Inconsistency Measurement for Improving Logical Formula Clustering

Yakoub Salhi
CRIL-CNRS, Université d'Artois, France
salhi@cril.fr

Abstract

Formal logic can be used as a tool for representing complex and heterogeneous data such as beliefs, knowledge and preferences. This study proposes an approach for defining clustering methods that deal with bases of propositional formulas in classical logic, i.e., methods for dividing formula bases into meaningful groups. We first use a postulate-based approach for introducing an intuitive framework for formula clustering. Then, in order to characterize interesting clustering forms, we introduce additional properties that take into consideration different notions, such us logical consequence, overlapping, and consistent partition. Finally, we describe our approach that shows how the inconsistency measures can be involved in improving the task of formula clustering. The main idea consists in using the measures for quantifying the quality of the inconsistent clusters. In this context, we propose further properties that allow characterizing interesting aspects related to the amount of inconsistency.

1 Introduction

Clustering is one of the main machine learning techniques for data analysis, which is used to divide data into meaningful groups of objects that share similar characteristics. It plays an important role in several domains, including information retrieval, social sciences, and biology (e.g. see [Aggarwal and Reddy, 2013]). The numerous application fields of clustering have resulted in a variety of data types, such as transactions, sequences, texts, graphs, - and consequently a variety of clustering methods (e.g. see [Berkhin, 2006]). In that regard, we can particularly point out symbolic datasets [de Carvalho *et al.*, 2009; de Souza and de Carvalho, 2004; Gowda and Diday, 1994] that are appropriate for dealing with complex objects (e.g. [Billard and Diday, 2012; Bock, 2000; Diday and Esposito, 2003]). We can also mention conceptual clustering proposed in [Michalski, 1980], which is a machine learning task that deals with a set of complex and heterogeneous objects and produces a classification scheme over them. In other words, conceptual clustering is a technique for explaining and summarizing data.

To the best of our knowledge, there is in the literature a unique work on clustering where propositional formulas are used to represent complex and heterogeneous data (beliefs, knowledge, preferences, etc.) [Boudane *et al.*, 2017]. In this work, it is mainly proposed adaptations of existing clustering methods to the case of sets of propositional formulas, in particular the two well-known k-means and hierarchical agglomerative clustering techniques. In fact, in the clustering framework introduced in [Boudane *et al.*, 2017] the propositional logic is only used as a tool for representing sets of objects in a compact manner: a propositional formula represents its Boolean models where each one represents a distinct object. In this context, we aim here at characterizing typical aspects of logical formula clustering that take into account that formal logic is more than a tool for compact representation, in particular investigating how it can be used in presence of inconsistency.

In this work, we introduce a framework for defining formula clustering methods that allow dealing with classical propositional logic as a tool of both representation and reasoning. There exist several important differences between the approach introduced in [Boudane *et al.*, 2017] and our framework. In particular, the definition of clustering methods in our framework is driven by rationality postulates, which allows us to have flexibility and capture different interesting aspects. In addition, the inconsistency measurement is involved in improving the quality of inconsistent clusters.

We first propose a basic postulate-based approach for defining formula clustering methods. The general idea behind our starting rationality postulates consists in preferring consistent clusters to inconsistent ones. For instance, the postulate $\mathsf{Consistency-Preference}$ states that the improvement of any cluster by making it consistent brings about the deterioration of at least one other cluster by making it inconsistent. We then introduce a specific form of formula clustering that satisfies the proposed postulates. It is important to mention that our basic approach makes only a distinction between consistent clusters and inconsistent ones, but no distinction is made between inconsistent clusters.

Then, we introduce our approach that involves the inconsistency measurement in formula clustering, which consists in using measures for quantifying the quality of the inconsistent clusters. Indeed, we present a specific type of formula clustering where we use a property that shows how to improve the inconsistent clusters by considering the amount of conflicts. This property says that this amount in every inconsistent cluster cannot be reduced without increasing the amount of inconsistency in at least one other cluster. We then propose additional properties that allow characterizing interesting aspects related to the amount of inconsistency such as a property for reducing as much as possible the value corresponding to the greatest amount of inconsistency, which in a sense may lead to balance the amount of inconsistency between clusters.

2 Preliminaries

The material in this section consists of a description of classical propositional logic, related notions, and notational conventions.

The considered language of classical propositional logic is inductively defined starting from a countably set of propositional variables, denoted $\mathcal{P}$, and the constant $\perp$ denoting $false$, and using $\wedge$, $\vee$, $\neg$ and $\rightarrow$ as logical connectives. The set of propositional formulas is denoted $\mathcal{F}$. Notationally, we use, possibly primed and/or with subscripts and/or with superscripts, the letters p, q and r to denote the propositional variables and the greek letters ϕ, ψ and χ to denote the propositional formulas. Further, given a formula ϕ (resp. a set of formulas S), we use $Prop(\phi)$ (resp. $Prop(S)$) to denote the set of propositional variables occurring in ϕ (resp. S). For a finite set S, we use $|S|$ to denote its cardinality.

A *Boolean interpretation* $\mathcal{I}$ of a formula ϕ is a function from a set of propositional variables P to $\{0, 1\}$ with $Prop(\phi) \subseteq P$. It is inductively extended to the propositional formulas as usual.

We say that a Boolean interpretation $\mathcal{I}$ of a formula ϕ is a *model* of the latter, written $\mathcal{I} \models \phi$, if we have $\mathcal{I}(\phi) = 1$. Given a set of propositional variables P with $Prop(\phi) \subseteq P$, we use $Mod(\phi, P)$ to denote the set of models of ϕ that are defined over P. A formula is said to be *consistent* if it admits a model. Further, we say that a set of formulas S is *consistent* if its associated formula $\bigwedge_{\phi \in S} \phi$ is consistent, otherwise it is *inconsistent*.

