REVSTAT – Statistical Journal
Volume 2, Number 2, November 2004

CENTRAL PARTITION FOR A PARTITION-DISTANCE AND STRONG PATTERN GRAPH

Authors: Joaquim F. Pinto da Costa
– Dep. de Matemática Aplicada & LIACC, Universidade do Porto, Portugal (jpcosta@fc.up.pt)

P.R. Rao
– Department of Computer Science and Technology, Goa University, India (pralhaad@rediffmail.com)

Received: January 2003 Revised: October 2004 Accepted: October 2004

Abstract:

- When several clustering algorithms are applied to a dataset E or the same algorithm with different parameters, we get several different partitions of the dataset. In this paper we consider the problem of finding a consensus partition between the set of these partitions. This consensus partition, called *central partition*, minimises the average number of disagreements between all of the partitions and has been considered for instance in [14, 5] in a different context from ours. We consider it in the context of partition-distance defined in [7]. We focus our attention in two particular distance functions between partitions and then do an experimental comparison between the two corresponding central partitions. In addition, by using the concept of strong patterns (maximal subset of elements that are always clustered together in all partitions), we define a new graph where the nodes are the strong patterns. This graph contains essentially the same information as the partition graph corresponding to the set E defined in [7], but is much simpler as the number of strong patterns is expected to be much smaller than the cardinal of E. Then, some properties of this new graph are proved.

Key-Words:

- *clustering; graph algorithms; node cover; combinatorial problems; strong pattern; central partition.*

AMS Subject Classification:

- 62-07, 62H30, 94C15.

1. INTRODUCTION

The concept of similarity between two partitions arises in several applications, such as molecular expression data in computational biology. When several different clustering methods are applied to the same data, or the same algorithm with different parameters, different partitions of the same data are produced. Also, if we have K qualitative variables describing our population, we might want to find a "central variable" which sumarizes these variables. These two problems are the same, because there is a one-to-one correspondence between qualitative variables and partitions. The problem of determining a central partition arises also in the case where the given partitions (qualitative variables) result from measurements at times $t, t+1, ..., t+K-1$ and we want to consider the notion of a moving consensus smoothing the partitions (or qualitative variables) at those times.

According to Barthelemy and Leclerc [2], there are three overlapping approaches that have been used to tackle the consensus problem:

(**i**) the *axiomatic approach*, where a central partition must satisfy some conditions that arise, for instance, from experimental evidence;

(**ii**) the *constructive approach*, where a way to construct the consensus is explicitly given, like the Pareto rule which states that two objects are linked in a consensus partition if and only if they are linked in all the K given partitions;

(**iii**) the *combinatorial optimization problem*, where we have some criterion measuring the *remoteness* (see equation (2.1)) of any partition to the given K partitions and we search for a partition that minimises this remoteness function.

This last approach, which goes back to Régnier [14], is the one we use in this work.

In order to find the best consensus, it becomes necessary to evaluate the closeness of the partitions produced. There are many distances that can be defined between two partitions of a dataset. The partition-distance is one such distance measure. This concept has been defined in [1], although Régnier [14] and Lerman (see p. 51 of [9]) had considered it before. This distance is further studied in [7], in which it is shown that the partition-distance between two partitions on a given set can be computed in polynomial time.

Further in [7], a new class of graphs called partition graphs has been defined. It is proved that the partition-distance between two partitions is equal to the size of the smallest node cover of the corresponding partition graph. By establishing the arrayed layout structure of the partition graph, it is shown in [7], that the partition graph is perfect.

Suppose $K \geq 2$ partitions of a nonempty set E consisting of n elements are given. In this paper, we define the notion of central partition with respect to the partition-distance used in [7]. The concept of central partition has been used in [5] in another context and with respect to a different measure of distance between partitions. The central partition is a partition that represents a consensus between all the initial K partitions obtained by different clustering algorithms or by the same algorithm with different parameters.

The computation of the central partition is hard. Hence, we have used an approximate algorithm (heuristic), described in [5], to compute an approximation to the central partition. In order to do this, we use the concept of strong patterns. A strong pattern is a maximal subset of elements of E that have been always clustered together in all of the K partitions. The heuristic consists in assuming that these elements should also be together in the central partition. In addition, by using this concept of strong patterns, we can define a graph where the nodes are the strong patterns, which contains essentially the same information as the partition graph corresponding to the K partitions, but is much simpler as the number of strong patterns is expected to be much smaller than n. The complexity is therefore dominated by determining the strong patterns.

