fig.2. In this matrix the first column shows the class names according to the known classification. In this zero order solution, 58.9% of the cells belonging to one of the three types considered are correctly classified. The histograms of $\Delta\eta$ for the three categories are also shown in Fig.2. Thresholds were assigned according to the arrows. The confusion matrix to the right shows the classification after the first iteration. The values of η corresponding to the situation before and after the first iteration are also given.

In this example, the solution stabilizes after iteration 4. Fig.3 shows the histograms of $\Delta\eta$, the confusion matrices and the values of η before and after this final iteration. In the final solution 90.4% of the cells belonging to one of the three types are classified correctly. All misclassifications could be attributed to failures in the process of feature extraction, in each case affecting at least one of the five parameters seriously. It is interesting to note that the right hand peaks in the histograms corresponding to iteration 4 are better separated than in the histograms corresponding to iteration 1.

On the same sample of cells the procedure was repeated, starting from more stringent initial selection criteria. (50.8% correct). This run yielded after 4 iterations the same solution as the previous one.

B. A classical data set to illustrate clustering algorithms is the set consisting of four measurements on a sample of 150 flowers of three species of iris. These data were listed by Fisher [5]. The observation space is four-dimensional. Fig.4 shows a two-dimensional projection. Fig.5 displays the same data as Fig.4 but with the three different class-memberships indicated by different symbols. The result of the program using the two variables only is given in Fig.6. This result corresponds to the confusion matrix shown in Fig.7. The percentage of correct classification is 90%. As can be expected from inspection of Fig.4, separation of the species versicolor and virginica presents a problem. Indeed, intuitively there is no reason to prefer the clustering as in Fig.5 over the one in Fig.6. Also, the η criterion does not favour

one of these. The clustering criterion takes the value -2.75 for the solution represented in Fig.5 and -2.74 for the solution of Fig. 6.

The same classification is obtained when using all four variables.

C. Finally, this clustering technique has been applied to measurements of particle interactions in a bubble chamber. The parameters used are the prism plot variables [6], i.e. in the 3-body final states studied here the total energies of two of the outgoing particles, the Van Hove angle [6] and a parameter related to the peripherality of the interaction.

Events for three types of interactions were generated artificially by a Monte-Carlo program.

1. $\pi^- p \rightarrow p+\pi^-+\pi^o$ at 10 GeV/c.

The data base consisted of 100 events of each of the following channels.

a) $N^*(1470) + \pi^o$ (N^*)

b) $p + \rho(765)$ (RHO)

c) $p + \pi^- + \pi^o$ (PHSP)

The objective was to identify clusters corresponding to the N^* and RHO channel on a background represented by PHSP events.

Figs.8 and 9 show the details of the first and last (fifth) iteration. The insets in Fig.8 are the parts of the histograms with positive $\Delta\eta$. Of the events generated as N^* or RHO, 61.5% are classified correctly after application of the initial selection criteria. These criteria were specified as regions in the invariant mass distributions of the $(p\pi^-)$ and $(\pi^-\pi^o)$ particles combinations for the definition of the N^* and RHO cluster-nuclei, respectively. The percentage of correctly classified events rises to 98% after iteration 5. At that point the situation is stable. The value of the clustering criterion η rises from 1.54 in the initial state to 4.91 at the end of the fifth iteration.

2. $K^+ p \rightarrow p+K^o+\pi^+$ at 5 GeV/c

The data base consisted of 2010 events, distributed among the resonant states as follows:

a) $K^*(890) + p$ 817 events

b) $K^*(1420) + p$ 130 events

c) $\Delta(1236) + K^o$ 1063 events

The percentages of events in the different classes were chosen so as to correspond most nearly to the experimentally observed cross-sections for the different channels. Practically no phase space is observed.

Fig.10 gives the confusion matrices both for the initial state (after application of appropriate mass cuts) and for the state after 9

iterations. The percentage of correctly classified events, which is 42.8% for the initial cluster nuclei, corresponding to $\eta = 3.51$ rises to 93.2% in the final state, corresponding to $\eta = 7.36$.

These numbers may seem rather satisfactory. However, if this procedure were used on measurements of real events, as opposed to Monte-Carlo events, the actual"true" class-memberships would not be known. In such a situation only 70% of all events assigned to the category $K^*(1420)$ are expected to stem from that intermediate resonant state. These percentages are much better for the $K^*(890)$ and the $\Delta(1236)$, namely 98% and 96%, respectively. The clustering criterion η was also calculated for the "correct" solution. This yielded a value of 6.92, i.e. lower than the final value of 7.36 found by the program. This apparent paradox is a consequence of the non-validity of the assumption on which the procedure is based, i.e. the assumption of Gaussian distributions within classes of the parameters. It is well known that the prism plot variables violate this assumption.

Although the results of this method are not entirely satisfactory, as discussed above, they compare favourably with those obtained with the prism plot technique. The percentages of true events in the $K^*(890)$, $K^*(1420)$ and $\Delta(1236)$ channels obtained with this latter technique are 97, 62 and 97, respectively, [8] as compared to 98, 70 and 96 with the clustering method. The number of unassigned events is 52 with the prism plot technique versus 32 with the clustering method. The computer time requirement for this analysis is in favour of the clustering method by at least a factor of 2 in this particular example.

