Towards Explainable Multi-Label Classification

Karim Tabia

Towards explainable multi-label classification

(Preprint version)

Karim Tabia

Univ Lille Nord de France, F-59000 Lille, France
Univ. Artois & CNRS, F-62300 Lens, France
tabia@cril.fr

***Abstract*—Multi-label classification is a very active research area and many real-world applications need efficient multi-label learning. During recent years, explaining machine learning predictions is also a very hot topic. A lot of approaches have been proposed for explaining multi-class classifier predictions. However, almost nothing has been proposed for multi-label and ensemble approaches. This paper brings two main contributions. It first proposes a natural framework consisting in reasoning with base classifier explanations in order to provide explanations for the multi-label predictions. The second contribution focuses on binary relevance, a widely used approach in multi-label classification, and distinguishes two kinds of explanations:** *common explanations* **shared by all base classifiers predicting positive labels and** *joint explanations* **combining explanations from each base classifier predicting a positive label. The paper proposes an efficient approach for deriving such explanations. Experimental studies show positive results that can be achieved on many multi-label datasets.**

***Index Terms*—Multi-label classification, explanation, binary relevance**

I. Introduction

Explaining machine learning predictions and AI systems is not a recent topic [1] but it has peaked after the rise and success of the last few years of machine learning techniques. Roughly, generating explanations consists of linking data and predictions in a way that is interpretable and understandable by the user [12]. Often, what can be generated as explanations depends on the model and the objectives of the application considered. Moreover, there is no consensus at the moment on the notions of quality of an explanation and no consensus on formal quality properties of explanations [6], [9].

There are now several explanation approaches in machine learning but they are all dedicated to the multi-class classification problem (where a data instance is associated with a single class). Almost nothing has been proposed to explain the multi-label classification techniques (where we associate a subset of labels to the data instance to classify) and ensemble approaches (except few works [7] on explaining Random Forest classifiers).

This work is intended to contribute filling this gap by proposing an approach to explain the predictions of a multi-label system. This is an attempt to extend a symbolic explanation approach from the case of multi-class classification to the multi-label case. First, we propose a natural framework for reasoning with explanations of base classifiers to derive explanations for multi-label predictions. Some interesting and natural properties to be satisfied in this context are also proposed. In a second step, we extend the approach based on knowledge compilation for the multi-class explanation case to the multi-label case. Our approach builds upon this symbolic approach to generate explanations for Binary Relevance (BR), a widely used multi-label classification technique. We start with extending the definition of explanations called PI (Prime Implicants) to the multi-label case. We then define two types of explanations: Common explanations (CE) based on selecting shared subsets of base classifiers explanations and joint ones (JE) obtained by joining base classifiers explanations. We propose an efficient procedure to derive the BR explanations guaranteeing interesting performances especially in terms of size of the target representations and number of explanations compared to base classifiers ones.

The rest of this paper is organized as follows: Section 2 fixes the notations that will be used along with this paper. Section 3 briefly recalls the multi-label classification problem then focuses on the Binary Relevance approach to multi-label classification. In Section 4, we present our approach for explaining multi-label BR predictions. Section 5 provides our experimental study while Section 6 provides discussions and concluding remarks.

II. Notations and Definitions

Along with this paper, we will use the following notations:

- A multi-label classification problem is defined by two sets of variables: feature space X=$\{X_1, .., X_n\}$ and label space Y=$\{Y_1, .., Y_k\}$ (the label space is interchangeably denoted as L=$\{l_1, .., l_k\}$) consisting of k binary variables encoding the presence/absence of the k labels. This is a common representation of multi-label classification problems. X_i denotes i^{th} feature while x_i denotes a value that can taken by X_i. l_j denotes the j^{th} label. We say that a label is positive if l_j=1 and we will use interchangeably *positive/1* and *negative/0*.
- For each input data instance x, a multi-label classifier is a function predicting $f(x)$=y. For the sake of simplicity, the features are also binary variables.

- An instance x is a complete assignation of values for all variables of X. A partial instance z is a subset of a complete instance x denoted $z \subseteq x$. For example if $X=\{X_1, X_2, X_3\}$ then a complete instance of X could be $x=(x_1=1, x_2=1, x_3=0)$ and a partial instance z could be $z=(x_1=1, x_3=0)$. Where there is no ambiguity, we'll simply write $x=110$.
- Also, the labels y of a data instance x will be compactly denoted when there is no ambiguity by a set instead of a vector. For is instance, instead of writing $y=(1,0,0,0,1)$ (here $|Y|=5$), we will write $y=\{y_0, y_4\}$.