Let S be a finite set of formulas and $\phi \in \mathcal{F}$. We say that S entails ϕ, written $S \vdash \phi$, if and only if for all Boolean interpretation $\mathcal{I}$ defined over $Prop(S \cup \{\phi\})$, if $\mathcal{I} \models \bigwedge_{\phi \in S} \phi$ then $\mathcal{I} \models \phi$. In particular, S is inconsistent if and only if $S \vdash \perp$ holds.

We define an *integrity constraint* as a consistent propositional formula, and a *knowledge base* as a finite set of propositional formulas. We use $\mathcal{K}_{\mathcal{F}}$ to denote the set of knowledge bases.

Definition 1 (MIS)**.** *Given a knowledge base K, a set M of formulas is said to be a* minimal inconsistent subset (MIS) *of K iff $M \subseteq K$, $M \vdash \perp$, and $\forall \phi \in M$, $M \setminus \{\phi\} \nvdash \perp$.*

Definition 2 (MCS)**.** *Given a knowledge base K, a set M of formulas is said to be a* maximal consistent subset (MCS) *of K iff $M \subseteq K$, $M \nvdash \perp$, and $\forall \phi \in K \setminus M$, $M \cup \{\phi\} \vdash \perp$.*

We use $\mathsf{MC}(K)$ and $\mathsf{MI}(K)$ to denote respectively the set of all maximal consistent subsets and that of all minimal inconsistent subsets of K.

Definition 3 (Free Formula)**.** *Given a knowledge base K and a formula ϕ in K, ϕ is said to be* free *in K iff $\phi \notin M$ for every $M \in \mathsf{MI}(K)$.*

Definition 4 (Problematic Formula)**.** *Given a knowledge base K and a formula ϕ in K, ϕ is said to be* problematic *in K iff there exists $M \in \mathsf{MI}(K)$ s.t. $\phi \in M$.*

We use $Free(K)$ and $Pb(K)$ to denote respectively the set of free formulas and that of problematic formulas in K. The set of inconsistent formulas in K is denoted $Inc(K)$.

The following notions are used for defining our approach for formula clustering. Given a knowledge base K, we use 2^K to denote the powerset of K, i.e., the set of its subsets. A subset $D \subseteq 2^K \setminus \{\emptyset\}$ is said to be a *partition* of K if (i) $K = \bigcup_{S \in D} S$ and (ii) $S \cap S' = \emptyset$ for every $S, S' \in D$ with $S \neq S'$. We say that D is an *m-partition* if in addition we have $|D| = m$. Furthermore, given a formula ϕ, a *consistent (m, ϕ)-partition* of K is a subset $D \subseteq 2^{K \cup \{\phi\}}$ where (i) $\phi \in S$ for every $S \in D$, (ii) $\{S \setminus \{\phi\} \mid S \in D\}$ is an m-partition of $(\bigcup_{S \in D} S) \setminus \{\phi\}$, and (iii) $S \nvdash \perp$ for every $S \in D$. We say that D is a *maximal consistent (m, ϕ)-partition* of K if there is no consistent (m, ϕ)-partition D' of K such that $\bigcup_{S \in D} S \subset \bigcup_{S' \in D'} S'$.

3 Motivation

Let us first consider the example of organizing a wedding dinner in order to illustrate the interest of using formula clustering; the aim is to propose a distribution of the guests around the available tables. In this context, we assume that we have m guests and n tables. We associate to each guest $i \in 1..m$ a distinct propositional variable denoted p_i. Each guest i expresses her/his preferences using a propositional formula, denoted ψ_i. For instance, if the guest 1 does not want to be with 2 and she/he accepts to be with the guest 3 if and only if the guest 4 is at the same table, then we can use the following formula $p_1 \wedge \neg p_2 \wedge (p_3 \leftrightarrow p_4)$ to represent her/his preferences. *Clustering* is a data mining task that consists in grouping a set of objects so that the objects in the same cluster are more compatible with each other than those in the other groups (e.g. see [Aggarwal and Reddy, 2013]). Thus, a clustering of the set $K = \{\phi_1, \ldots, \phi_m\}$ in n groups can be seen as a distribution of all the guests around the available tables. In this context, it is interesting to reduce the amount of conflicts in each table (cluster), this explains why we use here the notion of *inconsistency measure*. Moreover, it is appropriate to consider preferences that have to be taken into account in all the clusters. The formula representing such preferences is called here *integrity constraint*. For example, one can use the formula $(\sum_{i=1}^{m} p_i \leq 5) \wedge ((\bigvee_{i \in W} p_i) \wedge (\bigvee_{j \in M} p_j))$, where W represents the set of female guests and M that of male guests, to represent the fact that we can place up to five chairs per table and around every table there is at least one woman and at least one man. It is worth noting that the cardinality constraints $\sum_{i=1}^{m} p_i \leq n$ can be polynomially transformed into propositional formulas (e.g. see [Sinz, 2005; Marques-Silva and Lynce, 2007]).

