Cluster Analysis Based on the Central Tendency Deviation Principle

Julien Ah-Pine

Cluster analysis based on the central tendency deviation principle

Julien Ah-Pine

Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, France
julien.ah-pine@xrce.xerox.com

Abstract. Our main goal is to introduce three clustering functions based on the central tendency deviation principle. According to this approach, we consider to cluster two objects together providing that their similarity is above a threshold. However, how to set this threshold ? This paper gives some insights regarding this issue by extending some clustering functions designed for categorical data to the more general case of real continuous data. In order to approximately solve the corresponding clustering problems, we also propose a clustering algorithm. The latter has a linear complexity in the number of objects and doesn't require a pre-defined number of clusters. Then, our secondary purpose is to introduce a new experimental protocol for comparing different clustering techniques. Our approach uses four evaluation criteria and an aggregation rule for combining the latter. Finally, using fifteen data-sets and this experimental protocol, we show the benefits of the introduced cluster analysis methods.

Key words: Cluster analysis, clustering functions, clustering algorithm, categorical and real continuous data, experimental protocol.

1 Introduction

Clustering is one of the main tasks in data analysis and in data mining fields and has many applications in real-world problems. Given a set of N objects $\mathbb{O} = \{o^1, \ldots, o^N\}$, described by a set of P features $\mathbb{D} = \{D^1, \ldots, D^P\}$, the clustering problem consists in finding homogeneous groups of these objects. However, clustering is a NP-hard problem and one has to use heuristics for processing large data-sets. Reviews of such heuristics can be found in [1–3]. With respect to the usual taxonomy of clustering methods [1], the present paper aims at contributing to the family of hard partitional techniques. As it was underlined in [3, 4], one important factor in that case, is the partitioning function that a clustering algorithm attempts to optimize. Hence, the contributions of this paper are the following ones.

First, we introduce three clustering functions that have the following general formulation:

$$F(S, \mu, X) = \sum_{i,i'=1}^{N} (S_{ii'} - \mu_{ii'}) X_{ii'} \ . \tag{1}$$

where: S is a similarity matrix; μ is a matrix of central tendency measures of similarities given for each pair of objects; and X is a relational matrix which general term, $X_{ii'}$, equals 1 if o^i and $o^{i'}$ are in the same cluster; 0 otherwise. X is similar to an adjacency matrix, yet, as it must be a partition, it has to satisfy the following linear constraints [5, 6], $\forall i, i', i'' = 1, \ldots, N$:

$$\begin{array}{ll} X_{ii} = 1 \ . & \text{(reflexivity)} \\ X_{ii'} - X_{i'i} = 0 \ . & \text{(symmetry)} \\ X_{ii'} + X_{i'i''} - X_{ii''} \leq 1 \ . & \text{(transitivity)} \end{array} \tag{2}$$

Following the relational analysis method (RA in the following) in cluster analysis [6, 7], the related clustering problems that we want to solve can be formally stated as: $\max_X F(S, \mu, X)$ with respect to the linear constraints[1] given in (2).

Regarding (1), our main point of interest concerns the variable μ. By maximizing $F(S, \mu, X)$, we highly consider to group o^i and $o^{i'}$ together if their similarity $S_{ii'}$ is greater than $\mu_{ii'}$. However, how to set $\mu_{ii'}$? On which basis should we compute $\mu_{ii'}$? Several existing clustering methods are based on the same kind of objective function [9, 6]. However, most of them consider $\mu_{ii'}$ as a constant parameter which can be set by the user. *On the contrary, in this paper, we define new clustering functions for which the quantities $\mu_{ii'}$ are data-dependent and are interpreted as central tendency measures of pairwise similarities.*

Then, in order to rapidly find an approximate solution X, we also introduce a clustering algorithm that amounts to a local search based technique. This approach is quite similar to the leader algorithm proposed in [9] and is also studied in the RA context in [6, 2]. These techniques interesting as they don't require to fix the number of clusters.

The secondary purpose of this work, is to suggest a new experimental protocol for assessing different cluster analysis methods. Indeed, in most papers covering clustering techniques, only one or two assessment measures are used. *In this work, we propose to take into account four different evaluation criteria and a score aggregation method in our experimental protocol.* Hence, we propose to use this approach for benchmarking the introduced clustering methods.

The rest of this paper is organized as follows. In section 2, we introduce the central tendency deviation principle and define three new clustering functions. Next, in section 3, we detail the associated clustering algorithm. Then, we present our experimental protocol and the obtained results, in section 4. Finally, we give some conclusions and future work in section 5.

[1] Note that we can use integer linear programming to solve this clustering problem [5, 7] but only for very small data-sets. Indeed, as it was already mentioned, this optimization problem is a NP-hard one [8].

2 Maximal association criteria and the central tendency deviation principle

In this section, we introduce three clustering functions for partitioning continuous numerical data. Firstly, we recall the maximal association approach in the RA framework [7]. In this context, association measures are used in their relational representations in order to design objective functions for clustering categorical data. Secondly, we extend these clustering functions to the more general case of continuous data. In that perspective, we underline the central tendency deviation principle.

