Knowledge and Information Systems (2023) 65:5657–5694
https://doi.org/10.1007/s10115-023-01930-6

REGULAR PAPER

Concise and interpretable multi-label rule sets

Martino Ciaperoni[1] · **Han Xiao**[1] · **Aristides Gionis**[2]

Received: 28 January 2023 / Revised: 3 June 2023 / Accepted: 7 July 2023 /
Published online: 28 July 2023

Abstract
Multi-label classification is becoming increasingly ubiquitous, but not much attention has been paid to interpretability. In this paper, we develop a multi-label classifier that can be represented as a concise set of simple "if-then" rules, and thus, it offers better interpretability compared to black-box models. Notably, our method is able to find a small set of relevant patterns that lead to accurate multi-label classification, while existing rule-based classifiers are myopic and wasteful in searching rules, requiring a large number of rules to achieve high accuracy. In particular, we formulate the problem of choosing multi-label rules to maximize a target function, which considers not only discrimination ability with respect to labels, but also diversity. Accounting for diversity helps to avoid redundancy, and thus, to control the number of rules in the solution set. To tackle the said maximization problem, we propose a 2-approximation algorithm, which circumvents the exponential-size search space of rules using a novel technique to sample highly discriminative and diverse rules. In addition to our theoretical analysis, we provide a thorough experimental evaluation and a case study, which indicate that our approach offers a trade-off between predictive performance and interpretability that is unmatched in previous work.

Keywords Multi-label classification · Rule-based classification · Rule sampling · Interpretable machine learning

1 Introduction

Machine-learning algorithms are nowadays being used in almost every domain. While such algorithms are known to perform well in many tasks, they are often used as "black-boxes," i.e., the decision processes involved are too complex for humans to interpret. Black-box

✉ Martino Ciaperoni
martino.ciaperoni@aalto.fi

Han Xiao
han.xiao@aalto.fi

Aristides Gionis
argioni@kth.se

[1] Department of Computer Science, Aalto University, Espoo, Finland

[2] Division of Theoretical Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

machine-learning algorithms only offer a partial representation of the reality, and moreover, their performance is assessed using metrics that only capture particular aspects of the real world. The lack of interpretability considerably limits the level of trust humans put in black-box machine-learning algorithms and thus poses a barrier for the wide adoption of machine-learning techniques in real-world applications. In many applications, particularly when machine learning is used to aid in high-stake decision making, interest lies not only in having accurate predictions, but also in extracting information from the learned model that can be analyzed by human expert to produce valuable insight. In an attempt to overcome this barrier, interpretable and explainable machine learning has recently emerged as increasingly prominent topics, and in this context, classification problems hold a central position. In the standard classification setting, the goal is to learn a classifier that accurately maps data points to two or more *mutually exclusive* classes.

In this paper, we focus on a different setting, namely *multi-label classification*. In contrast to the standard single-label setting, in multi-label classification, a point can be associated with more than one class at the same time. Due to the increasing complexity of modern data, multi-label classification is becoming more and more popular in recent years, and it has been extensively studied. Nonetheless, the main focus is still on improving predictive performance [1]. Significantly less attention has been paid to interpretability.

Classification rules, due to their simple structures, are gaining popularity in interpretable multi-label classification literature. In rule-based approaches, the goal is to learn a set of rules that captures the most prominent associative patterns on the features and labels in the data. A rule usually takes the form "({a set of predicates} $\rightarrow$ {a set of labels})".

For a given data point, a rule would usually predict the associated labels to be present, if all the predicates in the rule evaluate to true.

Due to the structural simplicity of rules, classifiers based on a set of rules are generally considered more interpretable than other types of classifiers, such as neural networks or even decision trees.

The research question that we bring forward is whether we can design rule-based multi-label classification methods that are both accurate and interpretable. BOOMER [2, 3], a recently proposed rule-based multi-label classifier based on gradient boosting, gives promising results in accuracy. However, despite being a rule-based approach, its interpretability is limited because it often produces rules sets that are both *too large* and *redundant*. Such limitations pose increased cognitive load for humans, making it hard to interpret and use.

In this work, we propose CORSET, a rule-based method that significantly improves over the state-of-the-art BOOMER. The improvement is due to (1) reducing rule redundancy, which is achieved by incorporating a term in our objective that penalizes for rule overlap, and (2) explicitly limiting the complexity of rules via a suite of novel sampling schemes. As a result, our method produces a concise set of interpretable rules. An illustration of the concept of our approach is given in Fig. 1.

Example. To illustrate the improvement of CORSET over BOOMER, we consider as an example the bibtex dataset, where each data point represents a scientific article, with bag of words as features and topics as labels. We first consider predictive performance as a function of the number of rules. In Fig. 2, we show the (micro-averaged) balanced F_1 scores, a popular measure for multi-label classification used throughout this paper, for both CORSET and BOOMER. Due to the conciseness of its learned rule set, CORSET achieves a score close to 0.36 with about 100 rules, whereas BOOMER needs over 800 rules to achieve similar performance. Note that CORSET's performance starts to drop after about 100 rules, as there are no more good rules to learn. The drop indicates over-fitting, which can be addressed by standard methods, e.g., cross-validation.

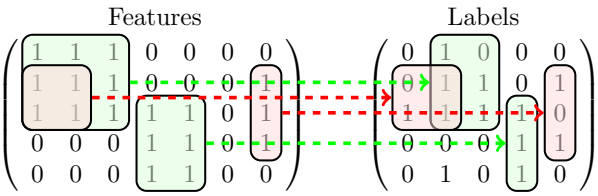

Fig. 1 Illustration of the concept of our approach for multi-label rule selection. A toy dataset is visualized as a feature matrix and a label matrix. Four rules are shown as colored regions. Regions covered by the same rule are connected by a dashed arrow. The rules in green are chosen because of accuracy, generality, and diversity. The rules in red are discarded

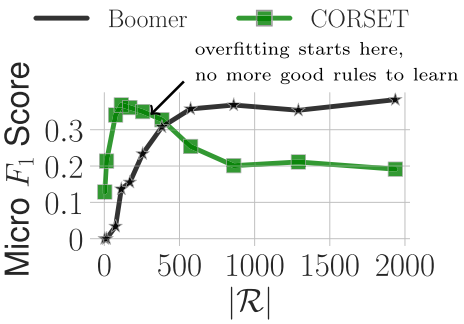

Fig. 2 Micro F_1, as a function of number of rules for CORSET vs. BOOMER in the bibtex dataset

In addition, Fig. 3 demonstrates the conciseness of the rules found by CORSET vs. the ones found by BOOMER. Here, we show a subset of rules as a bipartite graph, where nodes at the top represent labels and nodes at the bottom represent the predicates (features). Rules are represented by colors, and two nodes are connected if they are part of the same rule. CORSET uses fewer rules than BOOMER and rules tend to contain fewer predicates, resulting in a sparser graph.

Finally in Fig. 4, we reveal the underlying block structure in the subset of the feature matrix **F** and label matrix **L** of bibtex induced by the same set of rules generated by CORSET shown in Fig. 3. The rows and columns of the two matrices are re-ordered using a bi-clustering algorithm [4]. Specifically, each row in **F** (and **L**) corresponds to a training point, which is included in **F** and **L** if it is covered by at least one rule. An entry in **F** and **L** is plotted white if the corresponding value in the matrix is 0. An entry is plotted non-white if its value is 1. Further, the color is gray if the corresponding feature (label) is not selected in the rule set. Otherwise, the color of a block depends on the rule that covers it.

Contributions. Concretely, in this work we make the following contributions.

- We frame the problem of learning concise rule sets as an optimization problem. The objective function to be maximized over the space of possible rule sets is a linear combination of a quality function and a diversity function. The optimization problem is **NP**-hard and our proposed algorithm, CORSET, given a set of rule candidates, achieves an approximation ratio of 2, that is, it finds a rule set that is guaranteed to have value of the objective function which is at least half the value of the best possible rule set(s).
- The performance of CORSET depends on the quality of the rule candidates. To find good rules efficiently, we design a suite of fast sampling algorithms with probabilistic guarantees as well as an effective heuristic. The sampling algorithms allow to eschew the

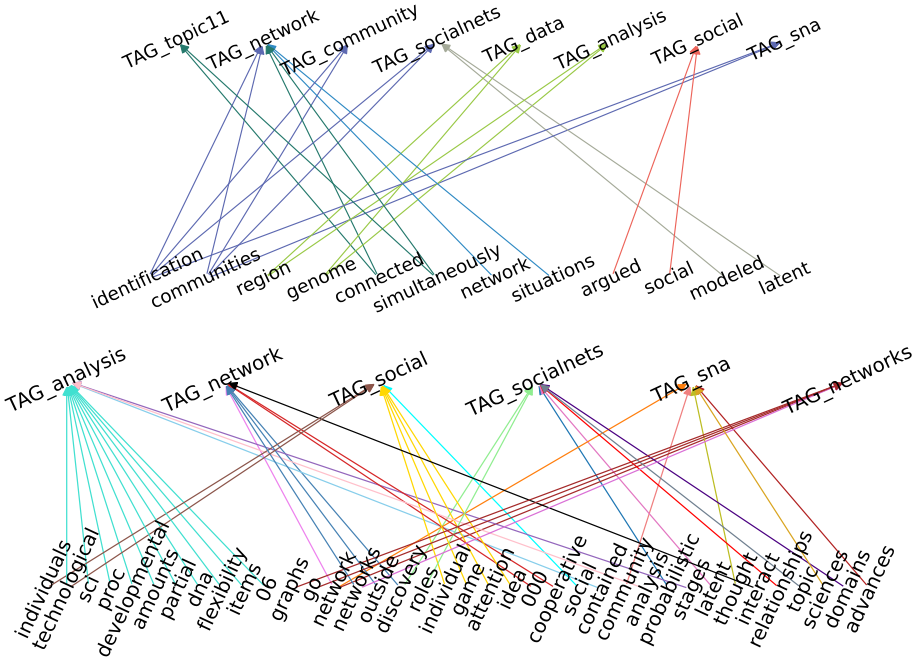

Fig. 3 An example set of rules returned by our algorithm (top) and BOOMER (bottom) on the bibtex dataset. We depict all rules for the set of labels { SNA, socialnets, social, networks, analysis}

combinatorial nature of the search space of rules and work with a considerably smaller space that is relevant to our objective.
- Our experiments show that CORSET achieves competitive predictive performance compared to state-of-the-art rule-based multi-label classifiers, while offering significantly better interpretability.

The rest of this paper is organized as follows: Section 2 discusses related work. Section 3 formalizes the problem we consider. Section 4 illustrates CORSET, omitting the details of the rule-sampling algorithms it relies on, which are described in Sects. 5 and 6. Section 7 analyzes the complexity of CORSET. Afterwards, Sect. 8 presents a thorough experimental evaluation of CORSET. Lastly, Sect. 9 demonstrates the practical applicability of our methods through a case study.

2 Related work

Multi-label classification. In multi-label classification, the goal is to learn a function that maps input points to one or more predefined categories. For instance, a song can be associated with multiple music genres. A plethora of algorithms have been proposed for this problem; interested readers may refer to a survey [1]. The simplest approaches for multi-label classification are the so-called transformation methods, which convert the original problem into multiple single-label classification problems. The main drawback of these approaches is that they fail to capture label correlations. To overcome this issue, label power-set approaches map each distinct set of labels to a unique meta-label, which serves as target label for a

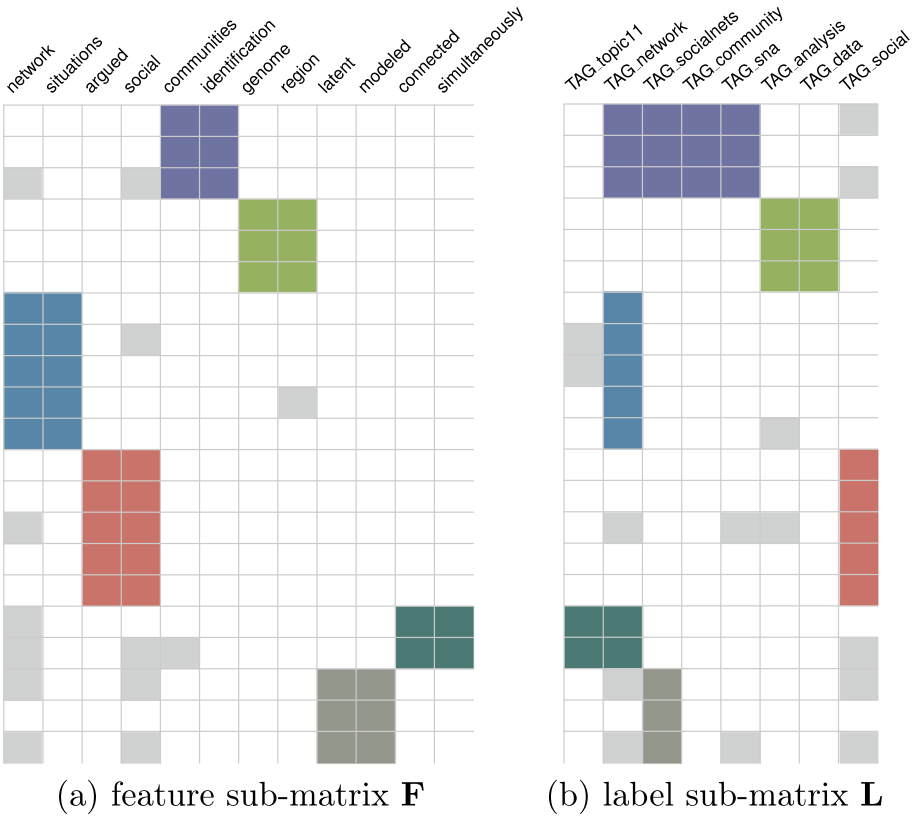

Fig. 4 Block structures of a feature sub-matrix (left) and a label sub-matrix (right) revealed by CORSET rules for the set of labels { SNA, socialnets, social, networks, analysis}

single-label classifier. Clearly, these approaches do not scale with the number of labels, and the pruned problem transformation method [5] has been proposed as a remedy. Another line of research focuses on designing ad hoc multi-label classification methods by extending existing single-label algorithms. Examples include adaption from support vector machines [6], k-nearest neighbor classifiers [7], and perceptrons [8].

Interpretable machine learning. There is no agreed formal definition of interpretability, but it can be loosely defined as the degree to which a human can understand the cause of a decision [9]. Broadly speaking, interpretability in machine learning can be achieved by imposing constraints or penalties to guarantee that the process behind the decision of the algorithm is understandable to humans.

