TR/10/85 April 1985

A.I. Approaches in Statistics

by

R.I. Phelps and P.B. Musgrove

A.I.Approaches in Statistics

R.I- Phelps, Brunel University, Uxbridge, U.K.

P.B. Musgrove, Polytechnic of the South Bank, London, U.K.

Abstract

The role of pattern recognition and knowledge representation methods from artificial Intelligence with in statistics is considered. Two areas of potential use are identified and one, data exploration, is used to illustrate the possibilities. A method is presented to identify and seperate overlapping groups within cluster analysis, using an A.I. approach. The potential of such 'intelligent' approaches is stressed.

1 .0 Introduction

The human brain in its development throug infancy faces a primary task of making sense of its environment. To achieve this it is adapted to extract structure from its environment using both inborn and learned mechanisms. Even lower animals use sophisticated strategies to structure their perceptual input: a pigeon, shown (repeatedly) only one example of a type of leaf, can generalize this concept to other leaves of the same type [HERRNSTEIN 84] . Such 'intelligent' mechanisms for structuring and processing information have proved very successful in evolutionary terms and it is the capabilities of this sort of processing which artificial intelligence (A.L.) wokers hope to capture . Although limited progress has been made using both rigid, unnatural forma1 methods and application - specific heuristics, the construetion of more generally intelligent programs has proved elusive.

This paper presents work which forms part of a lager A.I, research project on machine learning, in which an attempt is being made to model some of the basic processes involved in intelligent knowledge representation and processing, closely following experimental indications of how these processes operate in humans and animals [Phelps and Musgrove 35].We believe that this approach of following human behaviour in information processing is equally applicable in areas of statistics. We discuss the potential for A.I. approaches in these areas and present an example of such an approach applied to cluster analysis.

2.0 A.I. Approaches in Statistics

Two areas of statistics are identified in which we feel there is particular scope and need for A.I. approaches. These areas are exploratory data analysis and statistical modelling.

A commonly made distinction is between descriptive and inferential statistics. By descriptive is meant" the use of grouping processing to produce displays of data which can be easily processed for patterns by a human, i.e. the bringing out of structure inherent in the data.This is often regarded as a relatively elementary aspect of statistics best left to subjective considerations (e.g. what width should the intervals of a histogram be ?) about which formal statistical theory has little to say. Under the title of exploratory data analysis, formal attention has been paid to more complex and sophisticated techniques used on data where the structure that may exist is unknown or only vaguely appreciated (e .g cluster analysis, dimens ion reduet ion techniques) . Again, the aim is to find structure in the data that is comprehensible to the human brain.

Much of this activity is essentially concerned with finding and describing structure inherent in the data in the form of shapes, e.g. chocs ing interval widths to bring out the shape of a distribution from the data; looking Eor straight line or simple curve relationships in regression; finding relatively dense compact regions in clustering. Detecting shapes is something the human visual system is particularly well adapted to- it is a common complaint among non-statisticians that statistics only confirms the presence of relationships that are already 'obvious' and statistical tests are sometimes used merely as "objective" corroboration of an experimenter's beliefs. Pattern recognition techniques developed within the A.I. paradigm are not as yet anything 1 ike as good as human perception, but we feel that this approach of attempting to extract structure from data in the form of patterns that humans extract is a promising one. Exploratory data techniques in statistics tend either to be simple (e.g. stem and Leaf) and correspondingly limited in their pattern finding abilities, or mathematically complex but still only capable of detecting very limited types of pattern (e.g. existing clustering algorithms). Following the human approach to finding patterns holds the potential of more powerful and flexible techniques giving output in a form suitable for human assimilation. This approach means abandoning some of the mathematical rigour of existing statistical methods, *a* price we feel it is worth paying to obtain more informat ion – rich results. The cluster detection work presented below is an example of such possibilities.

By statistical model ling, the second area identified as suitable for an A.I. approach, we refer to models of complex inference systems, typically involving chains of reasoning to lead from pieces of evidence to conclusions regarding the validity of alternative hypotheses. This is the type of inference procedure often modelled as an expert system. The relative success of expert systems in such fields as medical diagnosis is well known [Szolovits, Ed, 82] . Again, the lesson to be drawn from the A.I. approach is the importance of finding and using structure in the problem domain in a way similar to humans.