2.1 Contingency and relational representations of association measures

Contingency table analysis aims at measuring the association between categorical variables. Let V^k and V^l be two categorical variables with category sets $\{D_s^k; s = 1, \ldots, p_k\}$ and $\{D_t^l; t = 1, \ldots, p_l\}$ (p_k being the number of category of V^k). Their associated contingency table denoted n of size $(p_k \times p_l)$, is defined as follows: n_{st} = Number of objects in both categories D_s^k and D_t^l. Then, it exists many association criteria based on the contingency table [10, 11]. We are particularly interested in the three following ones:

• The Belson measure (denoted B) introduced in [12] which is related to the well-known χ^2 measure. The former is actually a non-weighted version of the latter.

• The squared independence deviation measure (denoted E) which was introduced in [11] and studied in [13] for measuring similarities between categorical variables.

• The Jordan measure (denoted J), which is a coefficient based upon [14] but which was formally introduced in [11].

We recall below, the contingency representation of these criteria[2].

$$B(V^k, V^l) = \sum_{s=1}^{p_k} \sum_{t=1}^{p_l} \left(n_{st} - \frac{n_{s.} n_{.t}}{N} \right)^2 .$$

$$E(V^k, V^l) = \sum_{s,t} n_{st}^2 - \frac{\sum_s n_{s.}^2 \sum_t n_{.t}^2}{N^2} .$$

$$J(V^k, V^l) = \frac{1}{N} \sum_{s,t} \left(n_{st} \left(n_{st} - \frac{n_{s.} n_{.t}}{N} \right) \right) .$$

where $n_{s.} = \sum_{t=1}^{p_l} n_{st}$.

In the context of the RA approach in contingency table analysis [11, 15, 7], we can equivalently express the previous criteria using relational matrices. For V^k, its relational matrix is denoted C^k and its general term $C_{ii'}^k$ equals 1 if o^i and

[2] Notice that all of them are null when the variables V^k and V^l are statistically independent.

$o^{i'}$ are in the same category; 0 otherwise. Then, one can observe the following identities between contingency and relational representations [16, 15]:

$$\sum_{s,t} n_{st}^2 = \sum_{i,i'} C_{ii'}^k C_{ii'}^l \; ; \; \sum_{s} n_{s.}^2 = C_{..}^k \; ; \; \sum_{s,t} n_{st} n_{s.} n_{.t} = \sum_{i,i'} \frac{C_{i.}^k + C_{.i'}^k}{2} C_{ii'}^l \; . \quad (3)$$

where $\sum_i C_{ii'}^k = C_{.i'}^k$ and $\sum_{i,i'} C_{ii'}^k = C_{..}^k$.

Consequently, we obtain the following relational representations of the studied coefficients [15]:

$$B(C^k, C^l) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(C_{ii'}^k - \frac{C_{i.}^k + C_{.i'}^k}{N} + \frac{C_{..}^k}{N^2} \right) C_{ii'}^l \; . \quad (4)$$

$$E(C^k, C^l) = \sum_{i,i'} \left(C_{ii'}^k - \frac{C_{..}^k}{N^2} \right) C_{ii'}^l \; . \quad (5)$$

$$J(C^k, C^l) = \frac{1}{N} \sum_{i,i'} \left(C_{ii'}^k - \frac{C_{i.}^k}{N} \right) C_{ii'}^l \; . \quad (6)$$

The relational formulation of association measures is of interest for analyzing the differences between such coefficients [17]. In our context, this formalism allows to define clustering functions for categorical data. We recall and extend this aspect in the following subsection.

2.2 Maximal association criteria and their extensions

Let suppose M categorical variables $V^k; k = 1, \ldots, M$ and let Δ represent one of the three studied association measures (4), (5) or (6). Then, the maximal association problem introduced in [7], can be formally stated as follows:

$$\max_X \frac{1}{M} \sum_{k=1}^{M} \Delta(C^k, X) \; .$$

where X is a relational matrix which satisfies (2).

In other words, we want to find the consensus partition X that maximizes the mean average association with all categorical variables $C^k; k = 1, \ldots, M$. This amounts to a clustering model for categorical data or consensus clustering [18]. Thanks to the relational representation, we have the following property [7]:

$$\frac{1}{M} \sum_{k=1}^{M} \Delta(C^k, X) = \frac{1}{M} \Delta \left(\sum_{k=1}^{M} C^k, X \right) = \frac{1}{M} \Delta \left(\mathbf{C}, X \right) \; . \quad (7)$$

where $\mathbf{C} = \sum_{k=1}^{M} C^k$ and $\mathbf{C}_{ii'}$ is the number of agreements between o^i and $o^{i'}$ (the number of variables for which both objects are in the same category).

Therefore, it appears that the mean average of the association between X and the relational matrices $C^k; k = 1, \ldots, M$, is equivalent to the association between the former and an aggregated relational matrix[3], $\mathbf{C}$.

