STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **XLVII**, Number 2, 2002

MATHEMATICAL MODELS FOR ORGANIZING DATA COLLECTIONS

ILEANA TĂNASE

ABSTRACT. Mathematical organisation of data collections is based on three models: vector processing, logical and probabilistic. Vector processing model, materialised in the SMART system implementation has the best mathematical basis. In this model entities and queries have a vectorial representation and some similarities can be established between them based on the comparison of attached vectors. The similar entities will have answers for the same requests and will be searched together. On this observation the cluster hypothesis of van Rijbergen and Sparck is based. This hypothesis suggests detecting entities class as a way for increasing the efficiency of the search.

Key words: similarity measure, dissimilarity measure, clustering, criterion function.

1. Introduction

Mathematical organisation of data collections is based on three models: vector processing, logical and probabilistic. Vector processing model [4,5], materialised in the SMART system implementation has the best mathematical basis. In this model entities and queries have a vectorial representation and some similarities can be established between them based on the comparison of attached vectors. The similar entities will have answers for the same requests and will be searched together. On this observation the cluster hypothesis of van Rijbergen [7] and Sparck [6] is based. This hypothesis suggests detecting entities class as a way for increasing the efficiency of the search.

Consider a data collection $X = \{x^1, x^2, \ldots, x^d\}$. Each entity x^j is identified by one or more index terms. Each entity x^j is represented by a d-dimensional vector:

$$X^j = (x_1^j, x_2^j, \ldots, x_d^j),$$

where the values of x_i^j are restricted to 0 and 1 (x_i^j equals 0 if the i-th index terms is not assigned to the entity, and it equals 1 only if it is assigned to the entity).

2000 *Mathematics Subject Classification.* 62H30.

1998 *CR Categories and Descriptors.* I.5.3. [**Computing Methodologies**]: Pattern Recognition – *Clustering*.

For a better performance in retrieval of entities it is useful that entities be clusterised according to appropriate criteria. A space of entities could be represented as in Figure 1.

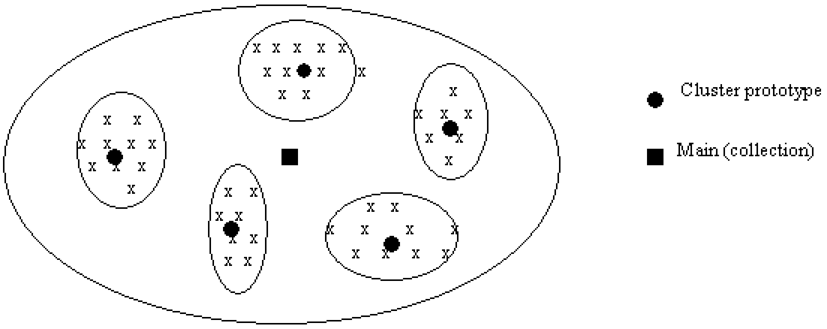

Figure 1. Collection and prototype representation

One must take into account that entities of the same class centre have similar characteristics. Thus, the best retrieval performance must be obtained for data collections consisting of individual compact classes, but with great distance between class prototypes.

Splitting a space of entities into classes can be done using several determinist classifying methods. These methods are mainly optimisation procedures of some criterion functions. In order to build a criterion function we may consider that each class is represented by a geometrical prototype. In the vector processing model [4, 5], in which classes have an approximately spherical shape, the prototypes will be points in the Euclidean $\mathbb{R}^d$ space.

2. Similarity measures

Let X be the space of entities to be classified. A similarity measure over X is a function $S : X \times X \to \mathbb{R}$, which satisfies the following axioms:

a) $S(x, y) \geqslant 0, \forall x, y \in X$,

b) $S(x, y) = S(y, x), \forall x, y \in X$,

c) $S(x, x) = S(y, y) > S(x, y), \forall x, y \in X, x \neq y$.

The most used similarity measure in vector processing model is considered the angle cossinus between two vectors:

$$S_1 = \frac{< x, y >}{\|x\| \cdot \|y\|} = \frac{x^T y}{\|x\| \cdot \|y\|}$$

But as shown before, the vectors x, y have binary components. When all the characteristics are binary, there is a set of well known similarity measures. These measures are based on the following values:

$$s = \sum_{i=1}^{d} x_i \cdot y_i,$$

which represents the number of index terms that simultaneously exist in x and y, in the same way:

$$t = \sum_{i=1}^{d} (1 - x_i)(1 - y_i),$$

represents the number of index terms which simultaneously miss from the x and y entities,

$$u = \sum_{i=1}^{d} x_i(1 - y_i),$$

represents the number of index terms that exist in x, but they miss from y,

$$v = \sum_{i=1}^{d} y_i(1 - x_i),$$

represents the number of index terms that exist in y, but they miss from x.

It is easy to show that:

$$s + t = x^T x$$

and

$$s + v = y^T y$$

Taking account of the prior features, the meaning of the following similarity measures is easy to understand [2]:

$$S_2 = \frac{s}{s + \frac{1}{2}(u + v)},$$

$$S_3 = \frac{s}{s + 2(u + v)},$$

$$S_5 = \frac{st - uv}{st + uv}.$$

3. The criterion function

Let $X = \{x^1, x^2, \ldots, x^p\}$ be the entities set that must be classified. Our aim is to find the cluster structure of the given set. The cluster structure of the set X can be done by a partition $P = \{A_1, A_2, \ldots, A_n\}$ of X. Each member of the partition P will correspond to an entity class. Using a similarity measure we can build a criterion function. The classification problem is reduced to an optimization problem.

