A Hybrid Approach to Identifying the Most Predictive and Discriminant Features in Supervised Classification Problems

Alexandre Bazin, Miguel Couceiro, Marie-Dominique Devignes, Amedeo Napoli

A Hybrid Approach to Identifying the Most Predictive and Discriminant Features in Supervised Classification Problems

Alexandre Bazin, Miguel Couceiro, Marie-Dominique Devignes, Amedeo Napoli
Université de Lorraine – CNRS – Inria, LORIA, F-54000 Nancy, France
Email: name.surname@loria.fr

***Abstract*—In this paper, we are interested in the predictive and discriminant nature of features in supervised classification problems. We discuss the notions of prediction and discrimination and propose a hybrid approach combining supervised classifiers, model explanation, multicriteria decision making and pattern mining for identifying the most predictive and discriminant features in a dataset. The explanation of models learned by supervised classifiers produces rankings of features according to various performance measures. Based on that, multicriteria decision making and pattern mining methods are used to, respectively, select the most important features and interpret their role in terms of prediction and discrimination. Finally, we present and discuss two experiments on public datasets illustrating the potential of the approach.**

I. Introduction

Biomedical sciences make increasing use of methods from computer science. For instance, biologists wishing to study diabetes now collect data, in the form of biomarkers, from both diabetic and healthy patients and then use machine learning techniques to discriminate classes of patients and predict the disease. Provided that the patients are sufficiently numerous and the data correctly collected, most modern supervised classification approaches [1] are able to build models capable of diagnosing, or predicting the onset of, diabetes in new patients with good performances. While useful for the patients themselves, simply applying such models is not sufficient for biologists who rather need to understand the underlying causes of diabetes, i.e. they need meaning to be assigned to the biomarkers in terms of their roles in the development of the illness and its diagnostic. Which biomarkers can best be used to predict that a patient has diabetes? Which biomarkers can best be used to predict that a patient does not have diabetes? Which biomarkers can best be used to discriminate between having and not having diabetes? Are the best biomarkers for these three tasks the same?

The problem of identifying the most important features (biomarkers) in a supervised classification setting (the diagnosis of diabetes by a model) belongs to the field of explainable machine learning [2], [3], which has now become one of the main research topics in artificial intelligence. However, existing methods only identify the features that are important for models with no explicit consideration for the different meanings this importance can have. Indeed, in supervised classification, models perform two subtly different tasks: prediction and discrimination. Features can thus be important for either the prediction or discrimination performance of models, i.e. a feature can be *predictive* or *discriminant*.

In this paper, we are interested in identifying predictive and discriminant features in supervised classification problems (i.e. in datasets) instead of in particular models. We first discuss the notions of prediction and discrimination, then introduce a general method for identifying the features that are the most predictive and/or discriminant in a supervised classification dataset. The proposed method combines a machine learning explanation approach with multicriteria decision making [4] and pattern mining [5]. Supervised classifiers are used on the data to create models. The model explanation approach produces rankings of features according to their importance w.r.t measures of performance on which background knowledge is expressed in terms of prediction and discrimination. The multicriteria decision making process uses the rankings of features to select the "most important" features. Pattern mining, through the formal concept analysis (FCA) formalism [6], [7], exploits the symbolic background knowledge about the measures of performance to transfer meaning to the selected features, and allows for an intuitive visual representation of the result. This workflow is illustrated in Fig. 1.

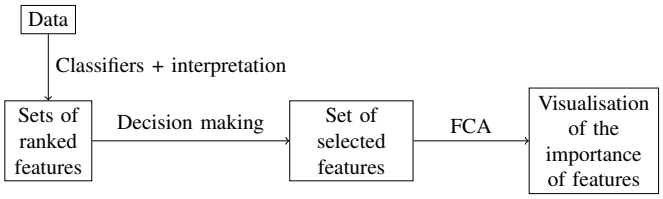

Fig. 1: Workflow of the hybrid approach.

As a practical case, we analyse two public biomedical datasets and show that the application of the proposed method allows the user to gain an understanding of the data and the classification problem. To the best of our knowledge, this is one of the few papers discussing the discriminant and predictive nature of dataset features in a supervised classification problem thanks to decision making and FCA.

This paper is organised as follows. In Section II, we propose a definition of discrimination and prediction and discuss how they can be measured using models. In Section III, we present

how to use multicriteria decision to identify the features we are interested in. In Section IV, we recall the necessary FCA background, and use it to assign meaning to features and present it in the form of a concept lattice. In Section V, we apply the proposed method to the public datasets to illustrate the capabilities of the framework introduced in this paper. In Section VI, we discuss the choices we made and present some directions for future work.

II. Prediction and Discrimination

We are interested in identifying predictive and discriminant features in a dataset. The notions of "predictiveness" and "discriminativeness" of features are rarely discussed in the literature so we will first introduce working definitions. *To predict* means to assert that something will happen, is true or, in a classification problem, belongs to a class. *To discriminate* means to be able to perceive the differences between two things. Regarding features in a classification problem, a feature is said to be *predictive* when its value can be used to assert that an individual belongs to a particular class, and as *discriminant* when its value can be used to differentiate between the classes. For instance, fevers are predictive of being ill because their presence can be used by doctors to diagnose illnesses but they are not discriminant as they do not allow to separate between several possible diseases. The definitions of predictive and discriminant are thus linked to the existence of an external process that uses the features to make decisions. Here, we chose to use models built with classifiers as external processes.