The problem for a purely statistical approach to such inference problems is *the* impossibly large number of combinations of interacting variables which need to be considered. This imposes impractical data requirements and leads in practice to the simplifying assumption of independence. From this viewpoint, the basic aim is to find a mapping from the input (evidence) onto the output (validity scores for the hypotheses, based on posterior probabilities). This is essentially a 'bLack box' process since no internal model of the problem domain is constructed: the essential structure which humans use to think about the problem is missing. (it should be noted that this criticism also applies to many expert systems which simply follow a set of expert-derived rules for a particular problem. This amounts to treating the expert as a 'black box' who applies various rules to the problem, without any attempts being made to understand the knowledge structure which produced and governs those rules.) The construction of a causal model representing human understanding of a problem as in [Kuipers and Kassirer, 84], would include the identification and modelling of interact ions between variables and provide a basis for simplifications, such as independence assumptions, where supported by the model. A further advantage of such a causal model is its potential use in producing an explanation of the inference process in terms similar to human explanation.

In both of these areas of statistics we have stressed the need to find the sort of structure in problems that the human brain is adapted to find. Modelling these forms of human thought, as in the A.I. disciplines of pattern recognition and knowledge representation, would provide improved achievement of statistical aims

3.0 CLUSTER DISCRIMINATION

Existing cluster analysis algorithms do not perform well across differing data sets. Some assume normal distribution forms, others assume spherical clusters and most do not perform well even on simple cases such as two distinct ellipsoids [Everitt, 80], Furthermore, they either need to be told the number of clusters to be found or they provide measures which it is hoped will show a marked 'jump' at the optirnum number of clusters, but which are unreliable in practice [Everitt, 80, Section 3.4] . Many involve parameters which need to be adjusted for each data set in order to provide good performance. Consequently there is a need for a method which does not require parameter adjustment, which can locate the 'correct' number of clusters and which will distinguish be tween clusters that humans can distinguish. The baseline for comparison in these tasks is human 'expert' performance since humans show much better performance than existing statistical techniques, at least in 2 -D.

The approach taken is the A.I. one of modelling the functions of human performance in this task, although not necessarily performing each function in quite the same way as a human. Human behaviour is thus followed as closely as possible while statistical techniques are used to optimize parameters within the algorithm. There is a problem here. With the, present state *of* cognitive research we do not know precisely how humans perform any pattern recognition task, and we thus have to rely on a combina*t*in of psyctiometric results (which are susceptible to several interpretations) together with our intuition of how we perform these tasks.

In the absence of firmly established experimental results, modelling from intuition is often an important component in the development of pattern recognition (and A.I.) techniques, and it is so here.

Two major factors that have been suggested in the cognitive science and A.I. literature which relate to visual perception e.g. [Narr 82] are a) texture and b) shape. In the case of clusters we specialize these to a) dot density relative to background dens i ty and b) convexity.

The first step is therefore to find areas of relatively high dot density. This is achieved statistically, using the approach of potentials [Devivjer and Kittler, 82] which models the human perception of clusters as darker areas, to find contiguous sets of point with a density at least M times greater than could be accounted for by random phenomena. M is in effect the ratio of cluster density to background density that is needed for human recognition of a cluster. From preliminary investigations of human discrimination this is set to 2 (but further psychological work is necessary). This first step leads to good agreement with human assessment of 2-0 data sets in the sense that cases causing problems for other clustering algorithms do not cause problems for this method , e.g. the distinct ellipsoidal clusters mentioned above are picked out as two seperate wholes , unless overlapping clusters are present . In that case the algorithm will indicate the presence of just one complicated cluster, whereas humans generally prefer the hypothesis of separate but overlapping clusters.

This identification of relatively dense areas in a necessary preliminary to considerations of shape. The algorithm outlined above is used but our analysis of cluster shape is applicable to any statistical cluster algorithm which detects such areas.

In illustration of our arguments for using pattern recognition techniques where shapes are involved, we examine the use of cluster shape to distinguish overlapping clusters. The basis for doing this is the human preferenee for convex shapes, in the sense of attempting to allocate points to separate classes which are roughly convex in shape, when dividing up a nonconvex shape. (Humans can, however, divide clusters up into nonconvex constituents if these exist, but to model this would require further A.I. work). To model this behaviour the initial need is for an idea of the 'shape' of the dot cluster. Results in line with human perception cannot be obtained simply by joining together exterior points of the cluster. The exterior points are typically of a 'noisy' and less characteristic nature than points toward the centre of the cluster. Thus apart from their joined shape often being erratic and heavily indented, even an smoothed version of this shape is general ly unrepresentative of percieved cluster shape .

