1. Publicação nº	2. Versão	3. Data	5. Distribuição
INPE-4436-PRE/1235		*Nov. 1987*	☐ Interna ☒ Externa ☐ Restrita
4. Origem *DPI*	Programa *SIAG*		

6. Palavras chaves - selecionadas pelo(s) autor(es)

ANÁLISE DE DADOS
CLASSIFICAÇÃO
AGREGAMENTO DE DADOS

7. C.D.U.: *528.711.7*

8. Título *INPE-4436-PRE/1235*

"UMA VARIANTE DO ALGORITMO ISODATA PARA APLICAÇÃO EM ALVOS AGRÍCOLAS"

10. Páginas: *16*

11. Última página: *15*

12. Revisada por

9. Autoria

Leonardo Sant'Anna Bins
Flávio Roberto Dias Velasco

Assinatura responsável

13. Autorizada por

Dr. Marco Antônio Raupp
Diretor Geral

14. Resumo/Notas

Este trabalho apresenta uma variante do algoritmo ISODATA aplicado a determinação automática de subclasses de alvos agrícolas em imagens multiespectrais. Tal como no ISODATA, o algoritmo apresentado dispensa o conhecimento a priori do número de subclasses existentes na imagem para uma dada classe. Em relação ao ISODATA, o algoritmo apresenta as seguintes vantagens: necessita do estabelecimento de menos parâmetros; tem convergência em geral mais rápida e evita a possibilidade de ciclo infinito. O algoritmo proposto neste trabalho supõe inicialmente a existência de uma única classe e, nas iterações seguintes, procede à abertura das classes com variança maior que C. O algoritmo termina quando todas as classes possuem variança dentro do limite desejado. Para minimizar o erro médio, a cada iteração é aplicado o algoritmo das k-médias, onde k é o índice da iteração. No trabalho é mostrado um exemplo da aplicação do algoritmo a alvos agrícolas em imagens LANDSAT, sensor TM.

15. Observações *Trabalho submetido para apresentação no II Simpósio Latino-americano sobre sensores remotos, 16-21 de novembro de 1987, Bogotá, Colombia*

UMA VARIANTE DO ALGORITMO ISODATA PARA APLICAÇÃO EM ALVOS AGRÍCOLAS

Leonardo Sant'Anna Bins
Flávio Roberto Dias Velasco

Ministério da Ciência e Tecnologia - MCT
Instituto de Pesquisas Espaciais - INPE
Departamento de Processamento de Imagens - DPI
Caixa Postal 515 - 12201 - São José dos Campos - SP - Brasil

RESUMO

Este trabalho apresenta uma variante do algoritmo ISODATA aplicado a determinação automática de subclasses de alvos agrícolas em imagens multiespectrais. Tal como no ISODATA, o algoritmo apresentado dispensa o conhecimento a priori do número de subclasses existentes na imagem para uma dada classe. Em relação ao ISODATA, o algoritmo apresenta as seguintes vantagens: necessita do estabelecimento de menos parâmetros; tem convergência em geral mais rápida e evita a possibilidade de ciclo infinito.

O algoritmo proposto neste trabalho supõe inicialmente a existência de uma única classe e, nas iterações seguintes, procede à abertura das classes com variança maior que C. O algoritmo termina quando todas as classes possuem variança dentro do limite desejado. Para minimizar o erro médio, a cada iteração é aplicado o algoritmo das k-médias, onde k é o índice da iteração.

No trabalho é mostrado um exemplo da aplicação do algoritmo a alvos agrícolas em imagens LANDSAT, sensor TM.

ABSTRACT

This paper presents a variant of the ISODATA algorithm for automatic clustering of agricultural targets in multispectral images. In the same way as ISODATA, the algorithm dispenses with the a priori knowledge of the number of classes. Compared to ISODATA, the algorithm has the following advantages: it depends on less parameters, it converges quicker (usually) and is guaranteed to terminate.

The proposed algorithm start with a single class and, in the following iteractions, splits classes for which the variance is greater than a given constant C. The algorithm terminates when all classes have variances within the desired limit. To minimize the mean square error, at each iteraction, a k-means algorithm is performed where k is the iteraction index.

In the paper is shown example of the application of the algorithm for LANDSAT/TM images.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO

INSTITUTO DE PESQUISAS ESPACIAIS

SOLITAÇÃO DE REFERÊNCIA

NÚMERO: 021/87

SOLICITO A REFERÊNCIA [X] C.D.U [ ] PALAVRAS - CHAVE

PARA A PUBLICAÇÃO Nº

TÍTULO

"Uma variante do algoritmo ISODATA para aplicação em alvos agrícolas".

AUTORES

Leonardo Sant'Anna Bins
Flávio Roberto D. Velasco.

SOLICITANTE: Janete da Cunha

ÓRGÃO: DPI

DATA: 29/10/87

REFERENCIA SOLICITADA

C.D.U.:

CÓPIAS DBD:

PALAVRAS - CHAVE

1 Análise de dados
2 Classificação
3 Agregamento de dados
4
5
6
7
8

BIBLIOTECÁRIA:

DATA: __/__/__

OBSERVAÇÕES

DBD RECEBIDA POR:

DATA PREVISTA: __/__/__

NÚMERO:

SOLICITANTE - RECEBI:

DATA: __/__/__

1 - INTRODUÇÃO

Agrupamento de dados, também conhecido como agregamento ("clustering"), é uma técnica estatística para o estudo da estrutura de dados em geral multivariados. O propósito do agrupamento de dados é identificar grupos no conjunto de dados de entrada. Estes grupos devem possuir interiormente elementos similares entre si e diferentes elementos em relação a elementos de outros grupos.

O algoritmo ISODATA ("Iterative Self Organizing Data Analysis Technique"), desenvolvido por Ball e Hall [1] no Instituto de Pesquisas de Stanford, se presta bem ao treinamento não supervisionado, estabelecendo os parâmetros de classificação do classificador.