3. $K^-p \rightarrow p+K^0+\pi^-$ at 11 GeV/c

For this reaction, events from each of the following channels were generated, the numbers of events being proportional to the experimentally observed cross-sections:

a)	$K^*(890) + p$	1250 events
b)	$K^*(1420) + p$	262 events
c)	$\Delta(1236) + K^0$	214 events
d)	$N^*(1470) + K^0$	108 events
e)	$N^*(1670) + K^0$	100 events
f)	$p + K^0+\pi^-$	66 events

Fig.11 describes the situation before the first iteration and after iteration 6 in terms of the confusion matrix and the clustering parameter η. The same remarks as for the K^+p example apply here. Although the procedure improves the overall correct classification result from 51% to 91%, corresponding to $\eta = 4.32$ and $\eta = 7.30$, respectively, if used in an unsupervised way, some of the reaction channels, especially the $N^*(1670)$ would be much contaminated. Also, the basic assumptions are violated in this case, as evidenced by the value of $\eta = 6.91$ for the "correct" solution, which is lower than the value of 7.30 achieved by this procedure.

5. Conclusions

The clustering algorithm described in this text has been applied to a variety of data sets. In some cases the results are very promising. In such cases one always arrives at a final situation where in the histograms of $\Delta\eta$ the peaks for high values of $\Delta\eta$ are well separated from the rest of the events. A small change in the cut-off values τ_i will therefore not affect the stability of the final result. This gives an indication of the applicability of the algorithm to any given data set, even if the procedure is run in an unsupervised way.

In the cases where this algorithm does not yield completely satisfactory results, it still does go a long way toward the correct solution. In such cases the assumption of Gaussian distributions of all the parameters within classes is violated. It is believed that even in such cases the solution provided by this procedure is a good starting point for further investigations. Such further work would then have to use parameter-free techniques such as nearest neighbour algorithms.

References

1. E.A. Patrick, Fundamentals of Pattern Recognition, Prentice-Hall, Inc. 1972.

2. R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, 1973.

3. E.A. Patrick, ref. 1, pp 334-340.

4. P.W. Neurath et.al., Computer Identification of White Blood Cells, Int. Comp. Symp. 1973, North-Holland Publishing Comp. Also available as CERN/DD/73/27.

