Comparison of Three Objective Functions for Conceptual Clustering

Céline Robardet* and Fabien Feschet

Laboratoire d'Analyse des Systèmes de Santé
Université Lyon 1
UMR 5823, bât 101, 43 bd du 11 nov. 1918
69622 Villeurbanne cedex
FRANCE

Abstract. Unsupervised clustering algorithms aims to synthesize a dataset such that similar objects are grouped together whereas dissimilar ones are separated. In the context of data analysis, it is often interesting to have tools for interpreting the result. There are some criteria for symbolic attributes which are based on the frequency estimation of the attribute-value pairs. Our point of view is to integrate the construction of the interpretation inside the clustering process. To do this, we propose an algorithm which provides two partitions, one on the set of objects and the second on the set of attribute-value pairs such that those two partitions are the most associated ones. In this article, we present a study of several functions for evaluating the intensity of this association.

Keywords. Unsupervised clustering, conceptual clustering, association measures.

1 Introduction

One of the main data mining process consists in reducing the dimension of a dataset to increase knowledge and understanding. When no prior information are available, unsupervised clustering can be used to discover the underlying structure of the data. Indeed, those algorithms aim to build a partition on the objects such that the *most* similar objects belong to a same cluster, and the *most* dissimilar belong to different clusters. Hence, those procedures synthesize the data into few clusters. There is however no consensus of the algorithm to use because there are many ways to evaluate the proximity between objects and the quality of a partition. Furthermore, the cardinality of the set of all possible partitions increases exponentially with the size n of the set of objects, which leads to use fastest but often rough approximated optimization procedures.

Among the algorithms, we can distinguish two main families. The first one gathers *numerical* algorithms. They can be characterized by the computation of

* Corresponding author, e-mail: robardet@univ-lyon1.fr

L. De Raedt and A. Siebes (Eds.): PKDD 2001, LNAI 2168, pp. 399–410, 2001.

a distance between pairs of objects. This synthesizes all the dimensions of the problem into a single one. The distance is used to construct the partition. For example in the K-MEANS algorithm [JD88], the distance is the Euclidean one computed between the descriptive vectors of the objects embedded in the metric space defined by the attributes. The objective function is equal to the sum over all the clusters of the intra-class variance. Unfortunately, this function favors over-cut partitions and it is necessary to fix the number K of clusters before using it. In the case of the EM algorithm [CS96], the distance is evaluated by a multivariate Gaussian density. At each step, the memberships of the objects to the different clusters are evaluated, just as the parameters of the Gaussian densities associated to each cluster. But this algorithm is still dependent on an *a priori* number of clusters.

Whereas *statistical* clustering methods are often constructed to process datasets described by continuous features, *conceptual* clustering methods mainly focus on symbolic features [TB01,BWL95]. They aim to provide a better integration between the clustering and the interpretation stages of the data analysis process. Each feature is an attribute of discrete type having several different values. Attribute-value pairs are used in the construction of the clustering. Those algorithms built a hierarchy of concepts using probabilistic representation based on a conditional probabilistic vector of the apparition of the several attribute-value pairs in the several clusters. In COBWEB [Fis87,Fis96], the objective function measures the average increase in the prediction of attribute-value pairs knowing the partition. The optimization procedure is incremental but order dependent.

Sharing the same aims than *conceptual* clustering (dealing with symbolic features and combining an interpretation with the obtained partition), we focus in this article on a non hierarchical approach of the problem. This type of methods consists in the construction of two linked partitions, called a bi-partition, one on the set of objects and the second one on the whole set of attribute-value pairs. The interest of such a method is to discover the underlying structure of the data on the point of view of the objects as well as the descriptors. Thus, we search a bi-partition such that a unique cluster of objects fits a unique cluster of attribute-value pairs and conversely. Consequently, each partition is an interpretation of the other one, making easier the understanding of the results. Such methods have already been built. The simultaneous clustering algorithm [Gov84] is an adaptation of the *nuées dynamiques* of Diday [CDG$^+$88] for symbolic data. It consists in searching a bi-partition with the partition of the set of objects in *a priori* K clusters and the partition of the attribute-value pairs in L clusters. An *ideal* binary table of dimensions $K \times L$ is constructed, such that the gap between the initial data table structured by the two partitions and the *ideal* table is minimized. This method needs to *a priori* fix the number of clusters (K, L) of the bi-partition and the iterative procedure leads to a local optimum. To avoid those drawbacks, we propose a new algorithm based on the optimization of an objective function which does not need to *a priori* fixed the number of clusters. We focus in this article on the construction of such a function.

