Validation of Very Large Data Sets Clustering by Means of a Nonparametric Linear Criterion

Israel Lerman[1], Joaquim Pinto da Costa[2], and Helena Silva[3]

[1] IRISA, University of Rennes I,
France, e-mail: lerman@irisa.fr
[2] DMA/FCUP, Faculdade de Ciências, Universidade do Porto,
Portugal, e-mail: jpcosta@fc.up.pt
[3] ISEP, Universidade do Porto,
Portugal, e-mail: helenabras@clix.pt

Abstract. In this paper we present a linearization of the Lerman clustering index for determining the number of clusters in a data set. Our goal was to apply the linearized index to large data sets containing both numerical and categorical values. The initial index, which was based on the set of pairs of objects, had a complexity $O(n^2)$. In this work its complexity is reduced to $O(n)$, and so, we can apply it to large data sets frequently encountered in Data Mining applications. The clustering algorithm used is an extention of the k-means algorithm to domains with mixed numerical and categorical values (Huang (1998)). The quality of the index is empirically evaluated on some data sets, both artificial and real.

1 Introdution

Due to the size of the data sets often used in Data Mining, the time consuming to evaluate a clustering index is a crucial problem. Therefore, if n is the number of data points to be clustered, a complexity redution from $O(n^2)$ to $O(n)$ is a very important task. Our aim is to write the Lerman index in such a way that its complexity becomes $O(n)$. We will first begin by remembering the earlier Lerman index expression (Lerman 1973, 1981, 1983); after that we will present the general principle of the redution and then, we will supply the explicite adaptation in the case where the data points can have an euclidean representation. We will finish by extending the new index to the case of mixed categorical and numerical variables.

2 Classical form of the index

Let us consider the simplified classic formula of the Lerman index. Let E be a finite set of cardinal n and S a similarity measure defined on it. Let $F = P_2(E)$ represent the set of pairs or subset with two elements of E. The values of S can be defined by the table

$$\{S(p)|p \in F\} \tag{1}$$

of dimension $n(n-1)/2$.

Let $\pi(E)$ be a partition of E into k clusters

$$\pi(E) = \{E_1, E_2, \ldots, E_l, \ldots, E_k\}, \tag{2}$$

whose adequacy with the similarity measure S is to be evaluated.

Regarding the table (1), we can calculate its mean $\mu(S)$ and its variance $\sigma^2(S)$:

$$\mu(S) = \frac{2}{n(n-1)} \sum \{S(p) | p \in F\} \tag{3}$$

$$\begin{aligned} \sigma^2(S) &= \frac{2}{n(n-1)} \sum \left\{ [S(p) - \mu(S)]^2 \, | p \in F \right\} \\ &= \frac{2}{n(n-1)} \sum \left\{ [S(p)]^2 \, | p \in F \right\} - [\mu(S)]^2 \end{aligned} \tag{4}$$

On the other hand, the partition (2) can be represented as a subset of F:

$$R(\pi(E)) = \sum_{1 \leq l \leq k} P_2 \left(E_l\right) \tag{5}$$

which defines the set of pairs put together by the partition $\pi(E)$. We can also designate by

$$S(\pi(E)) = \sum_{1 \leq l < l' \leq k} E_l * E_{l'} \tag{6}$$

the complementary set in F of $R(\pi)$. The expression $E_l * E_{l'}$ represents the set of non-ordered pairs $\{x, y\}$ where $x \in E_l$ and $y \in E_{l'}$, $1 \leq l < l' \leq k$.

Let us now designate by

$$r = card(R(\pi(E))) = \sum_{1 \leq l \leq k} \frac{n_l(n_l - 1)}{2} \tag{7}$$

and

$$s = card(S(\pi(E))) = \sum_{1 \leq l < l' \leq k} n_l \times n_{l'} \tag{8}$$

where $n_l = card(E_l), 1 \leq l \leq k$.

Before proceeding, the similarity S is normalized in the following way:

$$c(p) = \frac{S(p) - \mu(S)}{\sigma(S)} \tag{9}$$

In these conditions, the formula for the simplified form of the index is (see references above):

$$C(\pi, S) = \frac{1}{\sqrt{r \times s/f}} \sum \{c(p) | p \in R(\pi)\} \tag{10}$$

$$= \frac{1}{\sqrt{r \times s/f}} \sum \{\epsilon(p) c(p) | p \in F\} \tag{11}$$

where $f = r + s = card(F)$ and $\epsilon(p) = 1(p \in R(\pi))$.