The main goal of our work is first to make a summary of the works that have been done in the problem of consensus partitions. Then, the distance used in [5] and the partition-distance are compared using graph terminology. An experimental evaluation of the central partitions corresponding to these two distances is also presented. Next, a special graph, the strong pattern graph, is defined and some of its properties are given.

2. RELATED WORK

Suppose that we have K qualitative variables describing our set of objects E. Each such variable defines a partition of the set E. We can associate an equivalence relation on E with each variable: x and y are in the same equivalence class if the values of this variable are the same for x and y. Thus we obtain K equivalence relations on $E: R_1, R_2, ..., R_K$. In 1965 Régnier [14] proposes as a good clustering of E, a partition whose associated equivalence E_p minimises the quantity

$$\sum_{i=1}^{K} \delta(E_p, R_i) \ , \tag{2.1}$$

which is called a remoteness function. $\delta(R, E_p) = |R \cup E_p| - |R \cap E_p| = |R - E_p| + |E_p - R|$ is the number of non ordered pairs of points that are in the same cluster in one partition but not in the other. The partition which minimises equation (2.1) is called central partition by Régnier.

In 1981, Barthelemy and Monjardet [3] use the notion of median in order to unify the treatment of some problems which are based on the minimization of a remoteness function, like for instance aggregation problems in cluster analysis, social choice theory and paired comparisons methods. We will restrict and adapt the presentation of their median procedure to the case of clustering. These authors start by defining the partitions π_α (resp. π_β) to be such that two elements x and y are in the same cluster for this partition iff they are together in the same cluster for at least $K/2+1$ (resp. $K/2+0.5$) of the initial partitions. One can easily see that $\pi_\alpha \leq \pi_\beta$, which means that any cluster of π_α is included in a cluster of π_β. The authors define then the median interval of the K initial partitions to be $[\pi_\alpha, \pi_\beta]$. If K is odd, then $\pi_\alpha = \pi_\beta$ and so there is only one median partition; otherwise, every partition contained in that interval is a median partition. Barthelemy and Monjardet [3] then present some properties of this median procedure and survey some interesting mathematical problems related to the notion of median. In a later paper, Barthelemy and Leclerc [2], concentrate on the problem of finding a consensus partition that summarizes a K-tuple of partitions by using the median procedure. A detailed survey of the median procedure for partitions is given, from the axiomatic and the algorithmic points of view.

William Day [6] describes two models for the enumeration of metrics between partitions, focusing on the complexity of computing these metric distances. By doing so he rediscovers some metrics that already existed in the literature, but discovers some new metrics also. For some of them, there exist efficient algorithms with time complexities ranging from $\mathbf{O}(n)$ to $\mathbf{O}(n^3)$.

Strehl and Ghosh [15] propose three techniques for obtaining high-quality consensus partitions. The first one uses a similarity measure which is based on the given K initial partitions and then reclusters the objects using this new similarity measure. The second technique is based on hypergraph partitioning and the third technique collapses groups of clusters into meta-clusters which then compete for each object to determine the central partition. These authors claim that their techniques have low computational costs and so suggest further to use the three approaches for a given situation and then choose the best solution.

Monti *et al.* [10] use a resampling-based method to find the central (consensus) partition in the context of gene-expression microarray data. This type of data has the particularity of presenting many more variables (genes) than observations, which is a challenge for classical data analysis methods (see for instance [11]). Monti *et al.* [10] call their methodology consensus clustering which provides for a method to represent the consensus across multiple runs of a clustering algorithm and to assess the stability of the discovered clusters. They also provide a visualization tool to inspect and validate the number of clusters, membership and boundaries.

3. TWO DISTANCES BETWEEN PARTITIONS BASED ON THE PARTITION-GRAPH

Let E be a nonempty set consisting of n elements. A cluster of E is a nonempty subset of E. A partition of E is a collection of mutually exclusive clusters of E, whose union is E. Two partitions π and π' of E are identical if and only if every cluster in π is also a cluster in π'.

Given two partitions π and π', the *partition-distance*, $D_p(\pi, \pi')$, between π and π' is the minimum number of elements that must be removed from E such that the two induced partitions (π and π' restricted to the remaining elements) are identical.

