Explainable Ensemble Classification Model based on Argumentation
Nadia Abchiche-Mimouni, Leila Amgoud, Farida Zehraoui

To cite this version:

HAL Id: hal-03989890
https://hal.science/hal-03989890
Submitted on 26 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ABSTRACT
An ensemble classifier considers several base classifiers to make its predictions. It is generally seen as a black-box which, in addition, overlooks conflicts that may exist between base classifiers’ rules.

This paper proposes two novel ensemble classifiers that bridge the above gaps. They consider \( k \) base classifiers, each of which is a set of classification rules called theory, and a theory of domain knowledge. They build an argumentation system over the \( k + 1 \) theories for identifying and solving possible conflicts between classification rules, and use the winning rules for making predictions. We show that the two classifiers guarantee some desirable properties including explainability, compliance to knowledge, and a global compatibility of the rules they use for making predictions.

KEYWORDS
Ensemble Classification Methods, Argumentation, Explainability.

2 LOGICAL LANGUAGE
Throughout the paper, we assume a finite and non-empty set \( A = \{a_1, \ldots, a_n\} \) of attributes describing input data (e.g., age, gender) and a function \( D \) which returns the domain of every \( a \in A \). In what follows, \( L \) is a first order language whose variables and constants include the elements of \( A \) and \( D \) respectively. We call instance (or input data) an assignment of values to all attributes, i.e., a set \( \{a_1 = v_1, \ldots, a_n = v_n\} \) where \( v_i \in D(a_i) \), and denote the set of all such instances by \( I \). Let \( c \) denote the feature to learn (e.g., the diagnosis of a patient) and \( C \) be the set of its possible values. \( L' \) is a set of atomic formulas of the form \( c = v \) with \( v \in C \), and \( L \cap L' = \emptyset \). \( L'' \) is a set of constants \( r, r_1, r_2, \ldots \) used for naming rules and \( L'' \cap (L \cup L') = \emptyset \). The function \( \text{Rule}(r) \) returns the rule whose name is \( r \). We distinguish three kinds of information:

- **Facts** that are elements of \( L \).
- **Defeasible rules** \( x_1, \ldots, x_n \leadsto x \) s.t. \( x_1, \ldots, x_n \in L, x \in L' \).
- **Strict rules** \( x_1, \ldots, x_n \rightarrow x \) s.t. \( x_1, \ldots, x_n \in L \) and \( x \in L' \) or \( x \in L'' \) and \( \text{Rule}(x) \) is defeasible.

The body \( \{x_1, \ldots, x_n\} \) of both types of rules is assumed to be consistent. Facts are information about instances and domain knowledge. A defeasible rule \( x_1, \ldots, x_n \leadsto x \) is read as follows: if \( x_1, \ldots, x_n \) hold, then generally \( x \) holds as well. A strict rule \( x_1, \ldots, x_n \rightarrow x \) means if \( x_1, \ldots, x_n \) hold, then \( x \) always holds. We call classification rule any strict or defeasible rule whose head is an element of \( L' \), i.e., an atomic formula of the form \( c = v \). It gives conditions for assigning the class \( v \). A **blocking rule** is a strict rule whose head is \( x \in L'' \), i.e., the name of a defeasible rule. Its body provides circumstances in which the rule cannot be triggered. For \( r = x_1, \ldots, x_n \rightarrow /\rightarrow x \), Head(\( r \)) = \( x \) and Body(\( r \)) = \( \{x_1, \ldots, x_n\} \).
Definition 1. Two classification rules $r, r'$ are compatible iff: If \( \text{Body}(r) \cup \text{Body}(r') \) is consistent, then \( \{\text{Head}(r), \text{Head}(r')\} \) is consistent. They are incompatible otherwise.

Definition 2. A theory is a triple \( T = (F, S, D) \) where \( F \subseteq L' \) and \( S = \{r \in L'' | r \text{ is strict}\} \). \( D = \{(r, w) | r \in L'' , r \text{ is defeasible and } w \in [0, 1]\} \). For any \((r, w) \in D\), \( \text{Sc}(r) = w \).

Definition 3. Let \( T = (F, S, D) \) be a theory. The set of consequences of \( T \) is \( \text{CN}(T) = \{ x | F \vdash x \} \cup \{ \text{Head}(r) \cap L'' | r \in S \} \). This amounts to comparing them with the domain knowledge, since their classification rules may be incompatible. Arguments from every theory \( T \) is consistent iff \( \text{CN}(T) \cap D = \emptyset \).

Definition 4. A theory \( T = (F, S, D) \) is consistent iff \( \text{CN}(T) \). It is coherent iff \( \text{CN}(T) \cap D = \emptyset \).

3 ENSEMBLE CLASSIFICATION MODELS

We define two ensemble classification models \( M_a \) and \( M_b \), each of which is a function mapping every instance in \( I \) to a class from the set \( C \). They consider \( k \geq 1 \) base classifiers \( M_1, \ldots, M_k \). Each \( M_i \) is represented by a theory \( T_i = (F_i, S_i, D_i) \) where \( F_i \subseteq S_i = \emptyset \) and \( D_i \) is the set of rules extracted from \( M_i \) using existing algorithms (see [5, 17]). The weight associated to a defeasible rule represents the certainty degree with which the classifier has extracted it. Note that the same classification rule may appear in several theories and with may be different scores. The models \( M_a \) and \( M_b \) take as input another theory \( T = (F, S, D) \) where \( D = \emptyset \), containing domain knowledge. This theory is assumed to be consistent as it contains only certain information. It is also consistent with any instance in \( I \) as the latter are feasible.

