Explainable Ensemble Classification Model based on Argumentation

Extended Abstract

Nadia Abchiche-Mimouni
IBISC, Univ Evry, Université Paris-Saclay
Evry, France
nadia.abchichemimouni@univ-evry.fr

Leila Amgoud
CNRS – IRIT, France
Toulouse, France
amgoud@irit.fr

Farida Zehraoui
IBISC, Univ Evry, Université Paris-Saclay
Evry, France
farida.zehraoui@univ-evry.fr

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.

ACM Reference Format:
Nadia Abchiche-Mimouni, Leila Amgoud, and Farida Zehraoui. 2023. Explainable Ensemble Classification Model based on Argumentation: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Ensemble classification methods are based on the idea of combining predictions of several base classifiers [6, 7, 14]. The most efficient algorithms are seen as black-boxes which lack transparency; this opacity hampers their relevance in critical domains like healthcare, where decision systems are becoming very popular for making diagnoses and recommending treatments. Moreover, their prediction for an instance is obtained by selecting via a voting rule one of the predictions made by the base classifiers for the instance. This approach may lead to incorrect predictions since the classification rules intra- (resp. inter-) base classifiers may be incompatible. Finally, exiting models do not integrate available domain knowledge, which may be useful for improving the quality of predictions. Such knowledge exist for instance in the healthcare domain, and may even contradict classification rules of base classifiers.

This paper proposes two novel ensemble classifiers (sceptical and credulous) that bridge the above gaps. They consider k base classifiers, each of which is represented by a set, called *theory*, of classification rules that are extracted from the classifier using well-known algorithms (eg., [2, 3, 10, 11, 13, 15, 18]), and an additional theory containing domain knowledge. They use argumentation theory, and more precisely structured argumentation (eg., [1, 9]) for solving possible conflicts between rules. It is worth mentioning that argumentation is a powerful approach for reasoning about conflicting information (see [4, 12, 16] for more on argumentation and its applications). The two classifiers build an argumentation system over the $k + 1$ theories for identifying and solving possible conflicts between 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.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

2 LOGICAL LANGUAGE

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

Facts that are elements of $\mathcal{L}$,

Defeasible rules $x_1, \ldots, x_n \rightsquigarrow x$ s.t. $x_1, \ldots, x_n \in \mathcal{L}, x \in \mathcal{L}'$,

Strict rules $x_1, \ldots, x_n \rightarrow x$ s.t. $x_1, \ldots, x_n \in \mathcal{L}$ and $x \in \mathcal{L}'$ or $x \in \mathcal{L}''$ and $\mathtt{Rule}(x)$ 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 $x_1, \ldots, x_n \rightsquigarrow x$ 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 $\mathcal{L}'$, i.e., an atomic formula of the form $\mathtt{c} = v$. It gives conditions for assigning the class v. A **blocking rule** is a strict rule whose head is $x \in \mathcal{L}''$, i.e., the **name** of a defeasible rule. Its body provides circumstances in which the rule cannot be triggered. For $r = x_1, \ldots, x_n \rightarrow / \rightsquigarrow x$, $\mathtt{Head}(r) = x$ and $\mathtt{Body}(r) = \{x_1, \ldots, x_n\}$.

Definition 1. *Two classification rules r, r' are* compatible *iff: If* $\mathtt{Body}(r) \cup \mathtt{Body}(r')$ *is consistent, then* $\{\mathtt{Head}(r), \mathtt{Head}(r')\}$ *is consistent. They are* incompatible *otherwise.*

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

Definition 3. *Let* $\mathcal{T} = (\mathcal{F}, \mathcal{S}, \mathcal{D})$ *be a theory. The set of consequences of* $\mathcal{T}$ *is* $\mathtt{CN}(\mathcal{T}) = \{x | \ \mathcal{F} \vdash x\} \cup \{\mathtt{Head}(r) \cap \mathcal{L}'' \mid r \in \mathcal{S}, \mathcal{F} \vdash x, \forall x \in \mathtt{Body}(r)\}$.

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

3 ENSEMBLE CLASSIFICATION MODELS

We define two ensemble classification models $\mathbf{M}_s$ and $\mathbf{M}_c$, each of which is a function mapping every instance in $\mathbf{I}$ to a class from the set $\mathbf{C}$. They consider $k \geq 1$ base classifiers $\mathbf{M}_1, \ldots, \mathbf{M}_k$. Each $\mathbf{M}_i$ is represented by a theory $\mathcal{T}_i = \langle \mathcal{F}_i, \mathcal{S}_i, \mathcal{D}_i \rangle$ where $\mathcal{F}_i = \mathcal{S}_i = \emptyset$ and $\mathcal{D}_i$ is the set of rules extracted from $\mathbf{M}_i$ 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 $\mathbf{M}_s$ and $\mathbf{M}_c$ take as input another theory $\mathcal{T}_* = \langle \mathcal{F}_*, \mathcal{S}_*, \mathcal{D}_* \rangle$ where $\mathcal{D}_* = \emptyset$, containing domain knowledge. This theory is assumed to be **consistent** as it contains only certain information. It is also consistent with any instance in $\mathbf{I}$ 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 based on argumentation theory, which generates a set $\mathtt{Arg}(\mathcal{T}_x)$ of arguments from every theory $\mathcal{T}_x = \langle \mathcal{F}_x, \mathcal{S}_x, \mathcal{D}_x \rangle$, $x \in \{1, \ldots, k, *\}$.