2 - O ALGORITMO ISODATA

A idéia qual do algoritmo ISODATA é minimizar o erro médio quadrático (da representação das amostras pelas respectivas médias) sujeitos a duas restrições: a variança das classes deve ser menor que uma constante C e a distância entre as médias de duas classes quaisquer deve ser maior que uma constante d. O algoritmo ISODATA é composto de duas operações que são iteradas até a convergência: a abertura, na qual uma classe que não satisfaz à primeira restrição é dividida em duas outras e fechamento na qual duas classes que não satisfazem à segunda restrição são agregadas. A designação de cada elemento do conjunto de padrões a um agrupamento é feita com base na mínima distância Euclidiana. Após a designação de todos os elementos as médias e as variânças são recalculadas em cada agrupamento. Três parâmetros são fornecidos pelo analista para a execução do algoritmo: 1) número mínimo de pontos que um agrupamento deve possuir, abaixo do qual o agrupamento é eliminado, 2) desvio padrão máximo, no qual os agrupamentos que possuírem um desvio padrão acima deste, em pelo menos uma dimensão, são divididos em dois outros, 3) distância mínima, na qual se dois agrupamentos estiverem distanciados de um valor inferior, os dois são então agregados.

O algoritmo termina quando não há mais aberturas e fechamentos.

O algoritmo abaixo é uma apresentação informal da versão básica do ISODATA:

ALGORITMO ISODATA BÁSICO

0. INICIALIZE

1. CLASSIFIQUE E CALCULE ESTATÍSTICAS

2. ELIMINE AGRUPAMENTO I SE NPAGR (I) > NPMIN

3. SE DVP (I) > DPMAX

   REALIZE ABERTURA NO AGRUPAMENTO I
   VOLTE AO PASSO 1

4. SE DT (I,J) < DLMIN

   REALIZE AGREGAMENTO DOS AGRUPAMENTOS I E J
   VOLTE AO PASSO 1

5. FIM

onde, NPAGR(I) - número de pontos do agrupagmento I
NPMIN - número mínimo de pontos
DVP (I) - maior desvio padrão do agrupamento I
DPMAR - desvio padrão máximo
DT (I,J) - distância entre dois agrupamentos
DLMIN - distância mínima permitida entre dois agrupamentos

3 - VARIANTE ISODATA

A variante do algoritmo ISODATA proposta neste trabalho é uma decorrência do relaxamento da segunda restrição. Dado que, geralmente, em alvos agrícolas, duas subclasses da mesma classe podem estar próximas hão há necessidade de fechamento, devendo-se garantir somente que, para um dado número de subclasses, tenha-se o menor erro quadrático. O algoritmo supõe inicialmente a existência de uma única classe e, nas iterações seguintes, procede à abertura das classes com variança maior que C. O algoritmo termina quando todas classes possuem variança dentro do limite desejado.

VARIANTE DO ALGORITMO ISODATA

1. INICIALIZAÇÃO

2. REALIZE K-MÉDIAS

3. CALCULE MATRIZ DE COVARIANÇA

4. REALIZE ABERTURA

5. SE HOUVE ABERTURA RETORNE AO PASSO 2

6. FIM

A fase de inicialização corresponde à escolha dos limites para o número mínimo de elementos permitido nos agrupamentos e o desvio padrão máximo permitido.

A realização do k-médias consiste de um processo iterativo em que os padrões são designados ao centro do agrupamento mais próximo, seguido da atualização destes centros. Se houve mudança nos centros, o processo recomeça considerando os centros atualizados. Desta forma, após a realização do K-médias, chega-se ao mínimo erro quadrático para o número de agrupamentos existentes. Os padrões são classificados segundo a menor distância Euclidiana dos mesmos ao centro de cada classe (agrupamento) dada por:

$$d(X,M) = \sqrt{\sum_{i=1}^{m} (x_i - \mu_i)^2} \ ,$$

onde X = vetor padrão
M = vetor média da classe
m = dimensionalidade dos padrões

O padrão X pertence à classe j se:

$$d(X,M_j) \leq d(X,M_\ell), \text{ para todo } \ell = 1, \ldots c, \ j \in [1,c]$$

onde c = número de classes
j,ℓ = índice do agrupamento

ALGORITMO K-MÉDIAS

1. INICIALIZE

2. CLASSIFIQUE

3. RECALCULE OS CENTROS

4. SE HOUVER MUDANÇA NOS CENTROS VOLTE AO PASSO 2

5. FIM

A inicialização do k-médias corresponde a seleção dos centros iniciais. Entretanto, como o k-médias é utilizado como uma rotina na variante do ISODATA, os centros iniciais são os centros utilizados na iteração da abertura das

classes. Como a velocidade de convergência decresce rapidamente nas primeiras iterações, determinou-se que a realização do k-médias consistirá de 3 iterações.

A fase de abertura das classes procede a abertura de somente uma classe. A seleção da classe a ser dividida em duas outras obedece o seguinte critério:

i) Seleciona-se os agrupamentos que possuam pelo menos uma dimensão com variança superior ao valor permitido;

ii) Seleciona-se entre os agrupamentos escolhidos em (i) aquele que possuir a maior média das variânças nas dimensões.

A abertura é realizada sobre a dimensão que possui a maior variança. O cálculo dos subagrupamentos é efetuado a partir dos agrupamentos que os gerou segundo:

$$(\mu_{1,1}, \ldots, \mu_{1,i}, \ldots, \mu_{1,m}) = (\mu_1, \ldots, \mu_i + \sqrt{\sigma_i^2}, \ldots, \mu_m)$$

$$(\mu_{2,1}, \ldots, \mu_{2,i}, \ldots, \mu_{2,m}) = (\mu_1, \ldots, \mu_i - \sqrt{\sigma_i^2}, \ldots, \mu_m)$$

onde $(\mu_1, \ldots, \mu_m)$ = centro do agrupamento pai

$(\mu_{1,1}, \ldots, \mu_{1,m})$ e $(\mu_{2,1}, \ldots, \mu_{2,m})$ = centro dos agrupamentos filhos

m = dimensionalidade

i = dimensão onde se realizará a abertura

σ_i^2 = corresponde ao elemento da diagonal da matriz de covariança (variança).