5. R.A. Fisher, Contributions to Mathematical Statistics, John Wiley and Sons, 1950.

6. J.E. Brau et. al., Prism Plot: A New Analysis of Multibody Final States, Phys. Rev. Lett. 27, 1481, (1971).

7. L. van Hove, Phys. Lett., 28B, 429 (1969).

8. A. Ferrando, Private Communication.

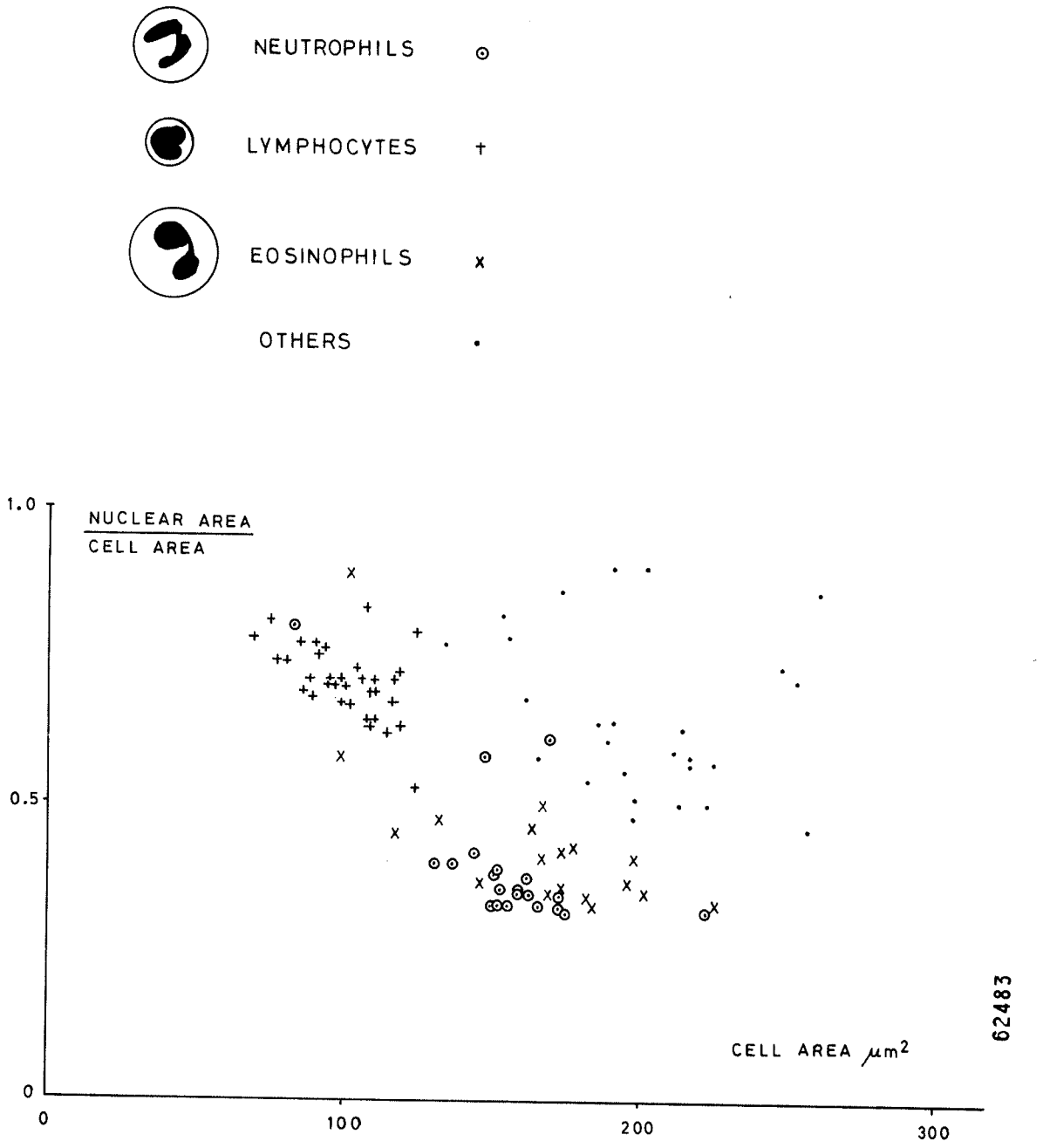
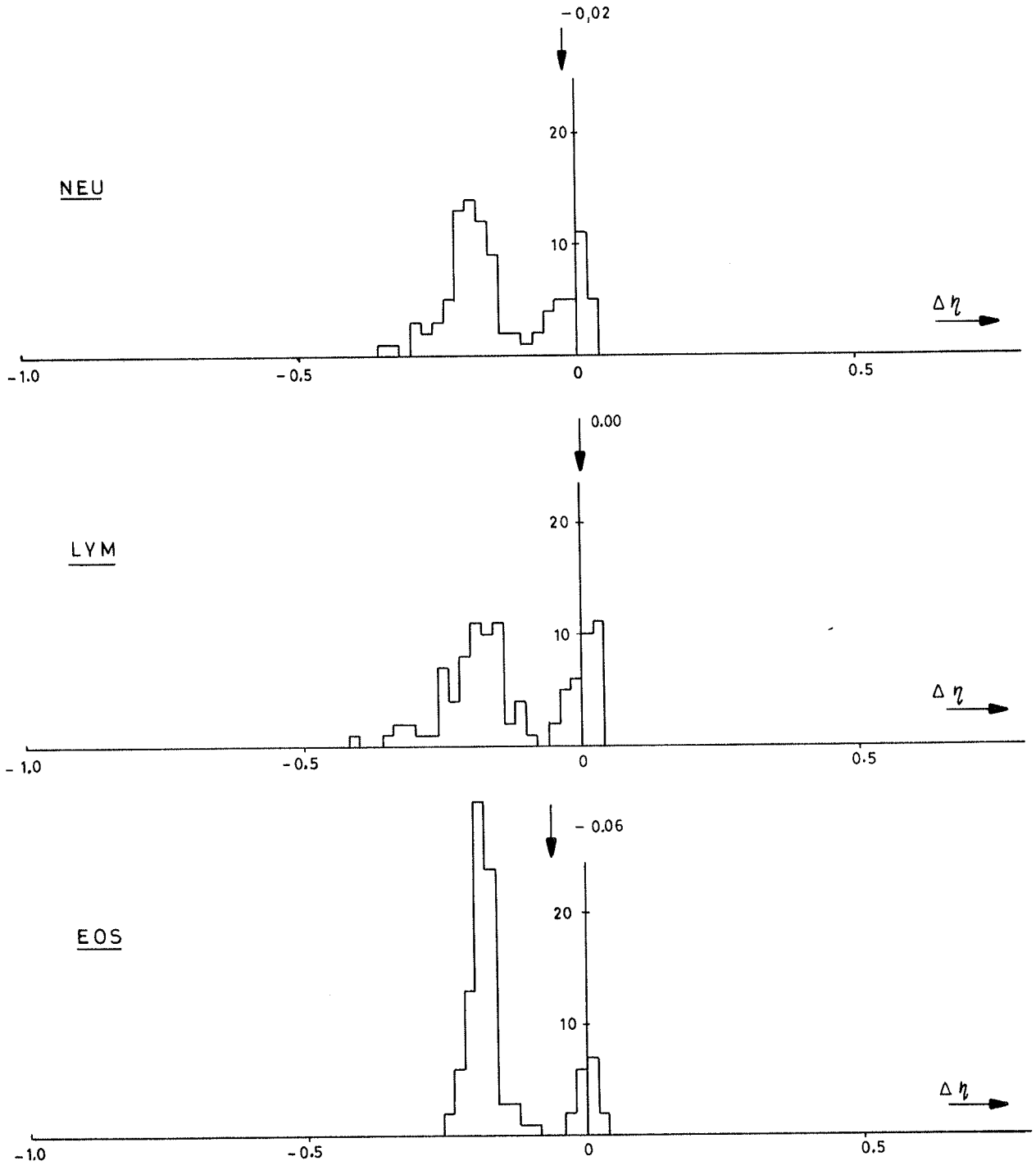


Figure 1 Scatter diagram of two parameters measured on a sample of 100 white blood cells. The objects indicated as "others" are mainly monocytes and basophils.