In the literature, there is a unique work on logical formula clustering [Boudane *et al.*, 2017], which proposes to use classical propositional logic as a tool for representing sets of objects in a compact manner. Indeed, a propositonal formula can have an exponential number of models and it can thus represents an exponential number of objects. In that regard, the previous work proposed essentially adaptations of existing clustering methods. In fact, it is especially proposed adaptations of k-means and hierarchical agglomerative clustering techniques. There are a number of relevant differences between our approach and that proposed in [Boudane *et al.*, 2017], and the major ones can be summarized in the following points:

- Our framework is defined through an approach based on rationality postulates. Differently, the prior framework is defined through a less flexible approach analogous to that used in standard clustering framework by mainly considering similarity measures.
- The use of the notion of inconsistency measure to deal with inconsistent clusters in our framework, which allows it to benefits from the numerous contributions on inconsistency measurement in the literature.
- The use of additional concepts in our framework such as integrity constraint.

4 Formula Clustering

In this section, we introduce our clustering framework. We first propose a basic definition of clustering using a postulate-based approach. Then, we introduce some interesting additional properties to characterize useful clustering forms.

4.1 Clustering and Consistency

The main idea behind the following definition of formula clustering is in the fact that consistent clusters are preferred to inconsistent ones.

Definition 5 (Formula Clustering). *Given a knowledge base K, a positive integer n and an integrity constraint ϕ, a (n, ϕ)-clustering of K is a set $\mathcal{C} \subseteq 2^{K \cup \{\phi\}}$ such that:*

- $\bigcup_{K' \in \mathcal{C}} K' = K \cup \{\phi\}$ *(*Completeness*);*
- $\phi \in \bigcap_{K' \in \mathcal{C}} K'$ *(*Integrity*);*
- $|\mathcal{C}| \leq n$ *(*Upper − Bound*);*
- $\forall K', K'' \in \mathcal{C}$*, if* $K' \neq K''$ *then* $K' \not\subset K''$ *(*Inclusion − Freeness*);*
- *if K admits a partition D s.t. $|D| \leq n$ and $S \cup \{\phi\} \nvdash \bot$ for every $S \in D$, then, $\forall K' \in \mathcal{C}$, $K' \nvdash \bot$ (*Consistent − Partition*);*
- *$\forall K' \in \mathcal{C}$ and $\forall K'' \subset K' \setminus \{\phi\}$ with $K' \vdash \bot$ and $K' \setminus K'' \nvdash \bot$, and $\forall \{K_1, \ldots, K_m\} \subseteq \mathcal{C} \setminus \{K'\}$ with $|K''| \geq m$, there exists no m-partition $\{K''_1, \ldots, K''_m\}$ of K'' s.t. $K_i \cup K''_i \nvdash \bot$ for $i \in 1..m$ (*Consistency − Preference*);*
- *if $\exists K' \in \mathcal{C}$ s.t. $K' \vdash \bot$, then $|\mathcal{C}| = min(|K|, n)$ (*Clusters − Number*).*

In a (n, ϕ)-clustering, n corresponds to the maximum number of clusters and the integrity constraint ϕ is used to represent properties that have to be shared by all the clusters. The postulate Completeness states that all the formulas of the considered knowledge base must occur in the clustering. The postulate Integrity says that the integrity constraint has to be in every cluster. The postulate Upper − Bound expresses that the number of clusters cannot exceed the fixed bound. The postulate Inclusion − Freeness states that there exist no two clusters where one contains the other. the postulates Consistent − Partition and Consistency − Preference are both used to express that consistency is preferred to inconsistency. In particular, Consistency − Preference is illustrated in Figure 1: the improvement of any cluster brings about the deterioration of another. Finally, Clusters − Number says that we have to use the maximum number of clusters in the presence of at least one inconsistent cluster, which can be seen as a manner to tend towards consistency.

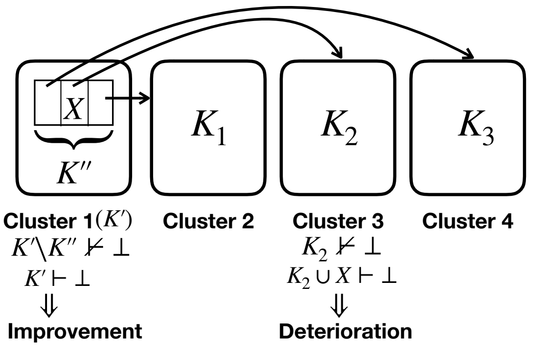

Figure 1: An illustration for Consistency − Preference.

Definition 6. *Let K be a knowledge base, n a strictly positive integer s.t. $|K| \geq n$, D a n-partition of K and $\phi \in \mathcal{F}$. Then, D is said to be ϕ-*minimal *iff, for all n-partition D' of K, $|\{S \in D \mid S \cup \{\phi\} \vdash \bot\}| \leq |\{S' \in D' \mid S' \cup \{\phi\} \vdash \bot\}|$.*

One of the important consequences of the following proposition is the fact that it shows that it is always possible to build a formula clustering from a knowledge base.

Proposition 1. *If D is a ϕ-minimal n-partition of a knowledge base K, then $\{K' \cup \{\phi\} \mid K' \in D\}$ is a (n, ϕ)-clustering of K.*

It is important to note that our definition of formula clustering makes only a distinction between consistent clusters and inconsistent ones, but no distinction is made between the consistent clusters, as well as between inconsistent clusters. In the case of consistent clusters, we are in possession of a strong tool to measure the quality of a cluster, namely the notion of model. In this context, a first idea consists in preferring the augmentation of the number of models. The intuition behind this idea is in the fact that formulas sharing a great number of models represent close objects. To take this point into account, one can replace (Consistency − Preference) with the following stronger property:

$\forall K' \in \mathcal{C}, \forall K'' \subset K' \setminus \{\phi\}$ with $|Mod(K', Prop(K))| < |Mod(K' \setminus K'', Prop(K))|$, and $\forall \{K_1, \ldots, K_m\} \subseteq \mathcal{C} \setminus \{K'\}$ with $|K''| \geq m$, there exists no m-partition $\{K''_1, \ldots, K''_m\}$

of K'' s.t. $Mod(K_i, Prop(K)) = Mod(K_i \cup K''_i, Prop(K))$ for $i \in 1..m$ ($\mathsf{Model-Augmentation}$).