Percept ion of the shape of a cluster seems to be affected by perception of its centre. To model this 'centre-edge' effect the shape of a cluster is derived from a consideration of the shape of the denser interior areas of the cluster as we11 as the outer surrounding dots . The first procedure is to apply a common pattern recognition technique to smooth out the Local dot densities within the cluster, FIG 1.

A grid is placed over the data points. The size of cells (hypercubes) comprising the grid is dtermined by giving each side of the cell length

$$L = \left[\frac{r_1 \; r_2 \cdots \; r_d}{3^d \; .N} \right]^{\frac{1}{d}}$$

where d is the mumber of diraens ions, N the number of points and r the range of values m dimension i. A 3^d window, consisting of a larger hypercube with each side length 3L, is then passed over the grid to obtain an array of counts, the count for each cell being the number of points falling in the window centred on that cell. The value of L is chosen so that for a uniform distribution of points the expected number per 3^d window is 1 , thus avoiding too fine or too coarse a grid. The result is shown in FIG 2. This grid is used to define a 'chessboard' metric: a point is contiguous to any of the 6 surrounding points, and the distance to a diagonal neighbour is one unit, the same as to a neighbour above, be low or at its side.

The second procedure is to locate the densest 'core' area of the cluster. T,o find this we start with cells containing the highest count value and include any forming contiguous regions of 3 or more cells. We then look at the next highest count and include cells contiguous to those already included or forming their own contiguous groups. This process continues until about 30% of cells have been included.

This core area is then grown outwards by utilizing a dynamic programming algorithm which starts with a value v = the highest cell count -1, and notes if cells outside the core with this value or higher form contiguous regions or are isolated (region of size 1). For each such region it calculates the path with maximum average cell count form the region to the core. If the average cell count taken over the region plus this path is at least v , then this region is added to the core. v is then set to v-1 and the process repeated.

This algorithm is used until a set percentage (usually 90%) of the original cluster area has been included, FIG 3. The contour around this area (or the average of several such normalized contours for different percentages) has been found informally to give an impression of cluster shape more in accord with human descriptions. The reader may personally judge the shape obtained from FIG.3.

Now that an idea of the shape has been formed, the next step is to model the human preference for convex parts rather than an irregularly shaped whole. There are two possible approaches to dividing up the whole into convex parts . One is to look for concavities in the shape contour and to cut the whole by lines joining pairs of concavities. This method has often been suggested e.g. [Marr and Nishihara, 78] but has problems regarding which pairs of concavities to Link to form cuts and problems due to digitization of the image, where a diagonal Line will be represented by a 'step function' pattern of points with corresponding concavities. The other approach , which is related to the medial axis transform [Blum , 73] is to note that for convex shapes, central points are most distant from all the edges. The shape is therefore 'contoured' so that interior points are given a height propertional to their distance from the nea rest edge, FIG 4 • Peaks of this contoured shape then form potential centres for subclusters. Peaks are found by dividing all cells of count L into contiguous regions. Any such region touching a cell of higher count is deleted. This process is repeated with cells of successively higher counts until all cells have been processed. The regions remaining form the peaks.

The whole shape is then divided into regions about each peak by starting with each peak and adding contiguous cells of equal or Lower count . Cells contiguous to this region are then added if they are of equal or lower count than an adjacent cell included in the region on or be fore the last iteration. This process is repeated until no further cells can be added. Cells which could be added to regions grown from two (or more) peaks are allocated on the basis of gradient. The gradient from each peak to the cell is calculated and the cell assigned to the peak with which it has the steepest gradient. This process results in FIG 5.

Contour irregularities often result in spurious peaks being formed, and so the regions that have been found are now tested to see if they can be joined to adjacent regions without significant Loss of convexity. Each region and each feasible union of regions is fitted with a convex hull.

The ratio of space (hull minus reg ion) volume to hull volume is calculate for each region and union of regions. A statistical test can now be performed in an approximate, incuitive manner to test whether the decreas in this ratio which occurs in going from individual regions to their unions is statistically significant, and to pick out the optimal unions if there are two or more conflic t ing possibilities.

