CHARACTERIZATION OF THE MAJORITY MATRICES OF PROFILES OF EQUIVALENCE RELATIONS [1]

Olivier HUDRY [2]

RÉSUMÉ – Caractérisation des matrices majoritaires des profils de relations d'équivalence
Dans le domaine de la classification, le problème de Régnier consiste à résumer une collection (appelée un profil) Π *de* p *relations d'équivalence définies sur un même ensemble fini* X *par une relation d'équivalence* E^* *à éloignement minimum de* Π. *L'éloignement considéré est fondé sur la distance de la différence symétrique et mesure le nombre total de désaccords entre* E^* *et* Π *; la relation* E^* *est alors dite relation d'équivalence médiane de* Π. *Il est habituel de résumer* Π *par sa matrice majoritaire. La matrice majoritaire de* Π *est une matrice de dimensions* $n \times n$, *où* n *représente le cardinal de* X, *dont les coefficients sont des entiers compris entre* $-p$ *et* p, *ayant tous la parité de* p *et telle que les coefficients diagonaux sont tous égaux à* p. *Le problème considéré ici est le problème inverse : quelles sont les matrices qui sont matrices majoritaires de profils de relations d'équivalence ? On montre qu'il est toujours possible de construire un profil de relations d'équivalence à partir d'une matrice* A *symétrique, paire ou impaire et dont les termes diagonaux sont tous égaux et suffisamment grands par rapport aux termes non diagonaux.*

MOTS CLÉS – Matrice majoritaire, distance de la différence symétrique, relation d'équivalence, classification, problème de Régnier.

SUMMARY – *In the field of classification, Régnier's problem consists in summarizing a collection (called a profile)* Π *of* p *equivalence relations defined on a same finite set* X *by an equivalence relation* E^* *at minimum remoteness from* Π. *The considered remoteness is based on the symmetric difference distance and measures the total number of disagreements between* E^* *and the equivalence relations of* Π *;* E^* *is then called a median equivalence relation of* Π. *It is usual to summarize* Π *by its so-called majority matrix. The majority matrix of* Π *is a* (n, n)*-matrix, where* n *denotes the cardinality of* X, *in which all the entries are integers between* $-p$ *and* p *and have the same parity as* p, *and such that all the diagonal entries are equal to* p. *We study the converse question : which matrices may be the majority matrix of a profile of equivalence relations ? We show that it is always possible to construct a profile of equivalence relations from any matrix* A *which is symmetric, and even or odd, when the diagonal entries of* A *are equal and are large enough with respect to the non-diagonal entries of* A.

KEYWORDS – Majority matrix, symmetric difference distance, equivalence relations, classification, clustering, Régnier's problem.

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

2. Telecom ParisTech, 46, rue Barrault, 75634 Paris Cedex 13, France, e-mail : Olivier.Hudry@telecom-paristech.fr

1. INTRODUCTION

A classic problem in classification or in clustering (for references on classification, see for instance [Arabie et al., 1996], [Barthélemy and Leclerc, 1995], [Barthélemy and Monjardet, 1981], [Brossier, 2003], [Brucker and Barthélemy, 2007], [Everitt et al., 2011], [Mirkin, 1996], [Romesburg, 2004]) 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 classes (the equivalence classes of E_k) such that the elements of each class 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 [Régnier, 1965] 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, 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 [Barthélemy, 1979]), 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 (see [Barthélemy and Monjardet, 1981]) :

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

Thus $\rho(\Pi, E)$ measures the total number of disagreements between Π and E. Then Régnier's problem [Régnier, 1965] consists in computing an equivalence relation which minimizes the remoteness from 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 $M = (m_{ij})_{(i,j) \in \{1,2,...,n\}^2}$ defined by $m_{ij} = 1$ if i and j are in relation according to R and $m_{ij} = 0$ otherwise.