	NEU	LYM	EOS	BGR
NEU	14	1	0	6
LYM	0	17	0	17
EOS	3	1	12	2
BGR	0	0	1	26

$\eta = -1.74 \times 10^{-1}$

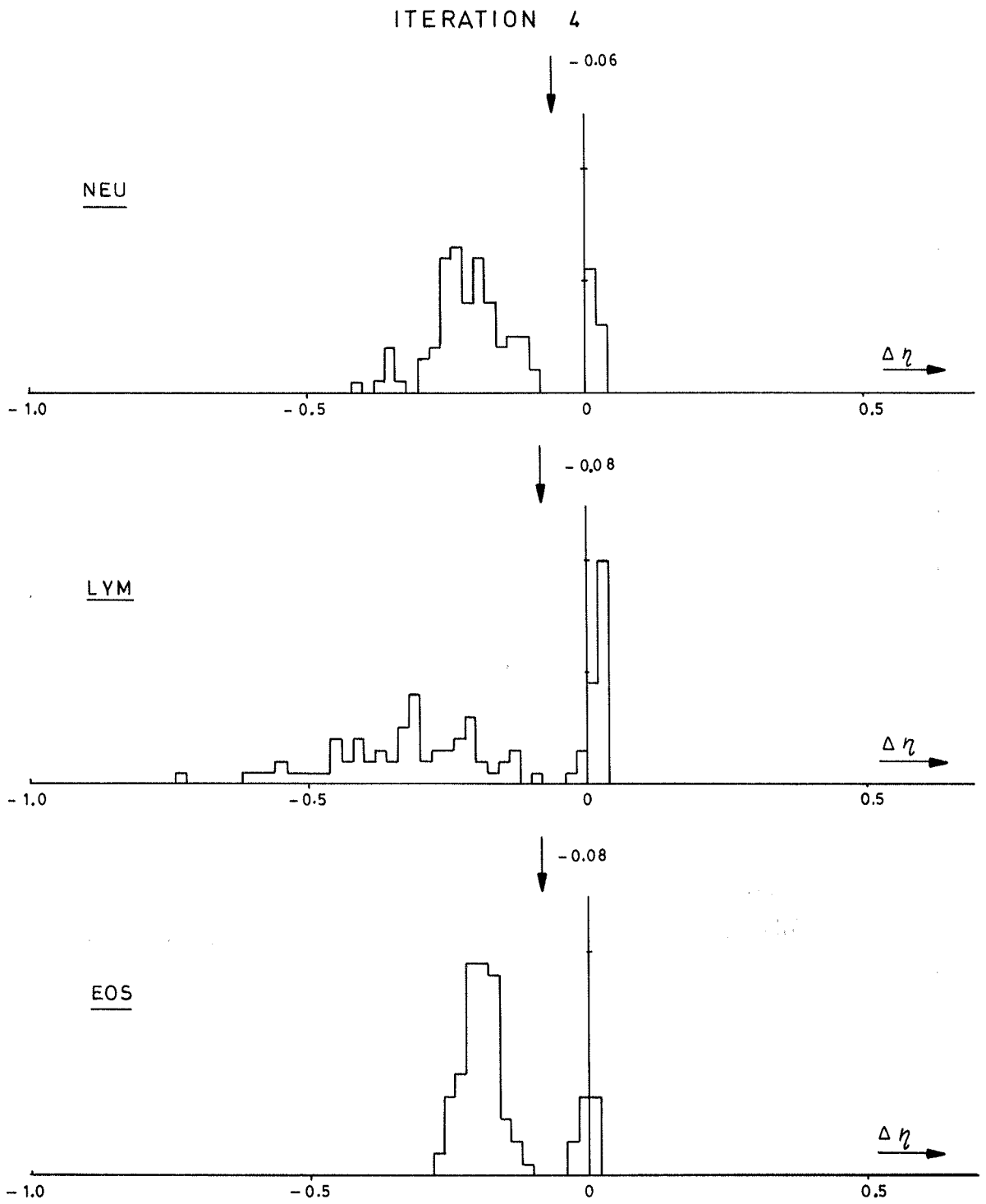
⟹

	NEU	LYM	EOS	BGR
NEU	17	0	0	4
LYM	0	18	0	16
EOS	0	0	16	2
BGR	0	0	1	26

$\eta = 6.84 \times 10^{-1}$

62484

Figure 2 Histograms of $\Delta\eta$ for three classes of white blood cells. Each object has an entry in each of the histograms. The value $\Delta\eta$ measures the similarity of the object to other objects of the indicated class. After application of the thresholds as shown, the confusion matrix changes as indicated below. The clustering quality is measured by η. The figure shows the first iteration.



	NEU	LYM	EOS	BGR
NEU	17	0	0	4
LYM	0	32	0	2
EOS	0	0	16	2
BGR	0	0	1	26

$\eta = 7.97 \times 10^{-1}$

⟹

	NEU	LYM	EOS	BGR
NEU	17	0	0	4
LYM	0	33	0	1
EOS	0	0	16	2
BGR	0	0	1	26

$\eta = 7.73 \times 10^{-1}$

62485

Figure 3 As figure 2, but for the fourth (final) iteration. After this iteration the solution is stable.