The summation property (7) is the basis of the extension of the studied objective functions, that we are going to introduce. Indeed, let represent categorical data as an indicator matrix T of size $(N \times P)$ where $P = \sum_{k=1}^{M} p_k$. In that representation, we consider all categories of all nominal variables. Accordingly, we denote $\mathbb{D} = \{D^1, \ldots, D^P\}$ the set of all categories. Then, the general term of T, T_{ij}, equals 1 if o^i is in category D^j; 0 otherwise. In that representation, any object o^i is a binary vector of size $(P \times 1)$ and the following observation becomes straightforward: $\mathbf{C}_{ii'} = \sum_{j=1}^{P} T_{ij} T_{i'j} = \langle o^i, o^{i'} \rangle$, where $\langle ., . \rangle$ is the euclidean dot product. *Given this geometrical view, we can consider the extension of the clustering functions recalled previously to the more general case of continuous numerical features. Indeed, we simply interpret $\mathbf{C}_{ii'}$ as a dot product between vectors o^i and $o^{i'}$ and we symbolically replace any occurrence of $\mathbf{C}$ with a generic notation S. Henceforth, we assume that S is a Gram matrix (also called similarity matrix) and the vectors $o^i; i = 1, \ldots, N$, are represented in a continuous feature space T of dimension P.*

2.3 The central tendency deviation principle

According to the previous subsection, the objective functions that we want to study are the following ones[4].

$$B\left(S, X\right) = \sum_{i,i'} \left(S_{ii'} - \left(\frac{S_{i.}}{N} + \frac{S_{i'.}}{N} - \frac{S_{..}}{N^2} \right) \right) X_{ii'} \ . \tag{8}$$

$$E\left(S, X\right) = \sum_{i,i'} \left(S_{ii'} - \frac{S_{..}}{N^2} \right) X_{ii'} \ . \tag{9}$$

$$J\left(S, X\right) = \sum_{i,i'} \left(S_{ii'} - \frac{1}{2} \left(\frac{S_{i.}}{N} + \frac{S_{i'.}}{N} \right) \right) X_{ii'} \ . \tag{10}$$

where $S_{i.} = S_{.i} = \sum_{i'} S_{ii'}$ (since S is symmetric) and $S_{..} = \sum_{i,i'} S_{ii'}$.

Given the previous equations, we can draw out the central tendency deviation principle. Indeed, one can observe that *all objective functions are based on a comparison between the similarity of two objects and a central tendency measure.*

In the case of B defined in (8), the transformation from $S_{ii'}$ to $\left(S_{ii'} - \frac{S_{i.}}{N} - \frac{S_{i'.}}{N} + \frac{S_{..}}{N^2} \right)$ is a geometrical one and is known as the Torgerson transformation [19]. Let $g = \frac{1}{N} \sum_{i=1}^{N} o^i$ be the mean vector. Then, we have: $\left(S_{ii'} - \frac{S_{i.}}{N} - \frac{S_{i'.}}{N} + \frac{S_{..}}{N^2} \right) = \langle o^i - g, o^{i'} - g \rangle$. For the Belson function, the *objects o^i and $o^{i'}$ could be clustered together providing that their dot product, centered with respect to the mean vector g, is positive.*

[3] Also called the collective Condorcet matrix in the RA approach [5].

[4] For the Jordan function, we drop the factor $\frac{1}{N}$.

Regarding E given in (9), the central tendency is the mean average over all pairwise similarities, $\frac{S_{..}}{N^2}$. This approach is also a global one as it considers all (pairs of) objects. In that case, *o^i and $o^{i'}$ are more likely to be in the same cluster if their own similarity is above the mean average of all similarities.*

Unlike the previous cases, the function J introduced in (10), is based on a local central tendency approach. For the Jordan function, *o^i and $o^{i'}$ have more chances to be grouped together if their similarity is greater than the arithmetic mean of the mean average of their similarity distributions $\frac{S_{i.}}{N}$ and $\frac{S_{i'.}}{N}$.*

However, one special case to consider is when the data are already centered. Indeed, if $S_{ii'} = \langle o^i - g, o^{i'} - g \rangle$, all three clustering functions become equivalent as $\frac{S_{i.}}{N} = \frac{S_{i'.}}{N} = \frac{S_{..}}{N} = 0$. Despite this point, we propose a version of the clustering functions that combines two kinds of central tendency approaches.

Following the previous observation and the Belson function, we first center the data. This leads to similarities $S_{ii'}$ that are either positive or negative. Next, *we focus on positive similarities only.* Indeed, the latter are related to pairs of vectors whose cosine index is positive which indicates that they are rather similar. Thus, let $\mathbb{S}^+$ be the set of pairs of objects having positive similarities: $\mathbb{S}^+ = \{(o^i, o^{i'}) \in \mathbb{O}^2 : S_{ii'} \geq 0\}$. Then, we compute the central tendency measures related to the clustering criteria, on the basis of pairs belonging to $\mathbb{S}^+$. More concretely, below are the clustering functions that we propose to define:

$$B^+(S, X) = \sum_{i,i'} \left(S_{ii'} - \left(\frac{S^+_{i.}}{N^+_{i.}} + \frac{S^+_{i'.}}{N^+_{i'.}} - \frac{S^+_{..}}{N^+_{..}} \right) \right) X_{ii'} \ . \tag{11}$$

$$E^+(S, X) = \sum_{i,i'} \left(S_{ii'} - \frac{S^+_{..}}{N^+_{..}} \right) X_{ii'} \ . \tag{12}$$

$$J^+(S, X) = \sum_{i,i'} \left(S_{ii'} - \frac{1}{2} \left(\frac{S^+_{i.}}{N^+_{i.}} + \frac{S^+_{i'.}}{N^+_{i'.}} \right) \right) X_{ii'} \ . \tag{13}$$

with, $\forall i = 1, \ldots, N$: $S^+_{i.} = \sum_{i':(o^i,o^{i'}) \in \mathbb{S}^+} S_{ii'}$; $S^+_{..} = \sum_{i,i':(o^i,o^{i'}) \in \mathbb{S}^+} S_{ii'}$; $N^+_{i.} = \#\{o^{i'} \in \mathbb{O} : (o^i, o^{i'}) \in \mathbb{S}^+\}$ and $N^+_{..} = \#\mathbb{S}^+$ (# being the cardinal).