Then, if $M^k = (m_{ij}^k)_{(i,j)\in\{1,2,\dots,n\}^2}$ denotes the characteristic matrix of E_k and if $M = (m_{ij})_{(i,j)\in\{1,2,\dots,n\}^2}$ denotes the characteristic matrix of E, we easily obtain :

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

As m_{ij}^k and m_{ij} takes binary values, $|m_{ij}^k - m_{ij}|$ is also equal to $(m_{ij}^k - m_{ij})^2$. Then we obtain, since we also have $(m_{ij}^k)^2 = m_{ij}^k$ and $(m_{ij})^2 = m_{ij}$:

$$\delta(E_k, E) = \sum_{1\leq i\leq n, 1\leq j\leq n} (m_{ij}^k - m_{ij})^2 = \sum_{1\leq i\leq n, 1\leq j\leq n} (m_{ij}^k - 2m_{ij}^k m_{ij} + m_{ij}).$$

Similarly, the remoteness $\rho(\Pi, E)$ becomes :

$$\rho(\Pi, E) = \sum_{k=1}^{p} \delta(E_k, E) = \sum_{k=1}^{p} \sum_{1\leq i\leq n, 1\leq j\leq n} (m_{ij}^k - 2m_{ij}^k m_{ij} + m_{ij}),$$

i.e.

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

where C is a constant (equal to $\sum_{k=1}^{p} \sum_{1\leq i\leq n, 1\leq j\leq n} m_{ij}^k$) 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. With this respect, we may consider that the matrix $A_\Pi = (2\alpha_{ij} - p)_{1\leq i\leq n, 1\leq j\leq n}$, that we shall call the *majority matrix of* Π in the following, utterly summarizes the profile Π.

Note that, for any integers i and j with $1 \leq i \leq n$ and $1 \leq j \leq n$, $2\alpha_{ij} - p$ is an integer between $-p$ (this happens if i and j are nether 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.

We may summarize these observations as follows. To be the majority matrix of a profile of p equivalence relations, a matrix A must fulfil the following properties :
1. A is symmetric ;
2. the entries of A are non-positive or non-negative integers with the same parity as p ;
3. the diagonal entries of A are equal to p (and thus are not equal to 0) ;
4. all the entries of A are between $-p$ and p.

In the next section, we study the converse question : if we consider a matrix A, what are the conditions on the entries of A so that there exists a profile Π with $A = A_\Pi$? For this, the following lemma will be useful.

Lemma 1. *Let* Π_1 *and* Π_2 *be two profiles and let* Π *denote the profile obtained by concatening* Π_1 *and* Π_2. *Then we have* $A_\Pi = A_{\Pi_1} + A_{\Pi_2}$.

PROOF. This equality comes from the expression of the entries of a majority matrix. □

2. MATRICES WHICH ARE THE MAJORITY MATRICES OF PROFILES OF EQUIVALENCE RELATIONS

B. Debord claimed to provide a characterisation of the majority matrices of the profiles of equivalence relations in his PhD thesis [Debord, 1987]. Unfortunately, the claimed characterisation given in [Debord, 1987] and its proof are not completely correct. We give below a sufficient condition for a matrix to be the majority matrix of a profile of equivalence relations.

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

Theorem 1. *Let n be an integer with $n \geq 3$ and let $A = (a(i,j))_{(i,j)\in X^2}$ be a matrix fulfilling the following properties :*

1. A is a symmetric matrix ;

2. all the entries of A are even (non-positive or non-negative) integers ;

3. for x belonging to $\{1, 2, ..., n\}$, all the entries $a(x,x)$ are positive and equal ; let p denote this common value of the entries $a(x,x)$;

4. $p \geq \sum_{i<j \text{ with } a(i,j)>0} a(i,j) + (2n-3) \sum_{i<j \text{ with } a(i,j)<0} |a(i,j)|$.

Then there exists a profile Π of p equivalence relations with A as its majority matrix ($A = A_\Pi$).

PROOF. Observe that we have $p > 0$ since the diagonal entries of A are assumed to be positive (and not only non-negative).

In order to prove this theorem, we need extra notation. For any integers i and j with $1 \leq i < j \leq n$, we define two matrices $A^+_{ij} = (a^+_{ij}(x,y))_{(x,y)\in X^2}$ and $A^-_{ij} = (a^-_{ij}(x,y))_{(x,y)\in X^2}$ as follows :

• A^+_{ij} contains only 0's except for $a^+_{ij}(i,j)$ and $a^+_{ij}(j,i)$, of which the values are equal to 2, and for the diagonal entries, which are also equal to 2 ;

• A^-_{ij} contains only 0's except for $a^-_{ij}(i,j)$ and $a^-_{ij}(j,i)$, of which the values are equal to -2, and for the diagonal entries, which 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 matrix is A^+_{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 matrix is A^-_{ij}.

• Step 3. We write A as a linear combination of the identity matrix I and of the matrices A^+_{ij} and A^-_{ij} for $1 \leq i < j \leq n$ and we apply the first two steps to build a profile Π with $A_\Pi = A$.

