On the Fitting and Consensus of Classification Systems

Bruno Leclerc

CAMS - EHESS
54 bd Raspail
75270 PARIS Cedex 06, France
(e-mail: leclerc@ehess.fr)

Abstract. Classification systems are families of subsets (classes) of a fixed set S that are closed for intersection and contain S and every single element subset of S. The main problem conidered here is that of the consensus of such systems. We first briefly mention results issued from lattice theory. Then, we consider the Adams approach for the consensus of hierarchies and point out its relation with closures, implications (as they appear in relational databases) and nestings. We show that Adams consensus correspond to the research of a particular subdominant nesting (or overhanging) relation, and generalize the corresponding fitting problem.

Keywords: Closure system, Classification system, Implication, Overhanging order, Lattice, Hierarchy.

1 Introduction

Let S be a finite set. We consider here the aggregation of a profile $\mathcal{F}^* = (\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_k)$ of classifications on S into a consensus classification $\mathcal{F} = c(\mathcal{F}*)$. A classification will be here a family of subsets (classes) containing the whole set S and every one-element subset of S (singleton), and closed under intersection. Equivalently, classification systems are the closure systems of the literature that include all the singletons.

There are two main purposes for the research of such a consensus. First, the classification of a set S described by variables of different types. Each qualitative or quantitative variable v induces a partition or a quasi-order on S, which in turn induces a classification system. With such a common formalization for various structures, a set of k variables leads to a profile $\mathcal{F}^*$ of k such systems. The idea is to aggregate the elements of $\mathcal{F}^*$ into a unique system $c(\mathcal{F}*)$ that summarize the profile in some useful sense (see [Domenach and Leclerc, 2004b] for more details).

The other reason is that several consensus problems already studied in the literature are particular cases of the consensus of closure systems. The basic example is provided by hierarchies, where, frequently in a purpose of phylogenetic reconstruction, many works have followed those of [Adams III, 1972] and [Margush and McMorris, 1981] (see the survey [Leclerc, 1998]). Other usual classification models correspond, directly or after straightforward completions, to closure systems. Thus, several other classical consensus problems are also particular cases, with restricted domains or codomains (or both), of the consensus of classification systems. An example is the aggregation of partitions [Régnier, 1965], [Régnier, 1983], [Mirkin, 1975], [Barthélemy and Leclerc, 1995].

2 Classifications and closure systems

Given a finite set S, and its power set $\mathcal{P}(S)$, a *classification system* on S is a family $\mathcal{F} \subseteq \mathcal{P}(S)$ of classes (subsets) of S. A class $C \in \mathcal{F}$ may be a set of elements sharing some common properties, or close to each other in some sense. Then the following conditions, although not always required, may appear as natural ones :

(C1) $S \in \mathcal{F}$;

(C2) $C, C' \in \mathcal{F} \Rightarrow C \cap C' \in \mathcal{F}$;

(C3) for all $s \in S, \{s\} \in C$.

Then, from (C2) and (C3), we have the empty class in $\mathcal{F}$. This property, although not usual, is appropriate to obtain structural coherence. A family $\mathcal{F}$ which satisfies only (C1) and (C2) is a so-called *closure system* (or *Moore family*).

The most usual classification models correspond to such classification systems, sometimes with the addition of some trivial classes. For instance, the addition of the empty class to a hierarchy $\mathcal{H}$, or the addition of S, the empty set and the lacking singletons to a partition provide classification systems. Pyramids (or quasi-hierarchies) and weak hierarchies, in their intersection-closed variants, are further examples.

We find in the literature three notions (among many others) which all are in one-to-one correspondence with closure systems (cf. [Caspard and Monjardet, 2003]).