The contributions of this paper are, first an algorithm without any prescribed number of clusters, second a modification of association measures on co-occurrence tables to increase their discrimination power, and third an empirical study showing the relevance of the approach.

2 An Algorithm for the Construction of a Bi-partition

A basic clustering algorithm consists in optimizing a function which rewards partitions with interested properties. To define our algorithm, we need to construct a function for evaluating the quality of a bi-partition. This function must favor bi-partitions which satisfy the following property,

Property The functional link, which allows to restore one partition on the basis of the knowledge of the second one, must be as strong as possible. Furthermore, both partitions of the bi-partition must have the same number of clusters.

We denote by f such a function over $\mathcal{P}_\mathcal{O} \times \mathcal{P}_\mathcal{Q}$, where $\mathcal{P}_\mathcal{O}$ is the set of partitions on the set of objects, and $\mathcal{P}_\mathcal{Q}$ is the set of partitions on the set of attribute-value pairs: $f : \mathcal{P}_\mathcal{O} \times \mathcal{P}_\mathcal{Q} \to \mathbb{R}$. Let us denote by P an element of $\mathcal{P}_\mathcal{O}$ and Q one of $\mathcal{P}_\mathcal{Q}$.

This function must satisfy some properties. Such properties have been defined in supervised clustering, where the function measures the agreement between two partitions of the same set: the one given by the *class variable* and the one constructed by the supervised clustering method. Those properties are the followings [BFOS84,Weh96]:

- The function is maximal when to each cluster of P (resp. Q) is associated one and only one cluster of Q (resp. P)
- When every clusters of P can be associated to each cluster of Q indiscriminately, then the objective function must be minimum.
- The function must be invariant under permutation of the clusters of $\mathcal{O}$ and under permutation of the clusters of $\mathcal{Q}$.
- The function must be able to compare two bi-partitions with different numbers of clusters.

Nevertheless, we must add two new properties due to the fact that in our problem none of the two partitions constituting a bi-partition is *a priori* fixed. The function must also check the two following ones when it is maximal:

- Each object of a cluster of P owns all the attribute-value pairs belonging to its associated cluster of Q.
- Each attribute-value pair of a cluster of Q is owned by all the objects of its associated cluster of P.

Having define the function f to evaluate the quality of a bi-partition, the clustering algorithm, we propose, is based on a *gradient like* optimization. We thus propose the following algorithm,

Let (P_0, Q_0) *be a randomized initial bi-partition*
Repeat
Q_i *is fixed, we search* $P_{i+1} = \min_{P_{i+1} \in \mathcal{P}_\mathcal{O}} f(P_{i+1}, Q_i)$
P_{i+1} *is fixed, we search* $Q_{i+1} = \min_{Q_{i+1} \in \mathcal{P}_\mathcal{Q}} f(P_{i+1}, Q_{i+1})$
Until a convergence criterion is met

To modify a given bi-partition (P_i, Q_i), several ways are possible; either computing (P_{i+1}, Q_{i+1}) in one step, or computing (P_{i+1}, Q_{i+1}) in two steps as in the proposed algorithm. We choose this method to have a more tractable optimization problem and also, it is a way to fix one partition as a reference so to optimize the functional link with only one unknown.

3 The Kind of Functions to Use

The previous properties are partially satisfied by association measures, which have been built to evaluate the link between two attributes of discrete type. Those coefficients are widely used in supervised clustering [LdC96], whereas few unsupervised clustering algorithms used them. RIFFLE [MH91] uses Guttman's λ to measure the link between the partition (considered as an attribute) and each original discrete attribute. Such association measures can be adapted to be used as objective function in the search of a bi-partition.

The research of criteria, on one hand sufficiently complex to discriminate the different situations encountered and in the other hand sufficiently simple to allow intuitive interpretation, led to the creation of a lot of measures. Those measures have been constructed on contingency tables. After presenting some of them, we show how we construct the co-occurrence table and we modify them in the clustering context.

3.1 Some Association Measures

In the following of the paper, p_{ij} is an estimate of the probability that the value i of an attribute X, and the value j of an attribute Y arise simultaneously. n is the cardinal of the set of objects. (p_{ij}) define the so-called contingency table with $p_{.j} = \sum_i p_{ij}$ and $p_{i.} = \sum_j p_{ij}$ the margins.