III. Multi-label explanations

A. Multi-label classification

Multi-label classification is a well-known predictive task in many real-world problems such as text categorization where each document can belong at the same time to several predefined topics (for example, a newspaper article may at the same time be classified as *sport* and *science*). We find this problem in different application areas such as objects recognition in images, sentiment analysis in social network data, etc. In multi-class classification, classes are mutually exclusive, while for multi-label problems, classes do not exclude each other allowing the same input instance to be classified into multiple classes at the same time. A multi-label classification problem is formally defined by a set of feature variables $X=\{X_1, .., X_n\}$ and label (binary) variables $Y=\{Y_1, .., Y_k\}$. A classifier is a function mapping each instance x of X to y, instance of Y. Abusing notation, we denote y the subset of Y set positively (only predicted labels). A dataset in multi-label classification problems is a collections of couples $<x,y>$ where x is an instance of X and y an instance of Y.

Example 1: Assume a multi-label classification problem where web pages are labeled in one or more categories (labels). For the sake of simplicity, assume that each web page is described by a set of keywords (content or metadata keywords for instance). Hence, using a binary bag-of-words representation, each web page will have a set of binary features where feature X_i=1 (resp. X_i=0) denotes that keyword X_i is present (resp. absent) in the web page content or metadata. Similarly, label variable Y_j=1 (resp. Y_j=0) denotes the fact the current web-page is positively labelled (resp. not labelled) in category Y_i. In the example of Table I, the feature space is $X=\{A,B,C,D,E,F\}$ composed of six binary variables A, B, C, D, E, F and three label variables $Y=\{Y_1,Y_2,Y_3\}$. A dataset can be in the form shown in Table I:

$X=\{A,B,C,D,E,F\}$	$Y=\{Y_1,Y_2,Y_3\}$
0 1 1 0 1 1	1 1 0
1 1 1 0 0 1	1 0 0
0 1 1 0 1 0	0 1 0
1 1 1 0 1 0	1 1 1
0 1 0 0 0 1	1 0 1
0 1 1 1 1 0	1 1 1
...	...
0 1 1 0 0 1	1 0 0

TABLE I
EXAMPLE OF MULTI-LABEL DATASET

Regarding multi-label classification techniques, there are three main categories:

1) **Problem transformation approaches** where the multi-label classification problem is transformed into a set of multi-class classification or mono-label regression problems. Examples of methods in this category are Binary Relevance (BR),) Classifier Chains (CC) and Label Powerset (LP). In general, problem transformation methods rely on binary classifiers to predict labels individually and then use a combination strategy to make the final prediction.
2) **Method adaptation approaches** based on extending multi-class techniques to predict instead of one single class a set of relevant labels. Examples of this category are ML-kNN [14], ML-C4.5 [4].
3) **Ensemble approaches** that combine ideas from the two first categories. RAndom k labEL sets (RAkEL), Hierarchy Of multi-label classifiERs (HOMER), Ensemble of Classifier Chains (ECC) and Ensemble of Binary Relevance (EBR). An ensemble approach is built upon a set of weak binary or multi-class classifiers, then the outputs of base classifiers are usually combined by weighted or unweighted averaging.

Another difference worth mentioning compared with the multi-class case is related to evaluation metrics used to assess the accuracy of multi-label techniques. Indeed, standard multi-class classification are no more enough and appropriate measures are specifically designed for this purpose (example of measure used in multi-label classification is the Hamming-Loss).

B. Binary relevance approach

Binary relevance is the main baseline for multi-label classification methods [10]. It is based on the label independence assumption which may be seen as a strong assumption and not verified in many domains. Despite this fact, a lot of studies highlighted the interesting properties and nice performances of this method [10]. The strategy of BR is to break the multi-label learning problem into a set of binary classification problems, one per label. The label independence assumption allows to learn each individual model f_i independently, using only the data of the label l_i. As said earlier, in spite of the fact that BR does not take into account label dependencies, BR has several obvious advantages such as linear complexity (w.r.t the number of labels), possibility of parallelization, good accuracy, etc. making it the main baseline method for assessing multi-label approaches.

C. Natural properties of BR explanations

Let us first focus on some natural properties that explanations should have in the context of multi-label classification.

- **Minimality**: Require only the minimal subset of x that will trigger the prediction y=$f(x)$.
- **Unanimity**: If an explanation e is provided for each predicted label then e should also be an explanation for the multi-label prediction. Namely, if $e \in exp(f_i(x)=1)$ then $e \in exp(f(x))$. Here $exp(f(x))$ denotes the set of explanations for data instance x using classifier f.
- **Decomposability**: In the context of multi-label classification, explaining a prediction y should lead to explain each label composing y.
- **Explanation independence**: If labels are (assumed) independent, then so should be individual label explanations. Namely, if a label l_i can be predicted with any other label l_j, then any base classifier explanation $e_i \in exp(f_i(x))$ could come with any other base classifier explanation $e_j \in exp(f_j(x))$.