A *closure operator* φ on S is a mapping on $\mathcal{P}(S)$ satisfying the three properties of *isotony* (for all $A, B \subseteq S$, $A \subseteq B$ implies $\varphi(A) \subseteq \varphi(B)$), *extensivity* (for all $A \subseteq S$, $A \subseteq \varphi(A)$) and *idempotence* (for all $A \subseteq S$, $\varphi(\varphi(A)) = \varphi(A)$). The elements of the image $\mathcal{F}_\varphi = \varphi(\mathcal{P}(S))$ of $\mathcal{P}(S)$ by φ are the *closed* (by φ) *subsets* of S, and $\mathcal{F}_\varphi$ is a closure system on S. Conversely, a closure operator $\varphi_\mathcal{F}$ on S is associated to any closure system $\mathcal{F}$ on S by $\varphi_\mathcal{F}(A) = \bigcap\{F \in \mathcal{F} : A \subseteq F\}$ (i.e., from (C1) and (C2), the smallest class of $\mathcal{F}$ containing A exists and is $\varphi_\mathcal{F}(A)$.

A *complete implication system* on S, denoted by I, $\rightarrow_I$ or simply $\rightarrow$, is a binary relation on $\mathcal{P}(S)$ satisfying, for all $A, B, C, D \subseteq S$:

(I1) $B \subseteq A$ implies $A \to B$;
(I2) $A \to B$ and $B \to C$ imply $A \to C$;
(I3) $A \to B$ and $C \to D$ imply $A \cup C \to B \cup D$.

An *overhanging order* (*nesting order* in some contexts) on S is a binary relation on P(S) too, denoted as Œ and satisfying, for all $A, B, C \subseteq S$:

(O1) A Œ B implies $A \subset B$;
(O2) $A \subset B \subset C$ implies A Œ $C \iff [A$ Œ B or B Œ $C]$;
(O3) A Œ $A \cup B$ implies $A \cap B$ Œ B.

It is not difficult to see that Œ is then a (partial) order on $\mathcal{P}(S)$. The sets of all closure systems, closure operators, complete implication systems and overhanging orders on S are respectively denoted as **M**, **C**, **I** and **O**. They are in one-to-one correspondence to each other. Besides the correspondence recalled above, we give hereunder two further correspondences, the first one due to [Armstrong, 1974], and the second pointed out in [Domenach and Leclerc, 2004]: for all $A, B \subseteq S$,

$$A \to B \iff B \subseteq \varphi(A)$$
$$A \text{ Œ } B \iff A \subset B \text{ and } \varphi(A) \subset \varphi(B)$$

So, in a classification system, $A \to B$ means that every class including the subset A of S also includes B, while A Œ B means that B properly includes A and, moreover, there exists at least one classs including A and not B.

Further conditions correspond to particular classes of systems. For instance, an overhanging order corresponds to a classification system if and only if it satisfies the following condition (OS) below, and to a hierarchy if, moreover, the following condition (OT) replaces (O3) [Adams III, 1986], [Domenach and Leclerc, 2004]: for all $A, B, C \subseteq S, s \in S$,

(OS) $s \notin A$ implies $\emptyset$ Œ $\{s\}$ Œ $A \cup \{s\}$;
(OT) A Œ C and B Œ C imply $A \cup B$ Œ C or $A \cap B = \emptyset$.

3 Consensus in the lattice of closure systems

The sets **M**, **C**, **I** and **O** are naturally ordered: **M** by set inclusion on $\mathcal{P}(\mathcal{P}(S))$, **I** and **O** by set inclusion on $\mathcal{P}(\mathcal{P}(S) \times \mathcal{P}(S)) = \mathcal{P}((\mathcal{P}(S))^2)$, **C** by the poinwise order on mappings: $\varphi \leq \varphi'$ if $\varphi(A) \subseteq \varphi'(A)$ for all $A \subseteq S$. The resulting orderings are either isomorphic or dually isomorphic: if φ, I and Œ (respectively φ', I' and Œ') are, respectively, the closure operator, complete implication system and overhanging order associated to a given closure system $\mathcal{F}$ (respectively to $\mathcal{F}'$), one has $\mathcal{F} \subseteq \mathcal{F}' \iff \varphi' \leq \varphi \iff I' \subseteq I \iff$ Œ $\subseteq$ Œ$'$ (cf. [Caspard and Monjardet, 2003] and [Domenach and Leclerc, 2004b] for the case of overhangings).