A first group of association measures gather *divergence measures between probability distributions*. Those coefficients evaluate the association between a couple of attributes by measuring the gap between the current contingency table constructed on the two attributes, and the one obtained in case of independence. The situation of independence is easily characterized by$p_{ij} = p_{i.} \times p_{.j}$. A well-known measure of divergence is the χ^2,

$$\chi^2 = n \left[\sum_i \sum_j \frac{p_{ij}^2}{p_{i.} p_{.j}} - 1 \right]$$

This measure doesn't allow to compare contingency tables of different sizes (with different numbers of row and/or column), that is why we prefer to use a normalized version of this measure, the Tschuprow coefficient,

$$T = \sqrt{\frac{\chi^2}{n\sqrt{(p-1)(q-1)}}}$$

where p and q are the numbers of different values of each attributes.

A second class of measures gather *connection indices*. Those coefficients evaluate the gap with the situation of functional dependency characterized by a function f linking Y to X, $Y = f(X)$.

Goodman and Kruskal [GK54] built a family of such indices denoted by *measure of proportional reduction in error* (or PRE). Those coefficients have an easy interpretation due to the fact that they evaluate a prediction rule in terms of probability of error. The construction of such measures requires the definition of three elements:

- A prediction rule (C) of Y when X is known
- A prediction rule (I) of Y when X is unknown
- A measure of the error associated to the prediction rules

The PRE measure is then equal to:

$$\frac{error(I) - error(C)}{error(I)}$$

Guttman's λ is a PRE measure. It consists in predicting the value of an attribute by the most frequent one:

$$\lambda = \frac{(1 - \max_j p_{.j}) - \left(1 - \sum_i p_{i.} \max_j \frac{p_{.j}}{p_{i.}}\right)}{1 - \max_j p_{.j}}$$

Goodman and Kruskal proposed another more accurate coefficient called τ_b. Whereas λ focuses only on the most frequent value, τ_b measure takes into account all the structure of the distribution:

$$\tau_b = \frac{\sum_i \sum_j \frac{p_{ij}^2}{p_{i.}} - \sum_j p_{.j}^2}{1 - \sum_j p_{.j}^2}$$

This way to define connection indices can be generalized using the notion of uncertainties. We call uncertainty measure a concave function $I()$ on probability distributions. For example, the Shannon entropy $(-\sum_i p_i \log p_i)$, and the quadratic entropy $\left(2\left[1 - \sum p_i^2\right]\right)$ belong to this class.

The gain of uncertainties allows to measure the reduction in error of the prediction of Y knowing X. It is equal to $\Delta I(Y \mid X) = I(P(Y)) - E_X[I(P(Y \mid X))]$.

We always have $\Delta I(Y \mid X) \geq 0$. Moreover, if X and Y are independent then $\Delta I(Y \mid X) = 0$. The converse is true if I is strictly concave. An index of connection is then defined by $C(Y \mid X) = \frac{\Delta I(Y|X)}{I(P(Y))}$.

The τ_b coefficient is obtained when $I(P(Y)) = 1 - \sum_j p_{.j}^2$ When using the Shannon entropy as $I(P(Y))$, we obtain the uncertainty coefficient:

$$U = \frac{\sum_i p_{i.} \log p_{i.} + \sum_j p_{.j} \log p_{.j} - \sum_i \sum_j p_{ij} \log p_{ij}}{\sum_j p_{.j} \log p_{.j}}$$

3.2 How to Build the Co-occurrence Table

In our problem, we search a bi-partition constituted of two partitions such that their association is maximal. However, whereas the previous measures are based on contingency table, i.e. a table crossing two partitions on a same set, the partitions considered in our problem are built on separate sets which have a semantic link expressed through the data table. This link allows us to construct a co-occurrence table (η_{ij}). We consider h attributes V_i with values in a discrete space dom_i ($V_i : \mathcal{O} \to \mathrm{dom}_i$) and denote by $\mathcal{Q}$ the set of attribute-value pairs $\left(\mathcal{Q} = \biguplus_{i=1}^{h} \mathrm{dom}_i\right)$. Using the previous notations, we built a co-occurrence table between a partition $P = (P_1, \ldots, P_K)$ on the set $\mathcal{O}$ of objects and a partition $Q = (Q_1, \ldots, Q_K)$ on the set $\mathcal{Q}$ such that the elements of this table equal the number of attribute-value pairs of Q_j taken by the objects of P_i. More precisely,