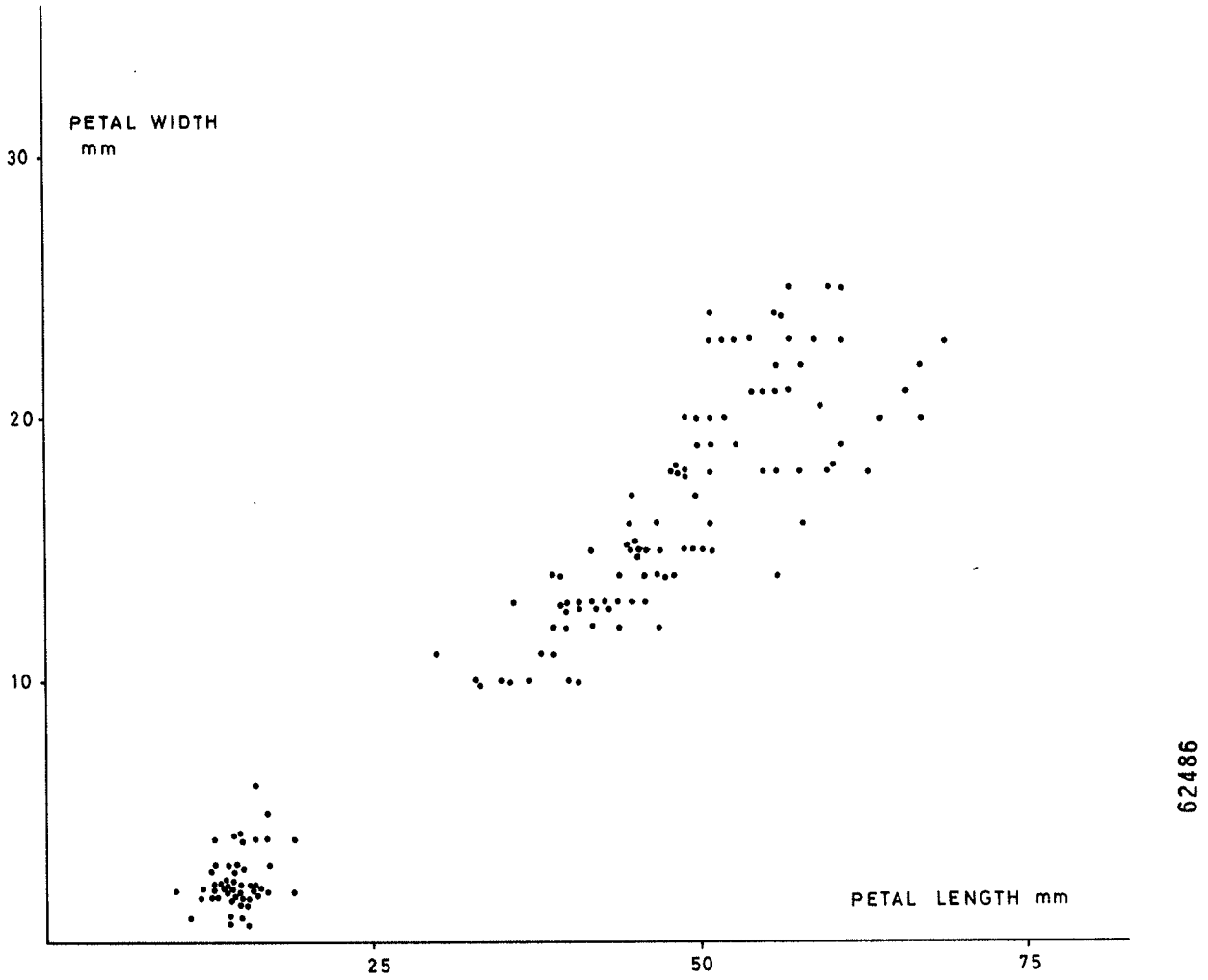


Figure 4 Scatter diagram of two parameters measured on three types of iris.

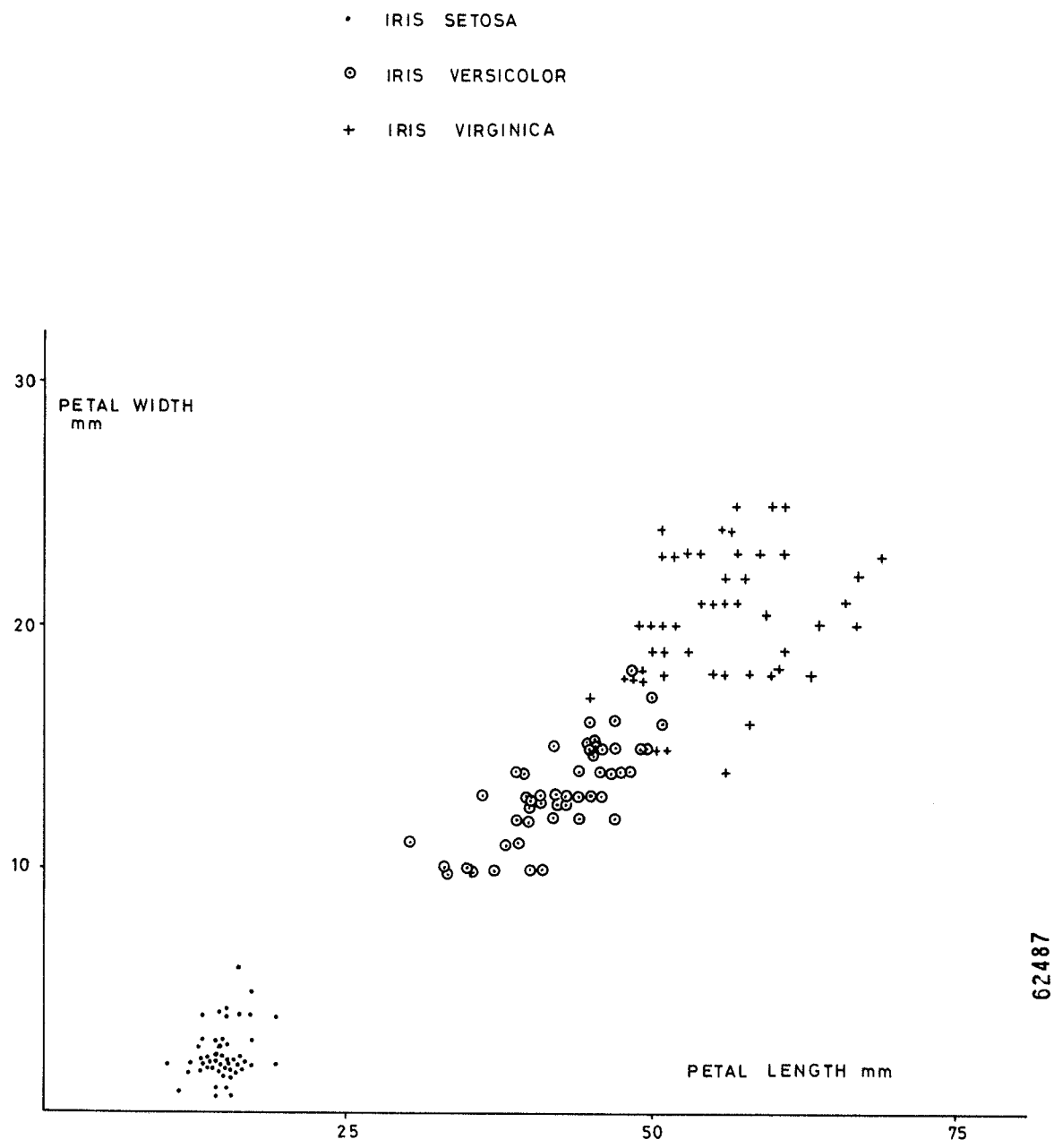


Figure 5 Same scatter diagram as in figure 4, now with the three types of iris identified.

