[1]

Majority graphs of profiles of equivalence relations and complexity of Régnier's problem

Olivier Hudry

Télécom ParisTech, Institut Mines-Télécom
46, rue Barrault, 75634 Paris Cedex 13, France
olivier.hudry@telecom-paristech.fr
http://www.infres.enst.fr/~hudry/

Abstract. A classic problem arising in classification consists in summarizing a collection Π, called a *profile*, of p equivalence relations defined on a finite set X by an equivalence relation E^* at *minimum remoteness* from Π. The remoteness is based on the *symmetric difference distance* and measures the total number of disagreements between E^* and Π, and then E^* is called a *median equivalence relation of* Π. It is usual to summarize Π by its *majority graph*. We study the converse issue. We give a sufficient condition for a graph to be the majority graph of a profile of equivalence relations. We then deduce from this that the computation of E^* is NP-hard when p is large enough.

Keywords: Majority graph, complexity, NP-hardness, symmetric difference distance, median relations, equivalence relations, classification, clustering, Régnier's problem

1 Introduction

A classic problem in classification or in clustering consists in gathering objects in clusters in such a way that objects belonging to a same cluster look like similar while the objects of two distinct clusters look like dissimilar. More precisely, given a finite set $X = \{1, 2, ..., n\}$ of n objects, we consider a collection, called a *profile*, $\Pi = (E_1, E_2, ..., E_p)$ of p equivalence relations (i.e. binary relations which are reflexive, symmetric and transitive) defined on X. Each relation E_k $(1 \leq k \leq p)$ may be interpreted as a criterion gathering the elements of X into clusters (the equivalence classes of E_k) such that the elements of each cluster share the same value according to E_k. For instance, if X contains geometric figures which are coloured, a first criterion may gather the objects with the same geometric shape (triangles, rectangles...), while a second criterion may gather the objects according to their sizes (big, medium, small...), a third criterion may gather them according to their colours (red, green, blue...), and so on. With this respect, Régnier's problem [7] consists in looking for an equivalence relation E^*

* Research supported by the ANR project "Computational Social Choice" ANR-09-BLAN-0305-03.

also defined on X which summarizes Π "as well as possible" (see below for the meaning of "as well as possible"); E^* is then called a *median equivalence relation of* Π (see [2]).

Section 2 shows how to associate a weighted graph to Π: the *majority graph of* Π. In Section 3, we study the converse problem: which weigthed graphs can be the majority graphs of profiles of equivalence relations? When such a profile Π exists, what is the minimum number of equivalence relations required in Π? Thanks to this, we may show (see Section 4) that the computation of a median equivalence relation of Π is an NP-hard problem.

2 From profiles to weighted graphs

To specify what "as well as possible" means, we consider the *symmetric difference distance* δ. This distance is defined between two binary relations R and S defined on X by:

$$\delta(R, S) = |R \Delta S|,$$

where Δ stands for the symmetric difference between sets. We may also state $\delta(R, S)$ as follows:

$$\delta(R, S) = |\{(x, y) \in X^2 \text{ s.t. } [xRy \text{ and not } xSy] \text{ or } [xSy \text{ and not } xRy]\}|,$$

where xRy (respectively xSy) means that x is in relation with y with respect to R (respectively S).

Thus the symmetric difference distance, which owns good axiomatic properties (see [1]), measures the number of disagreements between R and S. From this distance δ, we may define a *remoteness* (see [2]) $\rho(\Pi, E)$ between the profile $\Pi = (E_1, E_2, ..., E_p)$ and any equivalence relation E defined on X:

$$\rho(\Pi, E) = \sum_{k=1}^{p} \delta(E_k, E).$$

Thus $\rho(\Pi, E)$ measures the total number of disagreements between Π and E. Then Régnier's problem [7] consists in computing an equivalence relation E^* which minimizes the remoteness from Π. Such an equivalence relation E^* is called a *median equivalence relation of* Π.

In order to compute $\rho(\Pi, E)$, it is usual to consider the *characteristic matrices* of the relations E_k $(1 \leq k \leq p)$ and of E. Given a relation R defined on X, the *characteristic matrix* of R is the binary matrix $M = (m_{ij})_{(i,j)\in X^2}$ defined by $m_{ij} = 1$ if i and j are in relation according to R and $m_{ij} = 0$ otherwise. Then, if $M^k = (m^k_{ij})_{(i,j)\in X^2}$ denotes the characteristic matrix of E_k and if $M = (m_{ij})_{(i,j)\in X^2}$ denotes the characteristic matrix of E, we easily obtain:

$$\delta(E_k, E) = \sum_{1 \leq i \leq n, 1 \leq j \leq n} |m^k_{ij} - m_{ij}|$$

and, after some computations:

$$\rho(\Pi, E) = C - \sum_{1 \le i \le n, 1 \le j \le n} (2\alpha_{ij} - p) m_{ij},$$

where C is a constant (equal to $\sum_{k=1}^{p} \sum_{1 \le i \le n, 1 \le j \le n} m_{ij}^k$) and where α_{ij} is equal to $\sum_{k=1}^{p} m_{ij}^k$, i.e. to the number of equivalence relations E_k for which i and j are together.