The sets **M** and **I** are closed under set intersection in, respectively, $\mathcal{P}(\mathcal{P}(S))$ and $\mathcal{P}((\mathcal{P}(S))^2)$, and the set **O** is closed under set union in

$\mathcal{P}((\mathcal{P}(S))^2)$. The greatest elements of $\mathbf{M}$, $\mathbf{I}$ and $\mathbf{O}$ are, respectively, $\mathcal{P}(S)$, $\mathcal{P}(S))^2$ and $\{(A,B) : A,B \subseteq S, A \subset B\}$, whereas their lowest elements are, respectively, $\{S\}, \{(A,B) : A,B \subseteq S, B \subseteq A\}$ and the empty relation on $\mathcal{P}(S)$. So, $\mathbf{M}$ and $\mathbf{I}$ are themselves closure systems on, respectively, $\mathcal{P}(S)$ and $\mathcal{P}(S))^2)$.

Ordered by inclusion, any closure system $\mathcal{F}$ is a lattice $(\mathcal{F}, \vee, \cap)$, with $(F \vee F' = \varphi(F \cup F')$ for all closed subsets $F, F' \in \mathcal{F}$. The existence of such a lattice structure has important consequences for the consensus problem as described above, that is the aggregation of any profile $\mathcal{F}^* = (\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_k)$ of closure systems into a closure system $\mathcal{F} = c(\mathcal{F}^*)$. Previous results on the consensus in lattice structures may be found, among others, in [Monjardet, 1990], [Barthélemy and M.F., 1991] and [Leclerc, 1994], with significant issues in particular cases like those of hierarchies ([Barthélemy *et al.*, 1986]), partitions ([Barthélemy and Leclerc, 1995]) or orders ([Leclerc, 2003]). Results for the particular case of closure systems are given in [Raderanirina, 2001] and [Monjardet and Raderanirina, 2004].

A *federation* on K is a family $\mathcal{K}$ of subsets of $K = \{1, ..., k\}$ satisfying $[L \in \mathcal{K}, L' \supseteq L] \Rightarrow [L' \in \mathcal{K}]$. We then define a *federation consensus function* $c_{\mathcal{K}}$ associated to the federation $\mathcal{K}$ by $c_{\mathcal{K}}(\mathcal{F}^*) = \bigvee_{L \in \mathcal{K}}(\bigcap_{i \in L} \mathcal{F}_i)$. Especially, K is an *oligarchic consensus function* if $K = \{L \subseteq K : L \supseteq L_0\}$ for a fixed subset L_0 of K.

Another class of consensus functions consists of the so-called *quota rules* $c_q = c_{\mathcal{K}}$, where $\mathcal{K} = \{L \in K : |L| \geq q\}$ for a given number q $(0 \leq q \leq k)$. Equivalently, $c_q(\mathcal{F}^*) = \bigvee\{A \subseteq S : |\{i \in K : A \in \mathcal{F}_i\}| \geq q\}$ is the closure system generated by those classes that are present in at least q of the $\mathcal{F}_i$'s. Especially, for $q = k$, the quota rule is the same as the oligarchie rule obtained with $L_0 = K$.

The above definition of federation consensus functions needs the set K (and, so, the integer k) to be fixed. Such a constraint is easily removed for quota rules by replacing the number q with a proportion (see [Barthélemy and M.F., 1991]). Note also that, if all the closure systems in $\mathcal{F}^*$ are classification systems, then the federation consensus system $c_{\mathcal{K}}(\mathcal{F}^*)$ is is still a classification system, for any federation $\mathcal{K}$. The same remark holds for quota rules.