We consider binary classification problems [8] in which individuals belong to one of two groups, the *positive* (or *target*) and *negative* classes. A *classifier* is a process that uses a set of individuals for which the classes are known (the *training set*) to create a *model* that is able to assign classes to a set of individuals for which the true classes are hidden (the *test set*) or unknown. Explaining the way models use the features to assign classes to individuals is currently one of the main topic in artificial intelligence. Guidotti et al. [3] proposed a categorisation of explanation problems and approaches into three categories: *model explanation*, *outcome explanation* and *inspection*. The first category contains approaches that, given a target model, aim at providing another model that is understandable and mimics the behaviour of the target model. Such understandable models include sets of rules [9] or decision trees [10]. The second category contains approaches that aim at providing explanations of the predictions on individual instances. Noticeable examples of outcome explanation approaches include LIME [11] and SHAP [12]. LIME explains individual predictions by presenting simple, understandable models trained from randomly generated instances similar to the one being explained. SHAP explains individual predictions by evaluating the importance that each feature had in the model's decision, i.e, in the introductory biomedical example, the importance that each biomarker had in the diagnosis of a patient. The third category contains approaches that aim at providing a (visual or textual) representation of the work of models. Most of these approaches focus on explaining neural networks [13].

In this work, we want to identify the features that are used by the models to predict and/or discriminate. We thus need an explanation approach that highlights features w.r.t. their importance in these two tasks. Many *measures* have been proposed [14], [15], [16] to quantify the performance of models with regard to various views of what a good model should be doing. The relations between these measures, as well as the role that they play in the evaluation of models [17], [18], [19], have been extensively studied. The subject is of particular importance in biostatistics where researchers are notably interested in the relations between performance measures and prediction and discrimination [20], [21]. In this paper, we consider measures that are combinations of four different scores obtained by guessing the classes of individuals in a test set. Some of these measures are presented in Table I. *True Positive* (TP) is the number of individuals belonging to the positive class that were classified as positive. False Negative (FN) is the number of individuals belonging to the positive class that were classified as negative. *False Positive* (FP) is the number of individuals belonging to the negative class that were classified as positive. *True Negative* (TN) is the number of individuals belonging to the negative class that were classified as negative.

The *sensitivity* (or *recall*) measure, for example, is the ratio of the number of positive individuals that have been correctly recognised as such by the model to the total number of positive individuals in the test set, i.e. sensitivity equals $\frac{TP}{TP+FN}$. Sensitivity quantifies the ability of the model to recognise the positive class. The *precision* measure is the ratio of the number of positive individuals that have been correctly recognised as such by the model to the total number of positive guesses, i.e. precision equals $\frac{TP}{TP+FP}$. It quantifies the ability of the model not to make mistakes when identifying the positive class. The *accuracy* measure is the ratio of the number of individuals whose class has been correctly guessed by the model to the total number of individuals, i.e. accuracy equals $\frac{TP+TN}{TP+TN+FP+FN}$. As such, it quantifies the ability of the model to recognise both positive and negative classes while not making mistakes.

These three measures represent different priorities. Sensitivity is maximised when the model is always predicting the positive class. Many errors can be made (false positive) but all individuals belonging to the positive class are recognised as such. Conversely, the precision can be maximised by being overly cautious with positive predictions. The accuracy can be considered as a compromise between the predictive power and the error avoidance that perceives both classes as equally important. We observe that two important notions are at play here: recognising classes and not making mistakes. We will consider that measures quantify the *prediction* power of models when they focus only on the recognition part (e.g. sensitivity, specificity), the *correctness* of models when they focus only on not making mistakes (e.g. precision, negative predictive value), and the *discrimination* power of models when they mix both goals (e.g. accuracy, Fscore). In Fig. 2, we present a representation of these different measures' characteristics in the form of a partial ordering of terms that serves

	Real class 1	Real class 0		
Predicted class 1	True Positive (TP)	False Positive (FP)	Precision= $\frac{TP}{TP+FP}$	$FDR = \frac{FP}{TP+FP}$
Predicted class 0	False Negative (FN)	True Negative (TN)	FOR= $\frac{FN}{FN+TN}$	$NPV = \frac{TN}{FN+TN}$
Sensitivity= $\frac{TP}{TP+FN}$		FPR = $\frac{FP}{FP+TN}$	FScore= $2\frac{Precision \times Sensitivity}{Precision+Sensitivity}$	
FNR = $\frac{FN}{TP+FN}$		Specificity = $\frac{TN}{FP+TN}$	Accuracy= $\frac{TP+TN}{TP+TN+FP+FN}$	
Positive Likelihood Ratio= $\frac{Sensitivity}{FPR}$			MCC = $\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	
Negative Likelihood Ratio= $\frac{FNR}{Specificity}$				

TABLE I: Some possible measures of the performance of a model. Measures in yellow quantify the *prediction* power of models, measures in green quantify the *correctness*, and those in blue quantify the *discrimination* power.