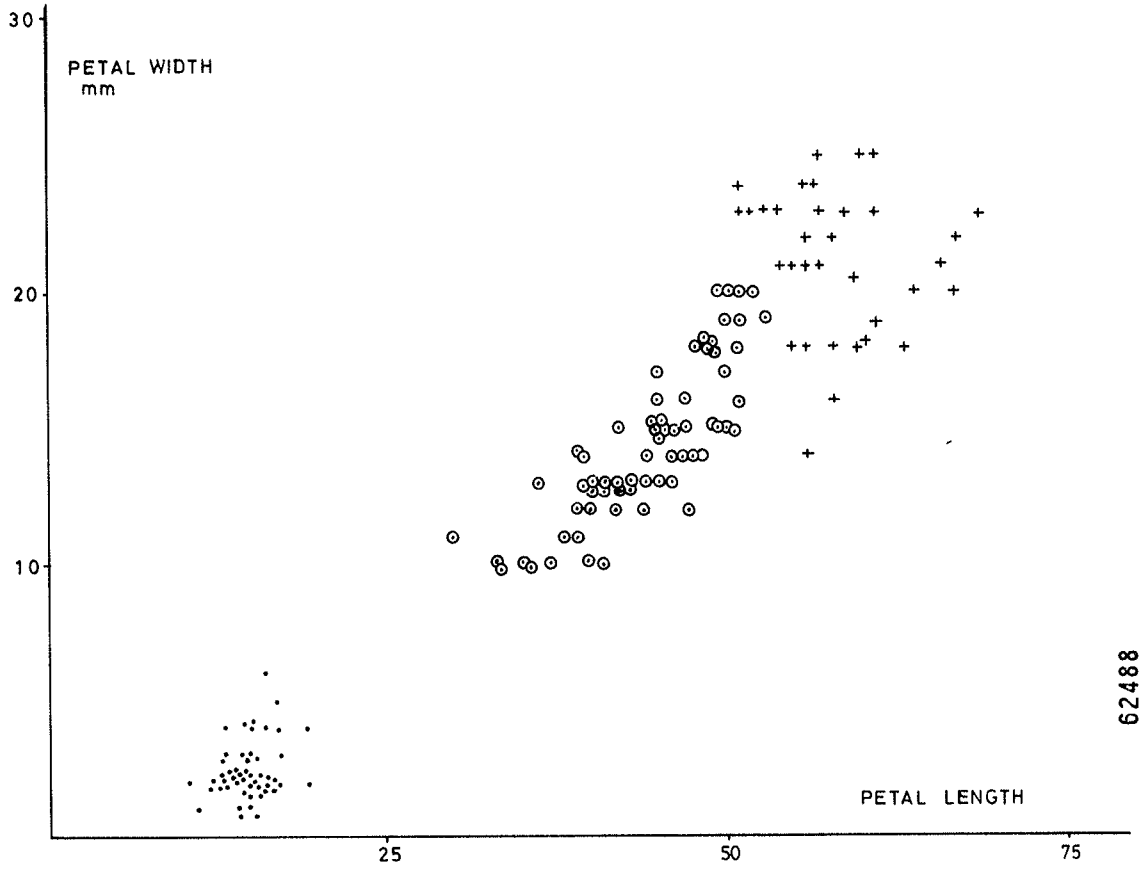


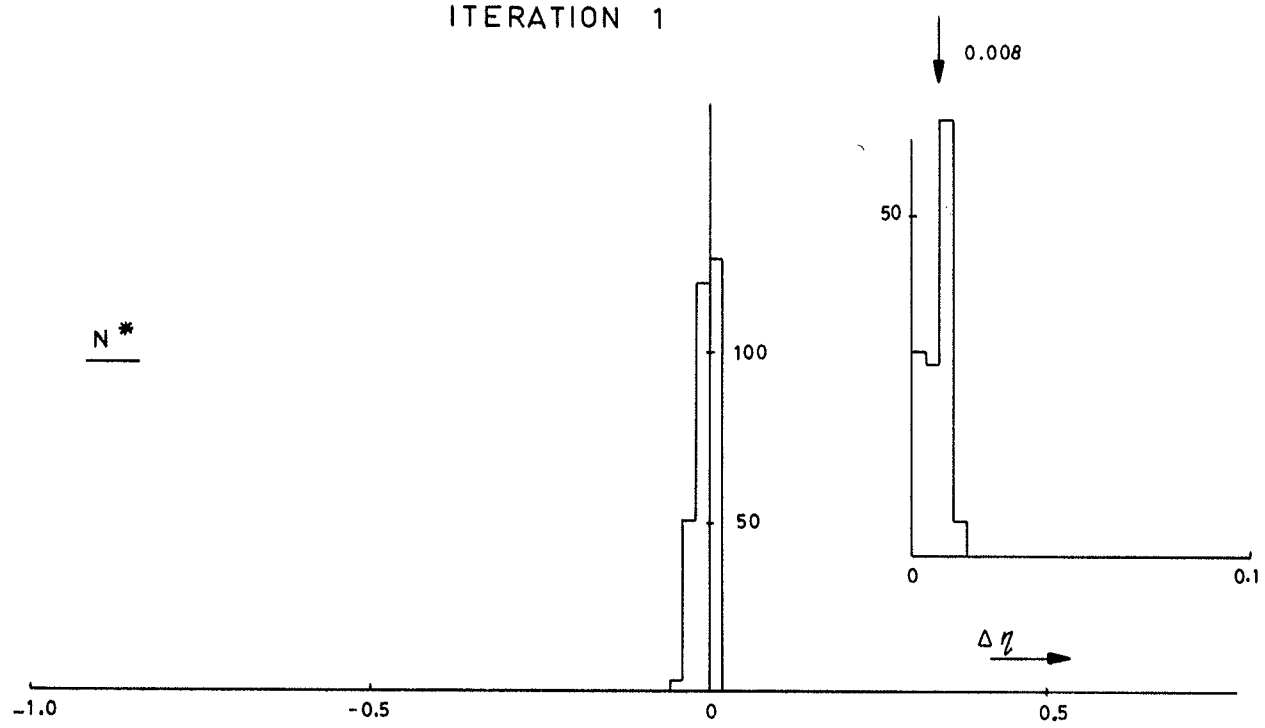
Figure 6 Same scatter diagram as in figures 4 and 5. The figure shows the three clusters as found by the clustering procedure.