A related topic is *explainable machine learning*, where the goal is to provide "explanations" to the predictions of black-box models. Although interpretable and explainable machine learning pursue the same general goal and are sometimes lumped together in the literature, whenever possible, interpretable models should be used rather than "explained" black-box models for high-stake decision making, because explanations for black boxes can be problematic, misleading, and error-prone [10]. By contrast, interpretable models, due to their inherent simplicity, are particularly simple to interpret and troubleshoot. Thus, in high-

stake domains, where it is crucial to understand the underlying mechanism leading to a given prediction, it is advisable to prioritize using interpretable models over explained ones.

Rule-based approaches to single-label classification. Research in interpretable machine learning has boomed in the last years. Rule-based (or associative) approaches have shown promising potential, because decisions are driven by a simple set of "if-then" rules. Liu et al. [11] are among the first to investigate association rule mining for single-label classification tasks, followed by extensions such as MCAR [12] and ADA [13]. These approaches are conceptually similar, but differ in their methodologies for rule learning, ranking, pruning, and prediction.

Concise single-label rule sets. Our work pursues for the first time the goal of designing a multi-label associative classifier for achieving a given classification performance with the smallest possible number of rules. A similar objective has been recently considered in the context of single-label classification. In particular, Zhang et al. [14] frame the problem of learning a set of classification rules as an optimization problem with an objective function combining rule quality and diversity. A 2-approximation algorithm is then proposed to solve this problem, which relies on existing frameworks for max-sum diversification and pattern sampling. In this paper, we investigate how to extend these ideas to the multi-label classification setting. The problem of controlling the number of rules used for prediction has also been studied for single-label rule boosting, where learned rules are combined additively [15]. An extension to multi-label classification represents a possible direction of future work. A different line of work, which has also addressed rule set conciseness [16, 17], relies on encoding single-label rule learning problems as SAT programs and solving them via SAT solvers. Finally, Wang et al. [18] propose to control the number of learned rules using a Bayesian approach.

In addition to the number of rules, different definitions of interpretability have been explored in association rule mining and single-label rule-based classification. For instance, rule set interpretability has been quantified by the numbers of conditions used by the rules [17] and by the *minimum description length* principle [19].

Rule-based approaches to multi-label classification. In general, adaptation from the single-label to the multi-label setting is not trivial and while single-label associative classification has been studied extensively, relatively few attempts have been made for associative multi-label classification. In an early work, Thabtah et al. [20] propose a label ranking-based assignment method. More recently, new approaches have been developed, and SECO [21] and BOOMER [2] are state of the art in the current literature of rule-based multi-label classification. The main limitation of the existing works, addressed in our paper, is that they use a very large set of highly redundant rules, which hinders interpretability. We compare our method against SECO [21] and BOOMER [2] in Sect. 8.

Pattern sampling. Association pattern discovery, which aims at discovering relevant association patterns between items in large datasets, is challenging due to the prohibitive size of the pattern space. This challenge is inherited by rule-based classifiers.

Deterministic approaches to association pattern discovery are based on exhaustive enumeration of the pattern space [22–24]. Furthermore, such approaches typically impose *hard* thresholds on some frequency-based measure (such as support and confidence) to distinguish *interesting* patterns from irrelevant ones.

Despite the remarkable progress made over the past years [22–24], deterministic approaches to association pattern discovery still incur several limitations [25–27]. First, the required enumeration (or partial enumeration) of the search space makes it hard to control the running time and consumed memory. Second, finding an appropriate threshold can be a non-intuitive and even cumbersome task. Third, these approaches generally do not allow

enough flexibility in the choice of the interestingness measure and often return enormous amounts of redundant patterns that describe the same local phenomenon.

To overcome the aforementioned limitations, efficient sampling methods have been proposed for various pattern mining tasks [25, 26]. These methods sample directly from the output space of patterns, circumventing the enumeration of the search space, they do not require setting an hard threshold, they are robust with respect to redundancy in the output patterns, and they allow great efficiency and flexibility in choosing the target distribution to sample patterns from.

In the multi-label classification setting, however, existing pattern samplers do not deliver satisfactory performance (Sects. 6 and 8). In this work, we extend existing methods to efficiently find high-quality candidate multi-label rules, as discussed in detail in Sect. 5 and 6.

3 Problem statement

At a high level, our objective is to capture the relevant patterns in the data that best discriminate sets of labels and are as concise as possible. Next we formally define the problem.

3.1 Preliminaries

We denote sets and multisets by uppercase letters, e.g., X. For a finite X, we denote by $\mathcal{P}(X)$ its power set. We consider a binary *dataset* $\mathcal{D}$ (all values are either 0 or 1) over a *feature set* $\mathcal{F}$ and a *label set* $\mathcal{L}$. The dataset $\mathcal{D}$ is a set of *data records*, $D_1, \ldots, D_n$. A data record or instance $D = (F, L)$ consists of a set of features $F \subseteq \mathcal{F}$ and a set of labels $L \subseteq \mathcal{L}$. We denote by F_D and L_D the feature set and label set of D, respectively. Furthermore, we denote by $|\mathcal{F}|$ and $|\mathcal{L}|$ the dimensions of the feature and label space, respectively, and we denote by $|\mathcal{D}|$ the total number of data records. We use $\|\mathcal{F}\|$ and $\|\mathcal{L}\|$ to refer to the total number of feature and label occurrences over all data records.

In multi-label classification, the goal is to learn a function mapping as accurately as possible the features F_D to one or more labels L_D. We use mappings consisting of *conjunctive rules*. A conjunctive rule $R = (B \rightarrow H)$ is composed of a non-empty feature set B (called *body*) and a non-empty label set H (called *head*). The body B can be viewed as a predicate $B : \{0, 1\}^{|\mathcal{F}|} \rightarrow \{\text{true}, \text{false}\}$, which states whether an instance contains all the features in B. If the predicate evaluates to true for some instance, the head H of R specifies that labels H should be predicted as present.

We say that a body B *matches* a data record D if $B \subseteq F_D$. Similarly, a head H matches D if $H \subseteq L_D$. We also say that a rule R *covers* a data record D if $B_R \subseteq F_D$ and similarly R matches a data record D if both B_R and H_R match D. For a dataset $\mathcal{D}$, we denote the *support set* of $X \in \{B, H, R\}$ by:

$$\mathcal{D}[X] = \{D \in \mathcal{D} \mid X \text{ matches } D\}.$$

The *space of all possible rules* we consider is $\mathcal{U} = \mathcal{P}(\mathcal{F}) \times \mathcal{P}(\mathcal{L})$, i.e., the Cartesian product of the power set of the feature set and the power set of the label set.

As explained in the remainder of this section, our ultimate goal is to find the most succinct set of rules to model the main dependencies in the data that are relevant for multi-label classification. Rule sets are denoted by $\mathcal{R}$. They are in disjunctive normal form, i.e., they are

"*OR* of *ANDs*", so that a label is predicted for a data record D if it belongs to the head H of at least one of the rules $R \in \mathcal{R}$ covering D.

Before introducing the details of the proposed problem formulation, we remark that more complex and structured rule-based models, such as rule lists or decision trees, have emerged as popular alternatives to rule set models [10]. In addition, *negative* rules, which predict the absence of labels rather than their presence, have also been investigated [28].

The choice of using simple rule sets in disjunctive normal form is a consequence of the fundamental principle underlying this work, that is, the design of a multi-label classifier that prioritizes interpretability while also offering accurate classification. Unlike rule lists and decision trees which introduce hierarchical structures among the rules, thereby complicating the process of interpretation, rule sets are easier to interpret due to the absence of hierarchy. Similarly, we do not consider negative rules because they may lead to labels simultaneously predicted as present and absent, which also hampers interpretability.

3.2 Problem formulation

We want to discover rules that are accurate and general, but also sufficiently different from each other. To capture this trade-off, we design an objective function that consists of a *quality term* $q : \mathcal{U} \to \mathbb{R}$ measuring the accuracy and generality of a single rule, and a *diversity term* $d : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ measuring the distance between pairs of rules.

Quality term. Given a rule R and a set of rules $\mathcal{R}$, the quality $q(R; \mathcal{R})$ of R with respect to $\mathcal{R}$ is the product of two values: the *uncovered area* $\text{area}(R; \mathcal{R})$, capturing the generality of R with respect to $\mathcal{R}$, and its *adjusted accuracy* $a(R)$,

$$q(R; \mathcal{R}) = \text{area}(R; \mathcal{R}) \cdot a(R). \tag{1}$$

Next we describe these two functions. To capture generality, we first define the *coverage* of R as:

$$\text{cov}(R) = \{(i, k) \mid R \text{ matches } D_i \in \mathcal{D} \text{ and } k \in H\}. \tag{2}$$

In other words, the coverage of a rule is the set of label occurrences it matches in a dataset. To incorporate what is already covered by a set of selected rules $\mathcal{R}$, we define *uncovered area* of R with respect to $\mathcal{R}$ as

$$\text{area}(R; \mathcal{R}) = |\text{cov}(R) \setminus \bigcup_{R \in \mathcal{R}} \text{cov}(R)|, \tag{3}$$

that is, the size of covered label occurrences by R after excluding those already covered by at least one rule in $\mathcal{R}$. Thus, a rule R is considered general with respect to $\mathcal{R}$ if $\text{area}(R; \mathcal{R})$ is large.

Before introducing the adjusted-accuracy function, we need some additional notation. Data records whose labels contain H are said to be positive with respect to H, whilst the remaining ones are negative. More formally, a head H bi-partitions a dataset $\mathcal{D}$ into two disjoint sets: a set of *positive data records* $\mathcal{D}_H^+ = \{D \in \mathcal{D} \mid H \subseteq L_D\}$ and a set of *negative data records* $\mathcal{D}_H^- = \{D \in \mathcal{D} \mid H \nsubseteq L_D\}$. Given a rule R, let $P_{\mathcal{D}[R]} = \frac{|\mathcal{D}[R]|}{|\mathcal{D}[B]|}$ be the *precision* of R and $P_{\mathcal{D}} = \frac{|\mathcal{D}[H]|}{|\mathcal{D}|}$ is the *base rate* of H in $\mathcal{D}$. We denote the corresponding binomial distributions as $\text{Bin}\big(P_{\mathcal{D}[R]}\big)$ and $\text{Bin}(P_{\mathcal{D}})$, respectively. Then, the *adjusted accuracy* of R is defined as:

$$a(R) = I(R) \cdot D_{\text{KL}}\big(\text{Bin}\big(P_{\mathcal{D}[R]}\big) \,||\, \text{Bin}(P_{\mathcal{D}})\big), \tag{4}$$

where $I(R)$ is 1 if $P_{\mathcal{D}[R]} > P_{\mathcal{D}}$ and 0 otherwise, and $D_{\mathrm{KL}}(\cdot \,||\, \cdot)$ is the KL divergence between two probability distributions. The underlying intuition is that if the precision of a rule is below its base rate, it is useless and receives a zero score. If instead the precision of a rule is larger than the base rate, the higher the precision is, the larger the score.

Diversity term. We measure the distance between two rules by how much their coverages overlap. Formally, given two rules R_1 and R_2, their distance is defined as

$$d(R_1, R_2) = 1 - \frac{|\mathrm{cov}(R_1) \cap \mathrm{cov}(R_2)|}{|\mathrm{cov}(R_1) \cup \mathrm{cov}(R_2)|}, \tag{5}$$

which is the *Jaccard distance* between $\mathrm{cov}(R_1)$ and $\mathrm{cov}(R_2)$.

Total quality and diversity. We extend the quality function and diversity function to rule sets, since we want to learn a set of rules. The extension of diversity is straightforward. Given a rule set $\mathcal{R}$, we define its total diversity as:

$$d(\mathcal{R}) = \sum_{R_i, R_j \in \mathcal{R}, i \neq j} d(R_i, R_j), \tag{6}$$

i.e., the sum of all pairwise diversity values in $\mathcal{R}$.

The extension of the quality function from single rules to rule sets requires careful judgment. The total quality of a rule set $\mathcal{R}$ is defined as:

$$q(\mathcal{R}) = \sum_{R_i \in \mathcal{R}} q\left(R_i; \mathcal{R}_{1,\ldots i-1}\right). \tag{7}$$

Thus, this definition implicitly defines an order on the rules.

We argue that an order is required to make the total quality function meaningful. Consider an alternative definition: $q'(\mathcal{R}) = \sum_{R \in \mathcal{R}} q(R; \mathcal{R}\backslash\{R\})$, which does not assume any order on $\mathcal{R}$. Recall that $q(R; \mathcal{R}\backslash\{R\}) = \mathrm{area}(R; \mathcal{R}\backslash\{R\}) \cdot a(R)$, where $\mathrm{area}(R; \mathcal{R}\backslash\{R\})$ counts the label occurrences uniquely covered by R. The issue with the definition of q' is that the area term in q' only considers the label occurrences that are *uniquely* covered by a single rule and neglects those that are covered more than once. In contrast, it is more desirable to consider the union of the label occurrences covered by all rules in $\mathcal{R}$. Our proposed definition in Eq. (7) fulfills this desideratum.

Problem definition. We frame the learning problem as a combinatorial optimization problem with budget constraint, where we set a budget on the maximum number of rules to discover, and rules should be selected to maximize a linear combination of the total quality and total diversity.

Problem 1 Given a dataset $\mathcal{D} = \{D_i\}_{i=1}^n$, a budget $\mathcal{B} \in \mathbb{Z}_+$, a space of rules $\mathcal{S} \subseteq \mathcal{U}$, and a parameter $\lambda \in \mathbb{R}_+$, find a set of $\mathcal{B}$ rules $\mathcal{R} = \{R_1, \ldots, R_{\mathcal{B}}\} \subseteq \mathcal{S}$, to maximize the following objective

$$f(\mathcal{R}) = q(\mathcal{R}) + \lambda d(\mathcal{R}). \tag{8}$$

This problem is known to be NP-hard [29]. In the next section, we present a greedy algorithm which finds a solution to Problem 1 with an approximation factor of 2, provided that the space of rules $\mathcal{S}$ can be visited in polynomial time.

4 The CORSET learning algorithm

In this section, we present a meta algorithm named CORSET (*concise rule set*) for Problem 1. CORSET greedily picks one rule at a time from a pool of candidate rules, so as to maximize a

modified version of the marginal gain for the objective in Eq. (8), i.e.,

$$f'(\mathcal{R} \cup \{R\}) - f'(\mathcal{R}) = \frac{1}{2} q(R; \mathcal{R}) + \lambda \sum_{R_j \in \mathcal{R}} d(R, R_j). \tag{9}$$

Here, $f'(\mathcal{R})$ is the same as $f(\mathcal{R})$ in our problem formulation with the sole exception that the quality term is multiplied by $\frac{1}{2}$. The difference can be ignored by adjusting λ accordingly and is only required for the proof of Proposition 1.