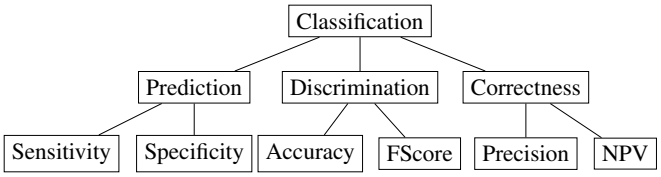

Fig. 2: Partial ordering of terms characterising performance measures and their role in a classification process.

as background knowledge for the interpretation of features. It states that sensitivity is related to prediction which is a possible characteristic of classification.

Let T be a test set, M a model, m a measure and f a feature used to describe the individuals in T. We denote by $m(M,T)$ the score of the model M for the measure m on the test set T, and denote by T_i^f the test set obtained by randomly permuting the values taken by f in T for a permutation i. The *impact* of the feature f on the score of the model M for the measure m is defined as the mean variation of the score of the model for m when the values taken by f in the test set are permuted [22], i.e., for a large enough k (number of permutations),

$$impact(f,M,m) \approx \sum_{i=1}^{k} \frac{m(M,T_i^f) - m(M,T)}{k}.$$

In other words, the impact reflects the importance that the feature has for the model w.r.t. the measure. This impact thus constitutes a form of explanation of the model. As it is a real number, the impact can be used to rank the features. A negative impact means that changing the values of the feature has a negative influence on the model's score, which means that the model makes use of the feature for whatever the measure is quantifying. Therefore, in this work, we define features as *predictive* or *discriminant* if they have negative impacts on measures of prediction or discrimination. Measures of correctness are not used in this work but are discussed again later.

III. Selecting Features

A. Prediction, Discrimination and Feature Importance

We have a dataset in which individuals are described by the values of a set $\mathcal{F}$ of features and a corresponding class (here we will assume that there are two classes). In Section 2, we defined predictive (resp. discriminant) features as those having a negative impact on the score of a model for a measure of prediction (resp. discrimination). All features in the dataset potentially have these characteristics but presenting them all to an expert would not help. Instead, we want to identify the features that are among *the most* predictive or discriminant.

We could argue that a feature f_1 is *more predictive* than a feature f_2 if it has a smaller impact value on a model's sensitivity score. However, if f_2 has a smaller impact value than f_1 on the same model's specificity score, which one is the most predictive? And if f_2 has a smaller impact value on another model's sensitivity score? As multiple measures are indicative of prediction or discrimination, and different models can result in different rankings of features, we represent the problem of identifying the most predictive and discriminant features as a multicriteria decision problem.

One of the goals of multicriteria decision making [4], [23] is to model preferences. In this paper, we take a utility-based approach. Let $V_1, ..., V_n$ be attributes, and let $\mathbb{R}$ be the set of real numbers that we use as evaluation space. By a criterion we mean a pair $\mathcal{V}_i = (V_i, \phi_i)$, $i \in \{1,\ldots,n\}$, where $\phi_i \colon V_i \to \mathbb{R}$ is an utility function. Such a criterion naturally defines a local preference relation (reflexive and transitive) $\preceq_i$ on V_i: for all $x, y \in V_i$, $x \preceq_i y$ if $\phi_i(x) \geq \phi_i(y)$. Let $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ be two alternatives in $V_1 \times \cdots \times V_n$. We say that *$a$ is preferred to b on the ith criterion* when $b_i \preceq_i a_i$. For instance, when buying a new car, two criteria could be based on the price and the maximum speed: price-wise, one would prefer a cheaper car whereas Speed-wise, one would prefer a faster car. However, preferences on those two criteria do not necessarily coincide and compromises must be made. When faced with a set of alternatives and multiple criteria, we will refer to the problem of identifying the "best" alternatives according to the criteria as a multicriteria decision problem.

Let $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ be two alternatives. Alternative a is said to *dominate* b, denoted by $b \preceq a$, if $b_i \preceq_i a_i$ for all $i \in \{1,\ldots,n\}$. The *Pareto front* of the multicriteria decision problem over a set $Crit$ of criteria and a set Alt of alternatives is denoted by $Pareto(Crit, Alt)$ and it is defined as the set of alternatives that are not dominated

by any other alternative. In other words, an alternative is in the Pareto front if it is better than all the others on at least one criterion. A car that is both slower and more expensive than another is surely not preferred and thus it does not constitute a better choice. Having excluded all the alternatives that are clearly worse than others, the Pareto front contains only alternatives for which one cannot improve on a criterion without losing on another. Notice that the Pareto front constitutes an anti-chain w.r.t. the overall preference relation $\preceq$.