4 - RESULTADOS

As figuras 1 e 2 ilustram os resultados obtidos com o algoritmo proposto. Os dados foram obtidos em um segmento agrícola através dos sensores TM 3 e 5 do satélite LANDSAT. Foram escolhidas só duas bandas para ser possível mostrar os resultados bidimensionalmente. As figuras 1.a e 2.a apresentam as distribuições dos pontos juntamente com suas frequências aproximadas de ocorrência. As figuras 1.b, 1.c, 1.d, 2.b, 2.c e 2.d mostram as sucessivas iterações do algoritmo. O valor do desvio padrão utilizado nos dois experimentos foi 6.0 e o número final de classes foi 3 em ambos os casos.

5 - CONCLUSÃO

Esta variante proposta do algoritmo ISODATA está sendo utilizada para a determinação de subclasses de alvos agrícolas em imagens multiespectrais no projeto SIAG de estatísticas agrícolas.

As vantagens que o algoritmo variante possui em relação ao ISODATA são:

i) Requer número menor de parâmetros a serem estabelecidos, o que facilita a operação;

ii) Possui em geral convergência mais rápida;

iii) Evita a possibilidade de ciclo infinito, originado pela alternancia dos processos de abertura e fechamento.

6 - REFERÊNCIAS

1 - BALL, G.H. and HALL, D.J. "ISODATA, An Iterative Method of Multivariate Data Analysis and Pattern Classification". In IEEE International Communications Conference, Philadelphia, June 1966.

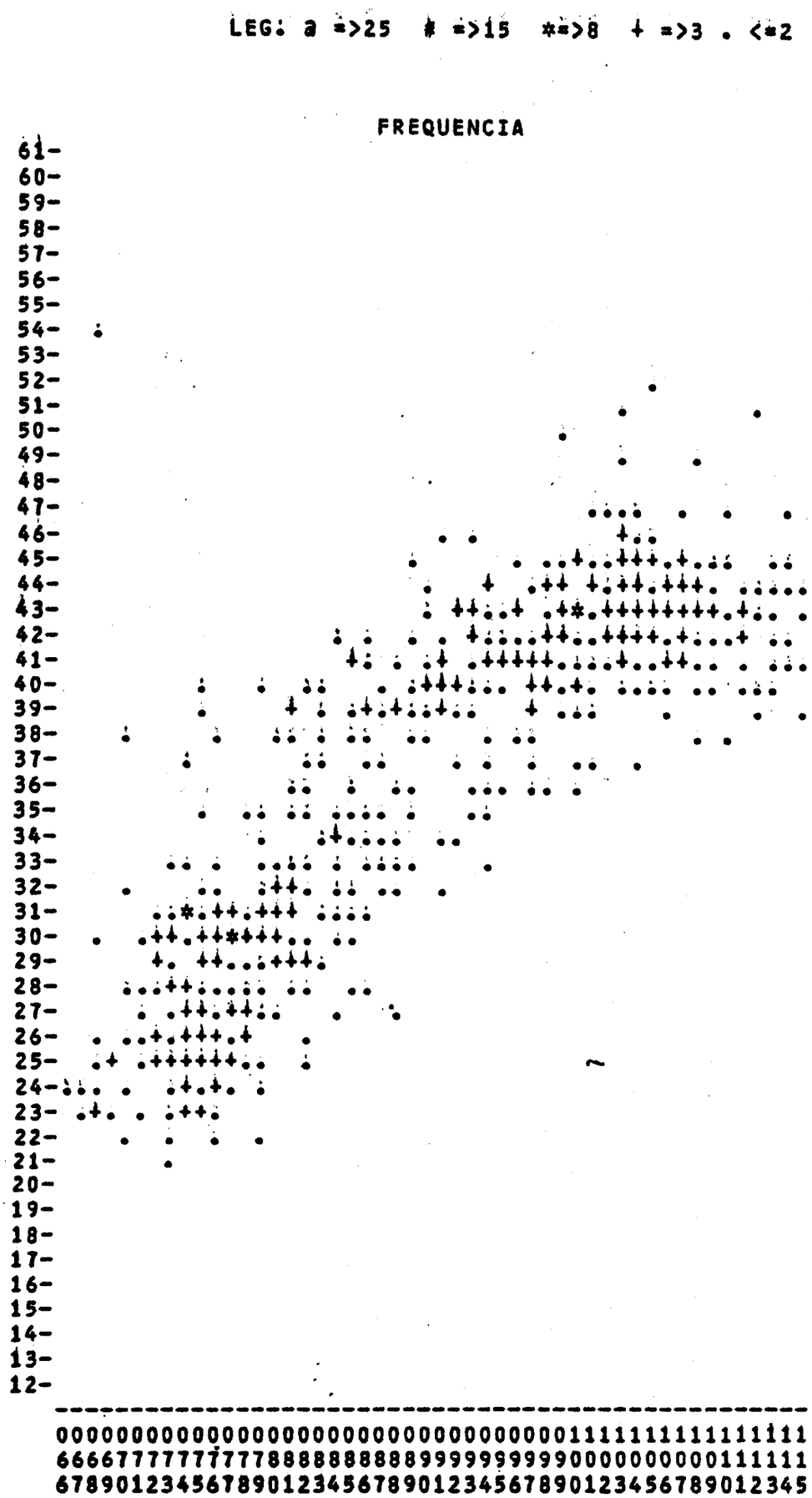

LEG: @ =>25 # =>15 *=>8 + =>3 . <=2
FREQUENCIA
61-
60-
59-
58-
57-
56-
55-
54-
53-
52-
51-
50-
49-
48-
47-
46-
45-
44-
43-
42-
41-
40-
39-
38-
37-
36-
35-
34-
33-
32-
31-
30-
29-
28-
27-
26-
25-
24-
23-
22-
21-
20-
19-
18-
17-
16-
15-
14-
13-
12-
000000000000000000000000000000000011111111111111111
666677777777778888888888999999999900000000001111111
678901234567890123456789012345678901234567890123456