The candidate rules CORSET picks from are generated by a procedure called GENCANDRULES. The effectiveness of GENCANDRULES heavily affects the predictive performance of the classifier. The goal is to sample high-quality rules in terms of generality, diversity, and accuracy (that is, tailored to our objective). This is a challenging goal since the size of rule space is exponential [30], and GENCANDRULES should therefore avoid exploring the whole space. We defer the description of the candidate generation to Sects. 5 and 6.

For now, we focus on the description of the main algorithm. CORSET maintains a set of selected rules, $\mathcal{R}$, which is initially empty. At each iteration, it considers a pool of candidate rules generated by GENCANDRULES. Within this pool, the rule R^* maximizing the marginal gain in Eq. (9) with respect to $\mathcal{R}$ is selected and added to $\mathcal{R}$. The process stops when the proportion of labels in $\mathcal{L}$ predicted by $\mathcal{R} \cup R^*$ and not by $\mathcal{R}$ falls below a user-specified tolerance level τ.

To ensure the aforementioned approximation guarantee, we need to repeat the greedy procedure a second time, over the union of the rules generated in all iterations, because using a different candidate set at each iteration does not offer an approximation guarantee. Denote as $\mathcal{R}'$ the new rule set obtained in this second round. We return as solution the set that yields the largest objective function value between $\mathcal{R}$ and $\mathcal{R}'$, since in practice $\mathcal{R}'$ is not necessarily better than $\mathcal{R}$.

The pseudo-code of CORSET is shown in Algorithm 1. Note that GENCANDRULES receives as input the current rule set $\mathcal{R}$, so as to generate rules different from $\mathcal{R}$, as explained in the next section.

We provide the approximation guarantee next.

Proposition 1 *For a fixed pool of candidate rules,* CORSET *is a* 2*-approximation algorithm for Problem 1.*

Proof Problem 1 is a special case of a general problem known as *max-sum diversification* [29], which has been studied in the context of *result diversification* [31]. In the general formulation, given a set of elements U in a metric space and a set-valued function $f : 2^U \to \mathbb{R}$, the problem asks for a subset $S \subseteq U$ that maximizes an objective function $f(S) + \lambda \sum_{u,v \in S} dist(u, v)$ under cardinality (or more general) constraints, where $dist$ is a pairwise distance function.

Setting f to be the quality function q in Eq. (7) and $dist$ to be the Jaccard distance d in Eq. (5), it follows that Problem 1 is a special case of the max-sum diversification problem.

A result due to Borodin et al. [29] shows that a simple greedy algorithm, which iteratively picks the element maximizing the marginal gain in the objective function, guarantees an approximation factor of 2 for the max-sum diversification problem and therefore for Problem 1, as long as f is monotone and submodular and d is a metric.

As a consequence, to show that the 2-approximation guarantee holds for CORSET, it is sufficient to show that the quality term $q(\mathcal{R})$ is monotone and submodular, since the Jaccard distance d is known to be a proper metric [32].

First, $q(\mathcal{R})$ is a sum of nonnegative terms. Hence, it is a monotone (and non-decreasing) function.

Algorithm 1 The CORSET algorithm.

Data: data $\mathcal{D}$, tolerance τ.
Result: a set of multi-label classification rules $\mathcal{R}$.

$\mathcal{R}, \mathcal{R}' \leftarrow \emptyset$;
$\mathcal{C}_R \leftarrow \emptyset, c \leftarrow \infty$;
while $c > \tau$ **do**
  $\mathcal{C} \leftarrow$ GENCANDRULES($\mathcal{R}$);
  $\mathcal{C}_R \leftarrow \mathcal{C}_R \cup \mathcal{C}$;
  $R^* \leftarrow \arg\max_{R \in \mathcal{C}} \left[f'(\mathcal{R} \cup \{R\}) - f'(\mathcal{R}) \right]$;
  $c \leftarrow (\sum_{(D; B_{R^*} \subseteq F_D, H_{R^*} \subseteq L_D)} |L_D \cap H_{R^*} \setminus \cup_{(R \in \mathcal{R} | B_R \subseteq F_D, H_R \subseteq L_D)} H_R|)/|\mathcal{L}\|$;
  $\mathcal{R} \leftarrow \mathcal{R} \cup R^*$;
end
for $i = 1, ..|\mathcal{R}|$ **do**
  $R^* \leftarrow \arg\max_{R \in \mathcal{C}_R} \left[f'\left(\mathcal{R}' \cup \{R\}\right) - f'\left(\mathcal{R}'\right) \right]$;
  $\mathcal{R}' \leftarrow \mathcal{R}' \cup R^*$;
end
if $f(\mathcal{R}') > f(\mathcal{R})$ **then return** $\mathcal{R}'$;
else return $\mathcal{R}$

Next, we show that $q(\mathcal{R})$ is also submodular. To prove the submodularity of q, consider two rule sets $\mathcal{R}'$ and $\mathcal{R}''$ such that $\mathcal{R}' \subseteq \mathcal{R}''$. Given a new rule R^*, the marginal gain in quality resulting from adding R^* to $\mathcal{R}'$ is:

$$\Delta' = q\left(\mathcal{R}' \cup \{R^*\}\right) - q\left(\mathcal{R}'\right) = a\left(R^*\right) \times |\text{cov}\big(R^*\big) \setminus \bigcup_{R \in \mathcal{R}'} \text{cov}(R)\,|.$$

Similarly, the marginal gain in quality for $\mathcal{R}''$ is:

$$\begin{aligned}\Delta'' &= q\left(\mathcal{R}'' \cup \{R^*\}\right) - q\left(\mathcal{R}''\right) = a\left(R^*\right) \times |\text{cov}\big(R^*\big) \setminus \bigcup_{R \in \mathcal{R}''} \text{cov}(R)\,| \\ &= a\left(R^*\right) \times |\text{cov}(R) \setminus \bigcup_{R \in \mathcal{R}'} \text{cov}(R) \setminus \bigcup_{R \in \mathcal{R}'' \setminus \mathcal{R}'} \text{cov}(R)\,|.\end{aligned}$$

Since

$$|\text{cov}\big(R^*\big) \setminus \bigcup_{R \in \mathcal{R}'} \text{cov}(R)\,| \geq |\text{cov}\big(R^*\big) \setminus \bigcup_{R \in \mathcal{R}'} \text{cov}(R) \setminus \bigcup_{R \in \mathcal{R}'' \setminus \mathcal{R}'} \text{cov}(R)\,|,$$

it immediately follows that $\Delta' \geq \Delta''$. Therefore, the quality function q is submodular.

We conclude that CORSET is a 2-approximation algorithm for Problem 1. This means that, given a space of candidate rules, CORSET is guaranteed to return rule sets achieving a value of the objective in Eq. (8) that is at least half of the maximum value of the objective function achieved by the optimal rule set(s). □

It is important to note that the approximation guarantee does not hold with respect to the entire rule space, but thanks to the second round of greedy selection, it holds with respect to the entire pool of rules sampled in all the iterations of the algorithm.

Prediction. At prediction time, given the set of selected rules $\mathcal{R}$, we return the set of predicted labels for a data record $D = (F_D, L_D)$ as $\bigcup_{R \in \mathcal{R} | B_R \subseteq F_D} H_R$, that is, the union of the heads of the rules such that F_D contains all attributes in the body B, or, equivalently, such that R covers D.

5 Rule sampling

In the next two sections, we present the main contribution of our work, a suite of rule-sampling algorithms used by GenCandRules. In this section, we first describe the technical basis of our proposal. Then, we formulate our sampling problem and present the algorithms for it. The devised sampling algorithms have, however, important limitations. In the next section, we discuss such limitations and describe practical enhancements.

5.1 Background: two-stage pattern sampling

Our sampling scheme builds on the pattern-sampling algorithms proposed by Boley et al. [25, 26]. These algorithms allow us to sample patterns according to a target distribution over the pattern space, without the need of exhaustive enumeration. The target distribution reflects a measure of interestingness for the patterns. Example measures include support, area, and, if the data are labelled, discriminativity. Sampling algorithms for a variety of measures share a two-stage structure, whilst the details depend on the measure under consideration.

The key insight brought by Boley et al. [25, 26] is that *random experiments reveal frequent events*. We use sampling by support and area for illustration. Consider a dataset $\mathcal{D} = \{D_1, \ldots, D_n\}$ over a finite ground set $\mathcal{E}$, with $D \subseteq \mathcal{E}$ for each $D \in \mathcal{D}$. Consider the problem of sampling an itemset (pattern) $F \subseteq \mathcal{E}$ with probability proportional to its *support* $q_{\text{supp}}(F) = |D[F]|$.

For each $D \in \mathcal{D}$, the set of itemsets including D in their support is $\mathcal{P}(D)$. It can be shown that sampling an itemset F uniformly from $\biguplus_{D\in\mathcal{D}} \mathcal{P}(D)$, where $\biguplus$ denotes the union operator of multi-sets, is the same as sampling F according to $|D[F]|$. To avoid materializing $\biguplus_{D\in\mathcal{D}} \mathcal{P}(D)$, Boley et al. use a two-step procedure:

1. sample a data record D with probability proportional to the weight $w(D) = \sum_{F\in\mathcal{P}(D)} 1 = 2^{|D|}$.
2. sample an itemset F uniformly from $\mathcal{P}(D)$.

To sample from the "area" distribution $q_{\text{area}}(F) = |F||D[F]|$, the above procedure is changed as follows: set weights $w(D) = \sum_{F\in\mathcal{P}(D)} F = |D|2^{|D|-1}$, and then sample F with weight $|F|$ from $\mathcal{P}(D)$.

The two-stage sampling idea can be generalized to a number of other measures. Some of them, such as discriminativity, which we use later, require sampling tuples of data records rather than a single one in the first stage.

Next we describe two sampling distributions and the corresponding sampling algorithms for our objective. The first distribution is a generalization of the area function (not discussed by Boley et al. [25, 26]) and is used for head sampling. The second distribution is discriminativity used for body sampling. For discriminativity, we propose an improved sampling algorithm, which is faster than the original version [26].

5.2 Sampling objectives

Our rule sampling objective can be expressed using the chain rule of probabilities as a product of two values, one reflecting the generality of a rule $R = (B \to H)$ given the current set of rules $\mathcal{R}$, and the other its discriminative power:

$$\Pr(B, H) \propto w(B, H; \mathcal{R}) = q_a(B; H) \cdot \text{area}(H; \mathcal{R}). \tag{10}$$

Note that the uncovered area function in Eq. (3) generalizes to tails, i.e., $\text{area}(H;\mathcal{R}) = |\text{cov}(H) \setminus \bigcup_{H' \in \mathcal{R}} \text{cov}(H')|$ and $\text{cov}(H) = \{(i,k) \mid H \text{ matches } D_i \in \mathcal{D} \text{ and } k \in H\}$. We use $\text{area}(H;\mathcal{R})$ because we want to extract heads that cover the largest possible area and are as diverse as possible.

For q_a, we choose the discriminativity measure studied by Boley et al. [25], which permits sampling in polynomial time.

Given a head $H \subseteq \mathcal{L}$, the *discriminativity* of B is defined as:

$$q_{\text{disc}}(B;H) = |\mathcal{D}_H^+[B]| \, |\mathcal{D}_H^- \setminus \mathcal{D}_H^-[B]|. \tag{11}$$

The goal is to sample bodies that have as large support as possible in $\mathcal{D}_H^+$ and as small support as possible in $\mathcal{D}_H^-$. Thus, discriminativity captures the ability of a body B in discriminating between the presence and absence of a given head H in data records.

To sample from the distribution in Eq. (10), we use the following steps:

1. sample H with probability proportional to $\text{area}(H;\mathcal{R})$;
2. sample B with probability proportional to $q_a(B;H)$.

We explain each sampling step next.

5.3 Head sampling

To sample from $\text{area}(H;\mathcal{R})$, we apply a similar two-step sampling procedure as in Boley et al. [25]: we first sample a data record D with probability proportional to its weight $w(D;\mathcal{R})$ and then sample H from D. The function $\text{area}(H;\mathcal{R})$ is a generalization of the area function considered in Boley et al. [25]. Adapting the original algorithm to our case requires to design a weight function $w(\cdot;\mathcal{R})$ appropriate for our target. To define $w(\cdot;\mathcal{R})$, a few new definitions are needed. Given a rule R and a data record D, the *D-specific coverage* of R is defined to be:

$$\text{cov}_D(R) = \begin{cases} L_D \cap H_R, & \text{if } R \text{ covers } D, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{12}$$

Extending D-specific coverage to a rule set $\mathcal{R}$, we have:

$$\text{cov}_D(\mathcal{R}) = \bigcup_{R \in \mathcal{R}} \text{cov}_D(R). \tag{13}$$

Given a label set H, its *marginal coverage* with respect to $\mathcal{R}$ is:

$$\text{cov}_D(H;\mathcal{R}) = (L_D \cap H) \setminus \text{cov}_D(\mathcal{R}), \tag{14}$$

that is, the covered label occurrences in D by H, excluding those by $\mathcal{R}$. As a shortcut, we define $\overline{\text{cov}}_D(\mathcal{R}) = \text{cov}_D(L_D;\mathcal{R})$, i.e., the set of label occurrences in D not covered by $\mathcal{R}$.

The weight of a label set H on a data record D is:

$$w(H,D;\mathcal{R}) = |\text{cov}_D(H;\mathcal{R})|. \tag{15}$$

We give a small example to illustrate these definitions:

Algorithm 2 Two-stage head sampling.

Data: a dataset $\mathcal{D}$, weights $w(D; \mathcal{R})$ (as in Equation (16)).
Result: a head $H \subseteq \mathcal{L}$ with $H \sim \text{area}(H; \mathcal{R})$.
1 draw $D \sim w(D; \mathcal{R})$;
2 **return** $H \sim w(H, D; \mathcal{R})$

$$
\begin{aligned}
&D : F_D = \{0, 1, 2\}, \quad L_D = \{a, b, c\} \\
&R_1 : (\{0, 1\} \to \{a\}) \\
&R_2 : (\{1, 2\} \to \{a, b\}) \qquad \mathcal{R} = \{R_1, R_2, R_3\} \\
&R_3 : (\{2, 3\} \to \{a, c\}) \qquad H = \{b, c\}
\end{aligned}
$$

For R_1 and R_2, the sets $\text{cov}_D(\cdot)$ are $\{a\}, \{a, b\}$, respectively. For R_3, the set $\text{cov}_D(R_3)$ is $\emptyset$ since R_3 does not cover D. Therefore, $\text{cov}_D(H; \mathcal{R}) = \{c\}$ and $w(H, D; \mathcal{R}) = 1$.