For this, we consider four kinds of equivalence relations defined as follows :

• the equivalence relation X^2 which contains only one class (which gathers all the elements of X) ;

• the equivalence relation U which contains n classes (each class contains only one element of X) ;

• for i with $1 \leq i \leq n$, the equivalence relation U_i which contains two classes : the first one contains only the element i of X, the other one contains all the other elements of X ;

• for i and j with $1 \leq i \leq n$, $1 \leq j \leq n$ and $i \neq j$, the equivalence relation U_{ij} which contains $n-1$ classes : the first one contains the two elements i and j of X, each other class contains only one element of X (different from i and j).

Step 1. For any pair of integers i and j with $i < j$, consider the profile Π_{ij}^+ which contains the two equivalence relations X^2 and U_{ij}. It is straightforward to check that its majority matrix is the matrix A_{ij}^+. Indeed, i and j are together twice, and for the other pairs $\{x, y\}$ (x and y are not simultaneously equal to i and j), x and y are together only once, except if x and y are equal (for the diagonal entries), in which case they are together twice.

Step 2. As n is greater than or equal to 3, consider, for any pair of integers i and j with $i < j$, the profile Π_{ij}^- which contains the following $4n-6$ equivalence relations : U_i, U_j, U_{ik} for $1 \leq k \leq n$ with $k \neq i$ and $k \neq j$, U_{jk} for $1 \leq k \leq n$ with $k \neq i$ and $k \neq j$, U_{ij}, $2n-5$ copies of X^2. It is straightforward to check that its majority matrix is the matrix A_{ij}^-. Indeed, as there are $4n-6$ equivalence relations in the profile Π_{ij}^-, all the diagonal entries are equal to $4n-6$. Moreover, i and j are together in $2n-4$ equivalence relations (namely, U_{ij} and the $2n-5$ copies of X^2). For $x \notin \{i, j\}$, i and x are together in $2n-3$ equivalence relations (namely, U_j, U_{ix} and the $2n-5$ copies of X^2) ; the same for j and x (by symmetry of the roles played by i and j). For two distinct elements x and y not simultaneously equal to i and j, x and y are also together in $2n-3$ equivalence relations (namely, U_i, U_j and the $2n-5$ copies of X^2).

Step 3. Let $A = (a(i,j))_{(i,j)\in\{1,2,\ldots,n\}^2}$ be a matrix with the properties of the statement of Theorem 1 : all the entries of A are even. Consider the two matrices defined from A by :

$$A^+ = \frac{1}{2} \sum_{i<j:\ a(i,j)>0} a(i,j) A_{ij}^+ \text{ and } A^- = \frac{1}{2} \sum_{i<j:\ a(i,j)<0} |a(i,j)| A_{ij}^-,$$

with the agreement that a matrix is equal to 0 if the associated sum is empty.

Set $q = \sum_{i<j:\ a(i,j)>0} a(i,j) + (2n-3) \sum_{i<j:\ a(i,j)<0} |a(i,j)|$. Observe that A can be written as :

$$A = A^+ + A^- + (p-q)I,$$

where I denotes the identity matrix. Indeed, consider two indices i and j with $i \neq j$; if the (i,j)-entry $a(i,j)$ of A is positive (respectively negative), then we recover this term from the (i,j)-entry of A^+ (respectively A^-) since the (i,j)-entry of A_{ij}^+ (respectively A_{ij}^-), i.e. $a_{ij}^+(i,j)$ (respectively $a_{ij}^-(i,j)$), is equal to 2 (respectively -2) ; conversely, the contribution to the entry $a(i,j)$ of the other matrices involved in the previous sums is equal to 0 ; so the non-diagonal entries of A and of $A^+ + A^-$ are the same. For the diagonal entries of A, the contribution of A^+ is equal to $\sum_{i<j:\ a(i,j)>0} a(i,j)$ since each diagonal entry of A_{ij}^+ is equal to 2, while the contribution of A^- is equal to $(2n-3) \sum_{i<j:\ a(i,j)<0} |a(i,j)|$ since each diagonal entry of A_{ij}^- is equal to $4n-6$. Hence the equality $A = A^+ + A^- + (p-q)I$, in which $p-q$ is even. Moreover, by the assumption 4 of the statement of Theorem 1, $p-q$ is non-negative.