However, this property does not take into account the structures of the clusters and, in particular, the number of formulas involved in each one. In order to clarify this point, consider for instance the $(2, \top)$-clustering $\mathcal{C} = \{K_1 = \{\top, p_1, \ldots, p_n\}, K_2 = \{\top, q_1 \wedge \cdots \wedge q_n\}\}$, where $p_1, \ldots, p_n, q_1, \ldots, q_n$ are pairwise distinct propositional variables. We use V to denote the set of variables occurring in $\mathcal{C}$. We clearly have $|Mod(K_1, V)| = |Mod(K_2, V)|$, but the number of formulas of K_1 can be significantly greater than that of K_2. A possible solution is the use of Jaccard similarity coefficient, as proposed in [Boudane *et al.*, 2017], instead of the number of models in $\mathsf{Model-Augmentation}$:

$$J_K(K') = \frac{|Mod(\bigwedge_{\psi \in K'} \psi, Prop(K))|}{|Mod(\bigvee_{\psi \in K'} \psi, Prop(K))|}$$

Consider again the previous example. We obtain in this case $J_K(K_1) = \frac{2^n}{2^{2n} - 2^n}$ and $J_K(K_2) = 1$, which means that the quality of K_2 can be much better than that of K_1.

The quality of a consistent cluster can be evaluated through the quantity of information. As stated in [Lozinskii, 1994], given a consistent set of formulas S, if a formula ψ is consistent with S, then the addition of ψ to S cannot decrease the amount of information: more formulas means more information. Thus, contrary to the previous measure, more models means less information. One of the well-known information measure is that proposed in [Lozinskii, 1994] and defined as follows:

$$L(K') = |Prop(K')| - log(|Mod(K', Prop(K'))|)$$

The choice between Mod, J_K and L depends strongly on the context. This explains why in the definition of formula clustering we only give advantage to the consistent clusters without any preference among them.

4.2 Additional Properties

We now describe some interesting properties on formula clusterings. Although these properties are not required for every clustering, they allow taking into consideration some important aspects and characterize useful clustering forms.

A (n, ϕ)-clustering $\mathcal{C}$ is said to be *overlap-free* if it satisfies the following property:

- $\forall K', K'' \in \mathcal{C}, K' \cap K'' = \{\phi\}$ ($\mathsf{Overlap-Freeness}$).

For instance, in the example of wedding dinner described previously, it is reasonable to consider $\mathsf{Overlap-Freeness}$ since each formula represents in a sense a distinct person. It is to be noted that the (n, ϕ)-clusterings built from the ϕ-minimal n-partitions, as described in Proposition 1, satisfy $\mathsf{Overlap-Freeness}$. In addition, in the case where overlaps between clusters are allowed, it would be interesting to restrict such overlaps to the formulas obtained through entailment in the case of the consistent clusters by using the following two properties:

- $\forall K' \in \mathcal{C}$ and $\forall \psi \in K$, if $K' \nvdash \bot$ and $K' \vdash \psi$, then $\psi \in K'$ ($\mathsf{Logical-Consequence}$);
- $\forall K', K'' \in \mathcal{C}$ with $K' \neq K''$, $K' \cap K'' \neq \{\phi\}$ iff $K' \nvdash \bot$, $K'' \nvdash \bot$, $(K' \setminus K'') \cup \{\phi\} \vdash \psi$ and $(K'' \setminus K') \cup \{\phi\} \vdash \psi$ for every $\psi \in K' \cap K''$ ($\mathsf{Intersection}$).

The property $\mathsf{Logical-Consequence}$ is mainly used to preserve classical reasoning under consistency by augmenting every consistent cluster with its logical consequences. Moreover, the property $\mathsf{Intersection}$ states that only the logical consequences of consistent clusters and the integrity constraint can occur more than once.

Let us now consider the following properties for a (n, ϕ)-clustering $\mathcal{C}$ of a knowledge base K:

- there exists a maximal consistent (n, ϕ)-partition D of K s.t. $\forall S \in D$ there exists $K' \in \mathcal{C}$ s.t. $S \subseteq K'$ and $S' \cap K' = \{\phi\}$ for every $S' \in D \setminus \{S\}$ ($\mathsf{MCS-BASE}$);
- $\forall K' \in \mathcal{C}$ and $\forall \psi \in Pb(K') \setminus \{\phi\}$, $K'' \cup \{\psi\} \vdash \bot$ and $\psi \in Pb(K'' \cup \{\psi\})$ hold for every $K'' \in \mathcal{C}$ ($\mathsf{PF-Reduction}$);
- $\forall K' \in \mathcal{C}$ with $K' \vdash \bot$, and $\forall \psi \in K' \setminus \{\phi\}$, there is no $K'' \in \mathcal{C}$ s.t. $K'' \cup \{\psi\} \nvdash \bot$ ($\mathsf{NbF-Reduction}$).