While the *Minimality* property is not specific to multi-label tasks, the *Unanimity*, *Decomposability* and *Explanation independence* properties are particularly relevant for BR approach and multi-label approaches more generally. For instance, the *Explanation independence* property naturally follows from the label independence assumption that is the basis of BR.

IV. From multi-class explanations to multi-label explanations

Our approach for explaining BR instance predictions is in line with the BR schema, namely we first explain individual label predictions by individual base classifiers then *use/combine* such explanations to build those explaining the multi-label prediction.

A. Explaining individual/base classifier predictions

Let us first recall the formal definition of classifier instance explanation our BR explanation approach is lying on and rephrase it in the context of multi-label tasks. As mentioned in the introduction, our approach for explanation in the multi-label case is based on the extension of a recently proposed symbolic approach for Bayesian classifiers [13]. In this approach, the authors propose two categories of explanations: Minimum Cardinality (MC) and Prime Implicant (PI) explanations. MC explanations are special case of PI explanations. Formally,

Definition 1 (PI explanation [13]): Let $f(X)$ denote the decision function associated to a classifier. A partial instance z of x is a PI-explanation of $f(x)$ if

- (a) $z \subseteq x$,
- (b) $f(x) = f(x^*)$ for every instance x^* such that $z \subseteq x^*$, and
- (c) no other partial instance $y \subset z$ satisfies (a) and (b).

Intuitively, an PI explanation identifies which part of the instance x suffices to give the prediction $f(x)$. Hence filling arbitrarily in the remaining attributes will not change the classifier prediction.

In [13], the authors propose i) first compiling a Bayes network classifier decision function into an equivalent decision function in the form of an Ordered Decision Diagram (ODD[1]) which is a tractable representation of a decision function and ii) use a polynomial algorithm in the size of the ODD to compute for each data instance x its PI explanations (we refer to this algorithm $PIAlgo$ in the following[2]).

The authors in [13] show experimentally that compiling Bayes network classifiers into ODDs can be handled efficiently and the number of PI explanations remains reasonable. In our work, we will rely on this symbolic approach for deriving PI explanations of base Bayes classifiers used by the BR method.

B. Explaining BR predictions

Our approach for explaining a BR multi-label classifier is to first derive base classifier explanations then reason with them to infer BR explanations. Basically, one can do two kinds of tasks with base classifiers explanation:

- i) selecting base classifier explanations that can explain all the predicted labels for a data instance x or
- ii) combining base classifier predictions to build the BR explanations.

Before extending and applying this approach to multi-label classification, let us first adapt Definition 1 for the case of multi-label task.

Definition 2 (BR PI explanation): Let $f(X)$ denote a BR multi-label classifier. A partial instance z of x is a PI-explanation of $f(x)$=y if

- (a) $z \subseteq x$,
- (b) $f(x) = f(x^*)$ for every instance x^* such that $z \subseteq x^*$, and
- (c) no other partial instance $z' \subset z$ satisfies (a) and (b).
- (d) for each positively predicted label l_i in y=$f(x)$, there exists a partial instance $z_i \subseteq z$ such that z_i is a PI explanation of $f_i(x)$ (in the sense of Definition 1).

Condition (b) aims to ensure that all (complete) instances x^* containing the partial instance z will be associated with the same multi-label prediction while condition (c) ensures the minimality of z in terms of the number of variables involved in z. Condition (d) ensures that a BR explanation e for an instance x should include a PI explanation e_i for each base classifier prediction $f_i(x)$=1.

Now given an input instance x to classify, one may be interested in the following explanations

- *Common explanations (CE):* A common (or universal) explanation is provided by all base classifiers predicting positively for the data instance x. This is a simple selection strategy for inferring BR explanations from base classifiers' ones.

[1] An ODD is a rooted directed acyclic graph representing a Boolean function. The nodes of an ODD consist of *variable nodes* depending on the modeled function and two *value nodes*. The value of the modeled function for a given variable instantiation is determined by traversing the ODD from its root to a value node.

[2] $PIAlgo$ refers to **Algorithm 5** encoding PI explanation given an ODD (see [13] for more details).

- *Joint explanations (JE):* A joint explanation involves exactly one explanation from each individual classifier f_i explanations where the label l_i is predicted positively. Here, it is a simple combination strategy for inferring BR explanations from base classifiers' ones.

By definition, both of CE and JE explanations can explain a BR prediction. While the number of JE explanations can be very large, the number of CE ones can be very small as we will see in our experimental study. In addition to CE and JE explanations, one may be interested in other types of explanations such most frequent or smallest ones (in terms of the number of involved variables). Let us focus more closely on CE and JE explanations.

