SÉLECTION DU MEILLEUR JEU DE CARACTÈRES PARMI UN EMSEMBLE

C.Y. Suen* et J. Quinqueton**

*Universite Concordia
Montreal (Quebec)

**I.N.R.I.A.
Rocquencourt, France

ABSTRACT

The problem can be stated as in operation research.

Let E be a set of points in a metric space X.

A partition of E into p clusters $\{C_i; i = 1 \text{ to } p\}$ is given.

Let d be the distance on X. We look for subset F of E such that:

- to F belongs one and only one point in each C_i (1)
- min $\{d(x_i,x_j)$; x_i and $x_j \in F\}$ is maximum (2)

This problem cannot be solved by a "Branch and Bound" algorithm.

We present a sub-optimal solution based upon the n nearest neighbors graph (in the sense of the distance d of the metric space X).

This technique is applied to the search of the most legible alphabet of characters (26 letters and 10 numbers) among a set of 121 dot matrix characters.

RÉSUMÉ

Le problème peut être posé comme une question de Recherche Opérationnelle.

Considérons un ensemble E de points d'un espace métrique X. On connait une partition de cet ensemble en p classes $\{C_i, i = 1 \text{ à } p\}$.

Soit d la distance sur X. On recherche un sous ensemble F de E vérifiant les deux conditions suivantes:

- F contient un et un seul élément de chaque classe (1)
- inf $\{d(x_i,x_j)$; x_i et $x_j \in F\}$ est le plus grand possible (2)

Ce problème ne peut être résolu par un algorithme de "Branch and Bound".

Nous proposons une solution sous optimale basée sur l'utilisation du graphe des n plus proches voisins (au sens de la distance d).

Cette technique est appliquée à la recherche du jeu de caractères (26 lettres et 10 chiffres) le plus lisible parmi un ensemble E de 121 caractères possibles d'imprimante à aiguilles.

1. INTRODUCTION

In recent years, matrix characters have been widely used for display in computer output system [1]. The most common size for such characters is 5 x 7 matrix.

These characters look like they were written which dotted lines.This fact makes them a little more difficult to read than if they were written with continuous lines. Then, as they become widely used, the problem of their legibility becomes very important, and must be stu died in a specific way. The conclusion obtained for continuous lines characters may not be available in our case [2].

This problem has been studied by several researchers. Most of them conducted experiments on human subjects [3, 4, 5, 6]. Recently, the problem was studied in a quantitative way, by developing evaluation methods to measure the distinctiveness of such dot matrix characters [7, 8].

We propose here a statement of the problem in terms of Operation Research. We show that it cannot be solved by "Branch and Bound" algorithms. Then we propose a sub-optimal algorithm based upon nearest neighbor graph.

2. STATEMENT OF THE PROBLEM

Let E be a set of points in metric space X. A partition of E into p clusters $\{C_i \ ; \ i = 1 \text{ to } p\}$ is given.

Let us consider the set $\mathcal{F}$ of all the subsets F of E which have p elements, one and only one in each C_i.

We can define an order between the elements of $\mathcal{F}$ in the following way.

For each $F \in \mathcal{F}$ we call $\Delta(F)$ the set of ordered distances between the elements of F (the distance measurement is here the distance d of the metric space X). All these sets of distances contain $\frac{p(p-1)}{2}$ elements $\{d_1,\ldots,d_i,\ldots\}$ with $i < j \implies d_i \leq d_j$.

Then, we define $F \prec F'$ in the following way : let d_i be the elements of $\Delta(F)$ and d'_i of $\Delta(F')$

$F \prec F' \iff \exists \ell \ ; \forall i < \ell \quad d_i = d'_i$ and $d_\ell < d'_\ell$ (1).

Then, our problem is the search of a set. $F_o \in \mathcal{F}$ such that, for each $F \in \mathcal{F}$, $F \prec F_o$.

3. THE EXACT SOLUTION

Let n_i be the number of elements of cluster C_i. Then, the number of sets F in $\mathcal{F}$ is :

$$N = \prod_{i=1}^{p} n_i$$

Such a number is generally very large, and makes impossible any combinatorial solution of the problem.

There is another approach of the solution, which consist in ordering all the distances between elements of E (except between elements of the same C_i).

Then, if we consider the pairs of points in increasing order of their distances, it is clear that the shortest must disappear.

To "eliminate" a short distance, we have to eliminate one of the two points.

But it is impossible to know what one exactly because we need to forecast the next eliminations. Then we obtain another combinatorial solution, which is prohibitive.

As it is impossible to find an upper bound of the possible subsets $F \in \mathcal{F}$ after the elimination of a given node, a "branch and bound" algorithm is impossible to be applied here.

Then, we propose to define heuristic rules for the choice of the node to be eliminated at each step. We obtain by such a way a sub optimal algorithm , but much quicker than the exact one.