We use a difference of two proportions test at the 5% significance level to successively merge two regions.

$H_0 \ : P_1 = P_2$

$H_1 \ : P_1 < P_2$

where P_1 - proportion of space averaged over the two separate hulls in a possible union

$P_2 =$ proportion of space in the hull around the union of the two regions.

The test statistic is $(P_1 - P_2)/\sigma$,

where $\sigma = \sqrt{pq\,(\frac{1}{n_1} + \frac{1}{n_2}}$

n = total number of cells in the two individual hulls

n = total number of cells in the hull around the union of the two regions

$$P = \frac{n_1\ P_1 + n_2\ P_2}{n + n}$$

If the test is not significant the union of the two regions is replaced by one region. The test is then repeated for further possible mergers. Of course, here the spaces are not distributed binomially as use of this test strictly requires, but it has proved effective in practice and we believe it to be preferable to an ad—hoc approach.

In this way we arrive at a division of the original cluster area into approximately convex parts. The outcome of this process is given in FIG 6. Here, the algorithm has ascribed the points in the intersection of the two original clusters to cluster 1, as this gives a slightly better convexity score . Of course, the method cannot pick out which points belong to which cluster from the overlap, but the presence of an overlap is indicated by the closeness of the two convexity scores.

4.0 DISCUSSION

It may be considered that such a result, obtained largely via pattern recognition techniques, cannot be thought of as a statistical one. Nevertheless, the aim of existing statistical cluster algorithms is to locate distinct clusters and the identificat ion and separation of overlapping clusters is an integral part of that aim. Thus we have a method which achieves statistical aims without using what are usually considered statistical techniques. Statistical techniques are classically based on as sumptions about distributions, but there has also been an acceptance of mathematical methods (e.g. graph theory in clustering Gower and Ross, 69) and display procedures (e.g. stem and leaf Hoaglin et.al., 83) *as* valid parts of statistics even though they are not based on distribution theory. Our contention is that the acceptance of 'non-standard' methodologies within statistics should increase, and embrace pattern recognition and A.I. methods based on models of human performance. The alternative approach, to keep statistics 'pure' , markedly restricts its applicability- There is a limit to what can be modelled in Che complex real world by rigorous mathemat ical or statistical models; this should be recognized and the information processing heuristics which have evolved to enable humans to cope with complexity should form the basis of methods applicable to a wide range of realistic problems .

References

[1] H Blum. Biological shape and visual science. J Theoretical Biology, 38, pp 203-287. 1973.

[2] P A Devivjer and J Kittler. Pattern Recoenition, A Statistical approach. Prentice da 11. *1962*. Appendix A.

[3] 3 Everitt. Cluster Analysis. Heinemann. 1930.

[4] J C Gower and G J S Ross. Minimum spanning trees & single linkage cluster analysis. Apni. Stat, 18, pp 54-64. 1969.

[5] R J Herrnstein. Objects, Categories and discriminative stimuli in Animal Cognition, Ed. H L Roitblat, T G Bever, H S Terrace, Lawrence Erlbaum, 1984 pp. 233,261.

[6] D C Hoaglin, F Mosteller, J W Tukey, Eds. Understanding robust & exploratory data analysis. Wiley. 1983.

[7] B Kuipers and J P Kassirer. Casual reasoning in medicine: analysis of a protocol. Cognit ive Science 8, pp. 363 —385. 1984.

[3] D Marr. Vision. Freeman, 1982.

[9] D Marr & H K Nishihara. Representation & recognition of the spat ial organization or 3-D shapes. Proc. Soc. London B200.

[10] K I Phelps & P B Musgrove. Brunei University Technical Report TR/02/85.

[11] P Szolovits, Ed, Artificial intelligence in medicine, Colorado: Westview Press. 1952.