Definition 5. *An* argument *is a tuple* $A = \langle H, h, x \rangle$ *verifying any of the following conditions:*

- $H \subseteq \mathcal{F}_x$, $h \in \mathcal{L}$ *and* H *is a minimal (for set inclusion) consistent subset of* $\mathcal{F}_x$ *s.t.* $H \vdash h$.
- $H = \{r\}$ *and* $h = r$ *with* $r \in \mathcal{S}_x \cup \mathcal{D}_x$ *and* $\mathtt{Head}(r) \in \mathcal{L}'$.
- $H = \{x_1, \ldots, x_j, r\}$, $h \in \mathcal{L}''$, $\forall i = 1, \ldots, j$, $\mathcal{F}_x \vdash x_i$, $r \in \mathcal{S}_x$, $\mathtt{Body}(r) = \{x_1, \ldots, x_j\}$ *and* $\mathtt{Head}(r) = h$.

Hence, we get $k + 1$ sets of arguments ($\mathtt{Arg}(\mathcal{T}_*)$, $\mathtt{Arg}(\mathcal{T}_1)$, …, $\mathtt{Arg}(\mathcal{T}_k)$). Each set of the k classifiers contains only arguments of type $\langle \{r\}, r, i \rangle$, where r is a classification rule, while $\mathtt{Arg}(\mathcal{T}_*)$ may include the three types of arguments. Every argument has a basic weight from the unit interval $[0, 1]$ defined as follows.

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

$$\sigma(A) = \begin{cases} 1 & \text{if } x = * \\ \mathtt{Sc}(r) & \text{if } x \neq * \text{ and } A = \langle \{r\}, r, x \rangle. \end{cases}$$

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 $\mathcal{T}_*$ 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* $A, A' \in \bigcup_{i \in \{1,\ldots,k,*\}} \mathtt{Arg}(\mathcal{T}_i)$. A attacks A' *iff one of the following holds:*

- $A, A' \in \bigcup_{i=1,\ldots,k} \mathtt{Arg}(\mathcal{T}_i)$ *and* h, h' *are incompatible.*
- $A \in \mathtt{Arg}(\mathcal{T}_*)$, $A' \in \bigcup_{i=1,\ldots,k} \mathtt{Arg}(\mathcal{T}_i)$ *and*
  - h *is a classification rule and* h, h' *are incompatible, or*
  - $h = h'$, *or*
  - $h \equiv \neg S$ *where* $S \subseteq \mathtt{Body}(h')$.

We introduce the notion of *argumentation system* as follows:

Definition 8. *An* argumentation system *(AS) defined over the theories* $\mathcal{T}_*, \mathcal{T}_1, \ldots, \mathcal{T}_k$ *is a tuple* $\mathbf{G} = (\mathcal{A}, \sigma, \mathcal{R})$ *where:*

- $\mathcal{A} = \mathtt{Arg}(\mathcal{T}_*) \cup \mathtt{Arg}(\mathcal{T}_1) \cup \ldots \cup \mathtt{Arg}(\mathcal{T}_k)$,
- σ *is a mapping from* $\mathcal{A}$ *to* $[0, 1]$ *(as in Definition 6)*
- $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ *is a defeat relation defined as follows: for* $A, B \in \mathcal{A}$, A defeats B *iff* A *attacks* B *(see Definition 7) and* $\sigma(A) \geq \sigma(B)$.

Arguments of $\mathbf{G}$ are evaluated using the *stable semantics* [8], which returns a set $\mathtt{Ext}(\mathbf{G})$ of acceptable **sets of arguments**. Each set is conflict-free and defeats any argument in $\mathcal{A}$ left outside. For $S \in \mathtt{Ext}(\mathbf{G})$, $\mathtt{Conc}(S) = \{r \in \mathcal{L}'' \mid \exists \langle \{r\}, r, x \rangle \in S\}$, i.e., it returns the set of *classification rules* supported by arguments in S.

The sceptical classifier $\mathbf{M}_s$ uses the classification rules which are supported by arguments in every extension. When an instance does not trigger any of the retained rules, $\mathbf{M}_s$ returns the symbol und meaning *undecided* classification.