The intuition of the definition of $w(H, D; \mathcal{R})$ is that H has large weight on D if it contains many label occurrences not covered by $\mathcal{R}$. Therefore, the weight of any data record D is simply the summation of the weights over all possible heads and has the following simple form:

$$
\begin{aligned}
w(D; \mathcal{R}) &= \sum_{H \subseteq L_D} w(H, D; \mathcal{R}) \\
&= \sum_{H \subseteq L_D} |\text{cov}_D(H; \mathcal{R})| \\
&= \sum_{\substack{(H_1 \cup H_2) \subseteq L_D \\ \text{s.t. } H_1 \subseteq \overline{\text{cov}}_D(\mathcal{R}), \\ H_2 \subseteq \text{cov}_D(\mathcal{R}), \\ H_1 \cap H_2 = \emptyset}} |H_1| \\
&= \sum_{H_1 \subseteq \overline{\text{cov}}_D(\mathcal{R})} |H_1| \sum_{H_2 \subseteq \text{cov}_D(\mathcal{R})} 1 \\
&= \left(|\overline{\text{cov}}_D(\mathcal{R})| \, 2^{|\overline{\text{cov}}_D(\mathcal{R})|-1} \right) 2^{|\text{cov}_D(\mathcal{R})|} \\
&= |\overline{\text{cov}}_D(\mathcal{R})| \, 2^{|L_D|-1}. \qquad (16)
\end{aligned}
$$

In the third equation, a head H is split into two disjoint parts, H_1 and H_2, such that H_1 belongs to the uncovered label occurrences, whereas H_2 belongs to the already covered ones. The remaining equations follow from simple algebra.

Using these weights, we adapt the sampling algorithm of Boley et al. [25, 26] as per Algorithm 2.

The sampling algorithm is supported by probabilistic guarantees, as formalized next.

Proposition 2 *Algorithm 2 returns $H \sim \text{area}(H; \mathcal{R})$.*

Proof We prove that Algorithm 2 returns $H \sim \text{area}(H; \mathcal{R})$.

$$
Pr(H \text{ is drawn}) = \sum_{D \in \mathcal{D}} Pr(H \text{ is drawn and } D \text{ is drawn})
$$

$$
\begin{aligned}
&= \sum_{D \in \mathcal{D}} Pr(D \text{ is drawn}) Pr(H \text{ is drawn from } \mathcal{P}(L_D)) \\
&= \sum_{D \in \mathcal{D}[H]} Pr(D \text{ is drawn}) Pr(H \text{ is drawn from } \mathcal{P}(L_D)) \\
&\propto \sum_{D \in \mathcal{D}[H]} w(D; \mathcal{R}) \times \frac{w(H, D; \mathcal{R})}{w(D; \mathcal{R})} \\
&= \sum_{D \in \mathcal{D}[H]} w(H, D; \mathcal{R}) = \sum_{D \in \mathcal{D}[H]} |\text{cov}_D(H; \mathcal{R})| = \text{area}(H; \mathcal{R}) .
\end{aligned}
$$

The first and second equations follow from the law of total probability and the chain rule of probabilities, respectively. The third equality is guaranteed because H can only be sampled from $D \in \mathcal{D}[H]$. If D is not in $\mathcal{D}[H]$, it has zero probability of generating H. In the fourth equality, we have used the definitions of $w(H, D; \mathcal{R})$ and $w(D; \mathcal{R})$, as well as the equality:

$$
\sum_{H \subseteq L_D} w(H, D; \mathcal{R}) = w(D; \mathcal{R}) .
$$

Finally, the last equality follows since $\text{area}(H; \mathcal{R})$ can be obtained by definition as the sum of the marginal coverage $|\text{cov}_D(H; \mathcal{R})|$ of H on D, given $\mathcal{R}$, over all data records $D \in \mathcal{D}[H]$. □

5.4 Body sampling

After a head H is sampled, we sample B according to $q_a(B; H) = q_{\text{disc}}(B; H)$, from Eq. (11). The two-stage sampling scheme by Boley et al. [25, 26] can be applied for this case. In contrast to the previous cases, the weight function is defined on *pairs* of data records:

$$
w(D^+, D^-) = 2^{|D^+|} - 2^{|D^+ \cap D^-|} - |D^+ \setminus D^-|, \tag{17}
$$

where $D^+ \in \mathcal{D}_H^+$ and $D^- \in \mathcal{D}_H^-$, and $|D^+|$ (resp. $|D^-|$) denotes the number of features present in D^+ (resp. D^-). Thus, pre-computing the weights leads to *quadratic* space complexity in $|\mathcal{D}|$, which limits the practicality of the sampling procedure.

The above limitation is addressed by Boley et al. [26] using the technique of *coupling from the past* (CFTP), which leads to *linear* space complexity. Unlike many Markov chain Monte Carlo (MCMC) methods, CFTP can guarantee that samples are generated according to the target distribution. It operates by simulating the Markov chain *backwards* by sampling from a proposal distribution, until all states coalesce to the same unique state. The main challenge of using CFTP is the design of the proposal distribution and the efficient monitoring of coalescence condition.

The proposal distribution should be (i) efficient to sample from; and (ii) an appropriate approximation to the target distribution to obtain fast convergence. Boley et al. [26] devise a "general-purpose" proposal distribution, which works for all target distributions they consider. For the case of discriminativity, the proposal distribution is defined as:

$$
\overline{w}(D^+, D^-) = \sqrt{w_1(D^+) \cdot w_2(D^-)}, \tag{18}
$$

where $w_1(D^+) = 2^{|D^+|} - |D^+| - 1$ and $w_2(D^-) = 2^{|\mathcal{F}|} - 2^{|D^-|} - |D^-| - 1$. Sampling from $\overline{w}(\cdot, \cdot)$ can be done efficiently by sampling separately from $w_1(\cdot)$ and $w_2(\cdot)$. However, we argue that the choice of $\overline{w}(\cdot, \cdot)$ is not a good approximation of the target, and therefore,

Algorithm 3 Two-stage body sampling.

Data: a dataset $\mathcal{D}$, a head H, weights $w_1(\cdot)$ and $w_2(\cdot)$.
Result: a body $B \in \mathcal{F}$ with $B \sim q_d(B; H)$.
1 initialize $i \leftarrow 1$, $\mathbf{D} \leftarrow \perp$;
2 **while** $\mathbf{D} = \perp$ **do**
3 $i \leftarrow i + 1$;
4 **for** $t = 2^i, \ldots, 0$ **do**
5 draw $u_t \sim u([0, 1])$ and $\mathbf{C}_t \sim \overline{w}(\mathbf{C}_t)$;
6 **if** $\mathbf{u}_t \leq \frac{\overline{w}(\mathbf{D})w(\mathbf{C}_t)}{w(\mathbf{D})\overline{w}(\mathbf{C}_t)}$ **then**
7 $\mathbf{D} \leftarrow \mathbf{C}_t$;
8 **end**
9 **end**
10 **end**
11 draw $B_1 \sim u(\mathcal{P}(D^+ \setminus D^-) \setminus \emptyset)$, $B_2 \sim u(\mathcal{P}(D^+ \cap D^-))$;
12 **return** $B = B_1 \cup B_2$

it suffers from slow convergence. We provide empirical evidence to support this claim in Sect. 8. The reason is that when a data record is high-dimensional but sparse, as is often the case in multi-label classification, $w_2\big(D^-\big)$, and hence, $\overline{w}\big(D^+, D^-\big)$ grow exponentially with the number of features, making the acceptance probability extremely low and, as a consequence, convergence is extremely slow.

To overcome the convergence issue, we use a different proposal distribution better suited for our setting. Our proposal is the same as in Eq. (18), except that w_2 is defined as a uniform function and the square root is removed:

$$\overline{w}'(D^+, D^-) = w_1\big(D^+\big) = 2^{|D^+|} - 1 - |D^+|. \tag{19}$$

It is easy to sample from this modified proposal distribution because we can sample D^+ weighted by $w_1\big(D^+\big)$ and D^- uniformly at random. Further, $\overline{w}'$ has an appealing property:

Proposition 3 *$\overline{w}'$ is a tight upper bound of the target weight distribution in w in Eq. (17).*

Proof First, $2^{|D^+|} - 1 - |D^+|$ is the value taken on by the weight function $w(D^+, D^-)$ for a tuple (D^+, D^-) when $D^+ \cap D^- = \emptyset$.

Also when $|D^+ \cap D^-| = 1$, we have $w(D^+, D^-) = 2^{|D^+|} - 2 - (|D^+| - 1) = 2^{|D^+|} - 1 - |D^+|$.

Finally, when $|D^+ \cap D^+| = |I| \geq 2$, we have $w(D^+, D^-) = 2^{|D^+|} - 2^{|I|} - (|D^+| - |I|) = 2^{|D^+|} - 2^{|I|} - |D^+| + |I|$. Because $2^{|I|} \geq |I| \; \forall I$ and $2^{|I|}$ grows at a faster rate than $|I|$, the target weight function $w(D^+, D^-)$ achieves its maximum $2^{|D^+|} - 1 - |D^+|$ at $|I| = 0$ and $|I| = 1$ and takes smaller values for $|I| \geq 2$. It follows that $\overline{w}'(D^+, D^-) = 2^{|D^+|} - 1 - |D^+|$ is a tight upper bound for $w(D^+, D^-)$, as claimed above. □

Proposition 3 ensures that our proposal distribution $\overline{w}'$ provides a better approximation of the target, as compared to $\overline{w}$ proposed by Boley et al. [26]. Therefore, it is expected that $\overline{w}'$ gives faster convergence than $\overline{w}$. We empirically evaluate the convergence speed in Sect. 8.

Body sampling is summarized in Algorithm 3. We first use CFTP (lines 1-9) to sample a pair $\big(D^+, D^-\big)$. Then, we sample a body in line 11. We denote $u(\cdot)$ as the uniform distribution over a set. For brevity, we use a boldface letter to denote a pair (tuple) of records, e.g., $\mathbf{D}$. We denote an empty pair by $\perp$ and define $\overline{w}(\perp)/w(\perp) = 1$.

6 Enhancements to the sampling scheme

In this section, we first discuss important drawbacks of the pattern sampling schemes described in Sect. 5 and then describe our enhancements to tackle these drawbacks. The benefits of such enhancements are thoroughly demonstrated via extensive experiments presented in Sect. 8.

6.1 Limitations of the two-stage pattern-sampling framework

While theoretically sound, in our setting, the two-stage sampling framework [25, 26] suffers from two limitations, as can be verified empirically. First, we observe that most of the sampled rules are very specific, with very low support. Second, rule interpretability is not explicitly considered.

Heavy-hitter problem for head sampling. Consider the head sampling part. Notice that the weight of a data record D in Eq. (16) is exponential in L_D. If there is a data record $D \in \mathcal{D}$ whose $|L_D|$ is moderately larger than the rest, its weight dominates, making it very likely to be sampled in the first sampling step. We refer to this issue as *the heavy-hitter problem*. For instance in bibtex, the largest label set of a data record D^* contains 28 labels, while the second largest contains 16. The probability of D^* being sampled is 99.97%. A head sampled from D^* has an expected length of 14.5. Empirically, heads of about this length match only a few data records. Thus, most of the sampled heads have low support, hampering the goal of sampling general rules.

Heavy-hitter problem for body sampling. A similar issue arises in body sampling. The weight function in Eq. (17) grows exponentially with $|D^+|$, so that CFTP most likely returns the positive data records with the highest number of present features. Therefore, sampled bodies tend to be very long and have small support (often 1). Thus, they may have high discriminativity but cannot generalize to unseen data.

Head interpretability. Interpretability of heads is a central focus in our work. Nonetheless, in the original pattern-sampling algorithms [25, 26] all elements in $\mathcal{P}(L)$ are considered possible heads, regardless of whether they are interpretable or not.

The original head-sampling algorithms described in Sect. 5.3 sample new heads based on *global* correlations. Some heads are sampled simply because the labels in them co-occur frequently rather than because the labels are strongly correlated. Such heads may contain overall prominent labels, but they may be hard to interpret.

The root of the above limitations is the enormous sample space under consideration, which contains a large amount of undesirable bodies and heads with small support and heads with limited interpretability. Next, we describe a strategy to eliminate undesirable bodies and heads, thus effectively addressing the above limitations of the original two-stage pattern-sampling framework.

6.2 Head sampling under interpretable label space

We propose to restrict the label sample space to a much smaller sample space $\mathcal{S}^- \subseteq \mathcal{P}(\mathcal{L})$ designed to contain only interpretable label sets so as to mitigate the heavy-hitter problem. We call $\mathcal{S}^-$ the *interpretable label space*. Before describing the construction of $\mathcal{S}^-$, we notice that pattern sampling under any subspace of $\mathcal{P}(\mathcal{L})$ is a slight generalization of the original sampling setting. Most importantly, the original sampling algorithms can be adapted to different sample spaces, such as $\mathcal{S}^-$, while preserving probabilistic guarantees.

Algorithm 4 Head sampling under $\mathcal{S}^-$ according to uncovered area.

Data: a dataset $\mathcal{D}$, sample space $\mathcal{S}^-$, and a rule set $\mathcal{R}$.
Result: a head $H \in \mathcal{S}^-$ with $H \sim \text{area}(H; \mathcal{R})$.
1 let $I[D] \leftarrow \left\{S \in \mathcal{S}^- \mid S \subseteq L_D\right\}$ for each $D \in \mathcal{D}$;
2 let $w(D) \leftarrow \sum_{S \in I[D]} |S \setminus \text{cov}_D(\mathcal{R})|$ for each $D \in \mathcal{D}$;
3 draw $D \sim w(D)$;
4 draw $H \in I[D] \sim |H|$;
5 **return** H

In Algorithm 4, we describe a procedure for sampling by uncovered area under $\mathcal{S}^-$. The algorithm can be easily adapted for other sampling objectives including discriminativity. Compared to sampling under $\mathcal{P}(\mathcal{L})$, we require the extra step of determining the set $I[D]$ of patterns in $\mathcal{S}^-$ contained by $D \in \mathcal{D}$ and computing the weight for D accordingly.