The property $\mathsf{MCS-BASE}$ is used to consider the clustering that are built over the maximal consistent partitions. In a sense, the notion of maximal consistent partition can be seen as the natural counterpart in the case of clustering of the notion of maximal consistent subset. The property $\mathsf{PF-Reduction}$ is used to capture the fact that reducing the number of problematic formulas in an inconsistent cluster allows improving it. Let us consider for instance the $(2, \top)$-clustering $\mathcal{C} = \{K' = \{\top, p \wedge q \wedge \neg r, \neg p \wedge r, \neg q\}, K'' = \{\top, p \wedge r, \neg r \wedge \neg p\}\}$. Clearly, $\mathcal{C}$ does not satisfy $\mathsf{PF-Reduction}$ since $\neg q \in Pb(K')$ but $\neg q \notin Pb(K'')$; however, $\mathcal{C}' = \{\{\top, p \wedge q \wedge \neg r, \neg p \wedge r\}, \{\top, p \wedge r, \neg r \wedge \neg p, \neg q\}\}$ satisfies $\mathsf{PF-Reduction}$ (the number of problematic formulas is reduced: $\neg q$ is not a problematic in any cluster). The property $\mathsf{PF-Reduction}$ allows us to partly characterize the fact that a cluster can be improved by reducing the number of MISes. The property $\mathsf{NbF-Reduction}$ states that reducing the number of formulas in an inconsistent cluster allows improving it. Thus, this property captures, in part, the fact that we have to avoid the inconsistent clusters as much as possible. It is worth mentioning that the inconsistent clusters are equally considered by the postulates characterizing clustering in Definition 5. Thus, the properties $\mathsf{PF-Reduction}$ and $\mathsf{NbF-Reduction}$ allow us to partially distinguish inconsistent clusters.

Example 1. *Consider a simple example of a dinner organized by a company for its employees, which is similar to that of wedding dinner described previously. The aim in this context is to propose at most two menus. The preferences of the employees are defined through the following formulas:*

- $\psi_1 = soup \wedge fish;$
- $\psi_2 = fish \wedge ice\_cream;$
- $\psi_3 = \neg meat \wedge cheese;$
- $\psi_4 = salad \wedge meat \wedge cheese;$ *and*
- $\psi_5 = \neg soup \wedge meat \wedge ice\_cream.$

The integrity constraint corresponds to the fact that every menu contains at most one starter, at most one course, and at most one dessert, which corresponds to the formula $\phi = (\sum_{s \in S} s \leq 1) \wedge (\sum_{c \in C} c \leq 1) \wedge (\sum_{d \in D} d \leq 1)$. Note that each conjunct of ϕ corresponds to an instance of the well-known at-most-one constraint that can be linearly encoded as a propositional formula (e.g. see [Sinz, 2005]).

One can easily see that the set $K = \{\psi_1, \psi_2, \psi_3, \psi_4, \psi_5\}$ cannot be partitioned into two sets S_1 and S_2 such that $S_1 \cup \{\phi\} \nvdash \bot$ and $S_2 \cup \{\phi\} \nvdash \bot$. Thus, every $(2, \phi)$-clustering of K contains an inconsistent cluster. For example, the following sets are $(2, \phi)$-clusterings of K: $\mathcal{C}_1 = \{\{\phi, \psi_1, \psi_2\}, \{\phi, \psi_3, \psi_4, \psi_5\}\}$ and $\mathcal{C}_2 = \{\{\phi, \psi_1, \psi_2, \psi_3\}, \{\phi, \psi_4, \psi_5\}\}$. The first cluster in $\mathcal{C}_1$ is consistent and represents the menu soup, $fish$, and ice_cream; the second cluster is inconsistent but any of its formulas does not reject $salad$ and $cheese$, and one can also say that $meat$ is more accepted than $fish$ since $meat$ is a logical consequence of both ψ_4 and ψ_5. Thus, using $\mathcal{C}_1$ one can propose the two menus $m_1 = \{soup, fish, ice\_crem\}$ and $m_2 = \{salad, meat, cheese\}$. Regarding the $(2, \phi)$-clustering $\mathcal{C}_2$, the two clusters are both inconsistent. However, the first cluster contains the formulas that do not reject $soup$ and $fish$, and the second one contains those that do not reject $salad$ and $meat$. Moreover, note that $\mathcal{C}_3 = \{\{\phi, \psi_1, \psi_2, \psi_3, \psi_4, \psi_5\}\}$ is not a $(2, \phi)$-clustering since it does not satisfy Clusters − Number, *and the $(2, \phi)$-clustering $\mathcal{C}_4 = \{\{\phi, \psi_1\}, \{\phi, \psi_2, \psi_3, \psi_4, \psi_5\}\}$ does not satisfy both* PF − Reduction *and* NbF − Reduction. *Furthermore, it is clear that the three $(2, \phi)$-clusterings described in this example satisfy* Overlap − Freeness, Logical − Consequence, *and* Intersection.

5 Improving Clustering by Using Inconsistency Measures

The aim of this section is to introduce our approach for using inconsistency measurement in formula clustering. Indeed, the inconsistency measures are used to improve the inconsistent clusters.

5.1 Inconsistency Measure

In the literature, an inconsistency measure is defined as a function that associates a non negative value to each knowledge base (e.g. [Konieczny *et al.*, 2003; Hunter and Konieczny, 2010; Grant and Hunter, 2013; Jabbour *et al.*, 2016; Thimm, 2016; Ammoura *et al.*, 2017; Bona *et al.*, 2018; Thimm, 2018]), which is used to quantify the amount of conflicts. The different works on inconsistency measures use postulate-based approaches to capture inconsistency-related aspects. In particular, in [Bona *et al.*, 2018], the authors proposed the following formal definition of inconsistency measure.