Let $\mathcal{M}o$ be a set of models and let $\mathcal{M}e$ be a set of performance measures quantifying either prediction or discrimination. We represent the problem of identifying the most predictive and discriminant features as the multicriteria decision problem in which the set of alternatives is the set $\mathcal{F}$ of features and the set of attributes is the set $\mathcal{M}o \times \mathcal{M}e$ of the pairs composed of a model and a measure. The value of the attribute $(m_o, m_e) \in \mathcal{M}o \times \mathcal{M}e$ for the feature $f \in \mathcal{F}$ is $impact(f, m_o, m_e)$. The criteria are the pairs $((m_o, m_e), id)$ where id is the identity function. A feature f_1 is preferred to a feature f_2 for a criterion $((m_o, m_e), id)$ if and only if $impact(f_1, m_o, m_e) \leq impact(f_2, m_o, m_e)$. In order to simplify the notations, we thereafter identify the criteria with their attributes and use "*the criterion* (m_o, m_e)" to refer to $((m_o, m_e), id)$, as well as $\mathcal{M}_o \times \mathcal{M}_e$ to refer to the set of criteria.

Definition 1. *(*IMPORTANT FEATURES*) A feature f is said to be* important *if*

$$f \in Pareto(\mathcal{M}o \times \mathcal{M}e, \mathcal{F}).$$

We provide an illustrative example in the next subsection.

B. Identifying Important Features

Important features are defined w.r.t. a multicriteria decision problem that involves models and measures. As it is impossible to consider all possible models and measures, we have to make choices and restrict ourselves to finite sets.

Let $\mathcal{C}$ be a finite set of classifiers (e.g. Random Forests [24], Naive Bayes, Neural Networks, Support Vector Machines [25]) and $\mathcal{M}e$ be a finite set of measures of a model's performance (e.g. accuracy, specificity, sensitivity). As a running example, we will use

- $\mathcal{F} = \{f_1, f_2, f_3, f_4, f_5\}$
- $\mathcal{C} = \{$Random Forests (RF), Neural Networks (NN)$\}$
- $\mathcal{M}e = \{Specificity, Sensitivity, Accuracy\}$

Our assumption is that different types of classifiers learn, and thus perceive and use features, differently so the most important features for accuracy are not necessarily the same in models learned with neural networks and random forests. With each classifier $C \in \mathcal{C}$, we create a model that represents the classifier. The training and test sets are of fixed sizes and randomly drawn from the dataset at each new training phase. This process results in the creation of the set $\mathcal{M}o$ of models. From there, for each model M and measure m, we associate to each feature f its impact on M's m score. This results in the creation of a $|\mathcal{C} \times \mathcal{M}e| \times |\mathcal{F}|$ matrix that quantifies the importance that each classifier's representative model assigns to each feature w.r.t. each measure. Let us suppose that, in our running example, the matrix is the one depicted in Table II.

	f_1	f_2	f_3	f_4	f_5
$(RF, Accuracy)$	-0.01	0	0.04	-0.015	0.002
$(RF, Sensitivity)$	0	-0.02	0.01	0.017	0.001
$(RF, Specificity)$	-0.005	-0.18	0.002	0.015	0.03
$(NN, Accuracy)$	0	-0.002	0.02	-0.01	0.01
$(NN, Sensitivity)$	0	0.006	0.009	0.01	-0.01
$(NN, Specificity)$	0	0.005	0.01	0.009	-0.009

TABLE II: Matrix of impacts of features on models' scores. A negative score means that a feature is correctly used by the model to perform what the measure is quantifying.

We use the matrix of impacts of features to compute $Pareto(\mathcal{M}o \times \mathcal{M}e, \mathcal{F})$, i.e. the important features. In our running example, as a feature is preferred to another if its impact value is lower, we have the following preferences:

$$\begin{aligned}
c_1 &= (RF, Accuracy): && f_4 \succ f_1 \succ f_2 \succ f_5 \succ f_3 \\
c_2 &= (RF, Sensitivity): && f_2 \succ f_1 \succ f_5 \succ f_3 \succ f_4 \\
c_3 &= (RF, Specificity): && f_2 \succ f_1 \succ f_3 \succ f_4 \succ f_5 \\
c_4 &= (NN, Accuracy): && f_4 \succ f_2 \succ f_1 \succ f_5 \succ f_3 \\
c_5 &= (NN, Sensitivity): && f_5 \succ f_1 \succ f_2 \succ f_3 \succ f_4 \\
c_6 &= (NN, Specificity): && f_5 \succ f_1 \succ f_2 \succ f_4 \succ f_3
\end{aligned}$$

The Pareto front of this multicriteria decision problem, and the set of important features, is $\{f_1, f_2, f_4, f_5\}$. The features f_2, f_4 and f_5 are important because they are the best for some criteria and f_1 is important because it is better than the others on at least one criterion. The feature f_3 is not deemed important as it is always worse than f_1 and f_2.

Once the important features are identified, we want to interpret their importance in terms of prediction and discrimination.

IV. INTERPRETING IMPORTANT FEATURES

Knowing that a feature is important is not enough. We would like to know *why* it is important. Is it because it is particularly discriminant? Is it predictive? If it is predictive, is it predictive of the positive or negative class? We would like to explain the reasons why a feature has been deemed important so as to provide more insight into its role in the classification problem. As importance is defined as membership to the Pareto front of a multicriteria decision problem, explaining the importance of a feature is linked to identifying the criteria responsible for its presence in the Pareto front.