According to the previous steps, for any given integers i and j with $i < j$, A^+_{ij} can be associated to the profile Π^+_{ij} depicted in Step 1 and which contains two equivalence relations ; then A^+ can be associated with a profile Π^+ of $\sum_{i<j:\ a(i,j)>0} a(i,j)$ equivalence relations : this profile is obtained as the concatenation of the $\frac{1}{2}\sum_{i<j:\ a(i,j)>0} a(i,j)$ profiles Π^+_{ij}. Similarly A^-_{ij} can be associated to the profile Π^-_{ij} depicted in Step 2 and with $4n-6$ equivalence relations ; then A^- can be associated with a profile Π^- of $(2n-3)\sum_{i<j:\ a(i,j)<0} |a(i,j)|$ equivalence relations : this profile is obtained as the concatenation of the $\frac{1}{2}\sum_{i<j:\ a(i,j)<0} |a(i,j)|$ profiles Π^-_{ij}. Moreover $2I$ can be considered as the majority matrix of the profile containing two equivalence relations : U and X^2 ; then $(p-q)I$ can be associated to a profile Π_I of $p-q$ equivalence relations : this profile is obtained as $(p-q)/2$ replications of the profile (U, X^2).

Thanks to Lemma 1, the concatenation of the previous three profiles Π^+, Π^- and Π_I shows that A is the majority matrix of a profile of p equivalence relations. □

We now turn to the case of an odd matrix.

Theorem 2. *Let n be an integer with $n \geq 3$ and let $A = (a(i,j))_{(i,j)\in\{1,2,...,n\}^2}$ be a matrix fulfilling the following properties :*

1. A is a symmetric matrix ;

2. all the entries of A are odd ;

3. for x belonging to $\{1, 2, ..., n\}$, all the entries $a(x,x)$ are positive and equal ; let p denote this common value of the entries $a(x,x)$;

4. $p \geq \sum_{i<j:\ a(i,j)>-1} (a(i,j)+1) + (2n-3)\sum_{i<j:\ a(i,j)<-1} |a(i,j)+1| - 1$.

Then there exists a profile Π of p equivalence relations with A as its majority matrix ($A = A_\Pi$).

PROOF. Let $1_{n,n}$ denote the (n,n)-matrix for which all the entries are equal to 1. Note that $1_{n,n}$ is the majority matrix of the equivalence relation X^2. Consider the matrix $B = (b(i,j))_{(i,j)\in X^2}$ defined by $B = A + 1_{n,n}$, i.e. $b(i,j) = a(i,j) + 1$ for any integers i and j. Thanks to the hypotheses of Theorem 2, B fulfils the hypotheses of Theorem 1 : B is symmetric, all its entries are even, all its diagonal entries have the same value, namely $p+1$, with :

$$p+1 \geq \sum_{i<j:\ b(i,j)>0} b(i,j) + (2n-3) \sum_{i<j:\ b(i,j)<0} |b(i,j)|.$$

Thus there exists a profile Π_B of $p+1$ equivalence relations associated with B. The proof of Theorem 1 shows that, among these $p+1$ equivalence relations, at least one is equal to X^2. By removing such an equivalence relation X^2 from Π_B, we obtain, by Lemma 1, a profile Π containing p equivalence relations and of which the majority matrix is A : $A = A_\Pi$. □

To finish, we may pay consideration to the special cases not considered above, i.e. $n = 1$ or $n = 2$.

Proposition 1.

1. If $n = 1$, $A = (a(1,1))$ is the majority matrix of a profile of $a(1,1)$ equivalence relations if and only if $a(1,1)$ is positive.

2. If $n = 2$, $A = (a(i,j))_{(i,j)\in\{1,2\}^2}$ is the majority matrix of a profile of $a(1,1)$ equivalence relations if and only if we have :

** $a(1,1)$ and $a(2,2)$ are positive and equal ;*

** $a(1,2)$ and $a(2,1)$ are equal and between $-a(1,1)$ and $a(1,1)$, with the same parity as $a(1,1)$.*

PROOF. We give only the sketch of the proof.

1. Obvious.

2. We know that the conditions of the statement are necessary. Then it is sufficient to show about to build a profile of $a(1,1)$ equivalence relations admitting A as its majority matrix.

• If $a(1,2)$ is positive, we consider the profile obtained by the concatenation of the profile containing $a(1,2)$ times the equivalence relation X^2 and the profile containing $(a(1,1)-a(1,2))/2$ times the equivalence relations X^2 and U (note that the difference $a(1,1) - a(1,2)$ is even).