Definition 9. *A* sceptical ensemble classifier *defined over the theories* $\mathcal{T}_*, \mathcal{T}_1, \ldots, \mathcal{T}_k$ *is a function* $\mathbf{M}_s$ *mapping every instance* $I \in \mathbf{I}$ *into a class from* $\mathbf{C}$ *such that:*

$$\mathbf{M}_s(I) = \begin{cases} \mathtt{Head}(r) & r \in \bigcap_{S_i \in \mathtt{Ext}(\mathbf{G})} \mathtt{Conc}(S_i) \text{ and } \mathtt{Body}(r) \subseteq I \\ \mathtt{Und} & \text{otherwise} \end{cases}$$

where $\mathbf{G} = (\mathcal{A}, \sigma, \mathcal{R})$ *is the AS built over* $\mathcal{T}_*, \mathcal{T}_1, \ldots, \mathcal{T}_k$.

The classification rules used by $\mathbf{M}_s$ are pairwise compatible. This property guarantees a global consistency of $\mathbf{M}_s$'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 $\mathcal{T}_*$, it constitutes a consistent and coherent theory. Finally, $\mathbf{M}_s$ is explainable since it provides a prediction and the rule behind it.

Theorem 1. *Let* $\mathbf{G} = (\mathcal{A}, \sigma, \mathcal{R})$ *be an AS and* $\mathcal{T}_* = \langle \mathcal{F}_*, \mathcal{S}_*, \emptyset \rangle$.

- *Rules in* $\bigcap_{S_i \in \mathtt{Ext}(\mathbf{G})} \mathtt{Conc}(S_i)$ *are pairwise compatible.*
- $\langle \mathcal{F}_*, \mathcal{S}_*, \bigcap_{S_i \in \mathtt{Ext}(\mathbf{G})} \mathtt{Conc}(S_i) \rangle$ *is both consistent and coherent.*

For choosing the winning classification rules, the credulous ensemble classifier $\mathbf{M}_c$ 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

[1] Leila Amgoud and Farid Nouioua. 2017. An argumentation system for defeasible reasoning. *International Journal of Approximate Reasoning* 85 (2017), 1–20.
[2] Robert Andrews, Joachim Diederich, and Alan B. Tickle. 1995. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems* 8, 6 (1995), 373 – 389. Knowledge-based neural networks.
[3] M. Gethsiyal Augasta and T. Kathirvalavakumar. 2012. Reverse Engineering the Neural Networks for Rule Extraction in Classification Problems. *Neural Processing Letters* 35, 2 (01 Apr 2012), 131–150.
[4] Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leon van der Torre (Eds.). 2018. *Handbook of Formal Argumentation, Volume 1.* College Publications.
[5] Guido Bologna. 2021. A Rule Extraction Technique Applied to Ensembles of Neural Networks, Random Forests, and Gradient-Boosted Trees. *Algorithms* 14, 12 (2021).
[6] Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.
[7] Harris Drucker, Corinna Cortes, Lawrence D Jackel, Yann LeCun, and Vladimir Vapnik. 1994. Boosting and other ensemble methods. *Neural Computation* 6, 6 (1994), 1289–1301.
[8] Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artif. Intell.* 77, 2 (1995), 321–358.
[9] A. García and G. Simari. 2004. Defeasible logic programming: an argumentative approach. *Theory and Practice of Logic Programming* 4, 1-2 (2004), 95–138.
[10] Eduardo R. Hruschka and Nelson F.F. Ebecken. 2006. Extracting rules from multilayer perceptrons in classification problems: A clustering-based approach. *Neurocomputing* 70, 1 (2006), 384 – 397. Neural Networks.
[11] Hongjun Lu, Rudy Setiono, and Huan Liu. 1996. Effective Data Mining Using Neural Networks. *IEEE Trans. on Knowl. and Data Eng.* 8, 6 (Dec. 1996), 957–961.
[12] Iyad Rahwan and Guillermo R. Simari. 2009. *Argumentation in Artificial Intelligence* (1st ed.). Springer Publishing Company, Incorporated.
[13] Emad W. Saad and Donald C. Wunsch. 2007. Neural network explanation using inversion. *Neural Networks* 20, 1 (2007), 78 – 93.
[14] Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1249.
[15] Makoto Sato and Hiroshi Tsukimoto. 2001. Rule extraction from neural networks via decision tree induction. In *IJCNN*, Vol. 3. 1870 – 1875 vol.3.
[16] Guillermo Simari, Massimiliano Giacomin, Dov Gabbay, and Matthias Thimm (Eds.). 2021. *Handbook of Formal Argumentation, Volume 2.* College Publications.
[17] Giulia Vilone and Luca Longo. 2021. A Quantitative Evaluation of Global, Rule-Based Explanations of Post-Hoc, Model Agnostic Methods. *Frontiers in Artificial Intelligence* 4 (2021).
[18] Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen. 2016. DeepRED – Rule Extraction from Deep Neural Networks. In *Discovery Science*, Toon Calders, Michelangelo Ceci, and Donato Malerba (Eds.). Springer International Publishing, Cham, 457–473.