Definition 7 (Inconsistency Measure)**.** *An* inconsistency measure *is a function $I : \mathcal{K}_{\mathcal{F}} \rightarrow \mathbb{R}^{+}_{\infty}$ that satisfies the two following properties:*

- *$\forall K \in \mathcal{K}_{\mathcal{F}}$, $I(K) = 0$ iff K is consistent (*Consistency*); and*
- *$\forall K, K' \in \mathcal{K}_{\mathcal{F}}$, if $K \subseteq K'$ then $I(K) \leq I(K')$ (*Monotonicity*).*

The set $\mathbb{R}^{+}_{\infty}$ corresponds to the set of positive real number augmented with a greatest element denoted ∞.

The postulate Consistency means that an inconsistency measure must allow distinguishing between consistent and inconsistent knowledge bases. Monotonicity means that the amount of conflicts does not decrease by adding new formulas. Note that these two rationality postulates were first introduced in [Hunter and Konieczny, 2008]. There are several postulates other than Consistency and Monotonicity that have been introduced in the literature to characterize particular aspects related to inconsistency. For instance, one can mention the following interesting postulates:

- $\forall K \in \mathcal{K}_{\mathcal{F}}$ and $\forall \phi \in Prob(K)$, $I(K) > I(K \setminus \{\phi\})$ (Penalty);
- $\forall K \in \mathcal{K}_{\mathcal{F}}$ and $\forall \phi \in Free(K)$, $I(K) = I(K \setminus \{\phi\})$ (Free − Formula);
- $\forall K, K' \in \mathcal{K}_{\mathcal{F}}$ with $K \cap K' = \emptyset$, $I(K \cup K') \geq I(K) + I(K')$ (Super − Additivity).

Let us now describe some simple and intuitive inconsistency measures from the literature:

- $I_M(K) = |\mathsf{MI}(K)|$ ([Hunter and Konieczny, 2008])
- $I_A(K) = |\mathsf{MC}(K)| + |Inc(K)| - 1$ ([Grant and Hunter, 2011])
- $I_{HS}(K) = min\{|S| \mid S \subseteq M \ and \ \forall \phi \in K, \ \exists \mathcal{I} \in S \ s.t. \ \mathcal{I} \models \phi\} - 1$ with $M = \bigcup_{\phi \in K} Mod(\phi, Prop(K))$ and $min\{\} = \infty$ ([Thimm, 2016])

The measure I_M quantifies the amount of inconsistency through minimal inconsistent subsets: more MISes brings more conflicts. The measure I_A is defined in the same way as I_M by using MCSes instead of MISes. The mesure I_{HS} is defined through an explicit use of the Boolean semantics: the amount of inconsistency is related to the minimum number of models that satisfy all the formulas in the considered knowledge base.

5.2 IM-based Formula Clustering

In this section, we describe our approach that shows how the inconsistency measures can be involved in improving the task of formula clustering. The main idea consists in using the measures for quantifying the quality of the inconsistent clusters. In this context, we propose additional properties that allow us to characterize interesting aspects related to the amount of inconsistency.

Definition 8 (IM-based Rational Clustering)**.** *Given a knowledge base K, an inconsistency measure I, a positive integer n and a formula ϕ, a (n, ϕ)-clustering $\mathcal{C}$ is said to be I*-rational *iff it satisfies the following properties:*
*$\forall K' \in \mathcal{C}$ and $\forall K'' \subset K' \setminus \{\phi\}$ with $I(K' \setminus K'') < I(K')$, $\forall \mathcal{C}' \subseteq \mathcal{C} \setminus \{K'\}$ with $|\mathcal{C}| \leq |K''|$ and $m = |\mathcal{C}'|$, and $\forall P$ an m-partition of K'' and $\forall K_0 \in P$, there exists $K^3 \in \mathcal{C}'$ s.t. $I(K^3 \cup K_0) > I(K^3)$ (*Inconsistency − Minimality*).*

In other words, a (n, ϕ)-clustering is I-rational if the amount of inconsistency in every inconsistent cluster cannot be reduced without increasing the amount of inconsistency in at least another cluster. In particular, the inequality $I(K' \setminus K'') < I(K')$ is used to express that reducing the amount of inconsistency in K' (by removing K'' from K') leads to an increase in the amount of inconsistency in at least one other cluster (K^3).

Let us consider again Example 1 and the inconsistency measure I_M described in Section 5.1. The $(2, \phi)$-clustering $\mathcal{C}_1$ is I_M-rational since $I_M(\{\phi, \psi_1, \psi_2\}) = 0$ and $I_M(\{\phi, \psi_1, \psi_2\} \cup \{\chi\}) > 0$ holds for every $\chi \in \{\psi_3, \psi_4, \psi_5\}$. However, the $(2, \phi)$-clustering $\mathcal{C}_4$ is not I_M-rational since $I_M(\{\phi, \psi_1\}) = I_M(\{\phi, \psi_1, \psi_2\}) = 0$ and $I_M(\{\phi, \psi_2, \psi_3, \phi_4, \phi_5\}) = |\{\{\phi, \psi_2, \psi_3\}, \{\phi, \psi_2, \psi_4\}, \{\phi, \psi_2, \psi_5\}, \{\psi_3, \psi_4\}, \{\psi_3, \psi_5\}, \{\phi, \psi_4, \psi_5\}\}| = 6 > I_M(\{\phi, \psi_3, \phi_4, \phi_5\}) = 3$.

It is important to point out that requirements on the considered inconsistency measure allow obtaining interesting properties on the IM-based rational clusterings. For instance, we show in the following proposition that Penalty and Free − Formula allows obtaining the property PF − Reduction provided in Section 4.

Proposition 2. *If I satisfies* Penalty *and* Free − Formula*, then every I-rational (n, ϕ)-clustering satisfies* PF − Reduction.