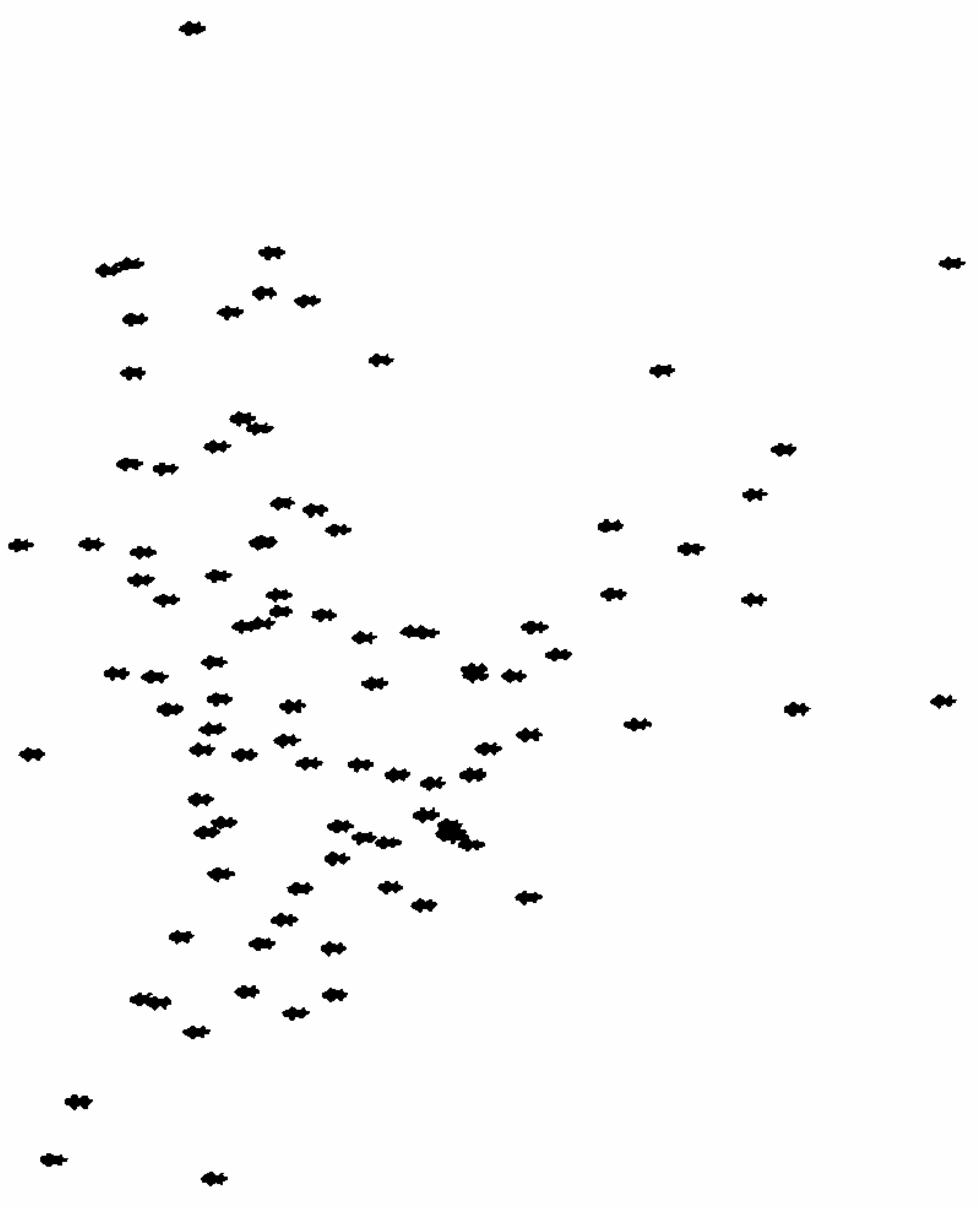
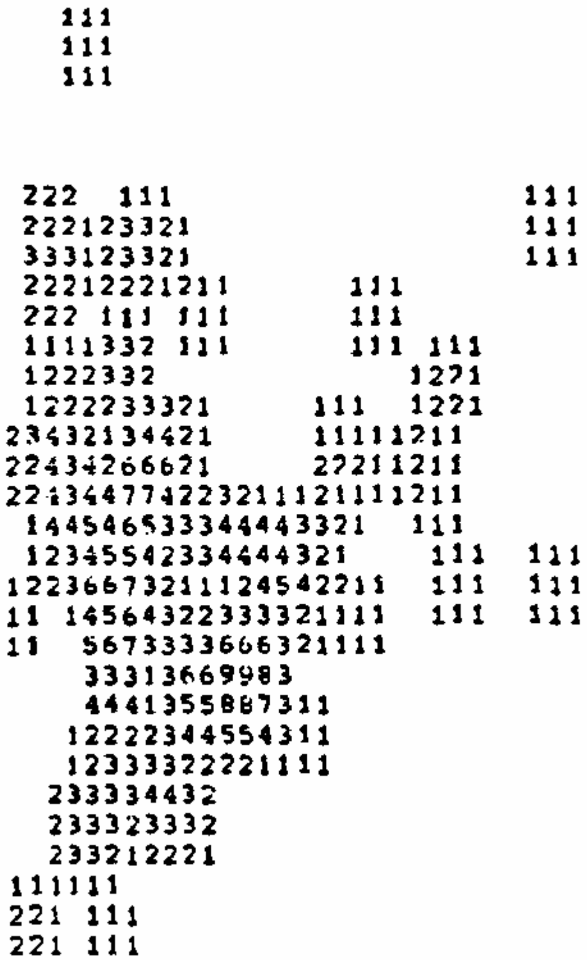