Let $\mathcal{G}$ be a set and $2^{\mathcal{G}}$ be the family of its subsets. *Closure* and *interior* operators on $\mathcal{G}$ are functions $f : 2^{\mathcal{G}} \mapsto 2^{\mathcal{G}}$ such that $X \subseteq Y \Rightarrow f(X) \subseteq f(Y)$ and $f(f(X)) = f(X)$. In addition, closure operators c are such that $X \subseteq c(X)$ and interior operators i are such that $X \supseteq i(X)$. A set $X \subseteq \mathcal{G}$ is said to be *closed* under a closure operator c if $X = c(X)$. It is said to be *open* under an interior operator i if $X = i(X)$. Let $\leq$ be a total order on the elements of $\mathcal{G}$. We call *lectic order* the partial order $\leq_l$ on $2^{\mathcal{G}}$ such that $X \leq_l Y$ if and only

if the smallest element of the symmetric difference of X and Y, according to $\leq$, is in Y. We then say that X is *lectically smaller* than Y.

Our set of criteria is $\mathcal{M}o \times \mathcal{M}e$ and our set of alternatives, or features, is $\mathcal{F}$. We define the interior operator $g : 2^{\mathcal{M}o \times \mathcal{M}e} \mapsto 2^{\mathcal{M}o \times \mathcal{M}e}$ such that, for a set X of criteria, $g(X)$ is the lectically greatest inclusion-minimal subset of X for which $Pareto(X, \mathcal{F}) = Pareto(g(X), \mathcal{F})$. Let

$$\mathcal{P} = \{P \subseteq \mathcal{F} \mid \exists X \subseteq \mathcal{M}o \times \mathcal{M}e, P = Pareto(X, \mathcal{F})\}$$

be the family of features sets P for which there exists a criteria set X such that $Pareto(X, \mathcal{F}) = P$. For $P \in \mathcal{P}$, we use $C(P)$ to denote the family of criteria sets X such that $Pareto(X, \mathcal{F}) = P$. We then have that, for any $P \in \mathcal{P}$, $G(P) = \{g(X) \mid X \in C(P)\}$ is the family of inclusion-minimal criteria sets for which P is the Pareto front. Hence, X is a minimal set of criteria for which a feature f appears on the Pareto front if and only if $X \in G(P)$ for some P that is inclusion-minimal such that $f \in P$. We use $M(f)$ to denote the family of such minimal criteria sets for the feature f.

In our running example, the feature f_2 appears in the Pareto front only when the criteria set contains c_2, c_3, both c_4 and c_5 or both c_4 and c_6. Hence, $M(f_2)$ = {{c_2}, {c_3}, {c_4,c_5}, {c_4,c_6}}. Similarly, for other features, we have $M(f_4)$ = {{c_1}, {c_4}}, $M(f_5)$ = {{c_5}, {c_6}} and, finally, $M(f_1)$ = {{c_1,c_2}, {c_1,c_3}, {c_1,c_5}, {c_1,c_6}, {c_2,c_5}, {c_2,c_6}, {c_3,c_5}, {c_3,c_6}, {c_4,c_5}, {c_4,c_6}}.

Once the minimal sets of criteria required for a feature f to appear in the Pareto front are identified, we interpret them in human-understandable terms. To a criterion $c_i \in \mathcal{M}o \times \mathcal{M}e$ we assign an *interpretation* $I(c_i)$, i.e. a set of terms, according to background knowledge. Using Fig 2's partially ordered set as background knowledge, the interpretation $I(c_i)$ is then the set of terms greater than or equal to the name of the measure in c_i. Indeed, Fig. 2 contains terms that represent the measures and tasks for which a feature can be good. As discussed in Section II, being good for accuracy is being good for *discrimination*. Furthermore, being discriminant is being good, in general, for *classification*. We know that the values of the criterion $(Random\ Forests, Accuracy)$ are the impacts of the features on the accuracy of the model that represents random forests. Hence, the best features for this criterion are good for the *accuracy*. The interpretation of $(Random\ Forests, Accuracy)$ is then $I(Random\ Forests, Accuracy)$ = {Accuracy, Discrimination, Classification}. In our running example, the interpretations of the criteria are:

$I(c_1)$ = {Accuracy, Discrimination, Classification}
$I(c_2)$ = {Sensitivity, Prediction, Classification}
$I(c_3)$ = {Specificity, Prediction, Classification}
$I(c_4)$ = {Accuracy, Discrimination, Classification}
$I(c_5)$ = {Sensitivity, Prediction, Classification}
$I(c_6)$ = {Specificity, Prediction, Classification}