Intuitively, this two-step approach allows to *obtain more compact clusters since pairs of objects that are finally clustered together must have very high similarities compared to central tendency measures based upon the most relevant similarities which are the positive ones and which actually correspond to the nearest neighbors of objects.*

To sum up, we give in Table 1, the parameters μ of the respective clustering functions. This table uses the notations provided in equation (1). Henceforth, we are interested in the following clustering problems: $\max_X \Delta^+(S, X)$ with respect to the constraints in (2), where Δ^+ is one of the three functions in Table 1.

Table 1. Clustering functions and their central tendency measures

Clus. func.	Central tendency measures
$B^+(S,X) = F(S,\mu^{B^+},X)$	$\mu^{B^+}_{ii'} = \frac{S^+_{i.}}{N^+_{i.}} + \frac{S^+_{i'.}}{N^+_{i'.}} - \frac{S^+_{..}}{N^+_{..}}$
$E^+(S,X) = F(S,\mu^{E^+},X)$	$\mu^{E^+}_{ii'} = \frac{S^+_{..}}{N^+_{..}}$
$J^+(S,X) = F(S,\mu^{J^+},X)$	$\mu^{J^+}_{ii'} = \frac{1}{2}\left(\frac{S^+_{i.}}{N^+_{i.}} + \frac{S^+_{i'.}}{N^+_{i'.}}\right)$

3 A Clustering algorithm based on transfer operations

The clustering problems that we have formally presented as integer linear programs, can be optimally solved but for very small data-sets. In this paper, we target large data-sets and our purpose is to introduce a clustering algorithm that allows to rapidly find an approximate solution X to these clustering problems. Regarding the notations, let $U = \{u_1, \ldots, u_q\}$ be the partition solution represented by the relational matrix X (q being the number of clusters of U). Hence, $u_l; l = 1, \ldots, q$, represents the set of objects that are within this cluster. As we have mentioned in section 1, the core of this heuristic is not new. Actually it can be seen as a prototype-based leader algorithm but employing similarities instead of euclidean distances. The basic ideas of this algorithm were already suggested in [9, 6, 2] for example. After introducing the mechanism of such an algorithm, we recall its particular weakness and propose a simple solution to overcome it.

3.1 The basic algorithm and its scalable version

Given U, a current partition of the objects, we want to find out if the transfer of an object o^i to any other clusters of U distinct from its current cluster, can increase the clustering function value. To this end, we introduce the following contribution measures. In the equations below, u_i symbolically represents the current cluster of o^i; u_l an existing cluster but different from u_i and $\{o^i\}$ the singleton constituted of o^i.

$$cont_{u_i}(o^i) = 2 \sum_{i':o^{i'} \in u_i} (S_{ii'} - \mu_{ii'}) - (S_{ii} - \mu_{ii}) \; . \tag{14}$$

$$cont_{u_l}(o^i) = 2 \sum_{i':o^{i'} \in u_l} (S_{ii'} - \mu_{ii'}) + (S_{ii} - \mu_{ii}) \quad \forall u_l \neq u_i \; . \tag{15}$$

$$cont_{\{o^i\}}(o^i) = (S_{ii} - \mu_{ii}) \; . \tag{16}$$

With respect to the objective function (1), and all other things being equal, one can observe that the quantities[5] measures given in (14), (15) and (16), allow to

[5] Notice that S and μ are symmetric in our context. In the case of $cont_{u_i}(o^i)$ for example, the non-symmetric formulation would be: $\sum_{i':o^{i'} \in u_i} (S_{ii'} - \mu_{ii'}) + \sum_{i':o^{i'} \in u_i} (S_{i'i} - \mu_{i'i}) - (S_{ii} - \mu_{ii})$.

decide whether object o^i should: stay in its current cluster; be transferred into another cluster; or generate a new cluster.

Regarding (15), let $cont_{u_{l^*}}(o^i)$ be the maximum contribution of o^i to an existing cluster (distinct from its current cluster). Then, in order to maximize the objective function (1), one can observe that the algorithm should:
• create a new cluster if $cont_{\{o^i\}}(o^i)$ is greater than $cont_{u_i}(o^i)$ and $cont_{u_{l^*}}(o^i)$,
• transfer o^i to u_{l^*} if $cont_{u_{l^*}}(o^i)$ is greater than $cont_{u_i}(o^i)$ and $cont_{\{o^i\}}(o^i)$,
• do nothing in all remaining cases.
Given this basic operation, the algorithm processes all objects and continue until a stopping criterion is full-filled. Typically, we fix a maximal number of iterations over all objects (denoted *nbitr* in the following).

Since we improve the clustering function value at each operation, this algorithm converges to a local optimum.

In order to have an efficient implementation of the algorithm, we need to compute the contribution quantities (14) and (15) efficiently. In the following, we start by discussing the computation of the central tendencies $\mu_{ii'}$. In a second time, we present an efficient way for computing the contribution quantities (14) and (15) using prototypes.