Thanks to this, we may define an undirected graph G which summarizes Π: the *majority graph of* Π. The set of vertices of G is X and all the possible edges belong to G: $G = (X, X^2)$; any edge $\{i, j\}$ of G has a weight $w(i, j)$ equal to $2\alpha_{ij} - p$. Observe that any weight is between $-p$ and p and that its parity is the one of p; moreover, the weight of any loop $\{x, x\}$ is equal to p because we consider reflexive relations. The majority graph of Π utterly summarizes the data characterizing Π. The next section is devoted to the study of the converse problem.

3 From weighted graphs to profiles

From the end of the previous section, we know that the weihgts of a majority graph of a profile of equivalence relations have the same parity. We first consider the case of a graph of which the weights are even, for $n \ge 3$.

Theorem 1. *Let n be an integer with $n \ge 3$ and let $G = (X, X^2)$ be a weighted undirected graph of which the weights fulfil the following properties:*
1. all the weights of G are even (non-positive or non-negative) integers;
2. for i belonging to X, all the weights $w(i, i)$ of the loops $\{i, i\}$ are positive and equal; let p denote this common value of the weights $w(i, i)$;
3. $p \ge \sum_{i<j \text{ with } w(i,j)>0} w(i, j) + (2n - 3) \sum_{i<j \text{ with } w(i,j)<0} |w(i, j)|$.
Then there exists a profile of p equivalence relations with G as its majority graph.

Proof. We give here only the principle of the proof (see [5] for a complete proof). This proof is based on Debord's works on equivalence relations [3]. In order to prove the theorem, we need extra notation. For any integers i and j with $1 \le i < j \le n$, we define two weighted graphs G_{ij}^+ and G_{ij}^- with all the possible edges as follows:
• all the weights of G_{ij}^+ are equal to 0 except for the edge $\{i, j\}$ of which the weight is equal to 2, and for the loops, of which the weights are also equal to 2;
• all the weights of G_{ij}^- are equal to 0 except for the edge $\{i, j\}$ of which the weight is equal to -2, and for the loops, of which the weights are equal to $4n-6$.

The proof is done in three steps:
• Step 1. For any pair of integers i and j with $i < j$, we build a profile Π_{ij}^+ of two equivalence relations such that its majority graph is G_{ij}^+.
• Step 2. For any pair of integers i and j with $i < j$, we build a profile Π_{ij}^- of $4n - 6$ equivalence relations such that its majority graph is G_{ij}^-.
• Step 3. We decompose G thanks to the graphs G_{ij}^+ and G_{ij}^- for $1 \le i < j \le n$ and we apply the first two steps to build a profile with G as its majority graph.

A similar theorem deals with the case when n is odd:

Theorem 2. *Let n be an integer with $n \geq 3$ and let $G = (X, X^2)$ be a weighted undirected graph of which the weights fulfil the following properties:*
1. all the weights of G are odd (positive or negative) integers;
2. for i belonging to X, all the weights $w(i,i)$ of the loops $\{i,i\}$ are positive and equal; let p denote this common value of the weights $w(i,i)$;
3. $p \geq \sum_{i<j:\ w(i,j)>-1} (w(i,j)+1) + (2n-3) \sum_{i<j:\ w(i,j)<-1} |w(i,j)+1| - 1$.
Then there exists a profile of p equivalence relations with G as its majority graph.

4 Complexity of Régnier's problem

From the previous theorems and a result due to M. Krivanek and J. Moravek [6], we obtain the following result about the computation of a median equivalence relation of Π (see [4] for its proof):

Theorem 3. *Given a profile Π of equivalence relations, the computation of a median equivalence relation of Π is an NP-hard problem.*

In the construction used to prove Theorem 3, the number p of equivalence relations involved in the profile Π is rather large with respect to n. Thus we can wonder what happens if p is a constant:

Problem 1. What is the complexity of Régnier's problem if p is assumed to be a constant?

More generally, we can wonder when Régnier's problem becomes NP-hard:

Problem 2. What is the minimum number p of equivalence relations with respect to n so that Régnier's problem is NP-hard?

References

1. Barthélemy, J.-P.: Caractérisations axiomatiques de la distance de la différence symétrique entre des relations binaires. Mathématiques et Sciences humaines 67, 85–113 (1979).
2. Barthélemy, J.-P., Monjardet, B.: The median procedure in cluster analysis and social choice theory. Mathematical Social Sciences 1, 235–267 (1981).
3. Debord, B.: Axiomatisation de procédures d'agrégation de préférences. PhD thesis, University of Grenoble, France (1987).
4. Hudry, O.: NP-hardness of the computation of a median equivalence relation in classification (Régnier's problem). Mathematics and Social Sciences 197 (2012).
5. Hudry, O.: Characterization of the majority matrices of profiles of equivalence relations, submitted to Mathematics and Social Sciences.
6. Krivanek, M., Moravek, J.: NP-hard problems in hierarchical-tree clustering. Acta Informatica 23, 311–323 (1986).
7. Régnier S.: Sur quelques aspects mathématiques des problèmes de classification automatique. I.C.C. Bulletin 4, 175–191 (1965). Reprint: Mathématiques et Sciences humaines 82, 13–29 (1983).