Constructing $\mathcal{S}^-$. To construct the interpretable label space, we first define interpretability in our setting. Humans are used to think in an associative manner [33]. The underlying cognitive process is called *associative activation*, which can be described as "ideas that have been evoked trigger many other ideas, in a spreading cascade of activity in your brain" [34, p. 51]. To accommodate such tendency, we argue that a label set is interpretable if the corresponding labels are sufficiently associated. The problem of constructing $\mathcal{S}^-$ is then framed as finding sufficiently associated label sets. We rely on a graph-based approach whereby we construct a suitable label graph and extract its dense subgraphs. Specifically, we construct a directed weighted graph $\mathcal{G} = (V, E, p)$. Each node represents a label. A node pair (u, v) is an edge in E if $\mathcal{D}[\{u\}] \cap \mathcal{D}[\{v\}] \neq \emptyset$. The corresponding weight is defined as $p(u, v) = \frac{|\mathcal{D}[\{u\}] \cap \mathcal{D}[\{v\}]|}{|\mathcal{D}[\{u\}]|}$, which can be interpreted as the conditional probability that label v occurs given that label u occurs.

The need of edge direction arises because in real-world multi-label datasets, the association between labels can be asymmetric. For instance in bibtex, general labels, such as statistical physics, and specific ones, such as simulation, co-exist. The asymmetry implies that a single value assigned to a pair of labels cannot fully capture their interaction. The specific label simulation is likely to co-occur with statistical physics in most of its occurrences, while the contrary is not true.

Probabilistic interpretation of the edge weights suggests that $\mathcal{G}$ can be viewed as a *probabilistic graph* [35]. From such point of view, the problem of finding sufficiently associated label sets can be seen as finding highly probable cliques in $\mathcal{G}$ [36], whose probability of forming is above a pre-specified threshold. To solve this problem, we adapt a recursive depth-first search (DFS) procedure similar to the one proposed by Mukherjee et al. [36].

Efficient preprocessing. Execution of line 1 in Algorithm 4 can be done efficiently by framing the problem appropriately. In this problem, we are given a set of subsets $\mathcal{S}^-$ and we are asked to find, for each $D \in \mathcal{D}$, the subsets in $\mathcal{S}^-$ that are contained in L_D. A naive solution checks the containment relations for all pairs of L_D and $\mathcal{S}^-$, and in practice can take hours for many datasets. However, the problem is an instance of the *the set containment problem*, extensively studied by the database community. Among several efficient solutions proposed for this problem, we resort to one well-established algorithm, PRETTI [37], built upon the idea of inverted index and prefix trees. The running time is effectively brought down to a few seconds.

6.3 Improved body sampling

To alleviate the heavy-hitter problem during body sampling, we consider two approaches. The first approach is based on reduced sample space, but may have scalability issues. The second is a greedy heuristic, which explicitly maximizes a modified version of discriminativity.

1. Using reduced sample space. We adapt a similar idea as in head sampling (Sect. 6.2) and use a reduced sample space $\mathcal{S}^-$ for body sampling. However, in practice, while in real multi-label classification tasks the label matrix is always sparse, the feature matrix can be dense. In this case, a scalability issue arises since the DFS procedure may take exponential time. For sufficiently sparse graphs, this is not a concern, whereas in denser graphs, constructing $\mathcal{S}^-$ becomes a key scalability bottleneck.

2. A greedy heuristic. To address the above scalability issue, we propose a greedy heuristic, which drops the probabilistic guarantee, but is highly effective in practice. We use CFTP as in Algorithm 3 to sample a tuple $\left(D^+, D^-\right)$. Then, we greedily select features in $D^+ \setminus D^-$ to maximize a modified version of discriminativity: for any B, we define the measure

$$\phi(B) = |\mathcal{D}[B] \cap \mathcal{D}_H^+| - \gamma |\mathcal{D}[B] \cap \mathcal{D}_H^-|, \tag{20}$$

where γ weighs the importance of positive and negative support, so smaller values of γ lead to more general but more error-prone bodies. Further, we use early stopping (controlled by ϵ) when $|\mathcal{D}[H]|$ is too small.

The algorithm is described in Algorithm 5. It iteratively picks a feature $h \in F_{D^+} \cup F_{D^-}$, which maximizes the marginal gain of ϕ. The best feature h^* is added to B and the support is updated accordingly. Finally, a linear sweep over B finds the body with the highest objective value (in Eq. (20)). In practice, we use a pre-computed inverted index to allow for efficient intersection of supports. Variations of Algorithm 5 have also been investigated in which the input is deterministic, the difference in line 4 is normalized by $\mathcal{D}[\{h\}]$ and the support is replaced by the portion of support not belonging to the support of previously chosen rules.

Summary. The second approach scales better for dense feature matrices than the first approach. However, the first approach has the following advantages: (1) body sampling has probabilistic guarantees, (2) it is much faster to run when the feature matrices are sparse, (3) bodies in the reduced sample space $\mathcal{S}^-$ are composed of highly correlated attributes and hence are easy to interpret. In the sequel, we use CORSET-SURS to denote the version where the first approach is used for body sampling, and CORSET-GH when the second approach is used.

7 Complexity analysis

Time complexity. Let T_f be the time complexity of evaluating the quality and diversity function. T_f is bounded by $|\mathcal{D}|(|\mathcal{F}| + |\mathcal{L}|)$. Let $\mathcal{S}_{\mathcal{L}}^-$ be the interpretable sample space for head sampling and $\mathcal{S}_{\mathcal{F}}^-$ be the reduced sample space for body sampling. The pre-processing times $T_S^{\mathcal{L}}$ and $T_S^{\mathcal{F}}$ to construct $\mathcal{S}_{\mathcal{L}}^-$ and $\mathcal{S}_{\mathcal{F}}^-$ are exponential in the worst case. It follows that the time complexity of CORSET is $\mathcal{O}(B|\mathcal{C}_R|T_f + T_S^{\mathcal{L}} + T_S^{\mathcal{F}})$. If $\mathcal{F}$ is sufficiently sparse, the exponential complexity is not a concern in practice. When $\mathcal{F}$ is dense, it is appropriate to use CORSET-GH, for which the time complexity is $\mathcal{O}(B|\mathcal{C}_R|T_f + T_S^{\mathcal{L}})$.

Space complexity. Space $S_R = \mathcal{O}(|\mathcal{D}| + |\mathcal{F}| + |\mathcal{L}|)$ is required to keep a single rule, as we store body, head, and support. Let S_S denote the space complexity of sampling. In both head and (owing to CFTP) body sampling, we only need to store a single weight value for

Algorithm 5 A greedy heuristic for body sampling.

Dataa dataset $\mathcal{D}$, sets $\mathcal{D}_H^+$, $\mathcal{D}_H^-$, parameters γ and ϵ.
Result: a body $B \in \mathcal{F}$.

let $F' \leftarrow F_{D^+} \setminus F_{D^-}$;
initialize $B \leftarrow$ an empty list, $Q_{\text{disc}} \leftarrow$ an empty list;
while $|B| < |F'|$ **do**
  $h^* \leftarrow \arg\max_{h \in F'} [\phi(H \cup \{h\}) - \phi(H)]$;
  add h^* to B;
  add $\phi(B)$ to Q_{disc};
  $F' \leftarrow F' \setminus h^*$;
  if $|\mathcal{D}[B]| < \epsilon\ |\mathcal{D}_H^+|$ **then**
    break
  end
end
$i^* \leftarrow \arg\max_{i=1,\ldots,|B|} Q_{\text{disc}}[i]$;
return $B[1 : i^*]$

each data record. Building $\mathcal{S}_{\mathcal{L}}^-$ and $\mathcal{S}_{\mathcal{F}}^-$ requires space $\mathcal{O}(|\mathcal{F}|^2)$ and $\mathcal{O}(|\mathcal{L}|^2)$, respectively. Furthermore, storing samples from $\mathcal{S}_{\mathcal{L}}^-$ ($\mathcal{S}_{\mathcal{F}}^-$) takes space $\mathcal{O}(|\mathcal{D}||\mathcal{S}_{\mathcal{L}}^-|)$ ($\mathcal{O}(|\mathcal{D}||\mathcal{S}_{\mathcal{F}}^-|)$). Despite this theoretical complexity, the graphs are very sparse in practice. Combining the above, we have that $S_S = \mathcal{O}(|\mathcal{L}|^2 + |\mathcal{F}|^2 + |\mathcal{D}||\mathcal{S}_{\mathcal{L}}^-| + |\mathcal{D}||\mathcal{S}_{\mathcal{F}}^-|)$ and the space complexity of CORSET is $\mathcal{O}(||\mathcal{F}|| + ||\mathcal{L}|| + |\mathcal{C}_R|S_R + S_S)$. When CORSET-GH is used, the greedy body sampler only takes space $\mathcal{O}(|\mathcal{F}|)$, and hence, S_S reduces to $S_{S+} = \mathcal{O}(|\mathcal{L}|^2 + |\mathcal{D}||\mathcal{S}_{\mathcal{L}}^-| + |\mathcal{F}|)$ so that the space complexity of CORSET-GH is $\mathcal{O}(||\mathcal{F}|| + ||\mathcal{L}|| + |\mathcal{C}_R|S_R + S_{S+})$.

8 Experimental evaluation

The main goals in this section are twofold: (i) we empirically verify that the sampling algorithms used by CORSET outperform prior approaches in a variety of settings; (ii) and CORSET (and in particular its two implementations CORSET-SURS and CORSET-GH) delivers a concise set of rules while still providing competitive performance in multi-label classification. We first present the experimental settings and then the results.'

8.1 Experimental setup

We describe the data, the baseline methods and the parameter settings used in the experiments. The design choices specific to the experiments comparing different sampling approaches are deferred to the dedicated section (Sect. 8.2).

Datasets. We use both synthetic and real-world datasets.

We use synthetic datasets to better understand the behavior of the methods with respect to different parameters. Data are obtained from a set of generating rules, and as a consequence, a notion of ground truth is available. For simplicity, bodies and heads are assumed to be composed of 3 attributes and labels, respectively. Moreover, different bodies do not share attributes, while heads may share labels. For each generating rule, we sample its support either (i) uniformly at random, or (ii) from a skewed distribution where a small subset of rules covers a large portion of the data, mimicking the typical behavior of real-world data. Supports of different rules may thus have arbitrary intersection. To obtain the synthetic dataset

from the generating rules, we initialize the feature matrix and the label matrix to contain all zeros. Then, for each generating rule, once its support is sampled, we set to 1 its attributes and labels over its support. The case of skewed support (ii) is expected to reflect more accurately real-world data, where there are typically a small set of very general rules and a larger number of specific rules.

Further, noise is injected by flipping each entry in the feature and label matrices with a fixed probability.

For real-world data, we use 9 heterogeneous benchmark datasets for multi-label classification.[1] Summary statistics of the datasets are shown in Table 1. Categorical and numerical features are converted to binary form. For simplicity, we convert numerical features into binary ones by setting to 0 all values lower than a given percentile p (90-th percentile by default) and by setting to 1 the rest of the values. A more refined pre-processing is advisable to improve performance.

Next, we provide a brief description of each of the datasets utilized in our experiments. The mediamill [38] dataset is used for generic (broadcast news) video indexing. The Yelp [39] dataset contains reviews of clients for various businesses that are used to classify the quality of the businesses. The corel-5k [40] dataset is a benchmark for image classification. The bibtex [41] dataset, introduced in Sect. 1, is composed of BIBTEX entries associated with publications, and the goal is to assign a list of tags to each publication. The enron [42] dataset consists of a collection of email messages that were categorized into topic categories. The medical [43] dataset is used to predict a number of diseases, given clinical free text, and it is described more in depth in Sect. 9. The birds [44] dataset is used to predict the set of birds species that are present from a ten-second audio clip. The emotions [45] dataset classifies music into emotions that it provokes. The CAL500 [46] dataset consists of songs and the classifier maps each song to a number of semantic concepts.

Metrics. To measure the quality of a classifier, we primarily use the popular balanced micro F_1 score, which micro-averages precision and recall. In addition, for real data, we report the balanced macro F_1 score, which macro-averages precision and recall, as well as the Hamming accuracy (or Hamming score), defined as the fraction of label occurrences that are correctly predicted.

To monitor rule diversity, we report the average pairwise intersection size between the coverage of different rules. To assess interpretability, we rely primarily on the number of rules $\mathcal{R}$, but we also consider the number of conditions in each rule, and the degree of association of the rule heads.

Baselines. We compare our classifier with three baselines.

SECO [47] is a rule-based classifier, which extracts new rules iteratively and discards the associated covered examples from the training data if enough of their labels are predicted by already learned rules. To learn rules, SECO starts from the most general body and then it proceeds by adding conditions. Given a rule body, SECO searches for the best possible head according to a metric, while pruning the search space by exploiting properties of the metric, and introducing bias towards heads with multiple labels.

BOOMER [2, 15] utilizes the gradient-boosting framework to learn ensembles of single-label or multi-label classification rules that are combined additively to minimize the empirical risk with respect to a suitable loss function.

SVM- BR [48, 49] is a linear support vector machine classifier based on the binary relevance approach, whereby each label is treated independently. SVM- BR, unlike CORSET, BOOMER and SECO is not a rule-based model and does not offer opportunities for interpretation.

[1] http://mulan.sourceforge.net, https://www.uco.es/kdis/mllresources/.

Table 1 Summary statistics of the datasets used in the experimental evaluation

Dataset	Instances	Attributes	Labels	Cardinality	Distinct
mediamill	43 907	120	101	4.38	6 555
Yelp	10 810	671	5	1.64	32
corel-5k	5 000	499	374	3.52	3 175
bibtex	7 395	1 836	159	2.40	2 856
enron	1 702	1 001	53	3.38	753
medical	978	1 449	45	1.24	94
birds	645	260	19	1.01	133
emotions	593	72	6	1.87	27
CAL500	502	68	174	26.04	502

The last two columns refer to the average number of labels per example, and the total number of distinct label sets

Therefore, in our experimental evaluation, SVM- BR is used as a representative black-box machine-learning algorithm. The goal of the comparison between CORSET and SVM- BR is to demonstrate that simple rule-based models, if appropriately learned, can be competitive with popular black-box machine-learning algorithms. In other words, we seek to demonstrate that if there is a price to pay in performance for pursuing interpretability, that price is consistently small, and hence the pursuit of interpretability is justified.

While we consider multiple baselines to compare CORSET against, BOOMER is the most important one because it achieves the state-of-the-art performance in rule-based multi-label classification.

However, BOOMER and CORSET are significantly different in nature. As an ensemble method, BOOMER gains its good performance by learning a large number of *weak* (i.e., poorly optimized) rules, which are aggregated for prediction. BOOMER only optimizes for classification performance and does not favor ensembles composed of a small number of rules.