• If $a(1,2)$ is negative, we consider the profile obtained by the concatenation of the profile containing $-a(1,2)$ times the equivalence relation U and the profile containing $(a(1,1) + a(1,2))/2$ times the equivalence relations X^2 and U (note that the sum $a(1,1) + a(1,2)$ is even).

• If $a(1,2)$ is equal to 0 (then $a(1,1)$ is even), we consider the profile containing $a(1,1)/2$ times the equivalence relations X^2 and U.

Details are left to the reader. □

3. CONCLUSION

We may summarize the results of the previous sections as follows.

- Necessary condition. If $A = (a(i,j))_{(i,j)\in X^2}$ is the matrix of a profile of p equivalence relations, then : A is symmetric ; all its entries have the parity of p ; all the entries $a(i,i)$ $(1 \leq i \leq n)$ are equal to p ; for i and j between 1 and n, the entry $a(i,j)$ is an integer between $-p$ and p.

- Sufficient condition. For $n \geq 3$, let $A = (a(i,j))_{(i,j)\in X^2}$ be a matrix fulfilling the following properties : A is symmetric ; all its entries have the same parity ; all the entries $a(i,i)$ $(1 \leq i \leq n)$ are positive and equal ; the non-diagonal entries of A are non-positive or non-negative integers of which the absolute value is not too large with respect to the diagonal entries of A (see above for the exact values) ; then there exists a profile of $a(1,1)$ equivalence relations with A as its majority matrix.

The only difference between this necessary condition and this sufficient condition relies on the relationship between the value of the diagonal entries, i.e. the number of equivalence relations of the profile, and the values of the other entries. In other words, if we do not consider the diagonal entries of A, we obtain a full characterization of the majority matrices of a profile of symmetric and transitive relations (it was the characterization claimed by B. Debord [1987]) : if we do not consider the diagonal entries, it is necessary and sufficient that a matrix A is symmetric and is even or odd (i.e. with all its entries sharing the same parity) to be the majoriy matrix of a profile of equivalence relations. Anyway, the diagonal entries are interesting because they give the number of relations of the profile. In the previous constructions, this number is rather large with respect to the entries of A.

Then an interesting question related to this subject is the following :

Open problem 1.
Given an even or odd symmetric matrix A, what is the minimum number $p^*(A)$ so that there exists a profile of $p^*(A)$ equivalence relations with A as its majority matrix ?

Theorems 1 and 2 provide upper bounds for this minimum number $p^*(A)$. It would be interesting to decrease these upper bounds. The last open problem pays attention to the same issue, but from the algorithmic point of view :

Open problem 2.
What is the complexity of the computation of $p^*(A)$?

REFERENCES

ARABIE P., HUBERT L. J., DE SOETE G. (eds) "Clustering and Classification," World Scientific Publishers, Singapore and River Edge NJ (1996).
BARTHÉLEMY, J.-P., *Caractérisations axiomatiques de la distance de la différence symétrique entre des relations binaires*, Mathématiques et Sciences humaines **67** (1979), 85–113.
BARTHÉLEMY J.-P., LECLERC B., *The median procedure for partitions*, in Partitioning data sets, I.J. Cox, P. Hansen, B. Julesz (eds), DIMACS Series in Discrete Mathematics and Theoretical Computer Science 19, Amer. Math. Soc., Providence, RI, 1995, p. 3-34.
BARTHÉLEMY, J.-P., and B. MONJARDET, *The median procedure in cluster analysis and social choice theory*, Mathematical Social Sciences **1** (1981), 235–267.
BROSSIER G., "Les éléments fondamentaux de la classification," in *Analyse des données*, G. Govaert (ed.), Hermès Lavoisier, Paris, 2003.
BRUCKER, F., and J.-P. BARTHÉLEMY, "Éléments de classification," Hermès, Paris (2007).
DEBORD, B., "Axiomatisation de procédures d'agrégation de préférences", PhD thesis, University of Grenoble, France, 1987.
EVERITT B. S., LANDAU S., LEESE M., STAHL D., "Cluster Analysis," John Wiley & Sons, 2011.
MIRKIN B., "Mathematical Classification and Ordination," Kluwer, Dordrecht, 1996.
RÉGNIER S., *Sur quelques aspects mathématiques des problèmes de classification automatique*, I.C.C. Bulletin **4**, p. 175-191, 1965. Reprint : Mathématiques et Sciences humaines **82** (1983), 13–29.
ROMESBURG C., "Cluster Analysis for Researchers," Lulu Press, North Carolina, 2004.