Structured Matrix Days 2019
May 23–24, 2019 – Limoges, France

Characterization of Régnier's matrices in classification

Olivier Hudry[1]

A classic problem in classification or in clustering (for references on classification, see for instance [2] or [3]) 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. *Régnier's problem* [4] consists in looking for an equivalence relation also defined on X which summarizes Π "as well as possible".

To specify what "as well as possible" means, it is usual (see [1]) to 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 measures the number of disagreements between R and S. From this distance δ, we may define a *remoteness* $\rho(\Pi, E)$ between the profile $\Pi = (E_1, E_2, ..., E_p)$ and any equivalence relation E defined on X:

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

This remoteness $\rho(\Pi, E)$ measures the total number of disagreements between Π and E. Seen as a combinatorial optimization problem, Régnier's problem consists in computing an equivalence relation which minimizes the remoteness from Π. An equivalence relation E^* minimizing ρ is called a *median (or cental) equivalence relation* of Π.

In order to compute a median relation, 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:

$$\delta(E_k, E) = \sum_{(i,j)\in X^2} |m^k_{ij} - m_{ij}|$$

and, after some computations:

$$\rho(\Pi, E) = C - \sum_{(i,j)\in X^2} (2\alpha_{ij} - p)m_{ij},$$

where C is a constant 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. With this respect, we may consider that the matrix $\mathcal{R}_\Pi = (2\alpha_{ij} - p)_{(i,j)\in X^2}$, that we shall call the *Régnier's matrix of* Π in the following, utterly summarizes the profile Π.

Note that, for any integers $i \in X$ and $j \in X$, $2\alpha_{ij} - p$ is an integer between $-p$ (this happens if i and j are never gathered by the relations of Π) and p (this happens if i and j are always gathered by the relations of Π; it is the case in particular when i and j are equal, because of the reflexivity of an equivalence relation) and fulfils the equality $2\alpha_{ij} - p = 2\alpha_{ji} - p$ (because of the symmetry of an equivalence relation). Moreover, these coefficients have the same parity, namely the parity of p.

The study of the complexity of Régnier's problem is based on these Régnier's matrices and, more precisely, requires to be able to reconstruct – in polynomial time – a profile from a given matrix, which, in its turn, requires to be able to characterize such matrices.

So, in this communication, we consider the following question: let $\mathcal{R}$ be a matrix, what are the conditions on the entries of $\mathcal{R}$ so that there exists a profile Π with $\mathcal{R} = \mathcal{R}_\Pi$? In other words: what could be a characterization of a Régnier's matrix? We provide such a characterization by proving the following result:

Theorem. A matrix $\mathcal{R}$ is the Régnier's matrix of a profile of p equivalence relations if and only if $\mathcal{R}$ fulfills the following properties:
1. $\mathcal{R}$ is symmetric;
2. the entries of $\mathcal{R}$ are (non-positive or non-negative) integers with the same parity as p;
3. the diagonal entries of $\mathcal{R}$ are equal to p;
4. all the entries of $\mathcal{R}$ are between $-p$ and p.

As the proof of this theorem is constructive (with a polynomial-time complexity), we may then prove that Régnier's problem is difficult to solve (more precisely, that Régnier's problem is NP-hard).

Keywords: Régnier's matrices, Régnier's problem, classification, combinatorial optimization

References

[1] J.-P. BARTHÉLEMY, B. MONJARDET, "The median procedure in cluster analysis and social choice theory, *Mathematical Social Sciences* 1, 1981, 235–267.

[2] G. BROSSIER, "Les éléments fondamentaux de la classification", in *Analyse des données*, G. Govaert (ed.), Hermès Lavoisier, Paris, 2003.

[3] F. BRUCKER, J.-P. BARTHÉLEMY, *Éléments de classification*, Hermès, Paris, 2007.

[4] S. RÉGNIER, "Sur quelques aspects mathématiques des problèmes de classification automatique", *I.C.C. Bulletin* 4, 1965, 175-191. Reprint: *Mathématiques et Sciences humaines* 82, 1983, 13–29.

[1]Télécom ParisTech, 46, rue Barrault, 75634 Paris Cedex 13, France
olivier.hudry@telecom-paristech.fr