3 Adaptation of the index for Euclidean data

Let us suppose that E can be represented by a set of points in an euclidean space. Let d be the metric distance, $I = \{1, 2, \ldots, i, \ldots, n\}$ the index set of E and I_l the subset of I containing the indexes of the cluster E_l, $1 \leq l \leq k$.

Instead of working with the table of similarity indices (1), we will consider from now on a table containing the distances between the elements of E:

$$\{d^2(i, i') | (i, i') \in I \times I\} \tag{12}$$

This table has dimension n^2. In these conditions, the expressions stated so far, wich were based on unordered pairs, can be now expressed by using the ordered pairs.

$$\mu'(d^2) = \frac{1}{n^2} \sum \{d^2(i, i') | (i, i') \in I \times I\} \tag{13}$$

$$\sigma'(d^2) = \frac{1}{n^2} \sum \{ \left[d^2(i, i') - \mu'(d^2)\right]^2 | (i, i') \in I \times I\} \tag{14}$$

The set of ordered pairs of objects that are clustered together is defined by:

$$R'(\pi(E)) = \sum_{1 \leq l \leq k} E_l \times E_l \tag{15}$$

and those pairs of objects that are in different clusters by:

$$S'(\pi(E)) = \sum_{1 \leq l \neq l' \leq k} E_l \times E_{l'} \tag{16}$$

In these conditions, we have:

$$r' = card(R'(\pi(E))) = \sum_{1 \leq l \leq k} n_l^2 \tag{17}$$

and

$$s' = card(S'(\pi(E))) = \sum_{1 \leq l \neq l' \leq k} n_l \times n_{l'} \tag{18}$$

For an ordered pair $q = (x, y)$, the normalized index related to $c(p)$ is:

$$c'(q) = \frac{d^2(x, y) - \mu'(d^2)}{\sigma'(d)} \tag{19}$$

and the index corresponding to (10) becomes:

$$C'(\pi, d^2) = \frac{1}{\sqrt{r' \times s'/n^2}} \sum \{c'(x, y) | (x, y) \in R'(\pi(E))\} \tag{20}$$

Let us remark that the index (20) can also be expressed by the formula:

$$C'(\pi, d^2) = \frac{1}{\sqrt{r' \times s'/n^2}} \sum_{1 \leq l \leq k} \sum \{c'(x, y) | (x, y) \in E_l \times E_l\} \tag{21}$$

4 Linear adaptation of the index for Euclidien data

We will start by expressing the basic formula that precisely allows the desired reduction in complexity. Let

$$\{M_i | i \in I\} \tag{22}$$

represent a cloud of n data points. We suppose, for simplification, that these points share the same weight; the generalization for different weights is immediate. Let G be the centroid of the cloud of points:

$$G = \frac{1}{n} \sum_{i \in I} M_i \tag{23}$$

It is known that

$$\frac{1}{n^2} \sum_{(i,i') \in I \times I} d^2(i, i') = \frac{2}{n} \sum_{i \in I} d^2(i, g) \tag{24}$$

where we replace M_i by i and G by g.

It is easy to discern that to calculate the left member takes $O(n^2)$ calculations of distances while the right member takes $O(n)$ calculations of distances.

The basic idea is then to replace the first two moments of the distribution

$$\{d^2(i,i')|(i,i') \in J \times J\} \tag{25}$$

by the first two moments of the distribution of

$$\{d^2(i,g_J)|i \in J\} \tag{26}$$

where g_J represents the centroid of the subset of data points indexed by J.