The two models start first by analysing the rules of the classifiers. This amounts to comparing them with the domain knowledge, solving possible conflicts, and identifying winning classification rules. In a second step, they query the winning rules for predicting the class of any instance. The approach is thus global and not instance-dependent like that followed by existing classifiers. It is solving possible conflicts, and identifying winning classification rules of rules, and if two incompatible rules have equal score, the model considers the number of sources providing each rule.

Definition 6. The basic weight of an argument is given by the function \( \sigma : \bigcup_{i \in \{1, \ldots, k\}} \text{Arg}(T_i) \rightarrow [0, 1] \) s.t. for any \( A \in \text{Arg}(T_i) \), \( \sigma(A) \).

Arguments of the same or distinct classifiers may be conflicting since their classification rules may be incompatible. Arguments from the domain knowledge do not attack each other since the theory \( T \) is consistent and does not contain defeasible rules. However, they may attack arguments of any classifier in three ways: 1) they may use a strict classification rule which is incompatible with the classifier’s, 2) they may argue in favour of blocking the classification rule of the classifier, and 3) they may argue that the preconditions of a classifier’s rule do not hold.

Definition 7. Let \( A = \langle H, h, x \rangle \), \( A' = \langle H', h', x' \rangle \) with \( A, A' \in \bigcup_{i \in \{1, \ldots, k\}} \text{Arg}(T_i) \). A attacks \( A' \) iff one of the following holds:

- \( A, A' \in \bigcup_{i \in \{1, \ldots, k\}} \text{Arg}(T_i) \) and \( h \) is a classification rule and \( h, h' \) are incompatible, or
- \( h = h' \), or
- \( h \equiv -S \) where \( S \subseteq \text{Body}(h') \).

We introduce the notion of argumentation system as follows:

Definition 8. An argumentation system (AS) defined over the theories \( T_1, \ldots, T_k \) is a tuple \( G = (A, \sigma, R) \) where:

- \( A = \text{Arg}(T_1) \cup \ldots \cup \text{Arg}(T_k) \),
- \( \sigma \) is a mapping from \( A \) to \([0, 1]\) (as in Definition 6)
- \( R \subseteq A \times A \) is a defeat relation defined as follows: for \( A, B \in A \), \( A \) defeats \( B \) if \( A \) attacks \( B \) (see Definition 7) and \( \sigma(A) \geq \sigma(B) \).

Arguments of \( G \) are evaluated using the stable semantics [8], which returns a set \( \text{Ext}(G) \) of acceptable sets of arguments. Each set is conflict-free and defeats any argument in \( A \) left outside. For \( S \subseteq \text{Ext}(G) \), \( \text{Cons}(S) = \{ r \in L'' | \exists (\tau, r, x) \in S \} \), i.e., it returns the set of classification rules supported by arguments in \( S \).

The sceptical classifier \( M_s \) uses the classification rules which are supported by arguments in every extension. When an instance does not trigger any of the retained rules, \( M_s \) returns the symbol undefined classification.

Definition 9. A sceptical ensemble classifier defined over the theories \( T_1, \ldots, T_k \) is a function \( M_s \) mapping every instance \( x \in I \) into a class from \( C \) such that:

\[
M_s(I) = \begin{cases} 
\text{Head}(r) & r \in \bigcap_{S_i \in \text{Ext}(G)} \text{Cons}(S_i) \text{ and Body}(r) \subseteq I \\
\text{Und} & \text{otherwise}
\end{cases}
\]

where \( G = (A, \sigma, R) \) is the AS built over \( T_1, \ldots, T_k \).

The classification rules used by \( M_s \) are pairwise compatible. This property guarantees a global consistency of \( M_s \)'s predictions as it avoids applying incompatible rules to distinct instances. Furthermore, the set of rules complies with the domain knowledge since, together with the sets of facts and strict rules of the theory \( T \), it constitutes a consistent and coherent theory. Finally, \( M_s \) is explainable since it provides a prediction and the rule behind it.

Theorem 1. Let \( G = (A, \sigma, R) \) be an AS and \( T_s = (F_s, S_s, 0) \).
- Rules in \( \bigcap_{S_i \in \text{Ext}(G)} \text{Cons}(S_i) \) are pairwise compatible.
- \( (F_s, S_s, 0) \) is both consistent and coherent.

For choosing the winning classification rules, the credulous ensemble classifier \( M_s \) takes into account first the certainty degrees of rules, and if two incompatible rules have equal score, the model considers the number of sources providing each rule.
ACKNOWLEDGMENTS

This work was partially supported by the ANR (ANR-19-PI3A-0004) through the AI Interdisciplinary Institute, ANITI, as a part of France’s “Investing for the Future — PIA3” program.

REFERENCES