What remains is to associate the terms used in the criteria's interpretations to important features and present the result to the user. To do this, we use notions from formal concept analysis [6], a mathematical framework based on lattice theory that aims at extracting meaningful classes from data and, as such, can be perceived as a form of clustering. A *formal context* is a triple $(\mathcal{O}, \mathcal{A}, \mathcal{R})$ in which $\mathcal{O}$ is a finite set of *objects*, $\mathcal{A}$ a finite set of *attributes* and $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{A}$ a relation between objects and attributes. We say that *the object o is described by the attribute a* when $(o, a) \in R$. Formal contexts are formalisations of binary datasets. Two operators can then be defined, both denoted by $\cdot'$:

$$\cdot' : 2^{\mathcal{O}} \mapsto 2^{\mathcal{A}}$$

$$O' = \{a \in \mathcal{A} \mid \forall o \in O, (o, a) \in \mathcal{R}\}$$

$$\cdot' : 2^{\mathcal{A}} \mapsto 2^{\mathcal{O}}$$

$$A' = \{o \in \mathcal{O} \mid \forall a \in A, (o, a) \in \mathcal{R}\}$$

They form a Galois connection and, as such, both $\cdot''$ compositions are closure operators. A pair (O, A), where $O \subseteq \mathcal{O}$ and $A \subseteq \mathcal{A}$, is called a *formal concept* if and only if $O = A'$ and $A = O'$. This implies that both O and A are closed sets, i.e. $O = O''$ and $A = A''$. The set of formal concepts of a formal context, ordered with the inclusion relation on either of their components, forms a complete lattice called the *concept lattice* of the formal context. The formal concepts (O, A) can be viewed as classes of objects in which O is the set of objects belonging to the class, called its *extent*, and A is the set of attributes or properties, called its *intent*, that the objects share and that describe the class. The concept lattice organises these classes in a structure that is easy to understand for a human, provided that the lattice is not too large, and allows for efficient algorithms to be applied. A formal concept (O, A) *introduces an object $o \in \mathcal{O}$* if $A = \{o\}'$. It is then called an *introducer concept* [26] and can be seen as the most specific class to which the object belongs.

We construct a formal context $(\mathcal{F}^p, \mathcal{I}, \mathcal{R})$ in which the objects ($\mathcal{F}^p$) correspond to the important features, the attributes ($\mathcal{I}$) correspond to the terms used to describe the criteria and $(f, i) \in \mathcal{R}$ if and only if there is a set $X \in M(f)$ such that $i \in \bigcap_{c_j \in X} \mathcal{I}(c_j)$. In other words, if a feature f is in the Pareto front of a set of criteria that are all interpreted as expressing being good for discrimination, then f will be described as being discriminant. The formal context constructed from our running example is shown in Table III. It states that the feature f_2 is good for sensitivity and specificity, which means that it is predictive, which in turns means that it is good for classification. The formal concepts of this context are composed of a set of features and a set of terms that describe all the features and are, in a sense, classes of important features. The introducer concepts allow for an intuitive visual representation of the hierarchy of these classes to be presented. To facilitate the understanding of the structure, only the features introduced by the concepts are depicted in the figures below.

The concept lattice corresponding to the formal context in Table III and restricted to the introducer concepts is shown in Fig. 3. The feature f_4 is depicted in the extent of the concept $(f_4, \{Accuracy,\ Discrimination,\ Classification\})$. This means that the feature f_4 is good for the accuracy measure, and so that it is discriminant. Similarly, the features f_2 and f_5

	Accuracy	Sensitivity	Specificity	Prediction	Discrimination	Classification
f_1						×
f_2		×	×	×		×
f_4	×				×	×
f_5		×	×	×		×

TABLE III: The formal context constructed from our running example.

are depicted in the concept $(f_2 f_5, \{Sensitivity, Specificity, Prediction, Classification\})$. This means that the features f_2 and f_5 are good for the sensitivity and specificity measures, and so that they are predictive. Lastly, the feature f_1 appears in the concept $(f_1, \{Classification\})$, which means that f_1 is a good compromise between discrimination and prediction and thus good for classification in general.

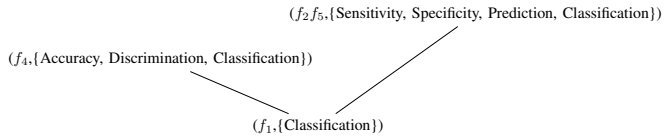

Fig. 3: Formal concepts presenting sets of important features from our running example together with their interpretation.

V. Experimental Results

In this section, we present the results of our method on real datasets. In order to illustrate the method's potential to provide an understanding of the dataset through the set of important features, we use public datasets with meaningful features.

A. Diabetes Data

1) Dataset and Experimental Setup: The Pima Indians Diabetes Database is a public dataset available on the Kaggle machine learning repository[1]. It contains 768 instances, 8 features and two classes : having diabetes (positive class) or not (negative class). We chose to consider four types of classifiers $\mathcal{C}$ = {Random Forests, Naive Bayes, Neural Networks, Support Vector Machines} and four measures $\mathcal{M}$ = {Sensitivity, Specificity, Accuracy, FScore}, the first two quantifying prediction and the last two discrimination so as to preserve a good balance of criteria. For each (classifier, measure) pair, we trained a model using Python's scikit-learn library[2]. For the random forest classifier, we set the number of trees to 20 and the minimum number of samples required to split a node to 1. For the neural network classifier, we used an architecture with two hidden layers containing respectively 10 and 5 neurons, *relu* as the activation function and batch sizes of 20. For the support vector machine classifier, we used a polynomial kernel function of degree 2. All other possible parameters were set to default values. The impacts of features were computed using 100 permutations.