On the other hand, to achieve high interpretability, CORSET is the first multi-label rule-based classifier which optimizes both classification performance and rule set conciseness. CORSET does not aggregate weak rules in an ensemble like BOOMER. Instead, CORSET builds a small set of carefully selected rules and predicts using the union of the predictions of the selected rules.

As a consequence of the illustrated differences between CORSET and BOOMER, BOOMER generally requires a larger number of rules than CORSET to achieve a similar level of classification performance, as we demonstrate through our experiments.

For the synthetic datasets, we focus on comparing our approach with BOOMER for increasing number of rules, whereas for the real-world datasets we consider all baselines.

Parameter setting. For the experiments with synthetic data, we explore the scalability of our algorithm with respect to the number of attributes and labels, as well as robustness with respect to noise. We vary the level of noise (the proportion of flipped entries in the feature matrix and the label matrix), and the number of attributes and labels by a geometric progression of ratio 1.5. When not varied, the number of attributes and labels is fixed to 100, and the noise level to 0.01. When the noise is varied there are 10 ground truth rules, otherwise the number of generating rules increases with the size of the data and it is given by $\lfloor \frac{\min(|\mathcal{F}|,|\mathcal{L}|)}{3} \rfloor$.

For the experiments with real-world data, we tune the hyper-parameters of BOOMER and CORSET via random search to minimize micro-averaged F_1 on a validation set.

As concerns BOOMER, we optimize the shrinkage parameter η, which controls the impact of individual rules, over [0.1, 0.3, 0.5] and the $L2$ regularization parameter λ in [0.0, 1, 10.0]. A key strength of BOOMER is its ability to optimize different *decomposable* and *non-decomposable* loss functions. For hyper-parameter tuning in our experiments, we consider all 4 loss functions provided in the available implementation,[2] namely a variant of logistic loss applied to each label individually and a variant of the logistic loss applied to all labels simultaneously (that are continuous surrogates for the standard Hamming loss and 0/1 loss, respectively), as well as a variant of the squared error loss and a variant of the squared hinge loss, both applied to each label individually. BOOMER additionally allows to learn rules with heads composed of a single label or all labels. In hyper-parameter tuning, we consider both single-label rule heads and rule heads composed of all labels.

CORSET does not support different loss functions as BOOMER does. Instead, CORSET is tailored to the optimization of the objective function in Eq. (8). Subsequently, the loss function implicitly minimized by CORSET is the objective function in Eq. (8) reversed in sign. For CORSET, the size of each sampled pool of rules $\mathcal{C}$ does not need to be tuned. Larger $\mathcal{C}$ improves performance at the cost of increased running time. While tuning we fix $\mathcal{C} = 150$, otherwise $\mathcal{C}$ is set to 500 by default. For hyper-parameter tuning, all hyper-parameters are searched in the range of (0, 1), except λ, which is searched in $(10^{-2}, 10^2)$.

Moreover, we also investigate the impact of λ on the diversity and accuracy of the set of chosen rules $\mathcal{R}$, by varying it in a geometric progression of ratio 10.

All reported experimental results are obtained as average over 10 repetitions.

Implementation. Experiments are executed on a machine with 2×10 core Xeon E5 2680 v2 2.80 GHz processor and 256 GB memory. Our implementation is available online. [3] To speed up and facilitate hyper-parameter tuning for CORSET, we have implemented two practical changes. First, we run only the first round of greedy selection in Algorithm 1. The second round guarantees the approximation factor, but often offers a modest increase in performance, not worth the increase in running time. Second, in the experiments with real data, we pass as input to CORSET the number of rules to be returned (at most 150) instead of the tolerance parameter (c in Algorithm 1) to reduce variability and simplify hyperparameter optimization.

8.2 Sampling performance

We evaluate the performance of our sampling algorithms and compare them against a few alternatives.

Head sampling. In this setting, we consider the task of sampling heads according to the area function. We compare two samplers: our sampler described in Alg. 4 and the original two-stage pattern sampler introduced by Boley et al. [25] (the baseline). They differ in the sample space they use: the baseline sampler operates in the original label space, i.e., the powerset of all labels, while ours uses a subspace of it, namely the interpretable label space, introduced in Sect. 6.2.

For each configuration of dataset and sampler, we sample 1000 heads. For each sampled head H, we evaluate its quality using the logarithm of its area and its support size, i.e., $\log(\text{area}(H))$ and $\log(|\mathcal{D}[H]|)$. In addition, we quantify the association of the labels in H (our proxy of interpretability) using the *edge density* of the subgraph induced by H.

[2] https://github.com/mrapp-ke/Boomer.

[3] https://github.com/DiverseMultiLabelClassificationRules/CORSET.

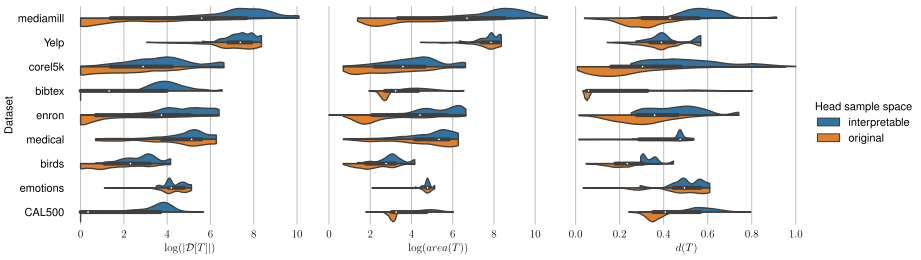

Fig. 5 Distributions of quality measures for the heads obtained by two samplers. Both samplers target at the area function, but operate under different sample spaces: the interpretable label space (Sect. 6.2) or the original space (used in [25]). We consider three measures: log of support size (left), log of area (middle), and edge density of the label subgraph (right). For all measures, the larger the values, the better

Specifically, given a label graph $\mathcal{G} = (V, E, p)$ and a set of labels (nodes) H, denote $\mathcal{G}[H]$ as the subgraph induced by H in $\mathcal{G}$. Then, the edge density of $\mathcal{G}[H]$ is defined as $d(\mathcal{G}[H]) = \left(\sum_{e \in \mathcal{G}[H]} p(e)\right) / (|H| \times (|H| - 1))$. We denote $d(H) = d(\mathcal{G}[H])$ for brevity.

In Fig. 5, we show the distributions of the metric scores, visualized using violin plots [50]. On the one hand, our sampler produces better heads than the baseline according to all measures, on average. On the other hand, the best head (by a certain metric) obtained by our sampler is no worse than the baseline. Further, it can be seen that on datasets where the heavy hitter problem is pronounced, such as bibtex, the baseline suffers the most and many heads have zero support size.

Body sampling. We then consider sampling bodies according to the discriminativity function and compare two samplers: the one used by CORSET, which samples from a reduced space (Sect. 6.3), and a baseline [26], which instead samples from the original feature space.

For each dataset, we consider the top-50 heads ranked by area. For each head and sampler, we sample 100 bodies and take the best one according to a specific goodness measure.

We consider two goodness measures for the bodies: the quality function q (as in Eq. (1)) used in our problem formulation and the micro-averaged F_1 score.

For each configuration of dataset and head, we report the relative difference between the best scores obtained by the samplers. For a given metric m, the bodies B obtained by our sampler, and the baseline B_{B}, the relative difference between B and B_{B} according to m is calculated as: $\Delta_m(B; B_{\mathrm{B}}) = (m(B) - m(B_{\mathrm{B}})) / (m(B_{\mathrm{B}}))$. For brevity, we use Δ_m instead. To bring down the numerical scales of different datasets to a similar level, we report the logarithmic version of Δ_m: $\log(\Delta_m) = \mathrm{sign}(\Delta_m) \log(|\Delta_m| + 1)$.

The distributions of $\log(\Delta_m)$ are summarized in Fig. 6 using letter-value plots [51].

The two measures give different impressions: according to the F_1 score, our sampler is almost always the best choice, while for q, our sampler still outperforms the baseline, but to a lesser extent than in the comparison based on the F_1 score (e.g., on birds and corel-5k).

Effect of proposal distributions on convergence speed. We examine the proposal distribution used by the CFTP subroutine (Algorithm 3) and study its effect on the convergence speed of the Markov chain. In particular, we compare the proposal in Eq. (18) (used By Boley et al. [26]) against the proposal in Eq. (19) (used by CORSET). Recall that the latter is a tight approximation of the target distribution, while the former is not. We measure the convergence speed of a Markov chain by N, the number of Monte Carlo samples that are drawn until the chain coalesces.

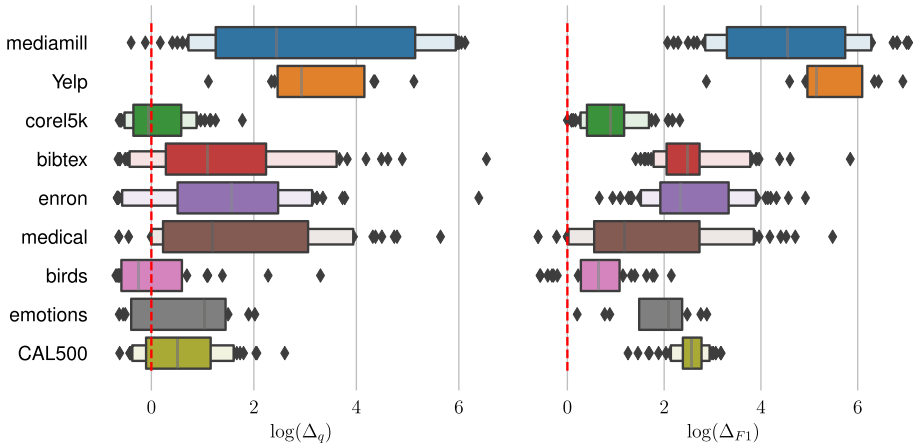

Fig. 6 Distributions of the relative difference for two samplers (our sampler and the baseline) evaluated on two metrics: the quality function q (left) and the micro-averaged F_1 score (right). Both samplers target at the discriminativity function, but operate under different sample spaces. Having a relative difference value greater (smaller) than zero indicates our sampler (the baseline) is better. Dashed vertical lines are drawn to illustrate such boundary

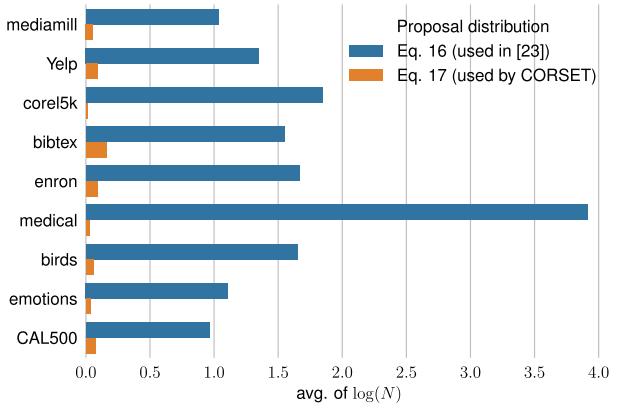

Fig. 7 Effect of the proposal distribution on the number of samples until coalescence. The smaller the value, the better

For each dataset and proposal distribution, we sample 50 random labels. Further, for each label, 100 bodies are sampled, resulting in a total of 5000 samples. Finally, we report the average of $\log(N)$.

The results are summarized in Fig. 7 and show clear superiority of our proposal distribution.

8.3 Classification performance

Synthetic datasets. Results on synthetic datasets, both for data generated from rules with uniform and skewed coverage, are shown in Fig. 8.

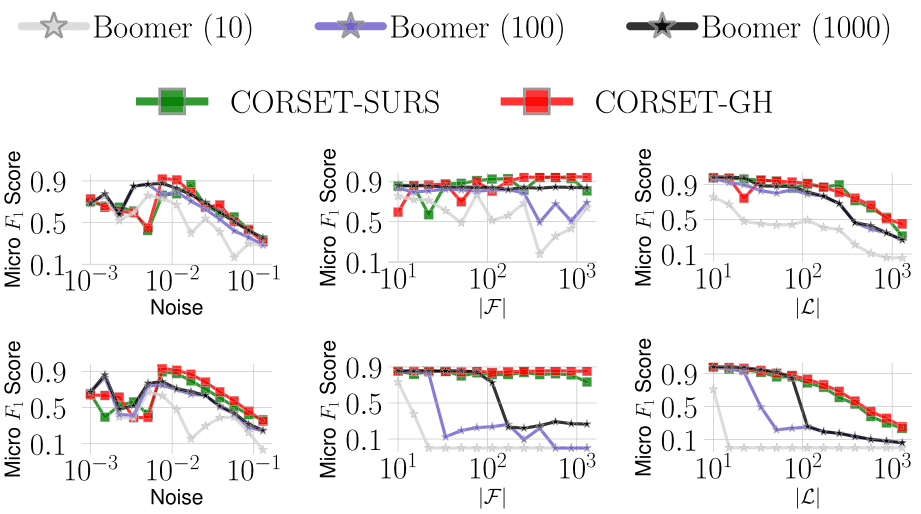

Fig. 8 Synthetic datasets generated from rules with uniform (top) and skewed (bottom) coverage. Micro-averaged F_1 score against proportion of noise (left), number of attributes (middle) and labels (right). The x-axis is in log scale

In general, when the noise level is not too large, CORSET tends to recover the generating rules.

In the experiment with synthetic data, we let CORSET run until the stopping condition is met, with $\tau = 0.01$. The number of rules retrieved by CORSET typically coincides with the number of generating rules, and it is always at most the maximum number of generating rules, 33. On the other hand, BOOMER based on 10 rules consistently offers poor performance. The classification accuracy of BOOMER increases when the number of rules increases, but even with 1000 rules, our method outperforms BOOMER while using a very concise set of rules. Unlike BOOMER, CORSET seeks to uncover the true set of generating rules and only use those for classification. Thus, the experiments with synthetic datasets clearly show the advantage of our approach. Also note that the performance of CORSET, unlike that of BOOMER, does not significantly deteriorate when $|\mathcal{F}|$ increases.

Real datasets: classification performance and interpretability. Results on real datasets, for classification performance and interpretability, are shown in Table 2, 3 and 4 where classification performance is measured by micro-averaged F_1 score (optimized in hyper-parameter tuning), macro-averaged F_1 score, and Hamming score, respectively. The tables also show the number of rules that is our main measure of interpretability. While we show results for three different classification performance metrics, the algorithms are tuned to optimize the micro-averaged F_1 score. Furthermore, caution is required in assessing the Hamming scores because they tend to be biased towards more conservative classifiers that predict less labels to be present.