An axiomatic approach (cf. [Day and McMorris, 2003]) of the consensus problem on $\mathbf{M}$ allowed to characterize oligarchic rules ([Raderanirina, 2001]), whereas a metric approach, based on the symmetric difference metric ∂ on $\mathbf{M}$ defined by $\partial(\mathcal{F}, \mathcal{F}') = |\mathcal{F} \triangle \mathcal{F}'|$ leads to the following result [Leclerc, 1994], where a median of $\mathcal{F}^*$ is a closure system $\mathcal{M} \in \mathbf{M}$ minimizing $\rho(\mathcal{M}, \mathcal{F}^*) = \sum_{1 \leq i \leq k} \partial(\mathcal{M}, \mathcal{F}_i)$.

Theorem. *For any profile $\mathcal{F}^*$of* $\mathbf{M}$*, and any median $\mathcal{M}$of $\mathcal{F}^*$, the inclusion $\mathcal{M} \subseteq c_{k/2}(\mathcal{F}^*)$holds.*

In other terms, any class of a median closure system belongs to at least half of the closure systems of the profile. It is not difficult to see that this result remains valid when considering classification systems.

4 A fitting result based on implications and overhangings

Federation consensus functions $c_{\mathcal{K}}$ take only in account the presence or absence of classes in a qualified part of the elements of a profile. But it has been observed, in the case of hierarchies, that we have there a limitation which can prevent us to recognize common features in the elements of the profile, even evident ones. Moreover, there is a risk that a consensus based on presence of entire classes lacks of interest. For instance, if no untrivial class (other than the empty class, the singletons, and S), appears in at least half of the elements of a profile, the approaches evoked in the previous section lead to a consensus classification system with only the trivial classes, that is providing no information. For reasons of this type, [Adams III, 1986] developed a consensus method on hierarchies based on intersection of classes, and caracterized it in terms of the overhanging orders (called there nestings) associated to the involved hierarchies. The following result is a generalization of an Adams one. It concerns the more general problem of the fitting of an overhanging order to a given binary relation Ξ on $\mathcal{P}(S)$. The only condition on Ξ is: $(A, B) \in \Xi$ implies $A \subset B$.

For the proof of the next results, we need some further definitions on lattices, especially those of closed sets. First, given two closed sets C, C' in a closure system $\mathcal{F}$, C is *covered by* C' (denoted by $C \prec C'$) if, for any $C'' \in \mathcal{F}$, $C \subseteq C'' \subseteq C'$ implies $C'' = C$ or $C'' = C'$. A closed set C is *meet irreducible* if it is covered by a unique closed set C^+ in $\mathcal{F}$. These meet-irreducibles generate the whole closure system $\mathcal{F}$, in the sense that every $C \in \mathcal{F}$ is obtained as an intersection of such elements. Now, the covering relation of the closure system $\mathbf{M}$ is characterized as follows: for $\mathcal{F}, \mathcal{F}' \in \mathbf{M}$, $\mathcal{F} \prec \mathcal{F}'$ if and only if $\mathcal{F} = \mathcal{F}' - \{C\}$ for some meet-irreducible C of $\mathcal{F}'$ (cf. [Caspard and Monjardet, 2003]).

Consider the following two properties of a closure system $\mathcal{F}$ and its overhanging order Œ:

(AΞ1) $\Xi \subseteq$ Œ, (preservation of Ξ)

(AΞ2) for any meet-irreducible C of $\mathcal{F}$, $(C, C^+) \in \Xi$. (qualified overhangings)

Theorem. *Let $\mathcal{F}, \mathcal{F}' \in \mathbf{M}$. If both $\mathcal{F}$and $\mathcal{F}'$satisfy Conditions* (AΞ1)*and* (AΞ2)*, then $\mathcal{F} = \mathcal{F}'$.*

Proof. Observe first that the set S is in both $\mathcal{F}$ and $\mathcal{F}'$. If $\mathcal{F} \neq \mathcal{F}'$, the symmetric difference $\mathcal{F} \triangle \mathcal{F}'$ is not empty. Let C be a maximal class in $\mathcal{F} \triangle \mathcal{F}'$. Then, $C \neq S$ and it may be assumed without loss of generality that C belongs to $\mathcal{F}$ (and, so, C does not belong to $\mathcal{F}'$). If C was not a meet-irreducible element of $\mathcal{F}$, it would be an intersection of meet-irreducibles, all belonging to both $\mathcal{F}$ and $\mathcal{F}'$ and, so, C would belong to $\mathcal{F}'$.