The terms used in the interpretation of criteria and features were *Sensitivity*, *Specificity*, *Accuracy*, *FScore*, *Prediction*, *Discrimination* and *Classification* with the same taxonomy as the one presented in Fig. 2.

[1]https://www.kaggle.com/uciml/pima-indians-diabetes-database
[2]https://scikit-learn.org/

2) Results: The four models have accuracies of 0.75 for random forest, 0.73 for naive Bayes, 0.61 for neural network and 0.79 for support vector machine. The 16 criteria created by the classifiers and measures produce a Pareto front, and thus a set of important features, of size 4. The interpretation of those 4 important features is presented in Fig. 4.

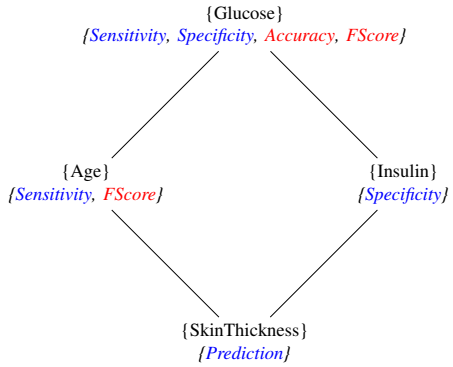

Fig. 4: The interpretation of the 4 important features in the Pima Indians Diabetes dataset. For the sake of legibility, only the most specific terms describing the sets of features are depicted. Terms are written in blue if they are related to prediction and in red if they are related to discrimination.

We observe that the feature *Glucose* is good for every measure and so is both predictive and discriminant. The feature *Age* is good for FScore and so it is discriminant. *Age* is also good for sensitivity, which is a measure of the ability of a model to predict the positive class (having diabetes), so we can say that *Age* is predictive of having diabetes. The feature *Insulin* is good for specificity, which is a measure of the ability of a model to predict the negative class (not having diabetes), so we can say that *Insulin* is predictive of not having diabetes. The feature *SkinThickness* is deemed good for prediction in general.

One can use these results to understand that, in the population described by this dataset, the plasma glucose concentration is the most important value to look at when trying to decide whether someone has diabetes. After the plasma glucose concentration, the insulin level is the most important value for diagnosing diabetes and the age is the most important value for diagnosing not having diabetes. If these features are not enough to decide, the triceps skin fold thickness can also be looked at.

B. Breast Cancer Data

1) Dataset and Experimental Setup: The Breast Cancer Wisconsin (Diagnostic) Data Set is a public dataset available on the UCI machine learning repository[3]. It contains 562 instances, 30 features and two classes characterising breast tumors : malignant (positive class) and benign (negative class). As explained in the description of the dataset, the features were extracted from images depicting cell nuclei. Ten real-valued features were computed for each cell nucleus and the

[3]https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+ %28Diagnostic%29

mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features identified by numbers ranging from 0 to 29.

- 0, 10, 20: radius (mean of distances from center to points on the perimeter)
- 1, 11, 21: texture (standard deviation of gray-scale values)
- 2, 12, 22: perimeter
- 3, 13, 23: area
- 4, 14, 24: smoothness (local variation in radius lengths)
- 5, 15, 25: compactness (perimeter^2/area-1.0)
- 6, 16, 26: concavity (severity of concave portions of the contour)
- 7, 17, 27: concave points (number of concave portions of the contour)
- 8, 18, 28: symmetry
- 9, 19, 29: fractal dimension ("coastline approximation" minus 1)

For instance, the feature 0 is the mean radius, the feature 10 is the standard error of the radius and the feature 20 is the largest radius in the image.

We used the same classifiers, measures, parameters and terms as in the preceding experiment with the diabetes dataset.

2) Results: The four models have accuracies of 0.96 for random forest, 0.89 for naive Bayes, 0.93 for neural network and 0.92 for support vector machine. The 16 criteria created by the classifiers and measures produce a Pareto front, and thus a set of important features, of size 13. The interpretation of those 13 important features is presented in Fig. 5.

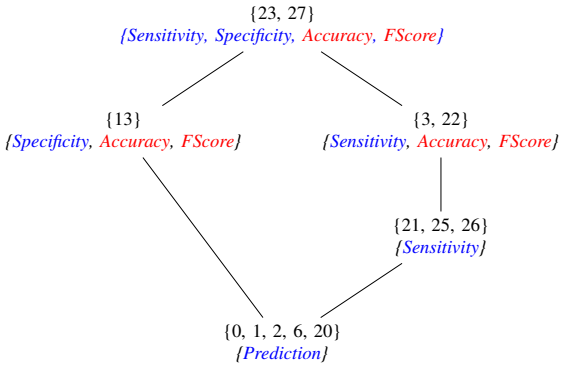

Fig. 5: The interpretation of the 13 important features in the Breast Cancer Wisconsin (Diagnostic) dataset. For the sake of legibility, only the most specific terms describing the sets of features are depicted. Terms are written in blue if they are related to prediction and in red if they are related to discrimination.