1) Common explanations (CE):

Definition 3: Let $f(X)$ denote a BR multi-label classifier where base classifiers are denoted f_i for i=1..k. The set of common explanations is defined as follows: $CE(x)$=$\{e \in \cap_{i|f_i(x)=1} PI_i(x)\}$ where $PI_i(x)$ denotes the set of PI explanations provided for the base classifier prediction $f_i(x)$=1.

Definition 3 defines a CE explanation of an instance x as the intersection of PI explanations of base classifiers predicting labels positively. The intersection operation here denotes the set intersection operation.

Lemma 1: Let $f(X)$ denote a BR multi-label classifier where base classifiers are denoted f_i for i=1..k. Let also $CE(x)$ be defined according to Definition 3, then $\forall e \in CE(x)$, e is a BR PI explanation and satisfies conditions (a)-(d) of Definition 2.

It is obvious that if e is a PI explanation of all base classifiers f_i predicting positively l_i then e is a PI explanation for the BR classifier prediction and conditions (a)-(d) of Definition 2 are satisfied.

As it will be shown in the experiments, the number of CE explanations could be small especially if labels are not overlapping. Indeed, if the labels do not overlap (share data instances), the predictions and explanations are very likely not to coïncide.

2) Joint explanations (JE):

Definition 4: Let $f(X)$ be the BR multi-label classifier. A prediction y for an instance x is provided by k binary base classifier $f_1(X)$,..,$f_k(X)$ (each label l_i is positively predicted or not by the corresponding binary classifier $f_i(X)$). Let $PI_i(x)$ be the set of PI explanations of $f_i(x)$=1. Let $JE(x)$ be the set of explanations obtained as follows: $JE(x)$=$\{\wedge_{i=1..k|y_i=1} e_i \in PI_i(x)\}$.

A joint explanation e is obtained by combining a PI explanation from each base classifier predicting positively using the logical conjunction operation. For instance, let e_1=$x_3\bar{x_6}x_7$ be an explanation provided by a classifier f_1 and e_2=$\bar{x_2}x_3$ be another explanation provided by a classifier f_2 then conjoining e_1 and e_2 gives e_1e_2=$\bar{x_2}x_3\bar{x_6}x_7$.

Example 2: Assume the multi-label classification problem of Example 1.

In the example of Table II, the classifiers f_1 and f_3 predicted positively for the instance x=$(1,0,1,1,0,0)$. Classifier f_1 has three PI explanations for predicting postively x, namely $PI_1(x)$=$\{\bar{b}cd, \bar{b}c\bar{f}, cd\bar{f}\}$ and f_3 has also three PI explanations for x that are $PI_3(x)$=$\{\bar{b}d, d\bar{e}, d\bar{f}\}$. Joining PI explanations of both classifiers f_1 and f_3 gives nine joint explanations $JE(x)$=$\{\bar{b}cd, \bar{b}cd\bar{e}, \bar{b}cd\bar{f}, \bar{b}cd\bar{f}, \bar{b}cd\bar{e}\bar{f}, \bar{b}cd\bar{f}, \bar{b}cd\bar{f}, cd\bar{e}\bar{f}, cd\bar{f}\}$.

X={A,B,C,D,E,F}	Y={Y_1,Y_2,Y_3}	$PI_i(x)$	JE(x)
1, 0, 1, 1, 0, 0	101	PI_1=$\{\bar{b}cd, \bar{b}c\bar{f}, cd\bar{f}\}$ PI_2=$\{\}$ PI_3=$\{\bar{b}d, d\bar{e}, d\bar{f}\}$	$\{\bar{b}cd, \bar{b}cd\bar{e}, \bar{b}cd\bar{f}, \bar{b}cd\bar{f}, \bar{b}cd\bar{e}\bar{f}, \bar{b}cd\bar{f}, \bar{b}cd\bar{f}, cd\bar{e}\bar{f}, cd\bar{f}\}$

TABLE II
EXAMPLE OF BASE CLASSIFIER EXPLANATIONS AND BR JOINT EXPLANATIONS

Clearly, if $PI_i(x)$ is the set of PI explanations provided for the classifier f_i for data instance x for each classifier f_i predicting positively, then the number of distinct joint explanations is at most $\prod_{i|f_i(x)=1} |PI_i(x)|$. This can be very large in multi-label problems with large feature sets.

Now, there remains two main questions: i) Are JE explanations BR PI ones and ii) how to compute them? Proposition 1 answers the first question:

Proposition 1: Let $f(X)$ be the BR multi-label classifier. A prediction y for an instance x is provided by k binary base classifier $f_1(X)$,..,$f_k(X)$ (each label l_i is positively predicted or not by the corresponding binary classifier $f_i(X)$). Let $PI_i(x)$ be the set of PI explanations of $f_i(x)$=1. Let $JE(x)$ be the set of explanations obtained using Definition 4, then explanations from $JE(x)$ are not guaranteed to satisfy condition (c) of Definition 2.