$$\eta_{ij} = \sum_{x \in P_i} \sum_{y \in Q_j} \sum_{i=1}^{h} \delta_{V_i(x),y}$$

where δ is the Kronecker[1] symbol. We also use the following notations: $\eta_{i.} = \sum_{j=1}^{K} \eta_{ij}$, $\eta_{.j} = \sum_{i=1}^{K} \eta_{ij}$ and $\eta_{..} = \sum_{i=1}^{K} \sum_{j=1}^{K} \eta_{ij} = \sharp(\mathcal{O}) \times h$.

Then, we compute the previous measures by substituting in the formulas p_{ij}, $p_{i.}$, $p_{.j}$ and n by respectively η_{ij}, $\eta_{i.}$, $\eta_{.j}$ and $\eta_{..}$. Nevertheless, the computation of the *Uncertainty* coefficient requires the normalization of the co-occurrence table. That is why we divide every η_{ij} by $\eta_{..}$ when computing this coefficient.

3.3 Adaptation of Functions

Recall that the association measures only partially check the necessary properties. On a contingency table the notion of *purity* has no meaning whereas it is a key point in our problem when wanting to compare tables of different sizes. A cluster of objects is all the more *pure* since the objects have similar description on all the attributes. Similarly, a cluster of attribute-value pairs is all the more *pure* since the attribute-value pairs are taken by the same set of objects.

For example, we consider a data table composed of three *pure* clusters on objects and attribute-value pairs, and the perfect bi-partition associated (see Fig. 1 left). In this case, the association measure is maximum. If we merge two classes of objects, and the two associated classes of attribute-value pairs (see Fig. 1 right), the association measure is still maximum. Consequently, those two different situations are not discriminated by the measure.

[1] $\delta_{V_i(x),y} = 1$ if $V_i(x) = y$, $\delta_{V_i(x),y} = 0$ *otherwise*

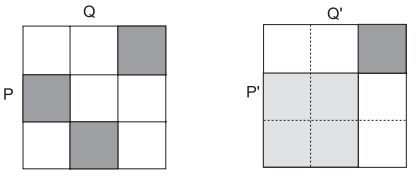

Fig. 1. A maximum value for two different situations

To overcome this drawback, we use a diversity measure between a cluster of objects and a cluster of attribute-value pairs to map η_{ij} into $[0,1]$ such that η_{ij} is maximum when the cluster is *pure*.

Since each object owns a unique value per attribute, η_{ij} is maximum when each object of P_i owns, for all the attributes represented in Q_j, a value belonging to Q_j. In this case, $\eta_{ij} = \sharp(P_i) \times \sum_i \left(1 - \delta_{V_i^{-1}(Q_j),\emptyset}\right)$, that is the number of objects times the number of different attributes belonging to Q_j. Thus in general η_{ij} divided by $\sharp(P_i) \times \sum_i \left(1 - \delta_{V_i^{-1}(Q_j),\emptyset}\right)$ belongs to $[0,1]$. However, this is not sufficient to discriminate the cases of Fig. 1. We must penalize Q_j with several values per attribute to solve the problem. We decide to penalize it by $\prod_{a \in H_j} \sharp(\mathrm{dom}_a \cap Q_j)$, with $H_j = \{a \in [1,h] \mid \mathrm{dom}_a \cap Q_j \neq \emptyset\}$. It is equal to one only when there is one value per attribute in Q_j and is greater than one otherwise.

Consequently to map η_{ij} into $[0,1]$, we replace η_{ij} by the following diversity measure in the co-occurrence table,

$$\frac{\eta_{ij}}{\sharp(P_i) \times \sum_i \left(1 - \delta_{V_i^{-1}(Q_j),\emptyset}\right) \times \prod_{a \in H_j} \sharp(\mathrm{dom}_a \cap Q_j)}$$