Fig. 1



```
   111
   111
   111



 222  111                  111
 222123321                 111
 333123321                 111
 22212221211      111
 222 111 111      111
 1111332 111      111 111
 1222332             1271
 1222233321     111  1221
23432134421     11111211
22434266621     27211211
221344774223211121111211
 144546533344443321  111
 12345542334444321    111  111
12236673211124542211  111  111
11 14564322333321111  111  111
11  5673333666321111
    33313669983
    4441355887311
   12222344554311
   12333322221111
  233334432
  233323332
  233212221
111111
221 111
221 111
```

Fig . 2.

```
3729
```

```
     222
22221222
222222222
2222222221
 2222222222
 2222221222
 222222222
 22222221
 22222222      2221 2 2
2222222222     2222 2 2
2222222222    12222 2 2
22222222221  222222 2 2
 222222222222222221 2 2
 222222222222222221 2 2
 22222222222222222
222222222222222222
222222222222222222
221222222222222
   22222222222
   22222222222
   222222222222
   222222222222
  1222222222222
 2222222221
 222222222
 22222222
2222221
2221222
222 22
```

Fig .3

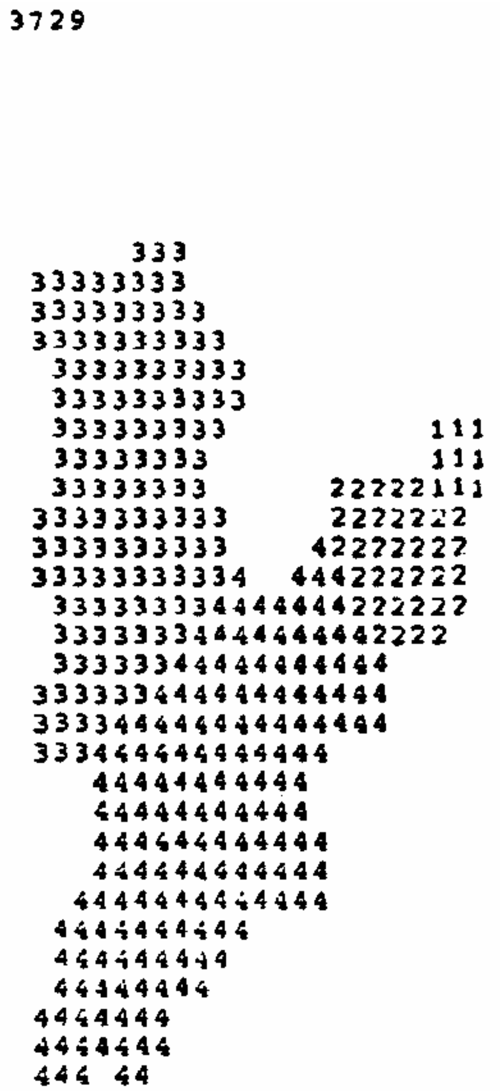
```
3729





     111
11111121
122222211
1123332211
 1234332211
 1234432211
 123443211          111
 12344321           121
 12344321      11111111
1123443211     1222221
1223443221    11233321
11234433211  112233221
 123444322111122222211
 12345433222222321111
 12344443333333321
112333444443322221
122223455543221111
111123456543211
   12345554321
   12344444321
   123443333211
   123433222221
  1123332211111
 1122332211
 122322211
 12332111
1122221
1211111
111 11
```