Now, in (21) let us consider the contribution of the cluster E_l for the value of the index. This is represented by:

$$\frac{1}{\sqrt{r's'/n^2}} \times \frac{1}{\sigma'(d^2)} \times \sum\{d^2(x,y) - \mu'(d^2)|(x,y) \in E_l \times E_l\} \tag{27}$$

But,

$$\sum\{d^2(x,y)-\mu'(d^2)|(x,y) \in E_l \times E_l\} = 2n_l \left\{ \sum_{x \in E_l} \left[d^2(x,g_l) - \frac{1}{n} \sum_{x \in E} d^2(x,g) \right] \right\} \tag{28}$$

Where g_l is the centroid of the cluster E_l. We remark that $d^2(x,g_l)$ is centered by the total moment of inertia. We can replace $\left[\sigma'(d^2)\right]^2$ by the variance of the distribution of $\{d^2(x,g)|x \in E\}$, which we designate by λ_g^2. The final expression is proportional to:

$$C_1(\pi,d^2) = \frac{1}{\sqrt{r's'}} \times \frac{1}{\lambda_g} \times \sum_{1 \leq l \leq k} n_l \sum\{d^2(x,g_l) - \mu_g|x \in E_l\} \tag{29}$$

where

$$\mu_g = \frac{1}{n} \sum_{x \in E} d^2(x,g)$$

and

$$\lambda_g^2 = \frac{1}{n} \sum_{x \in E} d^4(x,g) - \mu_g^2$$

5 The case of mixed numerical and categorical attributes

We are going to adapt our index for the metric case that is considered in (HUANG 1998) where the attributes may be numerical and/or categorical.

Each one of the data objects can be caracterized by $(v^1, \ldots, v^p, c^{p+1}, \ldots, c^m)$ where $v^1, v^2, \ldots, v^p$ represent the numerical attributes and $c^{p+1}, c^{p+2}, \ldots, c^m$ represent the categorical attributes $(m > p)$.

Here we consider $\{x_i | i \in I\}$ the data set $(I = \{1, 2, \ldots, n\})$. Let x_i^j be the value of the attribute j for the object x_i. This value is numerical if $1 \leq j \leq p$ and is categorical if $p + 1 \leq j \leq m$. The distance between x_i and $x_{i'}$ can be evaluate using the following formula:

$$d^2(x_i, x_{i'}) = \sum_{1 \leq j \leq p} \left(x_i^j - x_{i'}^j \right)^2 + \sum_{p+1 \leq j \leq m} \delta \left(x_i^j, x_{i'}^j \right) \tag{30}$$

where $\delta(x_i^j, x_{i'}^j) = \begin{cases} 0 \; if \; x_i^j = x_{i'}^j \\ 1 \; if \; x_i^j \neq x_{i'}^j \end{cases}$ for $p + 1 \leq j \leq m$

It is shown in (HUANG 1998) that the centroid of any set X is given by:

$$g_X = \left(g_X^1, \ldots, g_X^h, \ldots, g_X^p, f_X^{p+1}, \ldots, f_X^{p+j}, \ldots, f_X^m \right) \tag{31}$$

where g_X^h is the mean of the component h in X, $1 \leq h \leq p$, and where f_X^{p+j} represents the mode of the categorical variable $p + j$ (the most frequent category in X), $1 \leq j \leq m - p$.

Under these conditions, the adaptation of the index is immediate.

6 Application of the index

In this section the quality of the index is empirically evaluated on seven data sets. We tested on two artificial examples with two data sets each, and on three real data sets.

For each artificial example we have generated two data sets with 20,000 objects each. One of the data sets was defined on a bidimensional euclidean space, represented in Figures 1 and 5, and it contains five clusters. The second data set consisted on adding to the previous variables, four categorical attributes, having each four values. We determine in a random way the same distribution of these attributes on each cluster. But this common distribution of the four attributes is different from one cluster to another one. Thus, the probability distributions were the same for each categorical variable within a given cluster, but were different for different clusters. In Figure 2 and 7 it can be seen the five clusters identified by the k-prototype method (Huang 1998). The minimum value for our index is for $K = 5$ in both data sets of the first example , as can be seen in Figures 3 and 4. In Figures 8 and 9 it can be seen the variation of the index for the two data sets of the second example. For the mixed data set the number of clusters is correctly identified, whereas for the euclidean data set we got the minimum for $K = 3$ due to a weak efficience of the k-prototype identifying three and five clusters as it can be seen in Figure 6 and Figure 7.