4. A HEURISTIC ALGORITHM

Let us consider the set of the ordered inter-distances of the points of E (except between elements of the same C_i).

To each distance corresponds a pair of points. The short distances must disappear. Then, for each distance we have a choice between two nodes to be eliminated.

We define 3 rules for this choice :

Rule 1 : if one of the two nodes is the unique sample in its class, then the other node is eliminated.

Rule 2 : if we are not in the previous case, let d_1 and d_2 be the distances of the two nodes to their second nearest neighbour.

If $d_1 \neq d_2$, the node corresponding to the shortest distance is eliminated. If $d_1 = d_2$, apply rule 3.

Rule 3 : let n_1 and n_2 be the number of samples in the classes of the two nodes. Eliminate the node corresponding to the greatest number.

Before applying these rules, we have to ensure that both nodes are not unique samples of their class. In that case, the elimination is impossible, and we take the following edge in increasing order of length.

5. APPLICATIONS

We first test the method on a small set E of 10 elements, divided in 3 classes, shown on Figure 1. The distance is the Hamming distance.

From this set, the sequence of eliminations were

Ø3 (rule 3) ; D3 (rule 2) ; Ø1 (rule 2)
Ø2 (rule 2) ; D2 (rule 2) ; D1 (rule 3)
01 (rule 1) ;

The selected set F is {02, D4, Ø4} and is shown on Figure 2.

We also applied the algorithm to a set E of 121 elements [9] divided in 36 classes (26 letters and 10 numerals), and to a set of 60 elements, divided in 6 classes (6 colors of a cartographic image).

6. CONCLUSION

We think that the main interest of this method is its generality : it can work on any set of samples in a metric space.

The principle of the selection is rather logical, regarding the statement of the problem. We think that the elimination rules could be enhanced, so that the selection becomes more and more "intelligent" (in the sense of Artificial Intelligence).

We now investigate in that way.

7. REFERENCES

[1] R.A.Mc. LAUGHLIN
"Alphanumeric display terminal survey"
Datamation, 71-92, (1973).

[2] D.A. SHURTLEFF
"Studies of display legibility : XXV. The relative legibility of stroke generated symbols and a comparison of the legibility of stroke symbols and dot symbols"
Electronic Systems Division, Air Force Systems Command, Bedford, Mass., USAF-ESD-TR-70-203. (1970).

[3] R.W. HAMMING
"Error detecting and error correcting codes"
BSTJ, 29, 147-150, (1950).

[4] H.F. HUDDLESTON
"An evaluation of alphanumerics for a 5 x 7 matrix display"
Proceedings Conference on Displays, IEE pub n° 80, 145-147, (1971).

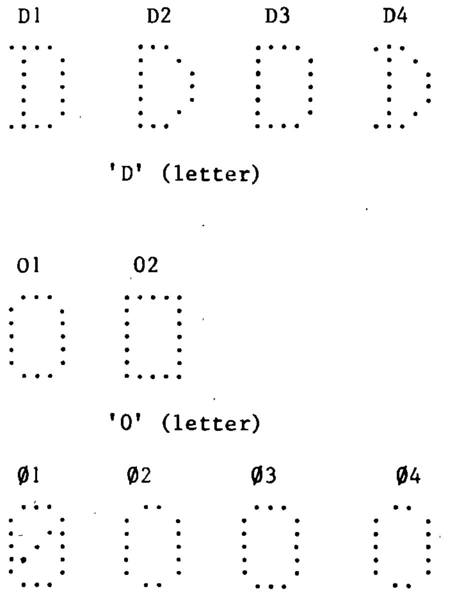
[5] M.E. MADDOX - J.T. BURNETTE - J.C. GUTMANN
"Font comparisons for 5 x 7 dot matrix characters"
Human Factors, 19 (1), 89-93, (1977).

[6] D.A. SHURTLEFF
"Studies of display symbol legibility : XXII. The relative legibility of four symbol sets made by a 5 x 7 dot matrix"
MITRE Technical Report (1969).

[7] C.Y. SUEN - C. SHIAU
"Optimum matrix character set for computer output system"
SID Intern. Symp. Digest of Tech. papers, 52-53, (1977).

[8] C.Y. SUEN - J. QUINQUETON
"Utilisation de graphes values pour la sélection automatique d'un jeu de caractères"
AFCET-IRIA, Congress on Pattern Recognition and Artificial Intelligence, Toulouse, (Sept. 1979).

[9] C.Y. SUEN - C. SHIAU
"An iterative technique of selecting an optimal 5 x 7 matrix character set for display in computer output systems"
To appear in Proc. Soc. Inf. Display, (1980).



'D' (letter)

'O' (letter)

'Ø' (numeral)

FIGURE 1 : Example of test set E

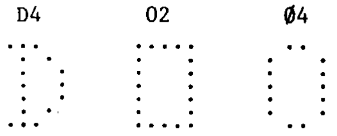


FIGURE 2 : Selected subset F of Figure 1