Proposition 1 states that combining base classifier PI explanations following Definition 4 does not guarantee to give BR PI explanations. Especially, the subset of features is not minimal (condition (c) of Definition 2). Following is a counter-example showing Proposition 1.

Example 3: Let us reuse joint explanations of Example 2. In this example both $\bar{b}cd$ and $\bar{b}cd\bar{e}$ are JE obtained by combining PI explanations of f_1("101100") and f_3("101100"). Clearly, $\bar{b}cd\bar{e}$ is not minimal since $\bar{b}cd$ is a JE with a smaller size.

This counter-example leads to the question of deriving only JE that are BR PI explanations. We provide in the following an efficient method for computing only BR PI explanations.

C. Computing BR PI explanations

The following proposition allows to derive only multi-label BR PI explanations from base classifiers PI explanations. The key idea is to take advantage of the nice properties of ODDs that can be manipulated efficiently through some (logical) operations such as conjunction, disjunction and negation. Then, in order to enumerate PI explanations of a set of classifiers predicting positively for an instance x, it is enough to first conjoin the ODDs of these base classifiers then apply the same PI explanation algorithm ($PIAlgo$) on the resulting ODD to output the BR PI explanations. Formally,

Proposition 2: Let ODD_i denote the ODD encoding the decision function of classifier f_i. Let ODD_f=$\wedge_{i=1..k|f_i(x)=1} ODD_i$ where $\wedge$ denotes the ODD conjunction operation. Then prime implicants of ODD_f obtained applying $PIAlgo$ are BR PI explanations.

Note that the ODDs that will be conjoined are only those corresponding to labels predicted positively (this number is in practice very low compared to the size of the label space) and that all the ODDs share the same variables (since all base classifiers share the same feature space). Hence conjoining a series of ODDs with n variables will result in an ODD with exactly n variables.

Proof 1 (Proof sketch): The idea of the proof is that the result of conjoining two ODDs is an ODD. This operation is associative and conjoining a series of ODDs will output an ODD. Thanks to the PI encoding of algorithm $PIAlgo$ proposed in [13], this algorithm applied on the obtained ODD_f will output only prime implicants of ODD_f. It is obvious that any prime implicant of ODD_f is also satisfying every used ODD_i (since ODD_f is the conjunction of a set of ODD_i). It can be easily shown that every prime implicant of ODD_f satisfies conditions (a)-(d) of BR PI explanations.

V. Experimental results

We report in this section some experimental results highlighting the main behavior of the BR explanation approach proposed in this paper. The experiments are carried out on a set of synthetic datasets such that we can easily vary some parameters and see the effects in terms of explanations, size of target representations, etc. The issues we want to highlight in particular are:

- *Number of explanations*: In particular, we want to compare the number of PI explanations of base classifiers with the number of BR explanations (joint explanations JE and common explanations CE).
- *Size of target ODD representations*: The aim here is to compare the size of ODDs (the size of an ODD is the number of its nodes) of base classifiers predicting positively and the size of their conjunction encoding the BR classifier.

A. Experimentation setup

- *Datasets*: The datasets used in this study are generated with different characteristics such as the number of features, number of labels, etc. In all the used datasets, all the features are binary. The cardinality[3] of the datasets is set to 25% of label set size (this is not a very common rate in multi-label learning datasets but our aim is to let the BR classifier to predict in average many labels per instance). Each experiment involves 5000 (resp. 10000) samples in datasets with 10 and 20 (resp. 30) features. For the explanations, 1000 data instances are randomly generated and their explanations enumerated. Larger test sets does not lead to significant variations in the results. Experiments with datasets with different cardinalities/densities may reveal some interesting behaviors but due to page limit such experiments are not covered in this paper.
- *Base classifiers*: The base classifiers used in our experimental study are naive Bayes classifiers [8] as algorithms compiling such classifiers to ODDs exist and they are very efficient [2], [13]. Of course, non naive Bayes classifiers such as latent-tree classifiers could also be used [13] but due to page limit, we focus only on the comparison between base classifiers explanations and the BR ones. Future works will deal with other types of base classifiers.

[3]Cardinality and density are among most relevant features in multi-label datasets. Cardinality (Card) refers to the average number of labels per data instance while the density denotes the cardinality divided by the size of the label set (namely Density=$\frac{Card}{|Y|}$).