In [7] this definition is extended to the case of $K > 2$ partitions. Also in [7], it is written that the partition-distance is equal to the minimum number of elements that must be moved between clusters in π, so that the resulting partition equals π'. This definition had already appeared before in the work of Régnier [14].

Example 3.1. Let $E = \{1, 2, 3, 4, 5, 6\}$. Consider the following partitions, π and π' of E:

$$\pi = \Big\{\{1, 2, 4, 6\},\ \{3, 5\}\Big\}\ , \qquad \pi' = \Big\{\{1, 2, 6\},\ \{3\},\ \{4, 5\}\Big\}\ ;$$

then the partition-distance between π and π' equals two, as the removal of two elements, namely 3 and 4, will make π and π' identical and no single element of E has this property.

Proposition 3.1. *The partition-distance, $D_p(\pi, \pi')$, between π and π' verifies the properties of a distance function.*

Proof: The first three properties are obvious. In fact, (**i**) $D_p(\pi, \pi) = 0$; (**ii**) $D_p(\pi, \pi') = D_p(\pi', \pi)$; (**iii**) $D_p(\pi, \pi') = 0 \ \Rightarrow\ \pi = \pi'$.

As for the triangular inequality, (**iv**) $D_p(\pi, \pi') \leq D_p(\pi, \pi'') + D_p(\pi'', \pi')$, let us start by denoting $D_p(\pi, \pi') = n_1$, $D_p(\pi, \pi'') = n_2$ and $D_p(\pi'', \pi') = n_3$. Suppose that $n_1 > n_2 + n_3$. If we remove n_2 elements from E, the two induced partitions of π and π'' become identical; the same happens between π'' and π' if we remove a certain set of n_3 elements. This means that if we remove at most $n_2 + n_3$ (corresponding to the union of the two previous sets to be removed) elements from E, the three induced partitions of π, π'' and π' become identical. This is absurd since by hypothesis, we need to remove at least n_1 elements, which is more than $n_2 + n_3$, in order to make the two induced partitions of π and π' identical. Therefore, we can not have $n_1 > n_2 + n_3$.

Given two partitions π and π' of the same set E, consider the graph $G(\pi, \pi')$ with one node for each element of the set E; two nodes are adjacent iff they are together in the same cluster of either π or π', but not in both. $G(\pi, \pi')$ is called a *partition graph* (see [7]). A *node-cover* of a graph is a subset of nodes Q such that every edge in the graph is incident with at least one node in Q.

As it is shown in [7], the partition-distance between two partitions π and π' is equal to the size of the smallest node-cover of the graph $G(\pi, \pi')$ (it has not been proved that the smallest node cover is unique). This means that the set of elements that must be removed so that the two induced partitions become identical is one of the smallest node covers. The distance used in [5] has also an interpretation in terms of this graph. For each partition π_l let v_l represent its associated equivalence relation: $v_l(i, i') = 1$ iff the two elements are in the same cluster. Then, the distance used in [5] is

$$D_C(\pi, \pi') = \frac{1}{2} \sum_{i,i' \in E} \left| v(i, i') - w(i, i') \right|$$

where the equivalence relations v and w correspond to the partitions π and π' respectively. It is easy to see that this distance is equal to the number of edges of the partition graph $G(\pi, \pi')$. □

4. THE CENTRAL PARTITION FOR A PARTITION-DISTANCE

In this section we start by defining the concept of strong pattern. Given K partitions of a dataset E, a strong pattern is a maximal subset of elements of E that have been always clustered together in all of the K partitions.

Now, in order to determine the strong patterns, we start by building a matrix R with n rows and K columns, where each column represents a partition. So, for instance, if the first partition has 5 clusters, the first column of R is composed of a sequence of numbers belonging to the set $\{1, 2, 3, 4, 5\}$. Thus, the element R_{ij} of this matrix is the cluster number attributed by partition π_j to the i^{th} observation.

From R we construct a square matrix A, of size n, such that $A_{ii'}$ is equal to the number of times that the objects i and i' are clustered together in the K partitions. The complexity of building the matrix A is therefore $n(n-1) \times K/2$.