Thus, C is a meet-irreducible, covered by a unique element C^+ of $\mathcal{F}$, with $C^+ \in \mathcal{F}'$. By (AΞ2), $(C, C^+) \in \Xi$ and, by (AΞ1), $C \ \text{Œ}' \ C^+$ (where $\text{Œ}'$ is the overhanging order associated to $\mathcal{F}'$). Set $C' = \varphi'(C)$ (where φ' is the closure operator associated to $\mathcal{F}'$). We have $C \subset C'$, since $C \in \mathcal{F}'$, and $C' \ \text{Œ}' C^+$, since $C' = \varphi'(C) = \varphi'(C') \subset \varphi'(C^+) = C^+$. But, according to the hypotheses, $C \subset C'$ implies $C' \in \mathcal{F}$, with $C \subset C' \subset C^+$, a contradiction with the hypothesis that C^+ covers C in $\mathcal{F}$.

In the particular case where $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_k$ are hierarchies on S, and $\Xi = \bigcap_{1 \leq i \leq k} \text{Œ}_i$ (where, for all $i = 1, ..., k$, Œ_i is the overhanging/nesting order associated with $\mathcal{F}_i$), we find a result implying the caracterization by Adams of his consensus method:

Corollary 1. *With the relation Ξ defined above, the Adams consensus hierarchy is the only closure system satisfying conditions* (AΞ1)*and* (AΞ2).

It is worth noticing that Adams results point out a case where it actually exists an overhanging order Œ satisfying conditions (AΞ1) and (AΞ2). Another case appears in [Semple and Steel, 2000] in the reseach of a "supertree". We exhibit other such cases in a work in preparation (for instance when Ξ is a relation satisfying conditions (O1) and (O2)). We end by the following result, where the solution to (AΞ1) and (AΞ2) appears, when it exists, to be actually an approximation of the given relation Ξ.

Corollary 2. *Let Ξ be a binary relation on $\mathcal{P}(S)$ and* Œ *an overhanging order satisfying conditions* (AΞ1) *and* (AΞ2). *Then, for any overhanging order* $\text{Œ}'$*, the inclusions $\Xi \subseteq \text{Œ}' \subseteq \text{Œ}$ imply* $\text{Œ}' = \text{Œ}$.

Proof. Assume $\Xi \subseteq \text{Œ}' \subset \text{Œ}$. Equivalently, if $\mathcal{F}'$ and $\mathcal{F}$ are the closure systems associated, respectively, to $\text{Œ}'$ and to Œ, there exists a meet irreducible C of $\mathcal{F}$ such that $\mathcal{F}' \subseteq \mathcal{F} - \{C\}$. It follows that $(C, C^+) \notin \text{Œ}'$, whereas, according to (AΞ2), $(C, C^+) \in \Xi$. This is a contradiction with the hypothesis $\Xi \subseteq \text{Œ}'$.

In the talk, we present examples where the data consist of a profile $\mathcal{F}^*$ of classification systems. In particular, profiles of hierarchies or phylogenies are considered. Now the above results prompt us to start from a relation Ξ obtained as another function of the Œ_i's than intersection. We are then able to obtain a consensus classification system which preserve more information from the profile than the Adams one, but is no longer a hierarchy.

References

[Adams III, 1972]E.N. Adams III. Consensus techniques and the comparison of taxonomic trees. *Systematic zoology*, pages 390–397, 1972.