B. Results

Fig. 1 shows main results of experiments done on multi-label datasets with 20 features. The number of labels is fixed to 5, 10, 15 and 20. The main findings are summarized in the following:

- *Average number of positive predictions*: The average number of positive predictions $Avg\_\#labels$ (upper left graph in Fig. 1) is low and this is in line with multi-label learning where datasets have low cardinality and density which is the case of most multi-label datasets. This parameter does not improve significantly as the number of labels is increased. Such rate is also observed almost over all the experiments conducted in this study. This result is important to understand the other results our approach.
- *Size of target ODD representations*: The aim here is to compare the size of ODDs of base classifiers predicting positively and the size of these ODDs conjunction encoding the BR classifier. The curves of upper right graph in Fig. 1 clearly shows two main findings:
  - First, the size of ODDs is very tractable for our 20 features datasets and does not explode as the number of labels is increased. Indeed, the ODDs size does not exceed 1400 nodes. This is mainly due to the fact that the number of base classifiers predicting positively does not grow significantly with the number of labels.
  - Second, the size of ODD encoding the BR classifier is always smaller than the sum of base classifier ODDs sizes. Indeed, conjoining ODDs can lead to smaller ODDs as size optimization operations on ODDs allow to take advantage of some redundancies in input ODDs to output an ODD with optimized size. This is particularly important when dealing with multi-label problems with very large labels set. This finding is also observed over all the experiments reported in this paper.
- *Number of explanations*: Here we comment and compare the number of explanations obtained form base classifiers, number of common explanations (CE) and finally the number of joint explanations (JE).

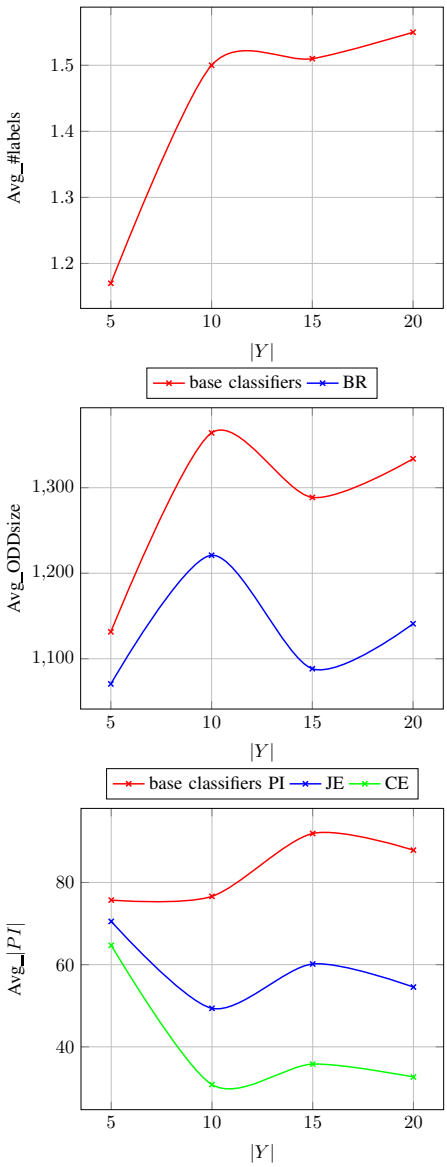

Fig. 1. Experiments on datasets with 20 features and different label set size (5, 10, 15, 20)

- The average number of all explanations provided by base classifiers (each base classifier predicting positively will provide its explanations for the instance in hand) is very tractables and does not seem to depend on the size of the label set. Indeed, this is more dependent on the number of features of the dataset.
- The number of common explanations CE is low as expected since it is the intersection of explanation sets provided by base classifiers. More importantly, this result is mainly due to the fact that the majority of test instances x are associated to only one label. Consequently, the number of common explanations is exactly all the set of PI explanations provided by the only base classifier having predicted positively the instance x in hand. This number is almost zero for all test instances where at least two labels are predicted positively. This also can be interpreted as due to labels that do not overlap.
- Regarding the number of BR explanations, the results of lower graph in Fig. 1 clearly show that their number is in average lower than the average number of cumulated PI explanations of base classifiers. The results are due to the fact that the majority of test data instances are associated with a small number of labels and that due to the fact that combining base classifiers PI explanations does not guarantee to give a BR PI explanations as stated by Proposition 1.

In addition to the three criteria highlighted above, one could be interested in other behaviors such as the size of explanations (number of features involved in explanations) and the computation times. The Discussions and concluding remarks section gives some insights into such issues.
Let us report other results carried out on datasets with larger feature and label sets.

The main findings of experiments of Fig. 2 are in accordance with those reported in Fig. 1 except for the size of target representations where the BR ODD is slightly bigger than the cumulated size of base classifier ODDs. Note that the graph denoting the number of explanations (lower graph in Fig. 2) shows that the curves of cumulated explanations number of base classifiers almost coincides with the BR explanations (more precisely, the former is slightly bigger than the latter). In the following, we report results where we fix the number of labels and vary the feature space size.