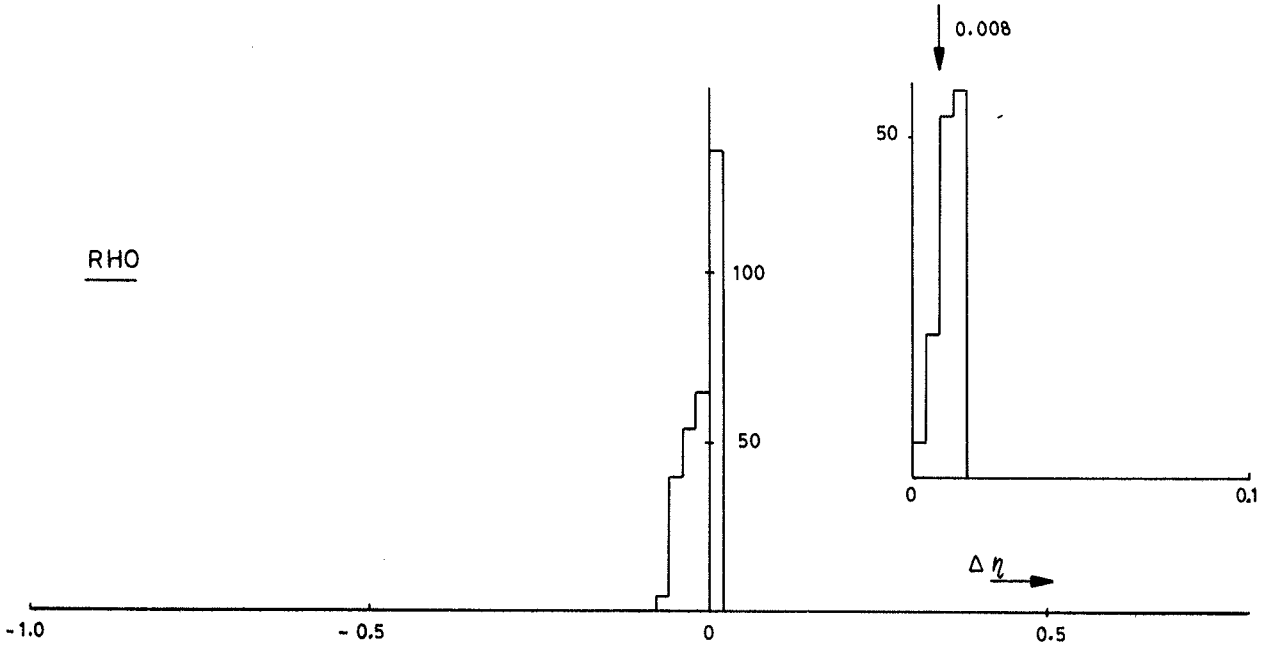
CONFUSION MATRIX AFTER 4 ITERATIONS

	SET	VERS	VIRG	BGR
SET	50	0	0	0
VERS	0	50	0	0
VIRG	0	15	34	0
BGR	0	0	0	0

62489

Figure 7 Confusion matrix for the solution shown in figure 6. The first column indicates the true classes, the first row the computed classes.





	N*	RHO	PHSP
N*	55	4	41
RHO	5	68	27
PHSP	6	1	93

η = 1.54

⟹

	N*	RHO	PHSP
N*	69	0	31
RHO	0	91	9
PHSP	0	19	81

η = 3.96

62490

Figure 8 Histograms of $\Delta\eta$ for the N* and RHO intermediate states in the reaction

$$\pi^- p \rightarrow p\pi^-\pi^0 \text{ at } 10 \text{ GeV/c}$$

The figure shows the first iteration. The insets are expansions of the parts with positive $\Delta\eta$.