We observe that the features 23 and 27 are good for every measure and, thus, are both predictive and discriminant. The feature 13 is good for specificity, which is a measure of the ability of a model to predict the negative class, so we can say that 13 is predictive of a tumor being benign. The features 3, 21, 22, 25 and 26 are good for sensitivity, which is a measure of the ability of a model to predict the positive class, so we can say that those features are predictive of a tumor being malignant. Additionally, the features 3 and 22 are discriminant. The features 0, 1, 2, 6 and 20 are good for prediction in general.

One can interpret these results to mean that the areas of cell nuclei, represented by the features 3, 13 and 23, are particularly important in diagnosing breast cancer. Additionally, the largest, or "worst-case", values extracted from the images, i.e. the features from 20 to 29, seem to be particularly important as most of them are represented in the Fig. 5. The standard errors, with the exception of the radius, seem to be less important in diagnosing breast cancer as they are not in the set of important features.

VI. Discussion and Conclusion

Our approach identifies important features in a dataset and labels them in terms of prediction and discrimination. To reach this result, we have made a number of choices that we discuss in this section.

First of all, we defined predictive and discriminant features through their usage by models learned from data. Whether it be from the selection of the training and test sets or the classifier algorithm itself, nondeterminism is introduced in the first step of the approach. The output is therefore not always the same. From our experiments, it appears that the set of important features is fairly stable while the interpretations are less so. The more precise the interpretation, the more the labels can change. A feature that is found to be particularly good for accuracy at the end of one application of the approach can be found to only be good for discrimination, without more precision, the next time. This randomness can be somewhat reduced by increasing the numbers of classifiers at the cost of computation time.

For the definition of "*most* predictive or discriminant features", we used the membership to the Pareto front of a multicriteria decision problem. We believe that this makes sense but other methods could be considered. In particular, instead of using only the greatest features in the partial order induced on the feature set by the Pareto dominance (i.e. the Pareto front), one could prefer to select all the features up to a given depth in this partially ordered set. With the depth as a parameter, the number of important features could be adjusted. Of course, other types of preferences aggregation could also be used.

Finally, our goal was to identify predictive and discriminant features but, in defining these notions, we also mentioned the existence of measures that quantify the *correctness* of models, i.e. the model's ability not to make mistakes. It would be interesting, as a future work, to integrate this notion into the approach and study the differences between features that are good for avoiding mistakes and features that are good for predicting classes. Similarly, only knowledge on the meaning of measures was used here even though the criteria in the multicriteria decision problem also involve a model that represents a classifier type. Background knowledge on the classifiers themselves could be used in the interpretation of the criteria and the features. For example, a feature could be considered as discriminant by neural networks and predictive by random forests.

ACKNOWLEDGMENT

This work has been partially funded by the LUE project GEENAGE (ANR-15-IDEX-04-LUE).

REFERENCES

[1] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. New York: Springer series in statistics, 2001, vol. 1, no. 10.
[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.
[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, p. 93, 2018.
[4] D. Bouyssou, D. Dubois, H. Prade, and M. Pirlot, *Decision Making Process: Concepts and Methods*. John Wiley & Sons, 2013.
[5] C. C. Aggarwal, M. A. Bhuiyan, and M. Al Hasan, "Frequent pattern mining algorithms: A survey," in *Frequent pattern mining*. Springer, 2014, pp. 19–64.
[6] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*. Springer - Verlag, 1999.
[7] J. Poelmans, D. I. Ignatov, S. O. Kuznetsov, and G. Dedene, "Formal concept analysis in knowledge processing: A survey on applications," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6538–6560, 2013.
[8] P. Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012.
[9] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-based systems*, vol. 8, no. 6, pp. 373–389, 1995.
[10] U. Johansson and L. Niklasson, "Evolving decision trees using oracle guides," in *2009 IEEE Symposium on Computational Intelligence and Data Mining*. IEEE, 2009, pp. 238–244.
[11] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.
[13] J. D. Olden and D. A. Jackson, "Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks," *Ecological modelling*, vol. 154, no. 1-2, pp. 135–150, 2002.
[14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
[15] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
[16] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2006, pp. 1015–1021.
[17] D. M. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
[18] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.
[19] N. Japkowicz and M. Shah, Eds., *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.
[20] M. J. Pencina and R. B. D'Agostino, "Evaluating discrimination of risk prediction models: the c statistic," *JAMA*, vol. 314, no. 10, pp. 1063–1064, 2015.
[21] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, "Assessing the performance of prediction models: a framework for some traditional and novel measures," *Epidemiology*, vol. 21, no. 1, pp. 128–138, 2010.
[22] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance," *arXiv preprint arXiv:1801.01489*, 2018.
[23] D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukias, and P. Vincke, *Evaluation and Decision Models with Multiple Criteria: Stepping Stones for the Analyst*. Springer Science & Business Media, 2006, vol. 86.
[24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[25] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.
[26] A. Berry, A. Gutierrez, M. Huchard, A. Napoli, and A. Sigayret, "Hermes: a simple and efficient algorithm for building the AOC-poset of a binary relation," *Annals of Mathematics and Artificial Intelligence*, vol. 72, no. 1-2, pp. 45–71, 2014.