The main finding in the results of Fig. 3 is the significant increase in the size of target representations (both base classifier ODDs and BR ones) while the curves almost coïncide over all the tested datasets. The second main finding concerns the number of explanations where the average number of explanations of BR classifier tends to coïncide with the cumulated number of PI explanations of base classifiers.

To summarize the results, it can be said that the approach proposed for BR provides the expected results both regarding the size of the representations and in terms of the number of explanations. This approach which is sound and complete does not induce experimentally significant extra computational costs compared with the used explanation approach for the base classifiers. As long as this latter provides explanations, our BR approach will provide explanations. Finally, the obtained results suggest that CE explanations are probably very exceptional when the number of predicted labels per instance is greater than 1, but this remains to be confirmed on benchmarks with different properties.

VI. Discussions and concluding remarks

This paper is the first attempt to extend a symbolic classifier explanation approach from the case of multi-class classification to the multi-label case. First, we propose a framework

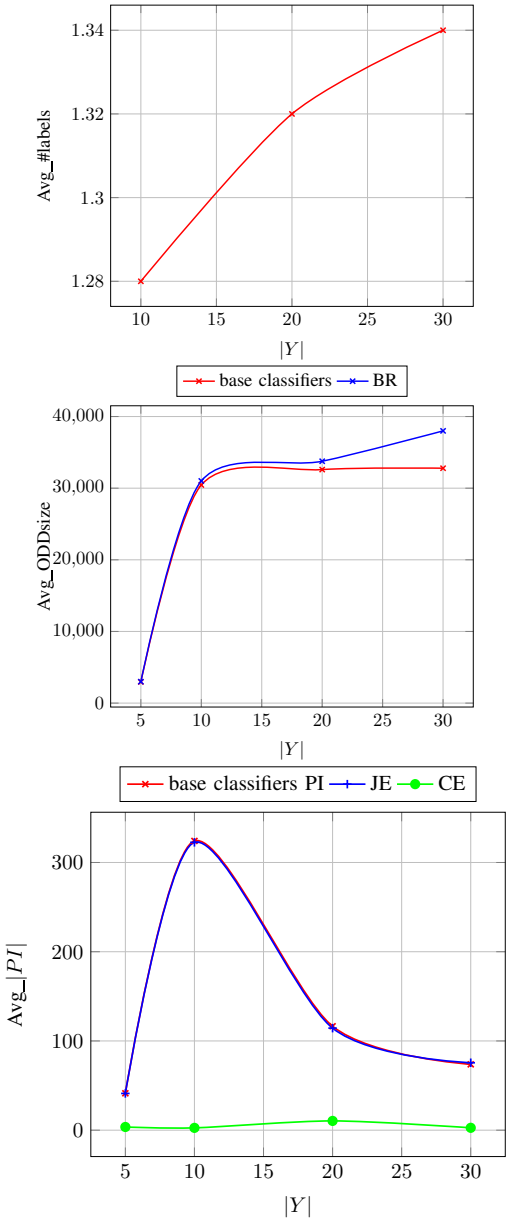

Fig. 2. Experiments on datasets with 30 features and different label set size (5, 10, 20, 30)

for reasoning with explanations of base classifiers' predictions to derive explanations for multi-label predictions. Some interesting properties to be satisfied in this context are also proposed. We then extend an efficient symbolic approach (based on knowledge compilation to tractable representations) for explaining multi-class classifier predictions to the Binary Relevance (BR) approach, one of the most widely used approaches for multi-label tasks. Indeed, the main baseline method for multi-label learning is BR. This latter is often criticized for its label independence assumption but a lot of works have shown its very interesting properties [10]. Our contribution provides a new interesting property which is the possibility of explaining BR predictions.

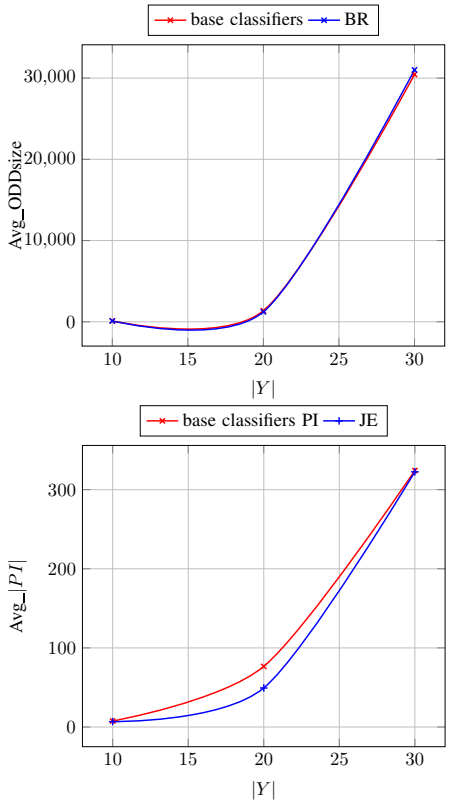

Fig. 3. Experiments on datasets with 10 labels and different features set size (10, 20, 30)