Figura 1.a

NRO DE CLASSES : 1

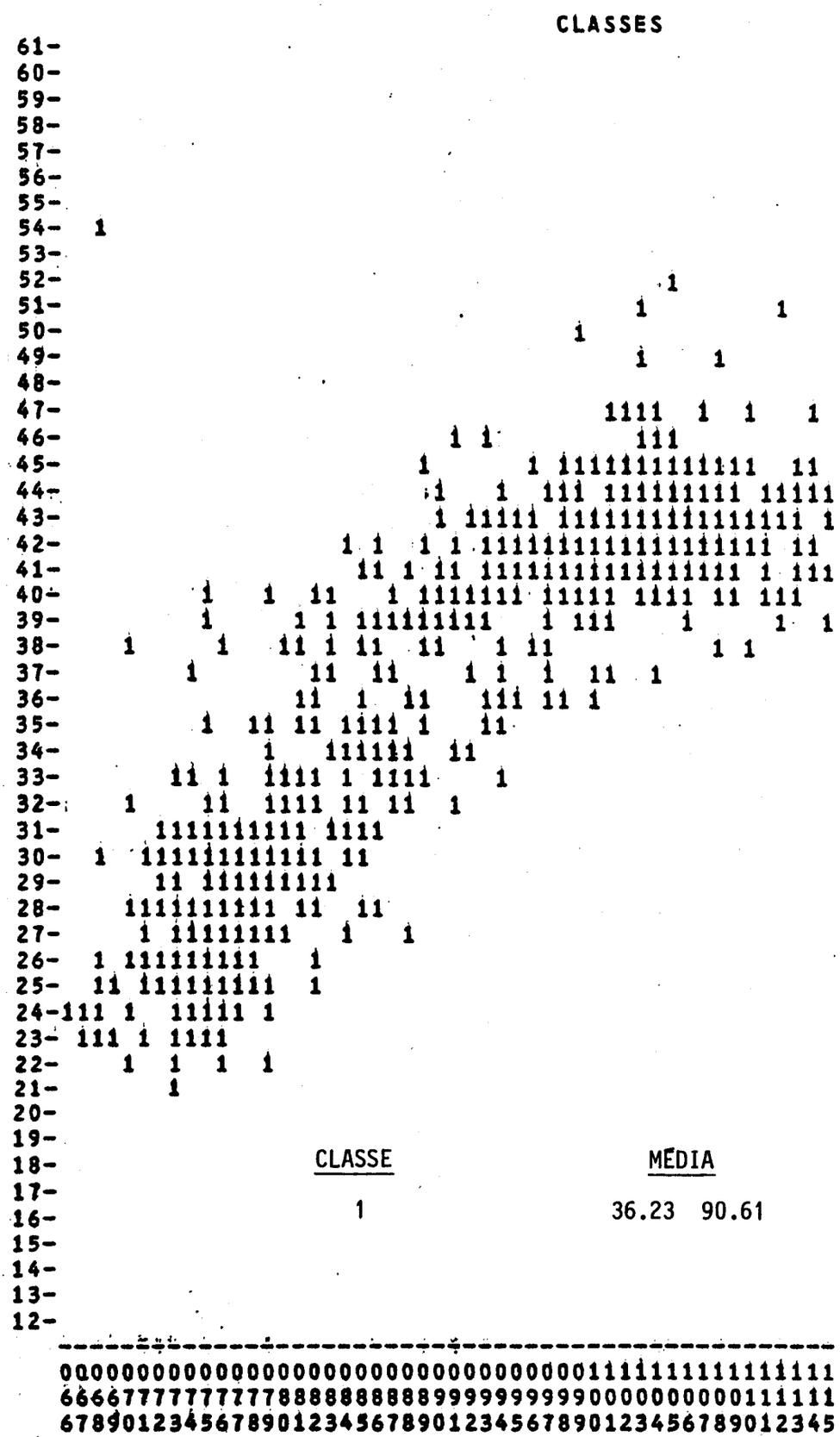
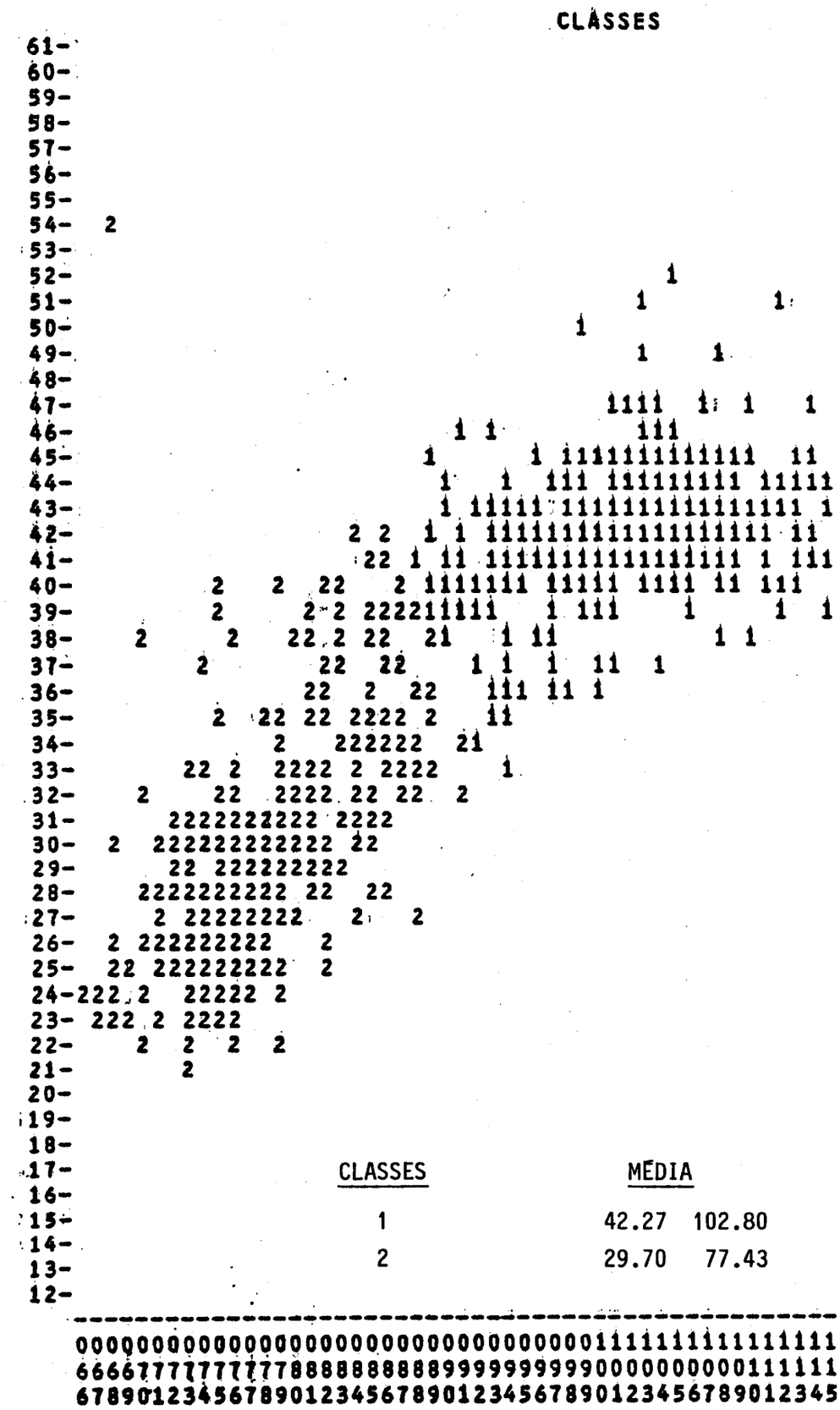