Consider now the equivalence relation

$$\forall\, (i, i') \in E \times E\ , \qquad w^K(i, i') = \begin{cases} 1 & \text{if } A_{ii'} = K \\ 0 & \text{otherwise .} \end{cases}$$

The partition of strong patterns corresponds to this equivalence relation. To find this partition we look at the elements of matrix A row by row, starting with the first row. First, to the first element is attributed the first cluster, which we can call cluster 1. Then, in the first row, everytime we find that $A_{1i'} = K$, we put the element i' in cluster 1 also, and we delete the row corresponding to i' from consideration. Then we go to the next row to be considered, and we do the same, this time attributing its elements to cluster 2. We proceed in the same manner until there are no more rows to be considered. The complexity of this step is at most $n(n-1)/2$. Therefore the complexity for determining the strong patterns is $O(n^2K)$.

Suppose we have K partitions of E, $(\pi^1, \pi^2, ..., \pi^K)$. We are going to consider now how to obtain from these K partitions a new partition which best represents a consensus between all of the initial K partitions. We call it Central Partition. First of all, the partition corresponding to the strong patterns represents an unanimous consensus between all the K partitions; nevertheless, it usually cannot be considered as a central partition because it has got too many clusters (strong patterns) and is therefore too refined.

Let us denote by π^* the central partition that we are looking for. We define the central partition as the one that minimises the following criterium:

$$C(\pi^*) = \sum_{k=1}^{K} D_p(\pi^*, \pi^k)$$

where $D_p(\pi^*, \pi^k)$ is the partition-distance between the partitions π^* and π^k, that is, the number of elements that have to be removed so that the two induced partitions become identical. Intuitively, the central partition minimises the average number of disagreements between the K partitions. The problem of finding π^* is NP-hard and so we are going to use an heuristic to find an approximation of it. This heuristic has already been used and justified in [5]; we will adapt it to our context. In [5], the distance between two partitions, $D_C(\pi, \pi')$, is equal to the number of edges of the partition graph $G(\pi, \pi')$ that has been defined in [7]. In our case we use the partition-distance, $D_p(\pi, \pi')$.

Let us denote by S the set of strong patterns and q ($q \ll n$) its cardinality. We define now a square matrix B of size q such that $B_{pp'}$ is the number of times that the strong patterns p and p' are together in all of the K partitions.

Theoretically, the partition corresponding to the strong patterns is associated with an equivalence relation u^K:

$$\forall\,(p, p') \in S \times S\ , \qquad u^K(p, p') = \begin{cases} 1 & \text{if } B_{pp'} = K \\ 0 & \text{otherwise .} \end{cases}$$

In a similar way, other relations u^j, $j = 0, 1, ..., K-1$, can be defined:

$$\forall\,(p, p') \in S \times S\ , \qquad u^j(p, p') = \begin{cases} 1 & \text{if } B_{pp'} \geq j \\ 0 & \text{otherwise .} \end{cases}$$

These relations u^j, are in general not transitive and so cannot represent a partition. Only u^0 and u^K represent partitions. To u^0 is associated the elementary partition, where there is only one cluster; to u^K is associated the partition of strong patterns. For $j = 1, ..., K-1$, u^j does not represent a partition, because it is generally not transitive, and the authors in [5] associate with each u^j an equivalence relation $\overline{u}^j$, which is the transitive closure of u^j. Let Γ^j represent the partition associated with $\overline{u}^j$. Let Γ^0 represent the partition with only one cluster and Γ^K the partition of strong patterns. It is then shown that the partitions $\Gamma^0, \Gamma^1, \Gamma^2, ..., \Gamma^K$ are nested, that is, Γ^j is obtained from Γ^{j+1}, by merging two of its clusters.

The heuristic that is then used in order to find the approximate central partition consists in restraining the search to the partitions Γ^j. Each such partition is composed of clusters of strong patterns. In 1984 Celeux [4] has shown that in practice the approximate central partitions obtained by this heuristic are the same or very close to the exact central partition. That is, the clusters corresponding to both partitions, the exact and the one found using the heuristic, are similar.

Let S be the set of strong patterns and define the distance index

$$d(p, p') = K - B_{pp'} \ , \qquad \forall\, (p, p') \in S \ .$$

Let us now prove that this measure is really a distance index. In fact, (**i**) $d(p,p) = 0$ because $B_{pp} = K$. Next, (**ii**) $d(p,p') = d(p',p)$ because the matrix B is symetric. Now, if (**iii**) $d(p,p') = 0$, we have $B_{pp'} = K$; this only happens if the two strong patterns p and p' are in fact one, that is, $p = p'$.