If the training does not terminate within 12 hours, we report NA in the corresponding table entry. Table 2, 3 and 4 show that BOOMER requires very large sets of rules to achieve competitive performance. Thus, interpretability of BOOMER rules is questionable. Similarly, SVM- BR performs well but it is not designed for interpretability. Instead, CORSET extracts a small set of rules, guaranteeing ease of interpretation, and yet it is consistently competitive with the baselines on all the datasets, according to all considered performance metrics. CORSET

Table 2 Micro-averaged F_1-scores on real datasets achieved by CORSET-SURS, CORSET-GH, SECO, BOOMER with increasing number of rules, and SVM- BR

Dataset	CORSET-SURS	CORSET-GH	SECO	BOOMER (10)	BOOMER (100)	BOOMER (1000)	SVM- BR	$\|\mathcal{R}\|$ CORSET-SURS	$\|\mathcal{R}\|$ CORSET-GH	$\|\mathcal{R}\|$ SECO
mediamill	0.44	0.51	NA	0.45	0.51	0.52	0.50	150	150	NA
Yelp	0.66	0.64	NA	0.50	0.63	0.73	0.70	67	82	NA
corel-5k	0.18	0.18	NA	0.07	0.07	0.08	0.16	142	150	NA
bibtex	0.36	0.40	NA	0.11	0.21	0.36	0.41	74	150	NA
enron	0.55	0.53	NA	0.39	0.51	0.56	0.52	41	48	NA
medical	0.81	0.83	0.63	0.35	0.64	0.91	0.99	27	88	199
birds	0.37	0.42	0.39	0.01	0.34	0.46	0.42	42	48	122
emotions	0.53	0.54	0.53	0.31	0.49	0.50	0.56	42	68	199
CAL500	0.29	0.32	NA	0.31	0.31	0.34	0.53	150	150	NA

The last three columns show the number of rules for CORSET-SURS, CORSET-GH, and SECO

Table 3 Macro-averaged F_1-scores on real datasets achieved by CORSET-SURS, CORSET-GH, SECO, BOOMER with increasing number of rules, and SVM- BR

Dataset	CORSET-SURS	CORSET-GH	SECO	BOOMER (10)	BOOMER (100)	BOOMER (1000)	SVM- BR	$\lvert\mathcal{R}\rvert$ CORSET-SURS	$\lvert\mathcal{R}\rvert$ CORSET-GH	$\lvert\mathcal{R}\rvert$ SECO
mediamill	0.051	0.071	NA	0.018	0.041	0.051	0.040	150	150	NA
Yelp	0.62	0.59	NA	0.24	0.48	0.64	0.64	67	82	NA
corel-5k	0.11	0.19	NA	0.00069	0.0032	0.0011	0.042	142	150	NA
bibtex	0.20	0.11	NA	0.0065	0.051	0.18	0.31	74	150	NA
enron	0.21	0.23	NA	0.045	0.086	0.15	0.22	41	48	NA
medical	0.88	0.51	0.35	0.050	0.23	0.69	0.99	27	88	199
birds	0.29	0.26	0.31	0.0062	0.16	0.33	0.30	42	48	122
emotions	0.60	0.53	0.53	0.27	0.43	0.49	0.60	42	68	199
CAL500	0.47	0.39	NA	0.071	0.14	0.17	0.49	150	150	NA

The last three columns show the number of rules for CORSET-SURS, CORSET-GH, and SECO

Table 4 Hamming scores on real datasets achieved by CORSET-SURS, CORSET-GH, SECO, BOOMER with increasing number of rules, and SVM- BR

Dataset	CORSET-SURS	CORSET-GH	SECO	BOOMER (10)	BOOMER (100)	BOOMER (1000)	SVM- BR	$\|\mathcal{R}\|$ CORSET-SURS	$\|\mathcal{R}\|$ CORSET-GH	$\|\mathcal{R}\|$ SECO
mediamill	0.96	0.95	NA	0.96	0.96	0.97	0.97	150	150	NA
Yelp	0.75	0.72	NA	0.74	0.79	0.83	0.82	67	82	NA
corel-5k	0.99	0.99	NA	0.98	0.99	0.99	0.99	142	150	NA
bibtex	0.99	0.98	NA	0.98	0.98	0.98	0.98	74	150	NA
enron	0.94	0.94	NA	0.94	0.95	0.95	0.95	41	48	NA
medical	0.99	0.99	0.98	0.96	0.98	0.99	0.99	27	88	199
birds	0.94	0.93	0.94	0.94	0.94	0.95	0.94	42	48	122
emotions	0.69	0.60	0.70	0.70	0.71	0.71	0.74	42	68	199
CAL500	0.87	0.84	NA	0.84	0.84	0.85	0.87	150	150	NA

The last three columns show the number of rules for CORSET-SURS, CORSET-GH, and SECO

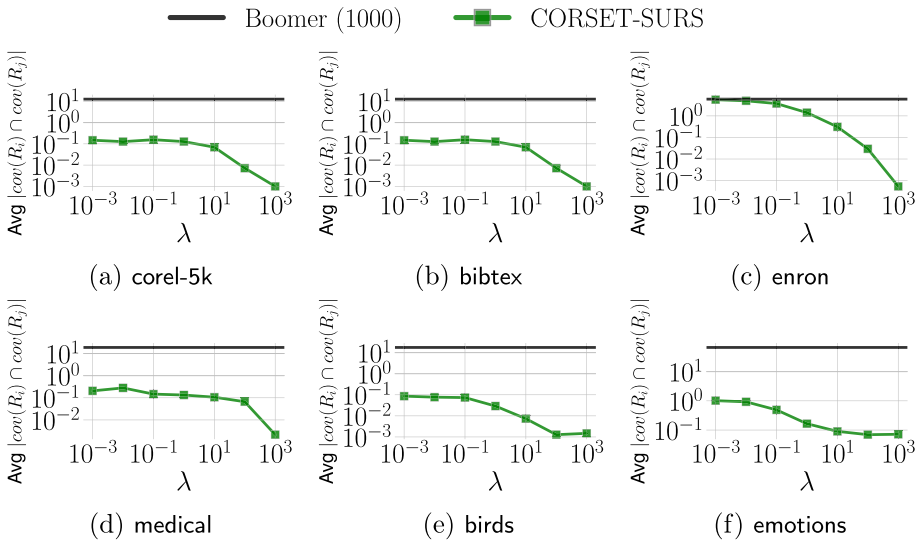

Fig. 9 Average coverage overlap between pairs of rules as a function of λ (lower values indicate higher diversity) on real datasets. Both axes are in log scale

generally requires fewer rules than rule-based alternatives to attain similar performance in multi-label classification. Further, it is never drastically worse than BOOMER with 1000 rules or even SVM- BR, suggesting that the *price of interpretability*, if there is one, is small when CORSET is used. Finally, note that CORSET-GH often outperforms CORSET-SURS but using a larger set of rules.

Real datasets: diversity and impact of λ. A fundamental characteristic of CORSET is that it allows to control the degree of diversity in the set of recovered rules via a single tunable parameter λ. In Fig. 9, we show for a subset of datasets that the shared coverage within rules is lower for CORSET than for BOOMER, and moreover that increasing the value of λ is very effective in reducing overlap between rules. As the impact of λ on overlap is not significantly different in CORSET-SURS and CORSET-GH, we only show results for the former.

In addition, since λ determines the weight assigned to the diversity term in Eq. (8), it also affects the performance of CORSET. Figure 10, however, shows that the performance of CORSET, as measured by the micro-averaged F_1 score, does not vary monotonically with λ, although there is evidence that very large values (e.g., $\lambda = 1000$) typically yield poor classification. The level of rule diversity leading to optimal classification performance varies in different datasets. Hence, in practice, λ must be carefully tuned to optimize the performance of CORSET. Note that the results in Fig. 10 are obtained by fixing the number of rules to 100 for both CORSET and BOOMER and by choosing the best between CORSET-SURS and CORSET-GH in terms of micro-averaged F_1 score.

Real datasets: body and head sizes. In addition to the number of rules considered before, it is possible to measure the interpretability of a model by the sizes of the bodies of its rules (i.e., the number of predicates of each rule in a rule set).

The number of rules within a rule set may be limited; however, if the body of the rules contains a substantial number of conditions, the interpretability of the rule set may be impaired. We show that this is not the case for CORSET.

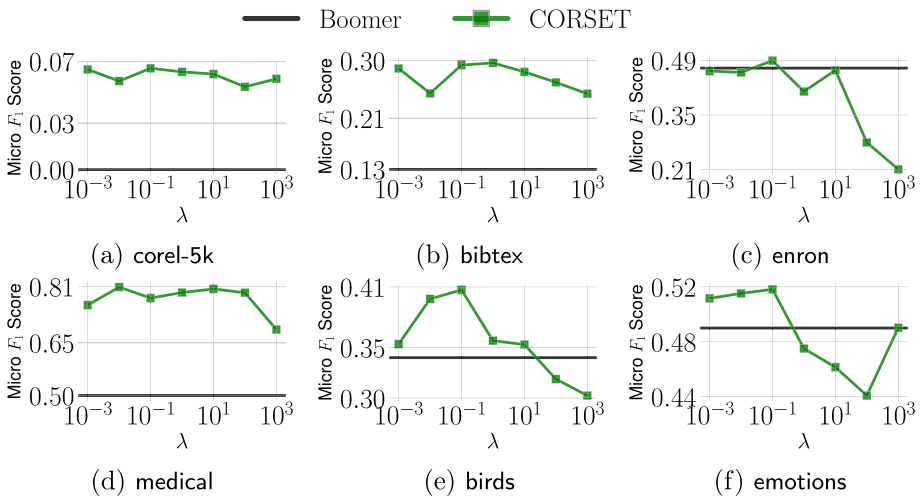

Fig. 10 Micro-averaged F_1 score as a function of λ (lower values indicate higher diversity) on real datasets. The x-axis is in log scale

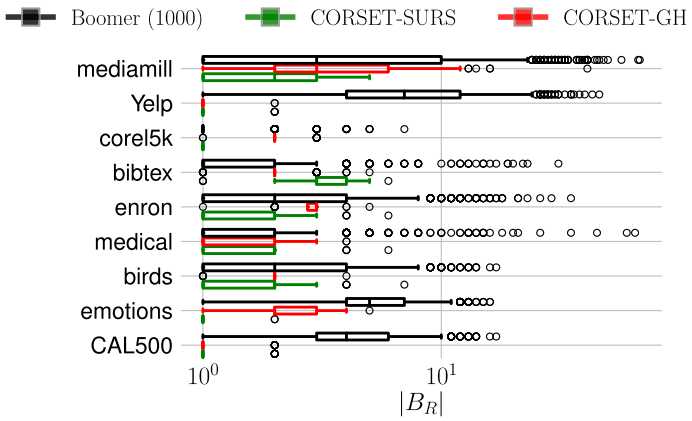

Fig. 11 Distributions of the number of attributes in the bodies of the rules extracted by CORSET-SURS, CORSET-GH, and BOOMER with 1000 rules. Smaller values indicate fewer attributes. The x-axis is in log scale

The number of attributes in bodies may vary considerably depending on whether CORSET-SURS or CORSET-GH is used. Figure 11 shows the distributions of body sizes over the rule sets obtained by CORSET-SURS, CORSET-GH and BOOMER (with 1000 rules). We consider the bodies of the rules whose results are summarized in Table 2. Clearly, the rules obtained by CORSET-SURS have consistently a small number of attributes (always ≤ 10). In general, the distribution of the sizes of the bodies of BOOMER is comparable to that of CORSET-SURS and CORSET-GH. Nonetheless, BOOMER produces a significant number of extremely long bodies in all datasets.

Although CORSET-SURS does not explicitly limit body sizes, the limited sizes are by-products of the reduced sample space constructed prior to sampling. As illustrated in Sect. 6.3, samples in the reduced space correspond to highly probable cliques in the feature graph. By

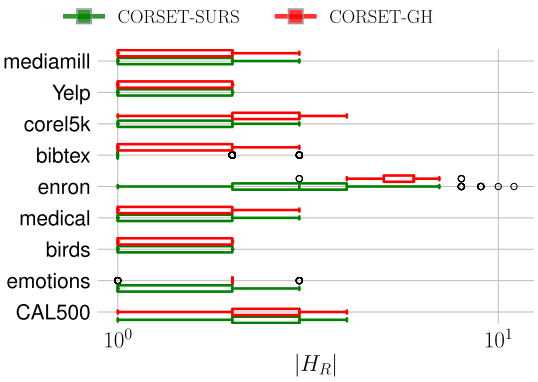

Fig. 12 Distributions of the number of labels in the heads of the rules extracted by CORSET-SURS and CORSET-GH. Smaller values indicate smaller heads. The x-axis is in log scale

the way that "probable" is defined, such cliques typically have a small number of nodes (attributes).

Similarly, CORSET-GH forms bodies greedily selecting attributes to maximize a modified version of discriminativity while controlling the balance between generality and precision with a user-specified parameter γ. Small values of γ favor generality, allowing for more mistakes and leading to bodies with a limited number of conditions, whereas larger values of γ generate more specific rules with a larger number of attributes. In Fig. 11, we show results for $\gamma = 0.5$. For this value of γ, bodies sampled by CORSET-GH tend to have more attributes than those sampled by CORSET-SURS. Nevertheless, remarkably large bodies are present only in the mediamill dataset, whereas BOOMER relies on rules with large bodies in all datasets.

Shorter heads do not necessarily correspond to more interpretable heads. However, small heads are typically easy to interpret. For completeness, in Fig. 12, we additionally show, with the same parameter settings, results concerning the distribution of the heads in the rules chosen by CORSET-SURS and CORSET-GH. BOOMER is not included in the comparison because BOOMER allows the user to decide whether to learn rules having heads composed of either a single label or all labels.

The distributions of the head lengths in the different datasets suggest that CORSET consistently learns rules with heads composed of a limited number of labels. Similarly to the case of body sampling, this experimental finding is largely a consequence of the construction of the interpretable label space prior to sampling, whereby long and hard-to-interpret heads are discarded.

Real datasets: running time. The main goal pursued in this work is to achieve a good balance between interpretability and classification quality. CORSET carefully chooses each rule to be used for prediction, which requires a significant amount of time.

Fast training is not a primary concern of our work.

Nonetheless, it is important to show that the execution of CORSET terminates within a reasonable amount of time in all the considered datasets. Table 5 reports the training time of CORSET and its competitors.