Figura 1.b

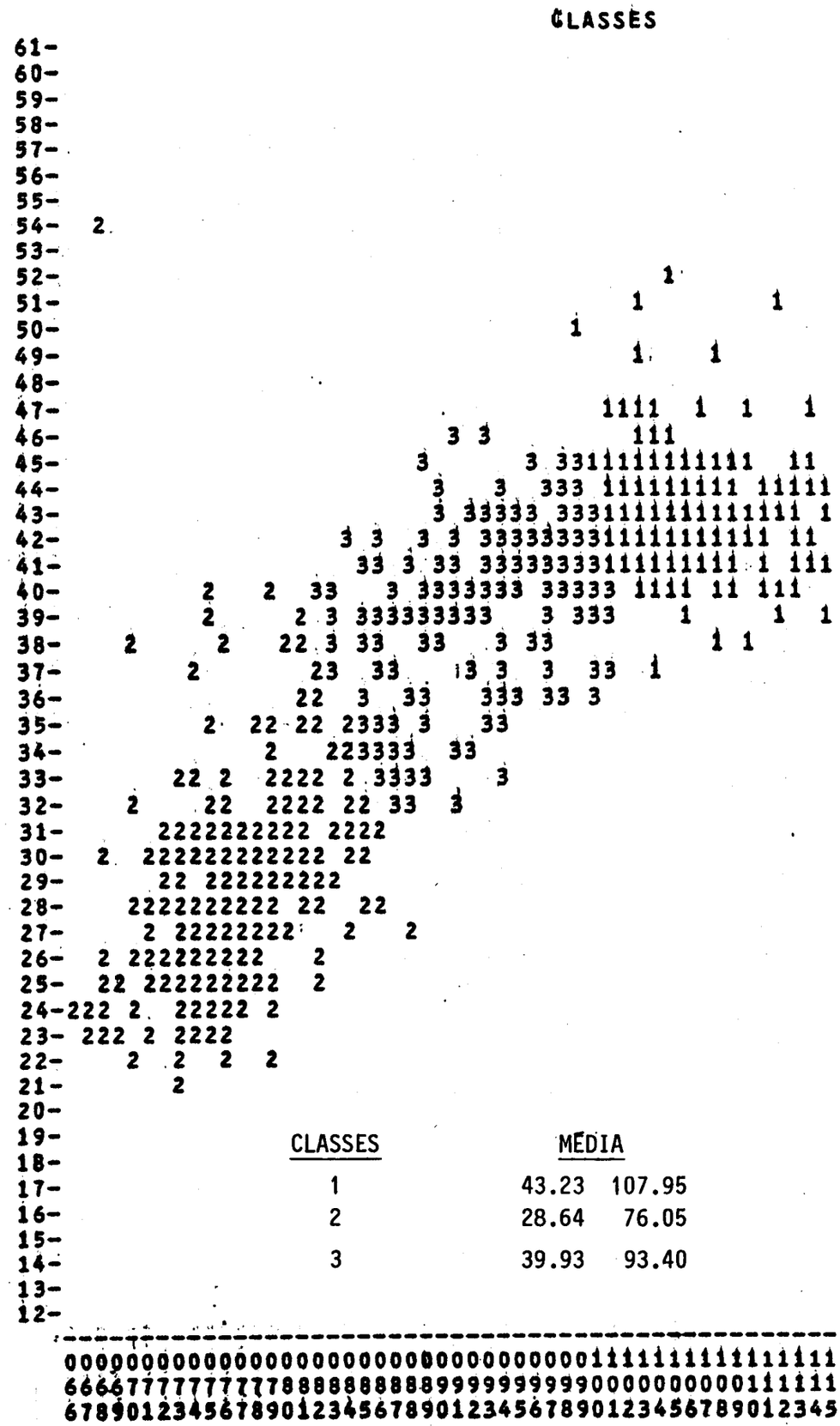
NRO.DE CLASSES : 2



CLASSES	MÉDIA	
1	42.27	102.80
2	29.70	77.43

Figura 1.c

NRO DE CLASSES : 3



CLASSES	MÉDIA	
1	43.23	107.95
2	28.64	76.05
3	39.93	93.40

Figura 1.d

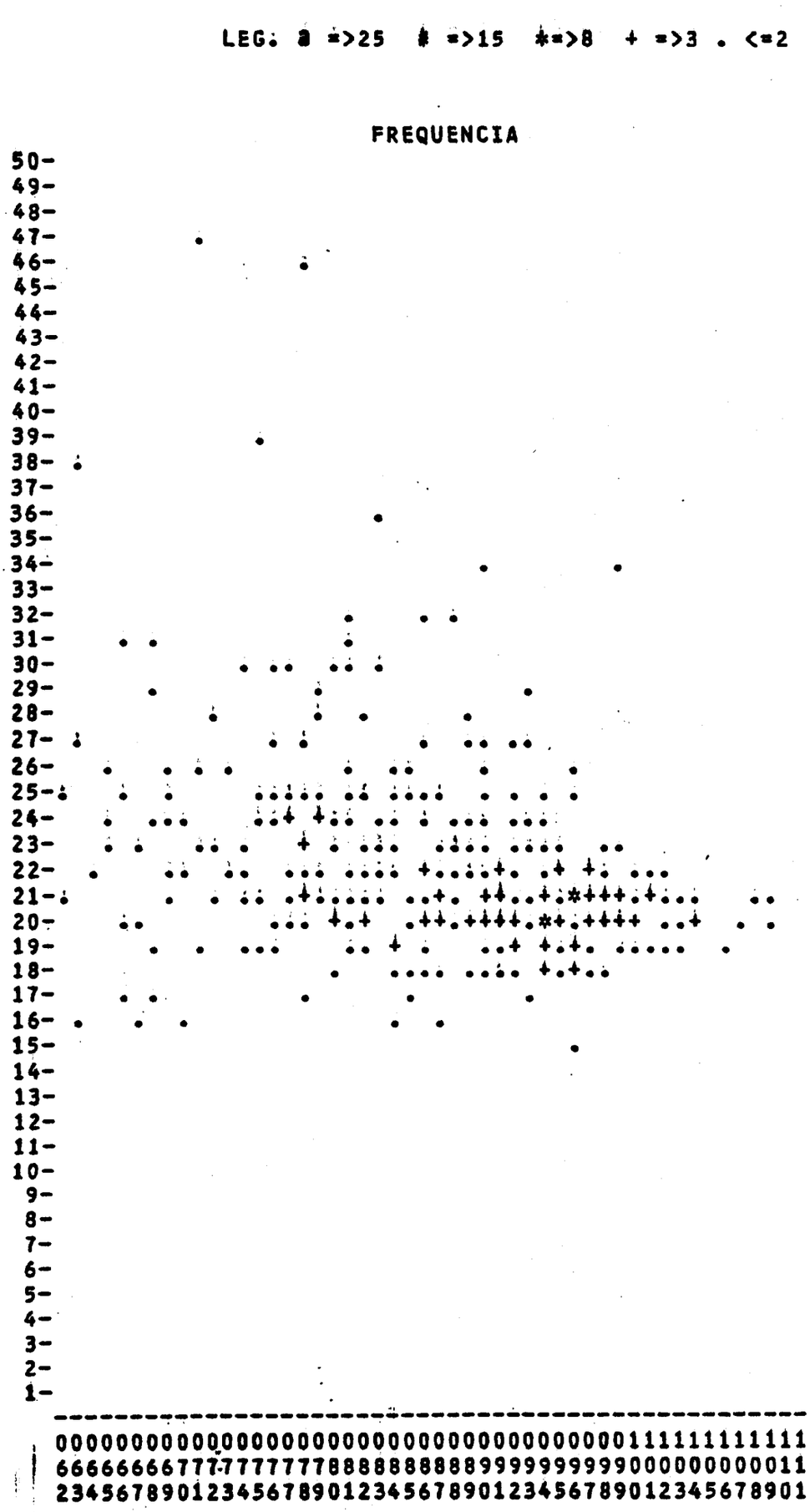

LEG: @ =>25 # =>15 *=>8 + =>3 . <=2
FREQUENCIA
50-
49-
48-
47-
46-
45-
44-
43-
42-
41-
40-
39-
38-
37-
36-
35-
34-
33-
32-
31-
30-
29-
28-
27-
26-
25-
24-
23-
22-
21-
20-
19-
18-
17-
16-
15-
14-
13-
12-
11-
10-
9-
8-
7-
6-
5-
4-
3-
2-
1-
000000000000000000000000000000000000000111111111111
666666667777777778888888888999999999900000000011
234567890123456789012345678901234567890123456789012345678901