In figures 10, 11 and 12 we can analyze the application of our index in the case of real data sets. These data sets belong to the Stalog database that are a subset of the datasets used in the European Statlog project. The data set , whose index values are represented in Figure 10 (Australian Credit Approval), concerns credit card applications. It has 690 objects with six numerical and eight categorical attributes and has 2 clusters. The data set, whose index values are represented in Figure 11, concerns image segmentation (Image Segmentation data). The instances were drawn randomly from a database of seven outdoor images, and the images were segmented to create a classification for every pixel. This data set has 2,310 objects with nineteen numerical attributes and has seven clusters. Finally in Figure 12 are the values of the index related to a data set (Shuttle Dataset) with 43, 500 objects with nine numerical attributes each and seven clusters.

7 Conclusions and Future Work

We have developped a criterion for the identification of the number of clusters present in a data set; and it can be applied to both numerical and categorical variables. This criterion has a linear computational complexity, which makes it very interesting to use in large data sets, as is common in Data Mining. We have seen empirically the importance of the new criterion on seven data sets, four artificial and three real ones. The criterion has been applied in conjunction with the k-prototypes algorithm (Huang 1998), and because of that, some interesting partitions that could have been identified by our criterion were not, because k-prototypes didn't find them. We are developping a new clustering method that incorporates the new criterion into the construction of the partitions, and hope that in this way the above disadvantage will be solved. We are also planning to consider the adaption of the index for unstructured data, because as it is now it can not be applied to the case of just one cluster. Finally, in the case of numerical data, the comparison of the behaviour of our criterion with some classical ones (Milligan and Cooper 1985) will be studied. For one of the most important of these, the "Cubic Clustering Criterion (CCC)" such comparison has been performed relative to the initial quadratic form of our criterion (Mollière 1986).

8 Acknowledgements

We thank Ross D. King for providing us with the real data sets.

References

HUANG, Z. (1998): Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery 2*, 283–304.

LERMAN, I.C.(1973): Étude distributionnelle de statisques de proximité entre structures finies de même type; application à la classification automatique. *Cahiers du Bureau Universitaire de Recherche Opérationnelle, 19*, Paris.

LERMAN, I.C.(1981): Classification et Analyse Ordinale des Données. Dunod, Paris.

LERMAN, I.C.(1983): Sur la signification des Classes Issues d'une Classiifcation Automatique de Données. In: J. Felsenstein (Eds.): *NATO ASI Series, Vol G1 Numerical Taxaromy.* Springer-Verlag.

MILLIGAN, G.W. and COOPER, M.C. (1985): An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50, 159-179.

MOLLIÈRRE, J.L. (1986): What's the real number of clusters. In: W. Gaul and M. Schader (Eds): *Classification as a tool research.*. North Holland.

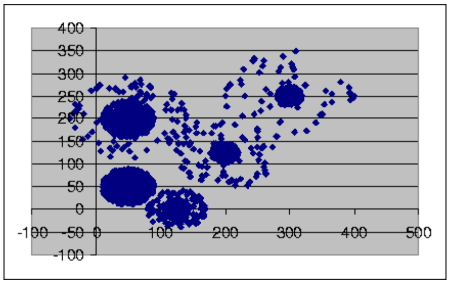

Fig. 1. Data set with 20,000 objects (two numerical attributes).

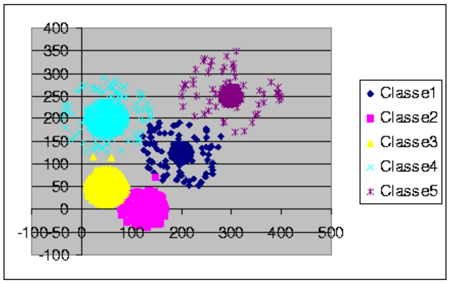

Fig. 2. Partition in five clusters identified by the k-prototype method, apllied to the data set of Figure 1.

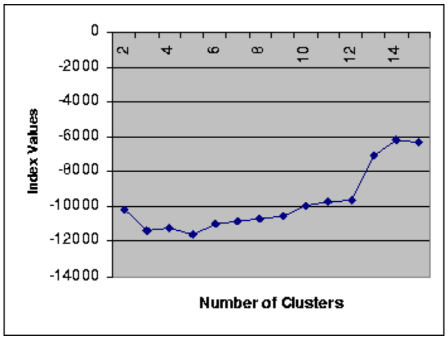

Fig. 3. Values of our index ($K = 2, \dots, 15$) for the data set represented in Figure 1.