Using this distance index, we build a matrix of distance indices between the strong patterns. The partitions Γ^j can be obtained in the following manner [5]. Start by building a minimal spanning tree (MST) containing q nodes (the strong patterns) and using the distance index $d(p, p') = K - B_{pp'}$ defined above. The edge joining two adjacent nodes p and p' has weight $d(p, p')$. Now, in order to determine the candidate central partitions, $\Gamma^0, \Gamma^1, ..., \Gamma^K$, we do the following: Γ^0 has just one cluster. Γ^1 is obtained from the MST by removing the edge of maximum weight and writing down the two obtained clusters. We continue by successively removing the edges of maximum weight, obtaining the other candidate central partitions $\Gamma^2, \Gamma^3, ..., \Gamma^K$. Everytime that we find two or more edges with maximum weight, we remove all of these at once. Celeux *et al.* [5] show that the candidate central partitions obtained by this methodology are the same defined above associated with $\overline{u}^j$.

For each candidate central partition, Γ^j, we compute the criterium defined above, that is,

$$C(\Gamma^j) \ = \ \sum_{k=1}^{K} D_p(\Gamma^j, \pi^k) \ ,$$

and we choose the partition which minimises this criterium. So, the central partition obtained is the one which minimises the sum of all the partition-distances

between the central partition and the initial K partitions.

Example 4.1. Let $E = \{1, 2, 3, 4, 5, 6\}$ and consider the following four partitions:

$$\pi^1 = \Big\{\{1,2\},\, \{3,4\},\, \{5\},\, \{6\}\Big\}\,, \qquad \pi^2 = \Big\{\{1,2,4\},\, \{3,5\},\, \{6\}\Big\}\,,$$
$$\pi^3 = \Big\{\{1,2,6\},\, \{3,4\},\, \{5\}\Big\} \quad \text{and} \quad \pi^4 = \Big\{\{1,2,5\},\, \{4,6\},\, \{3\}\Big\}\,.$$

This is a very small example with quite different partitions, but it serves to illustrate the determination of central partition. The strong patterns are therefore the subsets $\{1,2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$.

The symetric matrix B is:

	$\{1,2\}$	$\{3\}$	$\{4\}$	$\{5\}$	$\{6\}$
$\{1,2\}$	4	0	1	1	1
$\{3\}$		4	2	1	0
$\{4\}$			4	0	1
$\{5\}$				4	0
$\{6\}$					4

From B we construct the matrix of distance indices $d(p,p') = K - B_{pp'}$:

	$\{1,2\}$	$\{3\}$	$\{4\}$	$\{5\}$	$\{6\}$
$\{1,2\}$	0	4	3	3	3
$\{3\}$		0	2	3	4
$\{4\}$			0	4	3
$\{5\}$				0	4
$\{6\}$					0

Now, we build the minimal spanning tree (MST) between the strong patterns using for instance Prim's algorithm (see Figure 1).

Then, by starting to remove the edges of maximal weight, we get three candidate central partitions. Whenever two or more edges have maximum weight, we remove all of them at once.

The candidate central partitions are therefore:

$$\Gamma^0 = \Big\{1,\, 2,\, 3,\, 4,\, 5,\, 6\Big\}\,,$$
$$\Gamma^1 = \Big\{\{1,2\},\, \{3,4\},\, \{5\},\, \{6\}\Big\}\,,$$
$$\Gamma^2 = \Big\{\{1,2\},\, \{3\},\, \{4\},\, \{5\},\, \{6\}\Big\}\,.$$

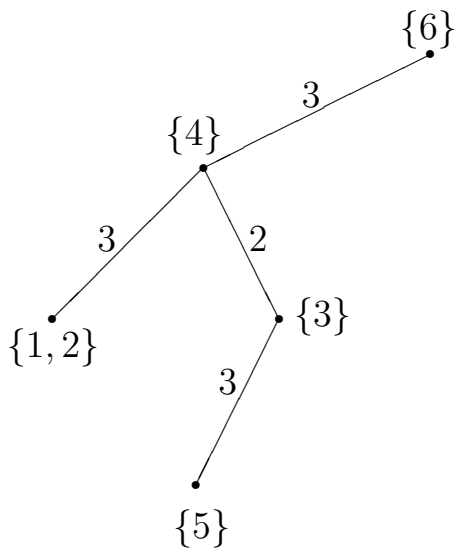

Figure 1: One possible MST between the strong patterns.