Figura 2.a

NRO DE CLASSES : 1

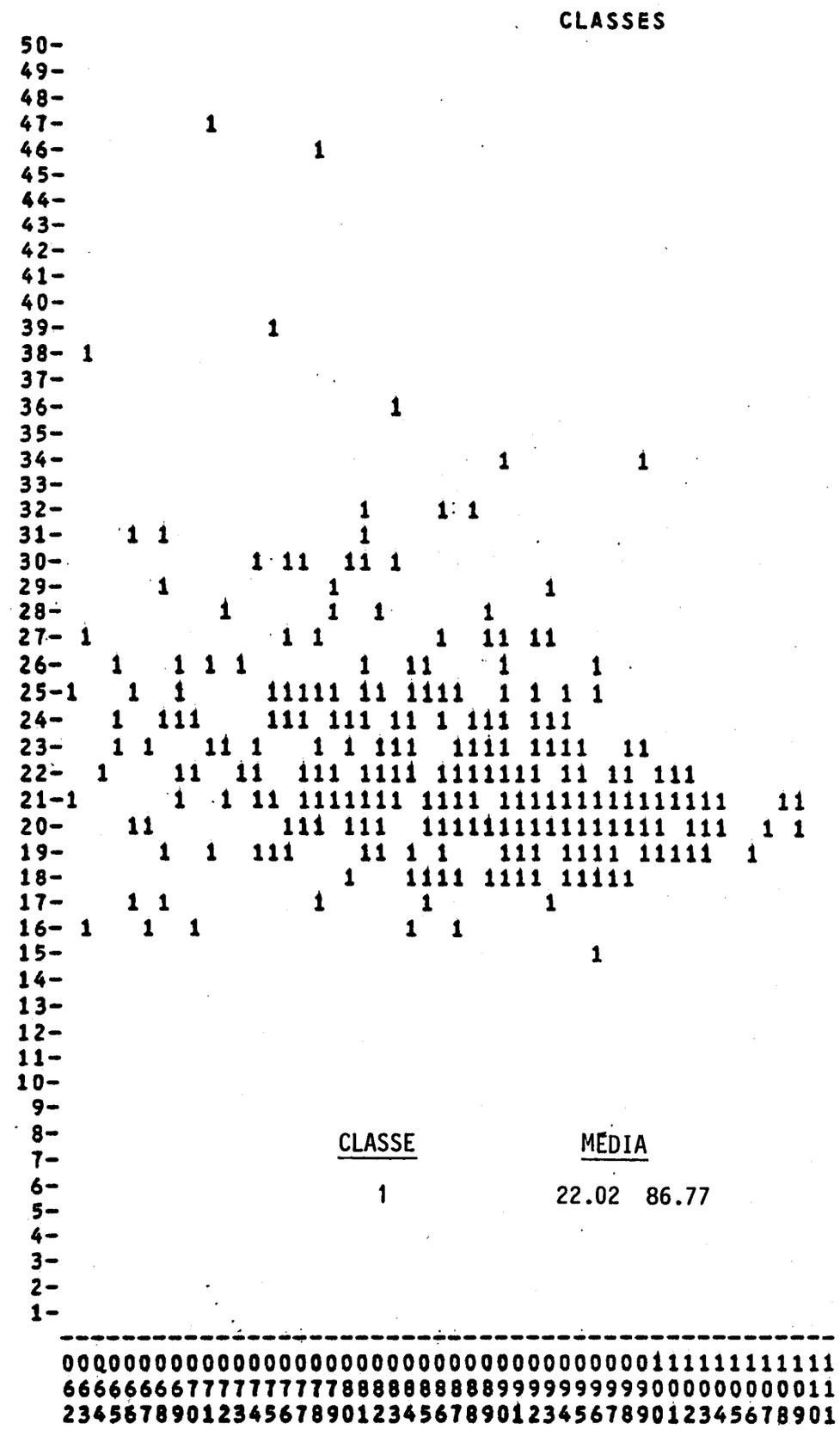


Figura 2.b

NRO DE CLASSES : 2