[Adams III, 1986]E.N. Adams III. N-trees as nestings: complexity, similarity and consensus. *Journal of Classification*, pages 299–317, 1986.

[Armstrong, 1974]W.W. Armstrong. Dependency structures of data base relationships. *Information Processing*, pages 580–583, 1974.

[Barthélemy and Leclerc, 1995]J.P. Barthélemy and B. Leclerc. The median procedure for partitions. In Cox I.J., P. Hansen, and B. Julesz, editors, *Partitioning data sets*, pages 3–34, 1995.

[Barthélemy and M.F., 1991]J.P. Barthélemy and Janowitz M.F. A formal theory of consensus. *SIAM Journal on Discrete Mathematics*, pages 305–322, 1991.

[Barthélemy *et al.*, 1986]J.P. Barthélemy, B. Leclerc, and B. Monjardet. On the use of ordered sets in problems of comparison and consensus of classifications. *Journal of Classification*, pages 187–224, 1986.

[Caspard and Monjardet, 2003]N. Caspard and B. Monjardet. The lattices of moore families and closure operators on a finite set: a survey. *Discrete Applied Mathematics*, pages 241–269, 2003.

[Day and McMorris, 2003]W.H.E. Day and F.R. McMorris. *Axiomatic Consensus Theory in Group Choice and Biomathematics*. SIAM, Philadelphia, 2003.

[Domenach and Leclerc, 2004]F. Domenach and B. Leclerc. Closure systems, implicational systems, overhanging relations and the case of hierarchical classification. *Mathematical Social Sciences*, pages 349–366, 2004.

[Domenach and Leclerc, 2004b]F. Domenach and B. Leclerc. Consensus of classification systems, with adams' results revisited. In D. Banks, L. House, F.R. McMorris, P. Arabie, and W. Gaul, editors, *Classification, Clustering and Data Mining Applications*, pages 417–428, 2004b.

[Leclerc, 1994]B. Leclerc. Medians for weight metrics in the covering graphs of semilattices. *Discrete Applied Mathematics*, pages 281–297, 1994.

[Leclerc, 1998]B. Leclerc. Consensus of classifications: the case of trees. In A. Rizzi, M. Vichi, and H.-H. Bock, editors, *Advances in Data Science and Classification*, pages 81–90, 1998.

[Leclerc, 2003]B. Leclerc. The median procedure in the semilattice of orders. *Discrete Applied Mathematics*, pages 285–302, 2003.

[Margush and McMorris, 1981]T. Margush and F.R. McMorris. Consensus n-trees. *Bulletin of Mathematical Biology*, pages 239–244, 1981.

[Mirkin, 1975]B. Mirkin. On the problem of reconciling partitions. In *Quantitative Sociology, International Perspectives on mathematical and Statistical Modelling*, pages 441–449, 1975.

[Monjardet and Raderanirina, 2004]B. Monjardet and V. Raderanirina. Lattices of choice functions and consensus problems. *Social Choice and Welfare*, pages 349–382, 2004.

[Monjardet, 1990]B. Monjardet. Arrowian characterization of latticial federation consensus functions. *Mathematical Social Sciences*, pages 51–71, 1990.

[Raderanirina, 2001]V. Raderanirina. *Treillis et agrégation de familles de Moore et de fonctions de choix, Ph.D. Thesis.* Université Paris 1, Paris, 2001.

[Régnier, 1965]S. Régnier. Sur quelques aspects mathématiques des problèmes de classification automatique. *ICC Bulletin*, pages 175–191, 1965.

[Régnier, 1983]S. Régnier. Sur quelques aspects mathématiques des problèmes de classification automatique (seconde publication). *Mathématiques et Sciences humaines*, pages 13–29, 1983.

[Semple and Steel, 2000]C. Semple and M.A. Steel. A supertree method for rooted trees. *Discrete Applied Mathematics*, pages 147–158, 2000.