Now, in order to choose one of these three candidate central partitions as a central partition, we need to compute the value of $C(\Gamma^j),\ j = 0, 1, 2$; we will do this using the partition-distance defined above:

$$\begin{aligned} C(\Gamma^0) &= 4+3+3+3 = 13\ , \\ C(\Gamma^1) &= 0+2+1+2 = 5\ , \\ C(\Gamma^2) &= 1+2+2+2 = 7\ . \end{aligned}$$

The final partition chosen, that is the one which minimises the criterium $C(\Gamma^j)$, is the partition $\big\{\{1,2\},\{3,4\},\{5\},\{6\}\big\}$, which, in this case, coincides with one of the initial partitions.

5. EXPERIMENTAL COMPARISON BETWEEN THE TWO CENTRAL PARTITIONS

In this section we will show the results of some experiences in order to compare the two central partitions corresponding to the partition-distance used in [7], $D_p(\pi, \pi')$, and the distance used in [5], $D_C(\pi, \pi')$. As was shown above, the partition-distance between two partitions π and π' is equal to the size of the smallest node-cover of the graph $G(\pi, \pi')$ and the distance used in [5] corresponds to the number of edges of $G(\pi, \pi')$. Since the first of these two distances is more complicated to compute, it is of interest to know if the corresponding central partition represents a better consensus between the initial K partitions; otherwise it would be better to use the other distance. To see which of the two central partitions represents a better consensus, we use the Rand index [13], which was

latter corrected for chance in [8]. We start by computing the value of this index between the central partition and each of the initial K partitions and then find the average. The formula for the corrected Rand index between two partitions, one with L clusters and the other with C clusters, is

$$CRI = \frac{\sum_{i=1}^{L}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{L}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{L}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1}\sum_{i=1}^{L}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}} \tag{5.1}$$

where n is the total number of objects, n_{ij} denotes the number of objects that are common to clusters u_i and v_j, $n_{i.}$ and $n_{.j}$ referring respectively to the number of objects in clusters u_i and v_j. This index takes values in the interval $[-1, 1]$ where the value 1 indicates a perfect agreement between the partitions, whereas values close to 0 correspond to cluster agreement found by chance.

We start by generating 19 random partitions of a dataset with 600 elements, with different numbers of clusters in each partition. We do not take into account the structure of the dataset underlying those partitions. In fact, the partitions were obtained by simulating an integer vector of size 600, where each component of this vector contains the cluster number attributed to the i^{th} element, $i = 1, 2, ..., 600$. This is because the two central partitions considered in this work only take into account the labels associated to each element of the dataset; that is, its cluster number, regardless of the structure of the dataset. So, the aim of this experiment is just to see which central partition best agrees with the initial partitions. Our aim is not to see if the initial partitions are a good clustering of any dataset. We suppose we are given K initial partitions and we want just to find the best possible consensus between them.

To generate the random partitions, we have used the code in [12], where it is also explained how the random partitions are generated. Then, we have written a program to compute the two central partitions. Let π^{*1} denote the central partition using the partition-distance $D_p(\pi, \pi')$ and π^{*2} the central partition using the distance $D_C(\pi, \pi')$. Now, we compute the corrected Rand index between each central partition and the initial 19 partitions and find the average. This procedure was repeated six times and the results are given in Table 1.

Table 1: *CRI* values for the two central partitions.

Dataset	Values relating to π^{*1}	Values relating to π^{*2}
1	.450616	.349814
2	.370913	.220207
3	.434782	.353463
4	.401694	.222835
5	.355193	.239976
6	.360283	.278106

As can be seen from these results, the central partition corresponding to the partition-distance presents higher *CRI* values, indicating therefore greater average similarity with the initial 19 partitions.

We have performed another controlled experiment that allows us to compare the two central partitions in the presence of noise. First, we partition a set with 500 elements into 10 clusters at random, as we did above, to obtain the original clustering. We duplicate this clustering 10 times, but, in each of these new 10 labelings, a fraction of the labels is replaced with random labels from a uniform distribution from 1 to 10 (number of clusters). Then, we find the two central partitions, π^{*1} and π^{*2}, for these 10 noisy partitions, and we compare each central partition with the initial partition which has no noise. The results, which are given in Table 2, contain the *CRI* values between π^{*1} and the initial partition, the average *CRI* values between π^{*1} and the given 10 partitions; and the same for π^{*2}.