Each A_i class can be represented by a prototype L_i, and denote by $L = \{L_1, L_2, \ldots, L_n\}$. Consider the representation of the P partition. In the vector processing model the classes have almost spherical shape and a class prototype will be a point in $\mathbb{R}^d$. This point is the same with the centre of the class, as shown in Figure 1.

A dissimilarity measure on X is a function $D : X \times X \to \mathbb{R}$, that satisfies the following axioms:

a) $D(x, y) \geqslant 0, \forall x, y \in X$,
b) $D(x, x) = 0, \forall x \in X$,
c) $D(x, y) = D(y, x), \forall x, y \in X$.

The criterion function (J) may be defined as [2]:

$$J(P, L) = \sum_{i=1}^{n} \sum_{x \in A_i} D(x, L_i), \tag{1}$$

where D is a dissimilarity measure (for instance, a distance on $\mathbb{R}^d$).

4. The n-mean algorithm

The n-mean algorithm is a very popular clustering technique. The following dissimilarity measure is considered:

$$D(x, y) = \|x - y\|^2.$$

The dissimilarity between a point x and the L_i prototype can be interpreted as error when the point x is approximated by the class prototype L_i. This dissimilarity can be written down as follows:

$$D(x, L_i) = \|x - L_i\|^2.$$

The criterion function will be in this case:

$$J(P, L) = \sum_{i=1}^{n} \sum_{x \in A_i} \|x - L_i\|^2. \tag{2}$$

Using the notation:

$$A_{ij} = \begin{cases} i, \ x^j \in A_i \\ 0, \ otherwise, \end{cases} \tag{3}$$

the criterion function will be:

$$J(P, L) = \sum_{i=1}^{n} \sum_{j=1}^{p} A_{ij} \left\| x^j - L_i \right\|^2 . \tag{4}$$

Taking into account that in Euclidian space, the scalar product has the form

$$\langle x, y \rangle = x^T M y,$$

where M is a symmetrical and positive defined matrix, the criterion function becomes:

$$J(P,L)=\sum_{i=1}^{n}\sum_{j=1}^{p}A_{ij}(x^j-L_i)^T M(x^j-L_i). \tag{5}$$

From the minimum condition

$$\frac{\partial J(P,L)}{\partial P}=0,\ i=1,\ldots,n \quad , \tag{6}$$

we have

$$-2\sum_{j=1}^{p}A_{ij}M(x^j-L_i)=0,\ i=1,\ldots,n \quad . \tag{7}$$

But the matrix M is nonsingular. Thus we obtain:

$$\sum_{j=1}^{p}A_{ij}x^j-\sum_{j=1}^{p}A_{ij}L_i=0,\ i=1,\ldots,n \quad . \tag{8}$$

From (8) we obtain:

$$L_i=\frac{\sum\limits_{j=1}^{p}A_{ij}x^j}{\sum\limits_{j=1}^{p}A_{ij}}\ i=1,\ldots,n. \tag{9}$$

But $p_i=\sum\limits_{j=1}^{p}A_{ij}$ represents the number of elements of the class A_i. L_i can also be written as:

$$L_i=\frac{1}{p}\sum_{x\in A_i}x. \tag{10}$$

We can now see that the prototype L_i is the mass centre of the A_i class. The representation $L=\{L_1,L_2,\ldots,L_n\}$, where L_i is given by (9) induces a new partition. This partition is obtained using *the nearest neighbour (NN) rule.*

If

$$\left\|x^j-L_i\right\|<\left\|x^j-L_k\right\|,\ k=1,\ldots,n,k\neq i \tag{11}$$

then x^{j}is assigned to the class A$_i$.

We may also write:

$$A_{ij}=\begin{cases}1,\ if\ \left\|x^j-L_i\right\|\leqslant\left\|x^j-L_k\right\|,\forall k\neq i\\ 0,\ otherwise\end{cases} \tag{12}$$

The n-mean algorithm consists of applying iteratively the equalities (9) and (12), starting from an initial partition of the set X. This partition can be arbitrarily chosen.

As a conclusion, we may say that determinist clustering methods allow the entities arrangement into classes which verify the following conditions:

a) the similarity between entities in a class is high;

b) the average similarity between class centres is low.

References

[1] Chang Y.K., Cirillo C., *"Evaluation of feedback retrieval using modified freezing, residual collection, and test and control groups",* Englewood Cliffs, Prentice Hall Inc., 1991

[2] Dumitrescu D., *"Mathematical principles of Classification Theory"*, Ed. Academiei Române, Bucuresti, 1999

[3] Popovici M., Rican G., *"Proiectare si implementare software",* Teora, 1998

[4] Salton G., *"Automatic Information Organization and Retrieval"*, McGraw-Hill, New York, 1975

[5] Salton G., Yang C.S., *"Contribution to the theory of indexing"*, American Elsevier, New York, 1980

[6] Sparck J., *"Automatic Keyword Classification for Information Retrieval"*, Butterworts, London, 1981

[7] van Rijsbergen C.J., *"Information retrieval"*, Butterworths, London, 1982

[8] Wong A., *"An investigation of the effects of differnt indexing methods on the document space configuration"*, Cambridge, England, 1987

[9] Veryard R., *"Information modelling – practical guidance"*, Prentice Hall, 1992