In order to present our approach, we start with extending the definition of explanations called PI (Prime Implicants) to the multi-label case, then we define two types of explanations: common explanations (CE) and joint ones (JE). We propose an efficient procedure to derive the BR explanations while guaranteeing interesting performances especially in terms of size of the target representations and number of explanations compared to base classifiers ones. Clearly, CE explanations are a special case of JE explanations. Among the properties defined in Section III-C, JE explanations obviously satisfy the properties of *Minimality*, *Decomposability* and *Unanimity*.

As stated earlier in this paper, the proposed approach for BR explanation is an extension of a symbolic approach for Bayes network classifiers [13]. Our approach for BR is though not limited to using Bayes network classifiers as base classifiers. Indeed, as long as the decision function of a classifier can be compiled into a tractable representation such as OBDD, ODD or SDD [5], our approach can be applied. Indeed, such representations can be conjoined efficiently and PI explanation encoding for these representations exists [13]. Recently, in [3] the authors propose an approach for compiling neural networks into a tractable representation that can be used in our BR approach. Note also that our approach can be applied for any discrete feature set thanks to the use of ODDs.

The paper focused only on explaining predicted labels for a given data instance. This is called *positive explanations* in

[13] where it is shown also how to obtain very explanations for non predicted labels (called *negative explanations*) just by negating the ODDs of classifiers predicting negatively and outputting the prime implicants of the negated ODDs. This remains also valid for our approach where the ODD encoding the BR classifier could be negated and explanations generated after this operation.

Our approach for BR explanations can be adapted and applied to explaining some other multi-label approaches and ensemble methods. For instance, Classifier Chains [11] which are one of the well-known multi-label techniques can be explained using our approach as chain classifiers model label (in)dependences as a Bayes network which can also be compiled into an ODD. Regarding ensemble methods used in multi-class problems, explanations provided by base classifiers can be used to help aggregating such predictions and output one final prediction by the ensemble approach.

Acknowledgement This work has been supported by the European project H2020, Marie Sklodowska- Curie Actions (MSCA), Research and Innovation Staff Exchange (RISE): Aniage project (High Dimensional Heterogeneous Data Based Animation Techniques for Southeast Asian Intangible Cultural Heritage Digital Content), project number 691215.

REFERENCES

[1] U. Chajewska and J. Y. Halpern. Defining explanation in probabilistic systems. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, UAI'97, pages 62–71, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[2] H. Chan and A. Darwiche. Reasoning about Bayesian network classifiers. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 107–115, San Francisco, California, 2003. Morgan Kaufmann Publishers.

[3] A. Choi, W. Shi, A. Shih, and A. Darwiche. Compiling neural networks into tractable Boolean circuits. In *AAAI Spring Symposium on Verification of Neural Networks (VNN)*, 2019.

[4] A. Clare and R. D. King. Knowledge discovery in multi-label phenotype data. In *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '01, pages 42–53, Berlin, Heidelberg, 2001. Springer-Verlag.

[5] A. Darwiche. Sdd: A new canonical representation of propositional knowledge bases. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 819–826. AAAI Press, 2011.

[6] D. Doran, S. Schulz, and T. R. Besold. What does explainable AI really mean? A new conceptualization of perspectives. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), Bari, Italy, November 16th and 17th, 2017.*, 2017.

[7] M. W. Dragutin Petkovic, Russ Altman and A. Vigil. Improving the explainability of random forest classifier - user centered approach. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 204–215, 2018.

[8] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Mach. Learn.*, 29(2-3):131–163, Nov. 1997.

[9] Z. C. Lipton. The mythos of model interpretability. *Queue*, 16(3):30:31–30:57, June 2018.

[10] O. Luaces, J. Díez, J. Barranquero, J. J. del Coz, and A. Bahamonde. Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, 1(4):303–313, Dec 2012.

[11] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Mach. Learn.*, 85(3):333–359, Dec. 2011.

[12] 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*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[13] A. Shih, A. Choi, and A. Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 5103–5111, 2018.

[14] M.-L. Zhang and Z.-H. Zhou. A k-nearest neighbor based algorithm for multi-label classification. In *2005 IEEE International Conference on Granular Computing*, volume 2, pages 718–721 Vol. 2, July 2005.