Proof. Let $\mathcal{C}$ be an I-rational (n, ϕ)-clustering of K, $K' \in \mathcal{C}$ s.t. there exists $\psi \in K' \setminus (Free(K') \cup \{\phi\})$ and $K'' \in \mathcal{C}$ s.t. $\psi \in Free(K'' \cup \{\psi\})$. Knowing that I satisfies Free − Formula, $I(K'' \cup \{\psi\}) = I(K'')$ holds. Further, using the fact that I satisfies Penalty, $I(K') > I(K' \setminus \{\psi\})$ holds. Thus, using the property of Inconsistency − Minimality described in Definition 8, we obtain a contradiction. □

Let us now consider the following IM-based requirement for a (n, ϕ)-clustering $\mathcal{C}$ of a knowledge base K:

- there exists no (n, ϕ)-clustering $\mathcal{C}'$ of K s.t. $max\{I(K') \mid K' \in \mathcal{C}\} > max\{I(K') \mid K' \in \mathcal{C}'\}$ (IM − Max).

The property IM − Max allows capturing the clustering with smallest maximum amount of conflicts. Consider for instance the $(2, \phi)$-clusterings $\mathcal{C}_1$ and $\mathcal{C}_2$ in Example 1. As said previously $\mathcal{C}_1$ is I_M-rational, and one can easily show that $\mathcal{C}_2$ is also I_M-rational. We have $max\{I_M(\{\phi, \psi_1, \psi_2\}) = 0, I_M(\{\phi, \psi_3, \psi_4, \psi_5\}) = 3\} = 3 > max\{I_M(\{\phi, \psi_1, \psi_2, \psi_3\}) = 1, I_M(\{\phi, \psi_4, \psi_5\}) = 1\} = 1$. Thus, the property IM − Max allows us to avoid $\mathcal{C}_1$ in favor of $\mathcal{C}_2$.

The property IM − Max can be generalized by considering an arbitrary fixed function F on the inconsistency values of the clusters :

- There exists no (n, ϕ)-clustering $\mathcal{C}'$ of K s.t. $F\{I(K') \mid K' \in \mathcal{C}\} \lhd F\{I(K') \mid K' \in \mathcal{C}'\}$

where $\lhd \in \{<, >\}$. For instance, one can use $minimum$ with the inequality operator $<$, $average$ with $>$, or $summation$ with $>$.

We now consider some aspects related to cluster sizes. In this context, consider the example of the following I_M-rational $(2, p \wedge q)$-clustering of $K = \{\neg p, \neg q, r_1, \dots, r_{100}\}$: $\mathcal{C} = \{\{p \wedge q, \neg p\}, \{p \wedge q, \neg q, r_1, \dots, r_{100}\}\}$, where $p, q, r_1, \dots, r_{100}$ are pairwise distinct propositional variables. Clearly, the two clusters are both inconsistent and have the same amount of inconsistency w.r.t. I_M, but the size of the second cluster is much greater than the size of the first one. If we consider that each formula in K corresponds to a piece of information related to a distinct agent, it would be appropriate to avoid the clustering $\mathcal{C}$ in favor of I_M-rational $(2, p \wedge q)$-clusterings where the clusters have close sizes, like $\mathcal{C}' = \{\{p \wedge q, \neg p, r_1, \dots, r_{50}\}, \{p \wedge q, \neg q, r_{51}, \dots, r_{100}\}\}$. To this end, we propose the two following requirements, for a (n, ϕ)-clustering $\mathcal{C}$ of a knowledge base K:

- $\forall K' \in \mathcal{C}$ with $K' \vdash \perp$, and $\forall \psi \in K' \setminus \{\phi\}$ and $I(K' \setminus \{\psi\}) = I(K')$, there is no $K'' \in \mathcal{C}$ s.t. $K'' \vdash \perp$, $I(K'') = I(K'' \cup \{\psi\})$ and $|K'| - |K''| > 1$ (Size − Balance);
- $\forall K' \in \mathcal{C}$ with $K' \vdash \perp$, and $\forall \psi \in Free(K') \setminus \{\phi\}$ and $I(K' \setminus \{\psi\}) = I(K')$, there is no $K'' \in \mathcal{C}$ s.t. $K'' \vdash \perp$, $I(K'') = I(K'' \cup \psi)$ and $|Free(K')| - |Free(K'')| > 1$ (FF − Balance).

The property Size − Balance allows prioritizing the IM-based rational clusterings where the sizes of the inconsistent clusters are balanced. However, FF − Balance allows us to focus on balancing the number of free formulas between the inconsistent clusters. Consider again the previous knowledge base K and the inconsistency measure I_M. Each of the previous two properties allows avoiding $\mathcal{C}$ in favor of $\mathcal{C}'$.

In the same way as the generalization of IM − Max, one can also generalize Size − Balance and FF − Balance by balancing the number of the formulas satisfying an arbitrary fixed property Pr:

- $\forall K' \in \mathcal{C}$ with $K' \vdash \perp$, and $\forall \psi \in Pr(K') \setminus \{\phi\}$ and $I(K' \setminus \{\psi\}) = I(K')$, there is no $K'' \in \mathcal{C}$ s.t. $K'' \vdash \perp$, $I(K'') = I(K'' \cup \{\psi\})$ and $|Pr(K')| - |Pr(K'')| > 1$ (Prop − Balance).

6 Conclusion and Perspectives

In this paper, we proposed a framework for defining clustering methods that allow dealing with data represented using classical propositional logic. The definitions of these methods are driven by different rationality postulates. In particular, we proposed an approach that uses the inconsistency measures to improve the quality of the inconsistent clusters. This allows benefiting from the numerous contributions on inconsistency measurement in the literature.