According to Table 1, we need to compute the following statistics to determine $\mu_{ii'}$: the $(N \times 1)$ vectors of general terms $S^+_{i.}$ and $N^+_{i.}$ and/or the scalars $S^+_{..}$ and $N^+_{..}$. All these statistics will be referred as the "components of μ". They are computed before applying the clustering heuristic and are considered as inputs of the algorithm. The computation cost of these statistics is $O(N^2 \times P)$. However, we only need one iteration to calculate them. Moreover, we can compute these different vectors incrementally.

Next, if we consider the formulation of the contributions given in (14) and (15) in terms of S, the computation cost of the algorithm is of order $O(N^2 \times P \times nbitr)$. When N is high, this implementation is costly. Hopefully, in our context, *we can reduce the complexity cost of these quantities*. Let us recall that, we are given the feature matrix T of size $(N \times P)$ as input. Furthermore, let us assume that the space dimension is much lower[6] than the number of objects, $P << N$. Then, since $S = T \cdot T'$, we can use the linearity properties of the dot products in order to quickly compute the contributions (14) and (15) by using prototypes. First, one can observe that:

$$\sum_{i':o^{i'} \in u_l} S_{ii'} = \sum_{i':o^{i'} \in u_l} \langle o^i, o^{i'} \rangle = \langle o^i, h^l \rangle \quad \text{where} \quad h^l = \sum_{i':o^{i'} \in u_l} o^{i'} \ . \tag{17}$$

h^l is the non-weighted mean vector of size $(P \times 1)$ representing the cluster u_l. Hence, using $h^l; l = 1, \ldots, q$, as prototypes allows to reduce the computation cost of $\sum_{i':o^{i'} \in u_l} S_{ii'}$ from $O(\#u_l \times P)$ to $O(P)$. Second, the computation of the aggregated central tendencies measures, $\sum_{i':o^{i'} \in u_l} \mu_{ii'}$, can also be reduced by

[6] For high-dimensional space we can assume a pre-processing step that reduces the dimension of the feature space.

Table 2. Clustering functions and aggregated central tendency measures of the contribution of object o^i to a cluster

Clus. func.	Aggregated central tendency measures
$B^+(S,X)$	$\sum_{i':o^{i'}\in u_l} \mu_{ii'}^{B^+} = \#u_l \frac{S_{i.}^+}{N_{i.}^+} + \nu_l - \#u_l \frac{S_{..}^+}{N^{2+}}$
$E^+(S,X)$	$\sum_{i':o^{i'}\in u_l} \mu_{ii'}^{E^+} = \#u_l \frac{S_{..}^+}{N^{2+}}$
$J^+(S,X)$	$\sum_{i':o^{i'}\in u_l} \mu_{ii'}^{J^+} = \frac{1}{2}\left(\#u_l \frac{S_{i.}^+}{N_{i.}^+} + \nu_l\right)$

keeping up to date the vector ν of size $(q \times 1)$ of general term:

$$\nu_l = \sum_{i':o^{i'}\in u_l} \frac{S_{i'.}^+}{N_{i'.}^+} \,. \tag{18}$$

Using ν, we can reduce the computation cost of $\sum_{i':o^{i'}\in u_l} \frac{S_{i'.}^+}{N_{i'.}^+}$, that is involved in the calculation of $\sum_{i':o^{i'}\in u_l} \mu_{ii'}$, from $O(\#u_l)$ to $O(1)$. Accordingly, we give in Table 2, the aggregated central tendency measures of the contribution of o^i to a cluster u_l, for the different clustering functions.

To sum up, if we pre-compute the components of μ and use the prototypes $h^l; l = 1, \ldots, q$, and ν then we can reduce the computation cost of the clustering algorithm to $O(N \times q \times P \times nbitr)$. In the meantime, the memory cost is kept to $O(N \times P)$. *These results are quite satisfying as they make the computation cost and memory cost of such an algorithm comparable to the popular k-means method with respect to the number of objects to be clustered.* We finally give in Algorithm 1, the pseudo-code of the proposed clustering algorithm.

3.2 Setting the scanning order of objects

One important issue related to Algorithm 1 is its dependency regarding the scanning order of the objects to cluster. To tackle this problem we propose to use one of the component of the central tendency μ that we have introduced beforehand. More precisely, *we propose to scan the objects according to the increasing order of* $(N_{1.}^+, \ldots, N_{N.}^+)$. For object o^i, $N_{i.}^+$ represents the number of objects with which it has a positive similarity (assuming centered data). Accordingly, we first process the less positively connected objects. This approach can be seen as a strategy for finding small and stable clusters rapidly. To some extent, it also can be viewed as a way for eliminating noise. Indeed, if we choose the decreasing order, the most positively connected objects will be processed first and they will bring in their clusters noisy objects.