Fig. 4



```
3729




     333
33333333
333333333
3333333333
 3333333333
 3333333333
 333333333          111
 33333333           111
 33333333      22222111
3333333333     2222222
3333333333    42222222
33333333334  444222222
 333333334444444222222
 33333334444444442222
 333333444444444444
333333444444444444
333344444444444444
333444444444444
   44444444444
   44444444444
   444444444444
   444444444444
  4444444444444
 4444444444
 444444444
 44444444
4444444
4444444
444 44
```

Fig. 5

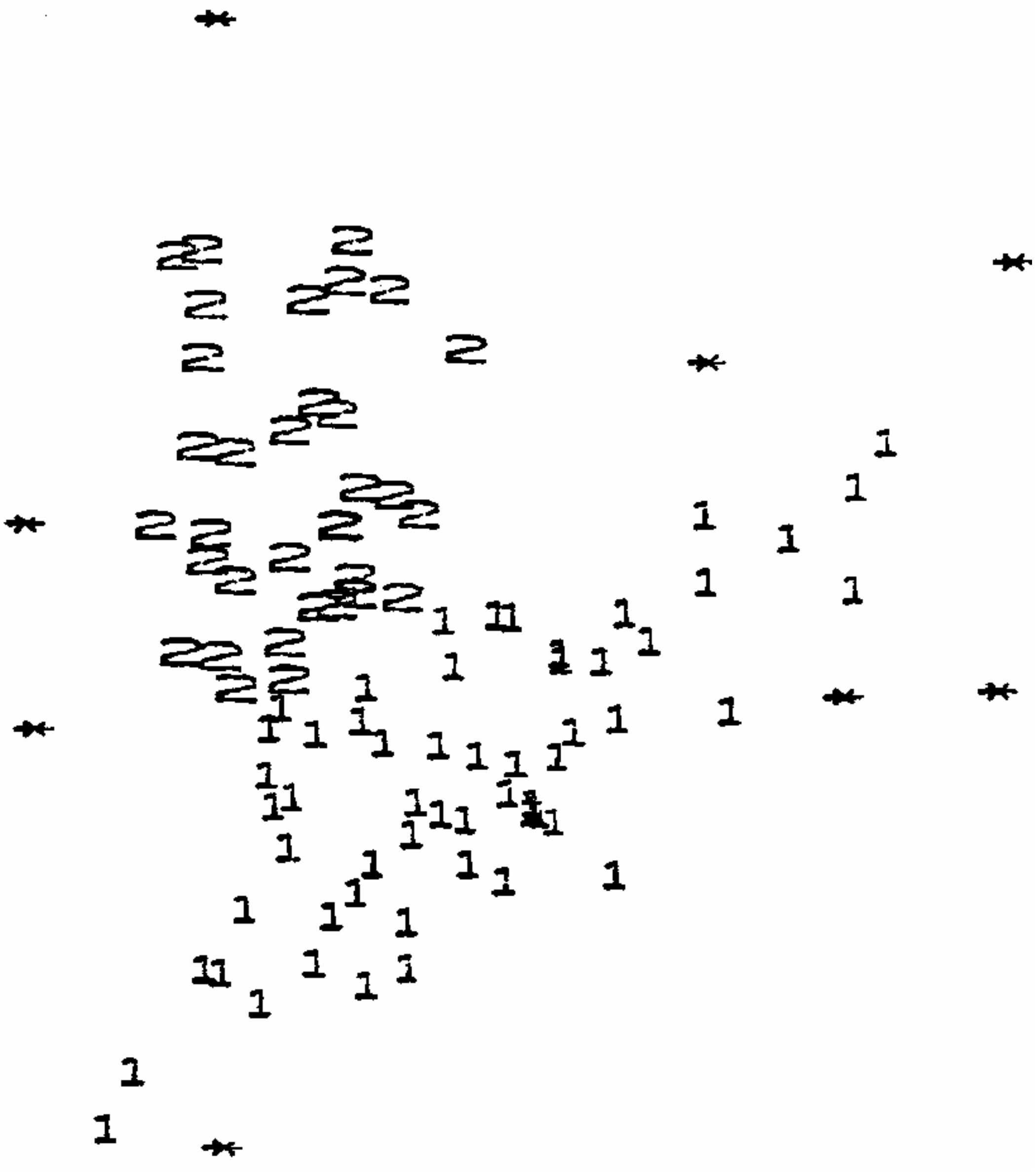


Fig. 6