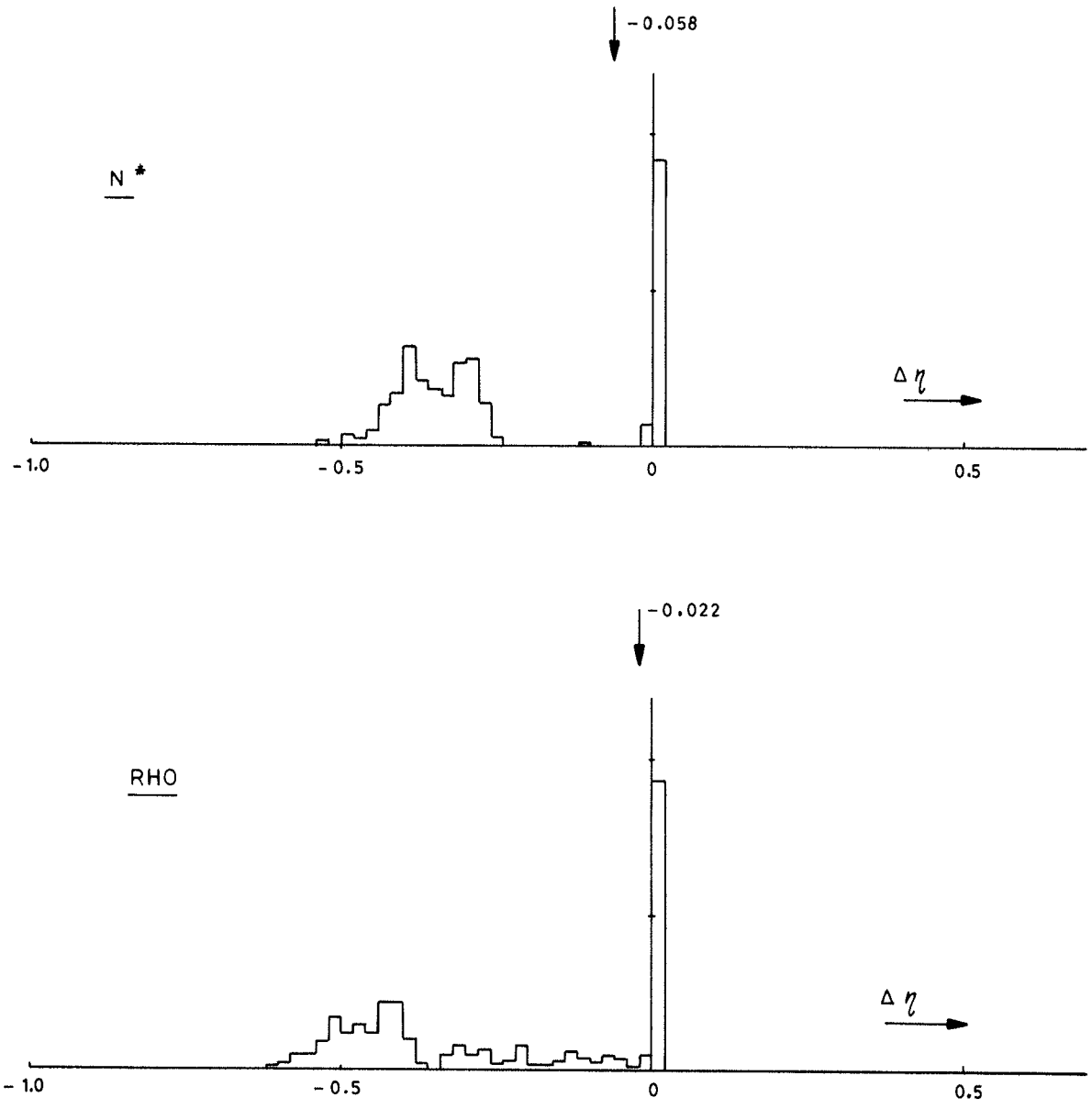


Figure 9 Histograms of $\Delta\eta$ as in figure 8, but for the fifth (final) iteration.

$K^+p \rightarrow K^0p\pi^+$ AT 5.0 GeV/c

CONFUSION MATRIX AFTER INITIAL SELECTION

	K*890	K*1420	DELTA	PHSP
K*890	373	0	7	437
K*1420	0	61	2	67
DELTA	0	12	426	625
PHSP	0	0	0	0

62482

$\eta = 3.51$

CONFUSION MATRIX AFTER 9 ITERATIONS

	K*890	K*1420	DELTA	PHSP	%
K*890	767	13	32	5	93.9
K*1420	2	109	13	6	83.8
DELTA	11	34	997	21	93.8
PHSP	0	0	0	0	
%	98.3	69.9	95.7		

62481

$\eta = 7.36$

η FOR "CORRECT" SOLUTION = 6.92

Figure 10 Three intermediate resonant states in the reaction $K^+p \rightarrow K^0p\pi^+$ at 5 GeV/c. The figure shows the confusion matrices before the first iteration (top) and after 9 iterations (bottom). The quality of clustering is measured by η. The "correct" solution is defined at the situation where each event is assigned its true identity.

$K^-p \rightarrow K^0p\pi^-$ AT 11.0 GeV/c

CONFUSION MATRIX AFTER INITIAL SELECTION

	K*890	K*1420	DELTA	N*1470	N*1670	PHSP
K*890	642	0	0	6	8	594
K*1420	1	128	2	1	3	127
DELTA	2	10	106	7	4	85
N*1470	3	7	2	56	7	33
N*1670	0	3	2	10	54	31
PHSP	1	1	2	1	5	56

62480

$\eta = 4.32$

CONFUSION MATRIX AFTER 6 ITERATIONS

	K*890	K*1420	DELTA	N*1470	N*1670	PHSP	%
K*890	1220	27	0	0	3	0	97.6
K*1420	6	218	7	13	18	0	83.2
DELTA	5	1	158	44	6	0	73.8
N*1470	6	0	4	84	12	2	77.8
N*1670	0	1	2	14	82	1	82.0
PHSP	4	15	2	4	9	32	
%	98.3	83.2	91.3	52.8	63.1		

62479

$\eta = 7.30$

η FOR "CORRECT" SOLUTION = 6.91

Figure 11 Five intermediate resonant states in the reaction $K^-p \rightarrow K^0p\pi^-$ at 11 GeV/c. Confusion matrices before the first iteration and after 6 iterations.