4 Experiments

In this section, we introduce an experimental protocol that aims to compare different clustering techniques. This protocol takes into account four different

Algorithm 1 Transfer based heuristic

Require: $nbitr$ = number of iterations; T = the feature matrix; μ = the central tendency components.
Take the first object o^i as the first element of the first cluster u_1:
$q \leftarrow 1$ where q is the current number of cluster
Update h^1 and ν_1
for $itr = 1$ to $nbitr$ **do**
  for $i = 1$ to N **do**
    if o^i hasn't been assigned a cluster yet **then**
      Set $cont_{u_i}(o^i) \leftarrow -\infty$ and compute $cont_{\{o^i\}}(o^i)$ using (16)
    else
      Compute $cont_{u_i}(o^i)$ using (14), (17), Table 2 and compute $cont_{\{o^i\}}(o^i)$ using (16)
    end if
    for u_l in the set of already constituted clusters **do**
      Compute $cont_{u_l}(o^i)$ using (15), (17), Table 2
    end for
    Find u_{l*} the cluster which has the highest contribution with object o^i
    if $cont_{\{o^i\}}(o^i) > cont_{u_{l*}}(o^i)$ and $cont_{\{o^i\}}(o^i) > cont_{u_i}(o^i)$ **then**
      Create a new cluster $u_{l'}$ whose first element is o^i:
      $q \leftarrow q + 1$
      Update $h^{l'}$ and $\nu_{l'}$ and update h^i and ν_i
    else
      if $cont_{u_{l*}}(o^i) > cont_{\{o^i\}}(o^i)$ and $cont_{u_{l*}}(o^i) > cont_{u_i}(o^i)$ **then**
        Transfer object o^i to cluster u_{l*}:
        Update h^{l*} and ν_{l*} and update h^i and ν_i
        if the cluster u_i where was taken o^i is empty **then**
          $q \leftarrow q - 1$
        end if
      end if
    end if
  end for
end for

evaluation criteria. The latter can be seen as four different "point of views" when ranking the clustering techniques. Therefore, we use an aggregation rule for combining the different rankings into one global ranking. In our experiments, we take the k-means algorithm as the baseline. We compare the results provided by the latter to the ones given by the clustering heuristic defined in Algorithm 1 and associated to the clustering functions mentioned in Table 1. We used fifteen data-sets of the UCI Machine Learning repository [20]. The results that we obtained show improvements over the k-means procedure.

4.1 Experimental protocol

We propose to use four evaluation criteria defined in the literature for assessing the clustering algorithms' results. Usually, only one or two assessment coefficients are used. In this paper, we argue that the more evaluation criteria we use in an experimental protocol, the more robust the conclusions we can draw out from the latter.

We assume that for a given data-set, we have at our disposal, the true label of all objects. Accordingly, we denote by $V = \{v_1, \ldots, v_k\}$ the true partition of the data-set. In that case, $v_m; m = 1, \ldots, k$, is called a class.

Then, the evaluation criteria we propose to use are the following ones: the entropy measure [4], the Jaccard index [21, 22], the adjusted Rand Index [23]

and the Janson-Vegelius coefficient [24, 15, 22]. They are formally defined below where X and C are the respective relational matrices (see subsection 2.1) of U and V:

$$Ent(U,V) = \sum_{l=1}^{q} \frac{\#u_l}{N} \left(\frac{-1}{\log(k)} \left(\sum_{m=1}^{k} \frac{\#(u_l \cap v_m)}{\#u_l} \log(\frac{\#(u_l \cap v_m)}{\#u_l}) \right) \right) .$$

$$Jac(X,C) = \frac{\sum_{i,i'=1}^{N} C_{ii'}X_{ii'} - N}{\sum_{i,i'=1}^{N} (C_{ii'} + X_{ii'} - C_{ii'}X_{ii'}) - N} .$$

$$AR(X,C) = \frac{N^2 \sum_{i,i'=1}^{N} C_{ii'}X_{ii'} - \sum_{i,i'=1}^{N} C_{ii'} \sum_{i,i'=1}^{N} X_{ii'}}{\frac{N^2}{2} \left(\sum_{i,i'=1}^{N} C_{ii'} + \sum_{i,i'=1}^{N} X_{ii'} \right) - \sum_{i,i'=1}^{N} C_{ii'} \sum_{i,i'=1}^{N} X_{ii'}} .$$

$$JV(X,C) = \frac{\sum_{i,i'=1}^{N} \left(C_{ii'} - \frac{1}{k}\right) \left(X_{ii'} - \frac{1}{q}\right)}{\sqrt{\sum_{i,i'=1}^{N} \left(C_{ii'} - \frac{1}{k}\right) \sum_{i,i'=1}^{N} \left(X_{ii'} - \frac{1}{q}\right)}} .$$

Each of these four coefficients allows to rank the different clustering functions. Except for the entropy measure, the higher the score, the better the clustering output U. Since we want to use several data-sets in the experiments, we have as many rankings as pairs in (evaluation criteria $\times$ data-sets). Consequently, we need to combine all these rankings. To this end, we propose to use Borda's rank aggregation rule. In that case, let assume that we need to aggregate r rankings of c clustering techniques' result. Then, for each ranking, Borda's rule consists in scoring the best result with $c-1$, the second one with $c-2$ and so on until the last one which has a null score. Then, the final ranking of the clustering techniques is obtained by summing the r different scores distributions. The best method is the one which has the highest aggregated score.

As for each clustering method, we have to aggregate rankings with respect to two dimensions (evaluation criteria $\times$ data-sets), we apply the previous aggregation method in a two-step framework. Given a data-set, we start by aggregating the ranks provided by the different evaluation criteria. Thus, for each data-set, we obtain a first aggregated ranking of the clustering techniques. Then, we aggregate these consolidated rankings given by each data-set which are now seen as criteria. This second aggregated score distribution provides us with the global ranking that allows to compare the clustering techniques from a global viewpoint.