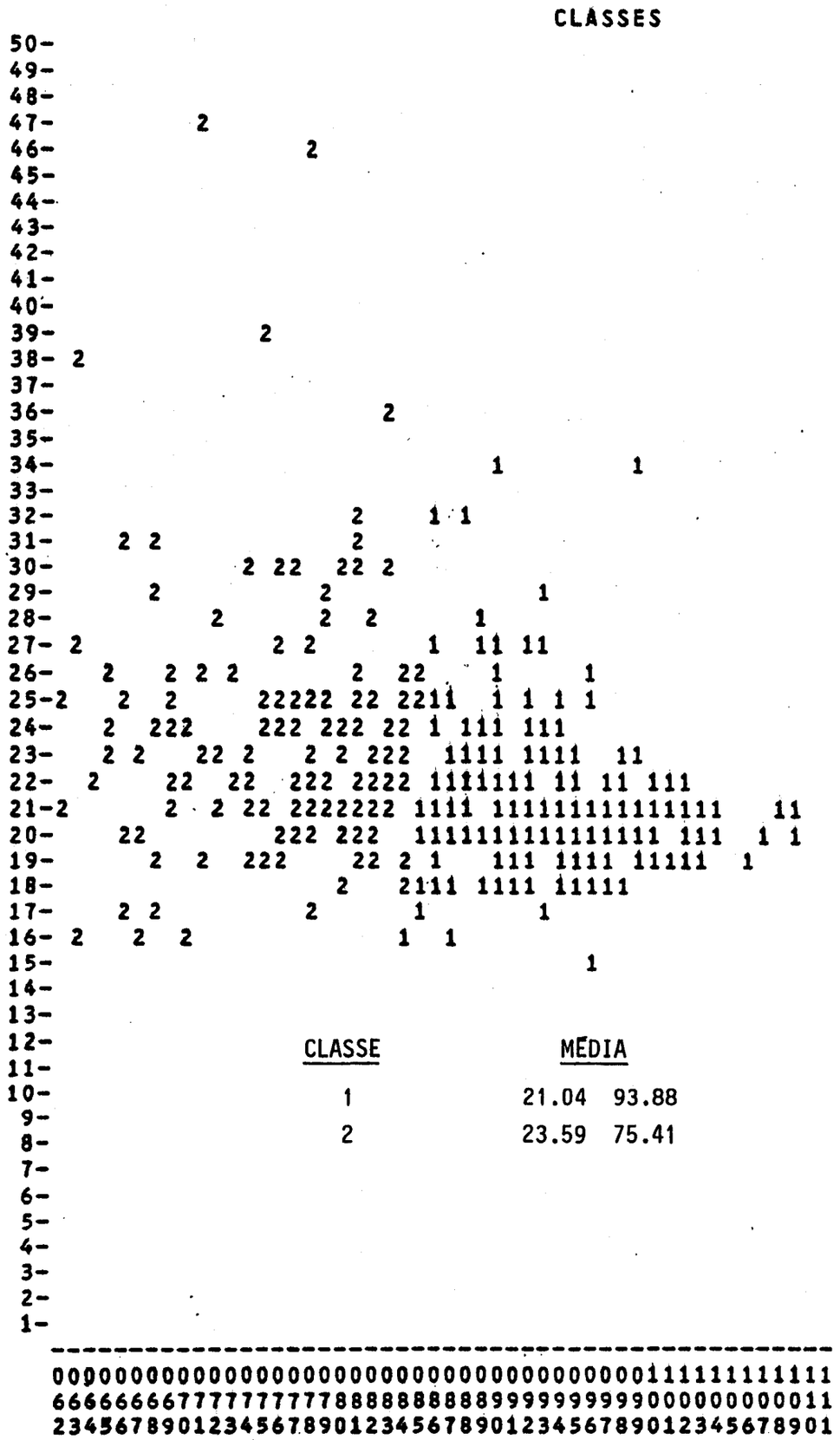


Figura 2.c

NRO DE CLASSES : 3

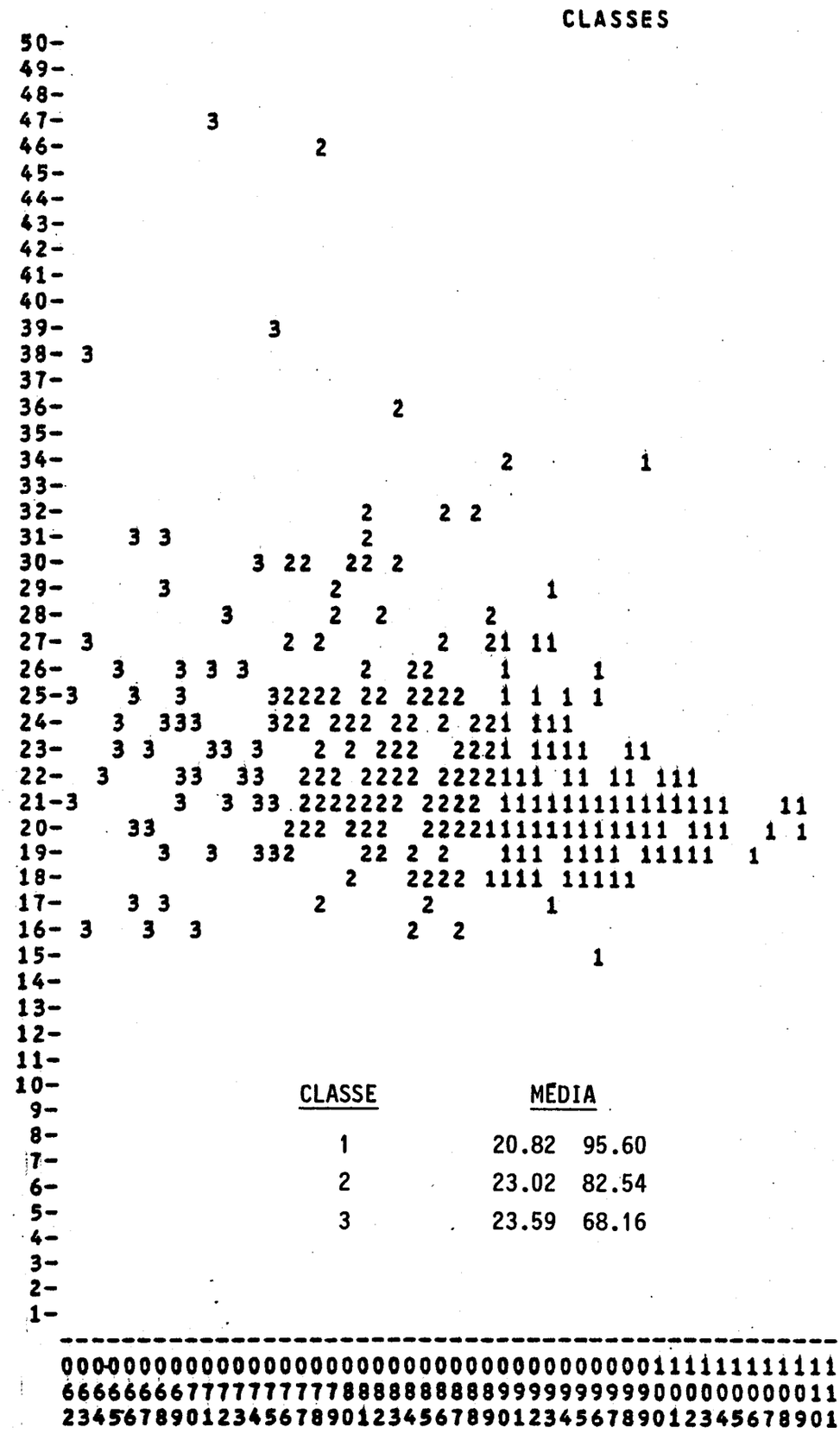


Figura 2.d