Nevertheless, modifying the values in the co-occurrence table does not influence the value of the association measure used. Indeed, association measures rely on the evaluation of the similarity between η_{ij} and $\eta_{i.}$ and/or $\eta_{.j}$. That is why in order to take into account the effect of the diversity measure we have to compute a global index of diversity. It consists in the embedding of the co-occurrence table in the set of assignment matrices to force a functional link between the elements of a bi-partition. The set of possible assignment matrices $A = (A_{ij})$ contains all matrices such that

$$\forall i, \exists! j \quad \textit{such} \quad \textit{that} \quad A_{ij} = \eta_{ij} \neq 0 \quad \text{and} \quad \forall j, \exists! i \quad \textit{such} \quad \textit{that} \quad A_{ij} = \eta_{ij} \neq 0$$

Among the set of assignment matrices, we choose the one whose coefficients average is maximum. The global index of diversity is this average. Notice that the association measures belong to $[0,1]$ and that the average of (A_{ij}) also belongs to $[0,1]$. We thus weight the association measures by multiplying it by this global diversity index.

4 An Experimental Study of the Functions

In this section, we empirically study the functions regarding their capabilities to discriminate associated partitions. Those functions are the τ_b, the *Tschuprow* and the *Uncertainty* coefficients. Whereas in the previous section we modify those functions to ensure their discrimination of the *pure* bi-partition, we study in this section the regularity of those functions over others bi-partitions.

4.1 Study of the Quality Indices of the Partitions

In [RF00] we have proposed a graph theoretical approach to define an order on the set of partitions. It consists in constructing a data table such that there exist a bi-partition whose clusters are all pure. This bi-partition is then used as a reference called *ideal* bi-partition. A distance between each partition of $\mathcal{P}_\mathcal{O}$ and the partition on $\mathcal{O}$ belonging to the *ideal* bi-partition is then computed to order $\mathcal{P}_\mathcal{O}$. On a same manner, a distance between each partition of $\mathcal{P}_\mathcal{Q}$ and the partition on $\mathcal{Q}$ belonging to the *ideal* bi-partition is also computed. The distances used [RF00] are well discriminant and evaluate the proximity between two partitions in terms of similarity on the variables and objects shared between the closest clusters of both partitions.

That is why we study the discrimination power of a function on a set of partitions through the link between function's values and those orders on the set of partitions.

The following graphs represent the variations of the measures regarding the distance of the partitions to the associated *ideal* one. The partitions on the set of objects $\mathcal{O}$ are ordered on abscissa axis (right). On the ordinate axis (left) are ordered partitions on the set of attribute-value pairs $\mathcal{Q}$. On the z-axis is plotted the value of the function computed with the two partitions. The partitions are based on a 12×12 data table composed of three *pure* clusters on each set. In a first step, we have generated all the partitions on each set. But given the fact the number of partitions on each set is huge (more than 4 millions), we could not plot the graphs over the whole set of couples of partitions. That is why we selected a subset of 100 partitions in each set. We did not choose those partitions randomly for two reasons. First, if we pick up partitions in a uniform manner, we almost obtain partitions with *worst* values regarding the function and the distance. Indeed, among the 4 millions of partitions there are lots of bad ones. Secondly, we do not know the distribution of *good* partitions. Consequently, we chose partitions among the exhaustive set such that the partitions are well spread over the distance.

The Tschuprow measure (Fig. 2 left) seems smooth and regular over the couples of partitions. The incline of the surface is correctly oriented, with high values for *good* partitions and slow decrease towards the *bad* ones. Nevertheless, the slope is not very important. This could be an obstacle for the optimization procedure. In fact, in case of an optimization based on local (with respect to the distance) descent, the bumpiness of the function might be an obstacle. However, when considering stochastic procedures, like genetic algorithms, the slope of the

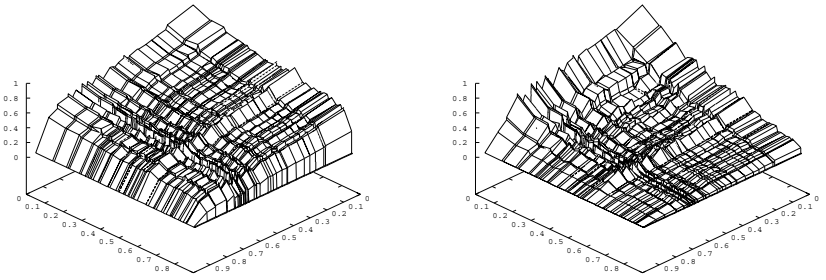

Fig. 2. Tschuprow and Tschuprow adapted

surface induces the survival of better individuals (with respect to the distance). There is a step in the border of the graph which means that a small increase of the quality of *very bad* partitions leads to an important increase of the objective measure. The modification of the measure (Fig. 2 right) globally increases the slope of the surface. Whereas the surface is a little bit more bumpy, the stochastic optimization should be more easy on this function.