4.2 Experiments settings

We report in Table 3, the description of the fifteen data-sets of the UCI Machine Learning repository [20], that we included in our experiments. These data-sets concern several domains with distinct numbers of objects, features and classes. *Our aim is to have a global evaluation of the introduced cluster analysis models rather than a specific evaluation restricted to a certain kind of data-sets.* For each data-set we centered[7] and standardized the feature matrix. We also normalized

[7] Following the comments given in subsection 2.3.

Table 3. Data-sets description

Name	iris	sonar	glass	ecoli	liv-dis	ionos	wdbc	synt-cont	veh-silh	yeast	mfeat-fou	img-seg	abalo	pag-blo	land sat
Nb. obj.	150	208	214	336	345	351	569	600	846	1484	2000	2310	4177	5473	6435
Nb. feat.	4	60	10	7	6	34	32	60	18	8	76	18	8	10	36
Nb. clas.	3	2	6	8	2	2	2	6	4	8	10	7	28	5	6

each object's vector in order to have a unit norm. The resulting feature matrix T is the input of all clustering techniques. Concerning the clustering functions we have introduced in this paper, this amounts to take the Gram matrix S as the cosine similarity matrix between centered vectors. Besides, when applying Algorithm 1, we set $nbitr = 10$.

We take the k-means algorithm as the baseline since this technique is pretty closed in spirit to our proposal (transfer operations) and since it was shown that it provides relevant results compared to other clustering techniques [25]. In our experiments, we set k as the number of true classes for the k-means algorithm while keeping free this parameter for Algorithm 1. Moreover, for each data-set, we launched 5 different runs with random seeds for the k-means method and took the mean average of the evaluation criteria measures.

4.3 Experiments results and discussions

We report in Table 4, the experiments results (in %) we obtained for each triple (evaluation measure × data-set × clustering algorithms). For each pair (evaluation measure × data-sets) we put in bold the score of the best method. We also give in the bottom lines of Table 4, the number of clusters found by the different methods. According to Table 4, one general comment we can make is that *the evaluation measures do not necessarily agree about their rankings*. Given a data-set, it happens that one clustering method is ranked first for one evaluation measure and last for another one. Besides, *some assessment measures are quite dependent on the number of clusters found*. This is typically the case for the entropy measure (better for larger number of clusters) and the Jaccard index (better for smaller number of clusters). *These observations justify our approach that supports the use of several evaluation criteria in order to have many "opinions"*. Accordingly, if we apply the experimental protocol we have described in subsection 4.1, we find the following global ranking: $J^+ \succ E^+ \succ B^+ \sim k$-means. As we can see, cluster analysis based on the central tendency deviation principle can outperform the k-means technique. Besides, the Jordan criterion J^+ seems to be the most consensual clustering approach. Interestingly, it is not the one that was most often ranked first. This indicates that the J^+ objective function is quite robust compared to the other approaches and with respect to the wide variety of data-sets we tested. However, it is worth noticing that regarding higher dimensional data-sets such as synt-cont or mfeat-fou, the B^+ function seems to perform better. Regarding the number of clusters found, E^+ gives, on average, the highest number of clusters; next comes B^+ (except for the iris and abalone data-sets); and then J^+. The number of clusters found using our proposals are

Table 4. Results according to the four evaluation criteria and number of clusters found.

Ent	iris	sonar	glass	ecoli	liv-dis	ionos	wdbc	synt-cont	veh-silh	yeast	mfeat-fou	img-seg	abalo	pag-blo	land sat
B^+	35.7	80.5	50.4	26.7	**94.1**	45.1	**17.4**	**17.0**	80.0	53.1	44.6	44.1	61.8	21.5	38.6
E^+	**30.6**	75.5	**43.4**	**23.0**	94.7	**33.6**	18.1	25.8	**78.8**	52.1	**42.5**	49.8	64.5	**16.0**	37.4
J^+	34.1	**72.3**	45.1	23.6	94.6	36.6	18.8	27.3	82.8	52.4	43.8	**41.4**	66.0	17.9	**36.5**
k-m	36.6	97.9	55.4	24.3	98.0	82.1	43.8	26.8	89.4	**50.3**	51.0	43.2	**59.3**	20.1	36.8
Jac															
B^+	36.7	12.4	22.7	32.1	11.8	23.8	28.5	**57.4**	14.4	16.2	**31.6**	36.0	10.7	23.2	**48.7**
E^+	54.3	12.2	**28.1**	**52.8**	13.2	23.9	34.5	43.0	17.1	16.2	26.1	31.2	11.5	24.9	44.1
J^+	42.0	10.8	27.8	48.6	11.4	24.7	30.7	44.5	17.0	16.2	30.1	38.4	**11.6**	23.5	47.3
k-m	**57.1**	**36.4**	27.7	37.6	**44.0**	**42.9**	**73.6**	51.8	**18.3**	**19.4**	27.6	**39.2**	5.5	**39.9**	44.5
AR															
B^+	36.4	**4.1**	22.3	37.2	**1.0**	18.0	26.2	**67.7**	8.0	14.3	**42.6**	44.9	**7.0**	3.6	**57.3**
E^+	57.8	3.0	26.2	**59.6**	0.1	18.5	32.2	51.0	8.8	14.2	35.7	37.5	5.9	7.4	51.8
J^+	45.4	3.9	**26.7**	55.4	0.8	**21.1**	28.5	52.8	**8.9**	14.3	40.7	**47.8**	6.1	5.6	55.5
k-m	**58.7**	1.9	22.1	43.6	-0.6	17.0	**66.8**	61.2	7.9	**18.0**	36.8	47.6	4.9	**10.7**	53.0
JV															
B^+	33.2	**6.3**	24.1	41.1	**1.8**	26.5	37.5	**67.5**	8.2	18.1	**42.3**	44.3	15.7	26.2	**57.0**
E^+	55.0	4.5	**31.8**	**64.5**	0.6	27.3	42.0	46.3	8.3	20.3	34.9	36.2	**18.0**	**40.0**	49.6
J^+	46.5	**6.3**	31.0	60.3	1.5	**31.8**	39.5	50.4	**8.5**	19.6	39.9	47.0	17.3	32.2	54.2
k-m	**58.7**	1.9	27.2	46.6	0.8	17.1	**67.0**	61.6	7.9	**21.5**	36.9	**48.2**	6.2	37.6	53.1
Nb. clus.															
B^+	24	11	7	9	10	13	8	7	9	12	16	11	146	9	10
E^+	7	22	15	15	13	48	21	66	21	19	42	16	69	20	49
J^+	6	17	9	12	11	17	15	14	13	16	27	11	19	12	16
k-m	3	2	6	8	2	2	2	6	4	8	10	7	28	5	6