Table 2: *CRI* values for the two central partitions in the presence of noise.

Fraction of noise	Average *CRI* values for π^{*1}	*CRI* between π^{*1} and initial part.	Average *CRI* values for π^{*2}	*CRI* between π^{*2} and initial part.
10%	.818964	.819189	.818964	.819189
20%	.672516	.667547	.666279	.651882
30%	.556944	.560007	.535590	.546835
40%	.454627	.487534	.398607	.414782
50%	.355307	.387085	.272208	.290658
60%	.274852	.298431	.167627	.174901
70%	.194300	.236023	..060229	.061390
80%	.119703	.149001	.024150	.028592

From these last results, we can see that the central partition corresponding to the partition-distance has higher *CRI* values with the initial partition than the other central partition; except for the case of 10% noise, where the results are the same. It seems also clear that the higher the presence of noise the larger the difference between the *CRI* values for the two central partitions. We can conclude therefore that in the presence of noise, the central partition using the partition-distance $D_p(\pi, \pi')$ is superior to the central partition using the distance $D_C(\pi, \pi')$. On the other hand, we can again see that the average *CRI* values are higher for π^{*1} than for π^{*2}, which confirms the results obtained above.

From this experimental study, we find that the partition-distance is more adequate to find a consensus partition.

6. STRONG PATTERN GRAPH

Having shown experimentally that the partition-distance is more adequate to find a consensus partition, we now present some independent results that were developped during the course of our investigation on the central partition. We start by defining a new graph based on the notion of strong pattern. This new graph contains essentially the same information as the partition-graph, but is much simpler. Then, some properties of this new graph are proved.

Let $U_1, U_2, ..., U_m$ be the strong patterns of K partitions on a set E of size n. *The strong pattern* graph $sp(G)$ consists of m nodes, $U_1, U_2, ..., U_m$ and any two nodes U_q, U_j are adjacent if the strong patterns U_q and U_j are together in the same cluster in at least one partition.

We will now prove that the smallest node-cover of $G(\pi, \pi')$, which is a subset of E, is the union of a set of strong patterns; that is, if an element of E belongs to the smallest node-cover, all of the elements belonging to the same strong pattern belong also to the smallest node-cover.

Proposition 6.1. *Any smallest node-cover of $G(\pi, \pi')$ is composed of a subset of strong patterns.*

Proof: In order to prove this proposition, consider two elements x and y belonging to the same strong pattern. Suppose now that x belongs to a smallest node-cover of $G(\pi, \pi')$. From the results above, x belongs also to a smallest set of elements that have to be removed so that the two induced partitions become identical. We want to prove that y belongs also to the same smallest node-cover; that is, that y has also to be removed. Suppose not; that is, after removing all the elements that have to be removed so that the two induced partitions become identical, y stays. This means that the cluster of the induced partition of π containing y and the cluster of the induced partition of π' containing y are the same. Hence, if we add x to these two clusters, these two clusters remain also the same, because x and y belong to the same strong pattern, that is, are always clustered together; and so x would not have to be removed, which is absurd by hypothesis. Therefore y has also to be removed. □

A *clique* in a graph is a subset of nodes which are pairwise adjacent; let $K(G)$ be the size of the largest clique in graph G. An *independent set* of nodes is a subset of nodes where no two nodes are adjacent; let $I(G)$ be the size of the largest independent set in graph G. If U is a non empty subset of the node set of graph G, then the subgraph H of G induced by U is the graph having the node set U and whose edge set consists of those edges of G incident with two distinct elements of U. The subgraph H is called a *node-induced* subgraph. A graph G is called *perfect* if $K(H) = I(H)$ for every *node-induced* subgraph H of G.