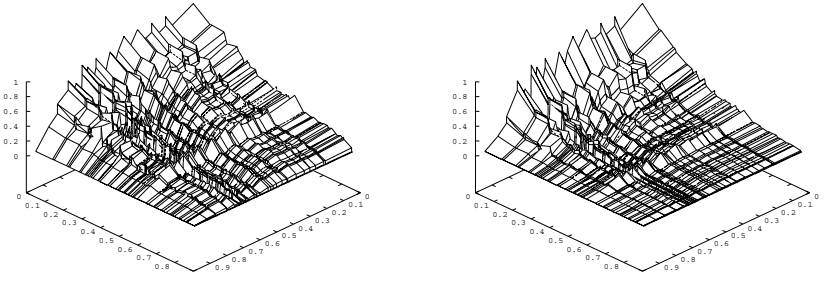

Fig. 3. τ_b (left) and τ_b adapted (right)

On Fig. 3 (left) we can see that the τ_b measure discriminates almost well the couples of partitions. The highest values of the function are obtained for *good* bi-partitions, and the values of the function decrease with the distance. The rough patches of the surface are more important than for the *Tschuprow* measure but the slope is much more important. The diversity coefficient (Fig. 3 right) flatten the surface partially erasing the bumps.

The results obtained with the *Uncertainty* measure (Fig. 4 left and right) are very similar to those given by the τ_b function. It is visually perceptible that

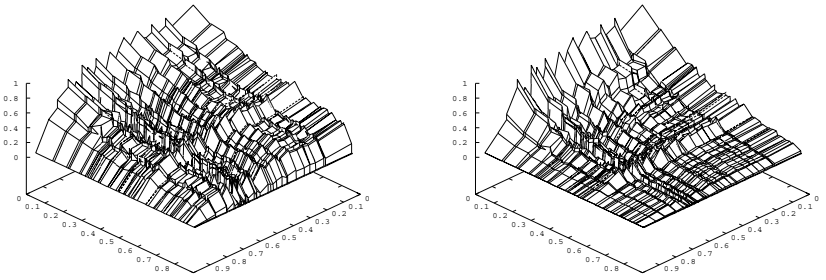

Fig. 4. Uncertainty coefficient original (left) and adapted (right)

those functions are similar each others and different to the *Tschuprow* measure. This is due to the fact that they do not evaluate the same property on the co-occurrence table. Whereas the *Tschuprow* evaluates the distance between the current table and the one corresponding to the independence situation, the two others evaluate the strength of the *functional* link between the two partitions.

4.2 Smoothness of the Functions

In this section, we study the behavior of the functions when the co-occurrence matrix is slightly modified. Those modifications are built by interpolation between two matrices. The following graphs represent the value of the functions at each of the 150 steps of the interpolation.

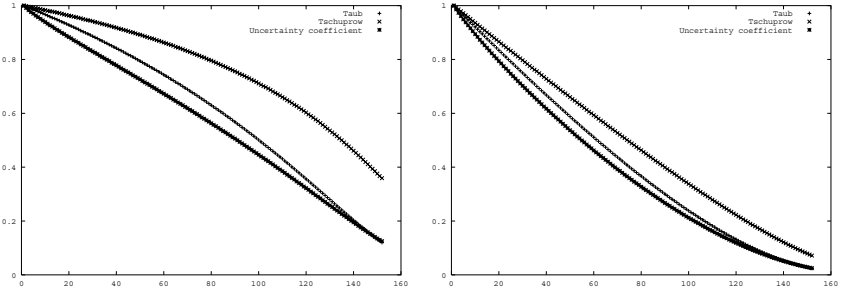

Fig. 5. Linear interpolation to Gaussian matrix non adapted vs adapted

In Fig. 5, we interpolate the matrix corresponding to the *ideal* bi-partition to a Gaussian modification of it. The interpolation is linear. We do this to measure the resistance of the function to a regular destruction of the functional link.

The three functions decrease quasi linearly (Fig. 5 left). This graph confirms that τ_b and *Uncertainty* coefficients have a similar behavior, which is rather different than those of the *Tschuprow* measure. Its decrease is slower regarding the two other functions. The diversity coefficient (Fig. 5 right) modifies the curves so that they have a similar linear slope.