distinct from the original number of classes. For most of the data-sets, the former is greater than the latter. This suggests that classes can contain several homogeneous subclasses that is interesting to find out in a knowledge discovery context.

5 Conclusion and future work

We have introduced three clustering functions for numerical data based on the central tendency deviation concept. These partitioning criteria can be seen as extensions of maximal association measures that were defined for clustering categorical data. We have presented an algorithm that approximately solves the clustering problems we have proposed. Moreover, we have defined a robust experimental protocol that involves four different assessment measures and an aggregation rule. We tested our clustering functions using fifteen data-sets and have showed that our proposals can perform better than the k-means algorithm.

In our future work, we intend to further analyze the geometrical properties of the clustering functions we have introduced in order to potentially apply them in more specific contexts such as high dimensional data or clusters with different shapes for example.

References

1. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**(3) (1999) 264–323
2. Grabmeier, J., Rudolph, A.: Techniques of cluster algorithms in data mining. Data Min. Knowl. Discov. **6**(4) (2002) 303–360
3. Xu, R., Wunsch, D.I.: Survey of clustering algorithms. IEEE Transactions on Neural Networks **16**(3) (2005) 645–678
4. Zhao, Y., Karypis, G.: Criterion functions for document clustering: Experiments and analysis (2001) Technical Report TR01-40, University of Minnesota.
5. Michaud, P., Marcotorchino, J.F.: Modèles d'optimisation en analyse des données relationnelles. Mathématiques et Sciences Humaines **67** (1979) 7–38
6. Marcotorchino, J.F., Michaud, P.: Heuristic approach of the similarity aggregation problem. Methods of operations research **43** (1981) 395–404
7. Marcotorchino, J.F.: Cross association measures and optimal clustering. In: Proc. of the Computational Statistics conference. (1986) 188–194
8. Wakabayashi, Y.: The complexity of computing medians of relations. Resenhas IME-USP **3** (1998) 323–349
9. Hartigan, J.: Clustering Algorithms. John Wiley and Sons (1975)
10. Goodman, L., Kruskal, W.: Measures of association for cross classification. Journal of The American Statistical Association **49** (1954) 732–764
11. Marcotorchino, J.F.: Utilisation des comparaisons par paires en statistique des contingences partie I (1984) Technical Report F069, IBM.
12. Belson, W.: Matching and prediction on the principle of biological classification. Applied statistics **7** (1959) 65–75
13. Abdallah, H., Saporta, G.: Classification d'un ensemble de variables qualitatives. Revue de statistique appliquée **46** (1998) 5–26
14. Jordan, C.: Les coefficients d'intensité relative de korosy. Revue de la société hongroise de statistique **5** (1927) 332–345
15. Marcotorchino, J.F.: Utilisation des comparaisons par paires en statistique des contingences partie II (1984) Technical Report F071, IBM.
16. Kendall, M.G.: Rank correlation methods. Griffin, Londres (1970)
17. Ah-Pine, J., Marcotorchino, J.F.: Statistical, geometrical and logical independences between categorical variables. In: Proceedings of ASMDA'07. (2007)
18. Lebbah, M., Bennani, Y., Benhadda, H.: Relational analysis for consensus clustering from multiple partitions. In: Proceedings of ICMLA'08. (2008) 218–223
19. Torgerson, W.: Multidimensional scaling : I. theory and method. Psychometrika **17** (1952) 401–419
20. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
21. Jaccard, P.: The distribution of the flora in the alpine zone. The New Phytologist **11** (1912) 37–50
22. Youness, G., Saporta, G.: Some measures of agreement between close partitions. Student **51** (2004) 1–12
23. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2** (1985) 193–218
24. Janson, S., Vegelius, J.: The J-index as a measure of association for nominal scale response agreement. Applied psychological measurement **6** (1982) 111–121
25. Jain, A.K., Topchy, A., Law, M.H.C., Buhmann, J.M.: Landscape of clustering algorithms. In: Proceedings of ICPR'04. (2004) 260–263 Vol.1