Proposition 6.2. *The strong pattern graph for two partitions of the same set is a perfect graph.*

Proof: The strong pattern graph corresponding to two partitions π^1 and π^2 is itself a partition graph. In fact we can form two partitions of the set of strong patterns: π^1_S is composed of clusters of strong patterns whose individual elements were clustered together in π^1; similarly for π^2_S. The strong pattern graph defined above corresponds to the partition graph for π^1_S and π^2_S. It is proved in [7] that any partition graph is a perfect graph. Therefore, the strong pattern graph, being a partition graph, is a perfect graph. □

7. CONCLUSIONS AND FUTURE WORK

We have considered in this paper the problem of finding a consensus partition (central partition) between a set of partitions corresponding for instance to the results of different clustering algorithms. The distance between partitions is the one defined in [7]. As the determination of the central partition is NP-hard, we have adapted an heuristic [5] which consists in assuming that if two elements are always clustered together in all of the initial partitions, they should also be together in the central partition. We have then shown experimentally that the central partition corresponding to the partition-distance represents a better consensus than the usual central partition, which uses the distance defined in [5]. By defining a strong pattern to be a maximal subset of elements which are always together, we have then defined a strong pattern graph where the nodes correspond to the strong patterns and two nodes are adjacent if the corresponding strong patterns are together in at least one partition. We have then proved that any smallest node-cover of a partition graph is composed of a subset of strong patterns and also that the strong pattern graph is a perfect graph.

As for the future work, we plan to implement a computer program to do some experiments in order to analyse the results of some clustering algorithms. This will serve as a way of summarising the results of several clustering algorithms, specially when we do not know which one is best suited to the particular problem at hand. Even if we do know which clustering algorithm to use, its results usually depend on a set of parameters which are not known. By trying different parameters, we will get different partitions and once again, it makes sense to find the central partition (corresponding to the partition-distance) as the one which minimises the average number of disagreements between the various outputs. We plan also to study more deeply the strong pattern graph which we introduce in this article.

ACKNOWLEDGMENTS

The second author would like to thank the Board of Directors of Fundação Oriente for awarding a Scholarship to undertake studies at LIACC, University of Porto. The encouragment given by Professor Pavel Brazdil is also gratefully acknowledged. The authors also acknowledge the suggestions from the referees.

REFERENCES

[1] Almudevar, A. and Field, C. (1999). Estimation of single generation sibling relationships based on DNA markers, *J. Agricultural, biological and environmental statistics*, **4**, 136–165.

[2] Barthelemy, J.P. and Leclerc, B. (1995). The Median Procedure for Partitions. In "Partitioning Data Sets" (J.J. Cox, P. Hansen and B. Julesz, Eds.), *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, **19**, Amer. Math. Soc., Providence, RI, pp. 3–33.

[3] Barthelemy, J.P. and Monjardet, B. (1981). The median procedure in cluster analysis and social choice theory, *Mathematical Social Sciences*, **1**, 235–267.

[4] Celeux, G. (1984). *Approximation rapide et interprétation d'une partition centrale pour les algorithmes de partitionnement*, Rapport de recherche INRIA n. 352.

[5] Celeux, G.; Diday, E.; Govaert, G.; Lechevalier, Y. and Ralambondrainy, H. (1989). *Classification Automatique Des Données*, Dunod, Paris.

[6] Day, W.H.E. (1981). The complexity of computing metric distances between partitions, *Mathematical Social Sciences*, **1**, 269–287.

[7] Gusfield, D. (2002). Partition-distance: A problem and class of perfect graphs arising in clustering, *Information Processing Letters*, **82**(3), 159–164.

[8] Hubert, L. and Arabie, P. (1985). Comparing Partitions, *Journal of Classification*, **2**, 193–218.

[9] Lerman, I.C. (1981). *Classification et Analyse Ordinale des Données*, Dunod, Paris.

[10] Monti, S.; Tamayo, P.; Mesirov, J. and Golub, T. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data, *Machine Learning*, **52**(1–2), 91–118.

[11] Pinto da Costa, Joaquim F. and Silva, Luis M.A. (2003). *Feature Selection in DNA Microarrays*. In "Actes du Xème Congrès de la Société Francophone de Classification", Neuchatel, Switzerland, 10–12 September.

[12] Rafill, T. http://www.soe.ucsc.edu/∼raff/ac-match/

[13] Rand, W. (1971). Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association*, **66**, 846–850.

[14] Régnier, S. (1965). Sur quelques aspects mathématiques des problèmes de classification automatique, *ICC Bull.*, **4**, 175–191, repr. (1983) *Mathématiques et Sciences Humaines*, **82**, 13–29.

[15] Strehl, A. and Ghosh, J. (2002). Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions, *Journal of Machine Learning Research*, **3**, 583–617.