Future work includes the definition of clustering methods by taking into account our postulates. We think that some well-known methods (e.g. k-means and hierarchical agglomerative clustering techniques) can be adapted to incorporate our properties. Other possible rationality postulates can be examined to consider specific application domains.

References

[Aggarwal and Reddy, 2013] Charu C Aggarwal and Chandan K Reddy. *Data clustering: algorithms and applications*. CRC Press, 2013.

[Ammoura *et al.*, 2017] Meriem Ammoura, Yakoub Salhi, Brahim Oukacha, and Badran Raddaoui. On an MCS-based inconsistency measure. *International Journal of Approximate Reasoning*, 80:443–459, 2017.

[Berkhin, 2006] Pavel Berkhin. A Survey of Clustering Data Mining Techniques. In Jacob Kogan, Charles K. Nicholas, and Marc Teboulle, editors, *Grouping Multidimensional Data - Recent Advances in Clustering*, pages 25–71. Springer, 2006.

[Billard and Diday, 2012] Lynne Billard and Edwin Diday. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons, 2012.

[Bock, 2000] Hans Hermann Bock. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2000.

[Bona *et al.*, 2018] Glauber De Bona, John Grant, Anthony Hunter, and Sébastien Konieczny. Towards a Unified Framework for Syntactic Inconsistency Measures. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA*, 2018.

[Boudane *et al.*, 2017] Abdelhamid Boudane, Saïd Jabbour, Lakhdar Sais, and Yakoub Salhi. Clustering Complex Data Represented as Propositional Formulas. In *Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, Proceedings, Part II*, volume 10235 of *Lecture Notes in Computer Science*, pages 441–452, 2017.

[de Carvalho *et al.*, 2009] Francisco de A.T. de Carvalho, Marc Csernel, and Yves Lechevallier. Clustering constrained symbolic data. *Pattern Recognition Letters*, 30(11):1037–1045, 2009.

[de Souza and de Carvalho, 2004] Renata M.C.R. de Souza and Francisco de A.T. de Carvalho. Clustering of interval data based on city–block distances. *Pattern Recognition Letters*, 25(3):353–365, 2004.

[Diday and Esposito, 2003] Edwin Diday and Floriana Esposito. An introduction to symbolic data analysis and the SODAS software. *Intelligent Data Analysis*, 7(6):583–601, 2003.

[Gowda and Diday, 1994] K. Chidananda Gowda and Edwin Diday. *New Approaches in Classification and Data Analysis*, chapter Symbolic Clustering Algorithms using Similarity and Dissimilarity Measures. Springer Berlin Heidelberg, 1994.

[Grant and Hunter, 2011] John Grant and Anthony Hunter. Measuring the Good and the Bad in Inconsistent Information. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain*, pages 2632–2637. IJCAI/AAAI, 2011.

[Grant and Hunter, 2013] John Grant and Anthony Hunter. Distance-Based Measures of Inconsistency. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 12th European Conference, ECSQARU 2013, Utrecht, The Netherlands*, pages 230–241. Springer, 2013.

[Hunter and Konieczny, 2008] Anthony Hunter and Sébastien Konieczny. Measuring Inconsistency through Minimal Inconsistent Sets. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference, KR 2008, Sydney, Australia*, pages 358–366. AAAI Press, 2008.

[Hunter and Konieczny, 2010] Anthony Hunter and Sébastien Konieczny. On the measure of conflicts: Shapley Inconsistency Values. *Artificial Intelligence*, 174(14):1007–1026, 2010.

[Jabbour *et al.*, 2016] Saïd Jabbour, Yue Ma, Badran Raddaoui, Lakhdar Sais, and Yakoub Salhi. A MIS partition based framework for measuring inconsistency. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa*, pages 84–93. AAAI Press, 2016.

[Konieczny *et al.*, 2003] Sébastien Konieczny, Jérôme Lang, and Pierre Marquis. Quantifying information and contradiction in propositional logic through test actions. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico*, pages 106–111. Morgan Kaufmann, 2003.

[Lozinskii, 1994] Eliezer L. Lozinskii. Information and evidence in logic systems. *Journal of Experimental & Theoretical Artificial Intelligence*, 6(2):163–193, 1994.

[Marques-Silva and Lynce, 2007] João P. Marques-Silva and Inês Lynce. Towards Robust CNF Encodings of Cardinality Constraints. In *Principles and Practice of Constraint Programming, 13th International Conference, CP 2007, Providence, RI, USA*, pages 483–497, 2007.

[Michalski, 1980] Ryszard Stanislaw Michalski. Knowledge Acquisition Through Conceptual Clustering: A Theoretical Framework and an Algorithm for Partitioning Data into Conjunctive Concepts. *Journal of Policy Analysis and Information Systems*, 4(3):219–244, 1980.

[Sinz, 2005] Carsten Sinz. Towards an Optimal CNF Encoding of Boolean Cardinality Constraints. In *Principles and Practice of Constraint Programming, 11th International Conference, CP 2005, Sitges, Spain*, pages 827–831, 2005.

[Thimm, 2016] Matthias Thimm. On the expressivity of inconsistency measures. *Artificial Intelligence*, 234:120–151, 2016.

[Thimm, 2018] Matthias Thimm. On the evaluation of inconsistency measures. In John Grant and Maria Vanina Martinez, editors, *Measuring Inconsistency in Information*, volume 73 of *Studies in Logic*. College Publications, February 2018.