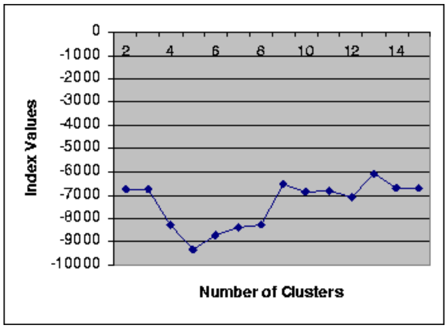

Fig. 4. Values of our index ($K = 2, \dots, 15$) for the data set with two numerical and four categorical attributes and five clusters.

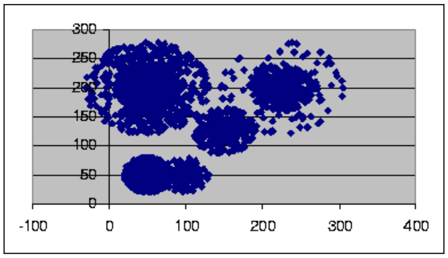

Fig. 5. Data set with 20,000 objects (two numerical attributes).

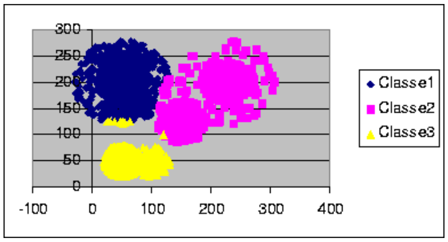

Fig. 6. Partition in three clusters identified by the k-prototype method, apllied to the data set of Figure 5.

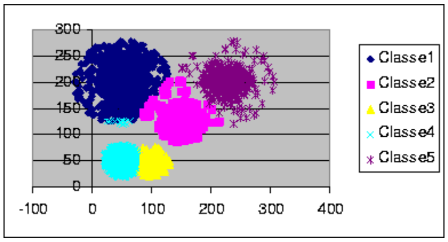

Fig. 7. Partition in five clusters identified by the k-prototype method, apllied to the data set of Figure 5.

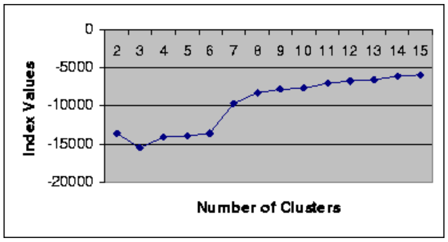

Fig. 8. Values of our index $(K = 2, \ldots, 15)$ for the data set represented in Figure 5.

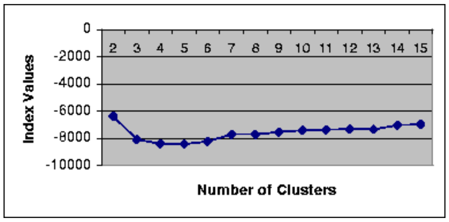

Fig. 9. Values of our index $(K = 2, \ldots, 15)$ for the data set with two numerical and four categorical attributes and five clusters.

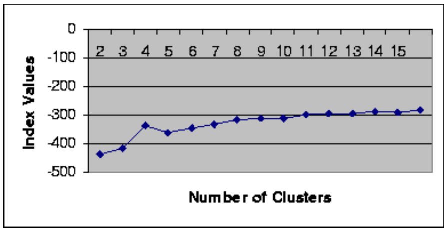

Fig. 10. Values of our index ($K = 2, \ldots, 15$) for the Australian Credit Approval data set.

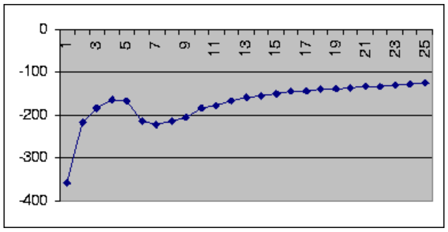

Fig. 11. Values of our index ($K = 2, \ldots, 25$) for the Image Segmentation data set.

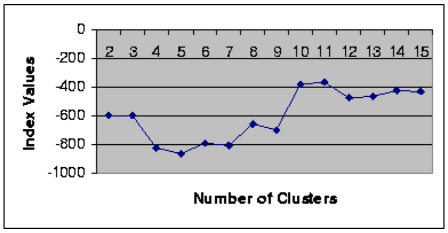

Fig. 12. Values of our index ($K = 2, \ldots, 15$) for the Shuttle Dataset.