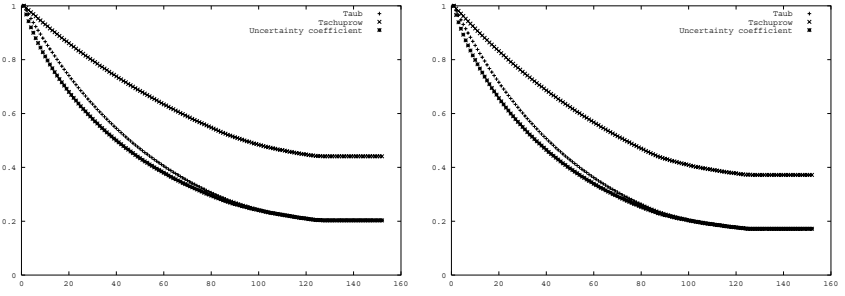

Fig. 6. Random interpolation to random matrix non adapted vs adapted

In Fig. 6, we interpolate the *ideal* co-occurrence matrix to a totally random one. The interpolation is linear for all cells of the matrix, but each cell has a randomly chosen number of steps. Consequently each cell of the matrix has its own speed of interpolation. This case is far less regular than the previous one.

Along the slight modifications of the matrix, the functions decrease more quickly than in the previous interpolation (Fig. 6 left). Nevertheless, the functions are still very smooth. The behavior of the functions is not affected by the diversity coefficient (Fig. 6 right).

Finally, the proposed modification of the association measures leads to an increase in their discrimination power but keeping their good behavior in resistance and regularity towards the measure of a functional link.

5 Conclusion and Perspective

In this article, we have presented an algorithm for finding bi-partition in unsupervised clustering. It is based on the search of a couple of the *most* associated partitions. Those partitions are based on the set of objects and the set of attribute-value pairs which are linked in the original dataset. In order to find this bi-partition, we propose three objective functions to optimize. We have adaptated the *Tschuprow*, the τ_b and the *Uncertainty* measures to the unsupervised clustering problem. The experimentation we provide, give two main information. First we notice that the τ_b measure and the *Uncertainty* coefficient have similar behaviors. The *Tschuprow* function is quite different, more smooth but may be less discriminant. Secondly, the application of the diversity coefficient we have

introduced, to allow the functions to check all the required properties, slightly modify the functions. Globally, the functions are smoother and discriminant. In a further work we will present optimization procedures.

References

[BFOS84] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth International, California, 1984.

[BWL95] G. Biswas, J. Weinberg, and C. Li. Iterate: a conceptual clustering method for knowledge discovery in databases. Technical report, Departement of Computer Science, Vanderbilt university, Nashville, 1995.

[CDG+88] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, and H. Ralambondrainy. *Classification automatique des données*. Dunod, paris, 1988.

[CS96] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, 13:195–212, 1996.

[Fis87] D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.

[Fis96] D. H. Fisher. Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research*, 4:147–180, 1996.

[GK54] L. A. Goodman and W. H. Kruskal. Measures of association for cross classification. *Journal of the American Statistical Association*, 49:732–764, 1954.

[Gov84] G. Govaert. Classification simultanée de tableaux binaires. In E. Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone, editors, *Data analysis and informatics III*, pages 233–236. North Holland, 1984.

[JD88] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice Hall, Englewood cliffs, New Jersey, 1988.

[LdC96] I.C. Lerman and J. F. P. da Costa. Coefficients d'association et variables à très grand nombre de catégories dans les arbres de décision : application à l'identification de la structure secondaire d'une protéine. Technical Report 2803, INRIA, février 1996.

[MH91] G. Matthews and J. Hearne. Clustering without a metric. *IEEE Transaction on pattern analysis and machine intelligence*, 13(2):175–184, 1991.

[RF00] C. Robardet and F. Feschet. A new methodology to compare clustering algorithms. In H. Meng K. S. Leung, L. Chan, editor, *Intelligent data engineering and automated learning-IDEAL 2000*, number 1983 in LNCS. Springer-Verlag, 2000.

[TB01] L. Talavera and J. Béjar. Generality-based conceptual clustering with probabilistic concepts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):196–206, 2001.

[Weh96] L. Wehenkel. On uncertainty measures used for decision tree induction. In *Info. Proc. and Manag. of Uncertainty*, pages 413–418, 1996.