The time incurred in obtaining predictions for new data records is generally modest for all algorithms, thus prediction times are not included in the measurements reported in Table 5.

CORSET always finishes training in less than 12 minutes. BOOMER and SVM- BR are faster to train than CORSET in most cases. Their better efficiency can be attributed to the implementation language (C/C++ against Python and Cython for CORSET) and, most importantly, to the continuous nature of the optimization problems they solve.

On the other hand, SECO is drastically slower than all the other algorithms.

Table 5 Execution time (in seconds) of CORSET-SURS, CORSET-GH, SECO, BOOMER with increasing number of rules and SVM- BR

Dataset	CORSET-SURS	CORSET-GH	SECO	BOOMER (10)	BOOMER (100)	BOOMER (1000)	SVM- BR
mediamill	640.081	713.28	NA	1.62	15.17	146.020	46.070
Yelp	398.22	92.74	NA	0.46	2.97	19.51	1.70
corel-5k	46.27	45.66	NA	0.62	6.44	60.20	2.32
bibtex	470.46	120.61	NA	0.64	5.56	56.22	7.71
enron	35.52	89.17	NA	0.15	0.64	5.96	1.28
medical	40.88	24.89	2011.83	0.29	1.50	9.78	2.64
birds	55.66	9.85	9163.79	0.18	0.21	1.24	0.19
emotions	5.44	10.51	2557.92	0.092	0.10	1.18	0.065
CAL500	572.33	634.22	NA	0.39	2.56	27.33	3.56

The reported times indicate the running time required to learn a multi-label classifier before it can be used for predictive purposes

For CORSET, a considerable amount of time is spent in the pre-processing stage to construct the interpretable label space for head sampling and, in CORSET-SURS, the reduced sample space for body sampling, which, as mentioned in Sect. 6.3, is a potential bottleneck.

Therefore, in practice, in case one is interested in learning multiple rule sets (e.g., corresponding to different values of λ), it is sufficient to carry out the pre-processing stage just once. While CORSET-GH overcomes the scalability issues that may arise in constructing the reduced sample space for body sampling, it requires significantly more time than CORSET-SURS to perform a single iteration. Subsequently, when the attribute space is sparse, the construction of the reduced sample space is fast and CORSET-SURS runs faster than CORSET-GH.

As explained in the end of Sect. 8.1, in our experimental evaluation with real data, by default we only execute the first round of greedy selection of CORSET. This follows since the gain in classification quality offered by the second round is generally not worth the increase in running time. In particular, the running time incurred by CORSET with both rounds of greedy selection is often 2 to 3 times as high as the running time incurred by CORSET with only one round of greedy selection.

9 Case study

We carry out a case study using the medical [10] dataset to showcase an application of CORSET in the medical domain. When machine learning models are used to assist medical decision making, interpretability of the models is of particular importance [10], since medical decisions are often high-stake and may deeply affect the lives of patients. In the context of rule-based models, human practitioners may have to assess the rules one-by-one to ensure that they have sufficient understanding of the models. Arguably, rule sets that are very large and highly redundant are costly to check, thus undesirable. Concise rule sets are instead particularly valuable. In order to effectively illustrate this point through a concrete example, we use the medical [43] dataset, which contains fully anonymized clinical free text in medical records, each labeled with one or more disease names. The disease names follow the *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM) standard, a globally used mapping of disease names to unique codes [52].

Because the medical practitioner wishes to evaluate the rules qualitatively using her expertise it is necessary to focus on a small set of rules, for instance, composed of 10 rules.

The first 10 rules picked by CORSET are general (as indicated by their coverage) and easy (often trivial) to interpret. All of them have coverage above 40, and the average coverage is 123. In contrast, the first 10 rules picked by BOOMER appear to be opaque and exhibit null coverage. In fact, CORSET using 10 rules achieves micro-averaged F_1 score above 0.7 while BOOMER with the same number of rules yields a drastically lower micro-averaged F_1 score. In particular, as shown in Sect. 8, BOOMER does not grant high-quality classification using 10 or 100 rules, but only using 1000 rules. A set of 1000 rules is impractical for medical practitioners to inspect.

Therefore, BOOMER, which does not impose constraints on model complexity, contradicts the principles of interpretable machine learning and may not be not be suitable for applications where interpretability is an important concern.

As regards the first 10 rules learned by CORSET, however, it is important to note that most of the remarkably accurate and general rules learned by CORSET do not necessarily generate novel insight. For example, the first rule selected by CORSET is $R_1 = (\{fever\ \&\ cough\} \rightarrow \{fever\ \&\ cough\})$. We argue that this is not a drawback of CORSET; rather it is a characteristic of the data.

The most general and accurate rules are necessary for CORSET to perform well in the multi-label classification task. For instance, R_1 has an uncovered area of 238 and adjusted accuracy of 1.06.

Even if the rules represent somewhat trivial associations between some words and some diseases, they are beneficial in revealing the salient patterns in the data, potentially confirming the medical practitioner's knowledge and helping build trust in the models.

It is also possible to run CORSET to extract a larger rule set, containing more specific and non-trivial rules. Further, if required, it is simple to filter out all rules capturing trivial associations during post-processing. As a demonstration, we run CORSET for 30 iterations and filter out all rules such that the attributes in the body are found as part of the labels in the tail (e.g., $R_1 = (\{fever\ \&\ cough\} \rightarrow \{fever\ \&\ cough\})$). In Table 6, we list the first 10 rules after the post-processing (in the order of rule selection by CORSET). The first rule maps the word "myelomeningocele" to the label "spina bifida without mention of hydrocephalus". This is a rule of broad coverage. As "myelomeningocele" is one of four types of "spina bifida," the rule is easy to interpret; "myelomeningocele" is likely to identify "spina bifida without mention of hydrocephalus" and since rules for other types of "spina bifida" are not present, we infer that this is likely to be the prevalent type of "spina bifida" in the data at hand. As another example, the fourth rule in the table, which is also general and easy to interpret, contains two attributes in its body and, in particular, it maps the words "lobe" and "atelectasis" to "pulmonary collapse." The rule suggests that when a lobe (i.e., a section) of the lung is affected by atelectasis, there is a significant chance that the lung will collapse. This is likely due to the fact that atelectasis can cause a decrease in the amount of oxygen that reaches the lung tissue and can lead to the accumulation of fluid or mucus in the lung, which can further decrease lung expansion and increase the risk of collapse. An atelectasis may also indicate the collapse of the entire lung, but the rule under consideration suggests that in the medical dataset the word *lobe* leads to a better discrimination of the subsequent occurrence of "pulmonary collapse."

As a final example, we consider a rule with a longer head. The eighth rule in the table maps the word "throat" to two diseases: "fever and physiologic disturbances of temperature regulation" and "acute pharyngitis." The rule indicates that there is a strong correlation in the data between the presence of symptoms in the throat and the subsequent occurrence of

Table 6 First 10 rules chosen by CORSET (body, head and coverage) after post-processing to filter out "trivial" rules

B_R	H_R	cov(R)
Myelomeningocele	Spina bifida without Mention of hydrocephalus	43
Pyelectasis	other specified disorders Of kidney and ureter	43
Uti	Urinary tract infection	23
Lobe & Atelectasis	pulmonary collapse	40
Enuresis	Urinary incontinence	65
left & Hydroureter & history	Hydroureter	6
Stones & Kidney & History	Calculus of kidney	6
Throat	Fever and physiologic disturbances of Temperature regulation & Acute pharyngitis	11
pneumonia & Lobe	Pneumonia	4
Turner	Gonadal dysgenesis & Other Specified anomalies of kidney	4

fever and physiologic disturbances of temperature regulation and acute pharyngitis. This is probably attributed to the fact that the pharynx in the throat is a sensitive area that can be easily infected by pathogens and viruses, which can cause acute pharyngitis, i.e., an inflammation of the pharynx, and fever as part of the immune response of the body. For instance, both acute pharyngitis and fever may be caused by the widespread influenza virus. Additionally, physiologic disturbances of temperature regulation may be due to the attempt of the body experiencing fever to fight off the infection.

All rules in Table 6, even the ones with smallest coverage, are biologically meaningful as they can find scientific explanations.

Some of the rules may describe associations that may look obvious for medical practitioners. In this case, the medical practitioner only acquires information on which diverse patterns in the data are the most relevant for multi-label disease classification.

The medical dataset is relatively small and does not contain a large number of surprising patterns.

More unexpected rules can be found in the medical dataset. However, they would have limited support and lead to a decrease in classification performance, if chosen.

Regardless of the knowledge generated by the extracted rules, our in-depth examination of the medical dataset shows an example of application of CORSET in a consequential domain, demonstrating that the rule sets provided by CORSET are more interpretable and more practical to assist in high-stake decision making than those provided by existing algorithms such as BOOMER.

10 Conclusion

We propose a novel rule-based classifier, CORSET, for multi-label classification tasks. Our training objective explicitly penalizes rule redundancy, thereby encouraging the algorithm to learn a concise and easy-to-interpret set of rules.

Furthermore, we design a suite of fast sampling algorithms, which can efficiently generate rules with good accuracy and interpretability. We show through extensive experiments that our sampling algorithms are highly effective and that CORSET achieves competitive performance comparable to strong baselines, while offering better interpretability. Thus, CORSET achieves an unmatched compromise between performance and interpretability in multi-label classification, and, as demonstrated through a case study, it is particularly valuable in multi-label classification tasks where interpretability is of primary interest.

Our work opens interesting questions for future research. Can we design training objectives that reflect popular multi-label classification metrics, while producing concise rule sets? Can we use the techniques in this work to address the interpretability issue of existing rule-based classifiers?

Acknowledgements This research is supported by the Academy of Finland project MLDB (325117), the ERC Advanced Grant REBOUND (834862), the EC H2020 RIA project SoBigData++ (871042), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Author contributions All authors contributed to writing and reviewing the manuscript.

Funding Open Access funding provided by Aalto University.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Tidake V, Sane S (2018) Multi-label classification: a survey. Int J Eng Technol 7(19):1045–1054
2. Rapp, M., Mencía, E.L., Fürnkranz, J., Nguyen, V.-L., Hüllermeier, E.: Learning gradient boosted multi-label classification rules. In: Proceedings of machine learning and knowledge discovery in databases: European conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III, pp. 124–140 (2021). Springer
3. Rapp, M., Mencía, E.L., Fürnkranz, J., Hüllermeier, E.: Gradient-based label binning in multi-label classification. In: Proceedings of machine learning and knowledge discovery in databases. Research track: european conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Part III 21, pp. 462–477 (2021). Springer
4. Colantonio A, Di Pietro R, Ocello A, Verde NV (2011) Visual role mining: a picture is worth a thousand roles. IEEE Trans Knowl Data Eng 24(6):1120–1133
5. Read J (2008) A pruned problem transformation method for multi-label classification. In: New Zealand computer science research student conference, p 41

6. Elisseeff A, Weston J (2001) A kernel method for multi-labelled classification. Adv Neural Inf Process Syst 14:681–687
7. Zhang M-L, Zhou Z-H (2007) Ml-knn: A lazy learning approach to multi-label learning. Pattern Recogn 40(7):2038–2048
8. Crammer K, Singer Y (2003) A family of additive online algorithms for category ranking. J Mach Learn Res 3:1025–1058
9. Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 267:1–38
10. Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C (2022) Interpretable machine learning: Fundamental principles and 10 grand challenges. Stat Surv 16:1–85
11. Liu B, Hsu W, Ma Y, et al (1998) Integrating classification and association rule mining. In: Kdd, vol. 98, pp 80–86
12. Thabtah F, Cowling P, Peng Y (2005) Mcar: multi-class classification based on association rule. In: The 3rd ACS/IEEE international conference on computer systems and applications, p 33
13. Wang X, Yue K, Niu W, Shi Z (2011) An approach for adaptive associative classification. Expert Syst Appl 38(9):11873–11883
14. Zhang G, Gionis A (2020) Diverse rule sets. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1532–1541
15. Boley M, Teshuva S, Bodic PL, Webb GI (2021) Better short than greedy: Interpretable models through optimal rule boosting. In: Proceedings of the 2021 SIAM international conference on data mining (SDM), pp 351–359
16. Yu J, Ignatiev A, Stuckey PJ, Le Bodic P (2021) Learning optimal decision sets and lists with SAT. J Artific Intell Res 72:1251–1279
17. Ghosh B, Malioutov D, Meel KS (2022) Efficient learning of interpretable classification rules. arXiv preprint arXiv:2205.06936
18. Wang T, Rudin C, Doshi-Velez F, Liu Y, Klampfl E, MacNeille P (2017) A Bayesian framework for learning rule sets for interpretable classification. J Mach Learn Res 18(1):2357–2393
19. Fischer J, Vreeken J (2019) Sets of robust rules, and how to find them. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 38–54
20. Thabtah FA, Cowling P, Peng Y (2004) Mmac: A new multi-class, multi-label associative classification approach. In: Fourth IEEE international conference on data mining (ICDM'04), pp 217–224
21. Klein Y, Rapp M, Loza Mencía E (2019) Efficient discovery of expressive multi-label rules using relaxed pruning. In: Discovery sience, pp 367–382
22. Zhao Q, Bhowmick SS (2003) Association rule mining: a survey. Nanyang Technological University, Singapore vol 135
23. Fournier-Viger P, Lin JC-W, Vo B, Chi TT, Zhang J, Le HB (2017) A survey of itemset mining. Data Mining Knowl Discov 7(4):1207
24. Luna JM, Fournier-Viger P, Ventura S (2019) Frequent itemset mining: a 25 years review. Data Mining Knowl Discov 9(6):1329
25. Boley M, Lucchese C, Paurat D, Gärtner T (2011) Direct local pattern sampling by efficient two-step random procedures. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 582–590
26. Boley M, Moens S, Gärtner T (2012) Linear space direct pattern sampling using coupling from the past. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 69–77
27. Fournier-Viger P, Gan W, Wu Y, Nouioua M, Song W, Truong T, Duong H (2022) Pattern mining: Current challenges and opportunities. In: Proceedings of database systems for advanced applications. DASFAA 2022 international workshops: BDMS, BDQM, GDMA, IWBT, MAQTDS, and PMBD, Virtual Event, April 11-14, 2022, pp 34–49. Springer, Berlin. https://doi.org/10.1007/978-3-031-11217-1_3
28. Wu X, Zhang C, Zhang S (2004) Efficient mining of both positive and negative association rules. ACM Trans Inf Syst 22(3):381–405
29. Borodin A, Lee HC, Ye Y (2012) Max-sum diversification, monotone submodular functions and dynamic updates. In: Proceedings of the 31st ACM SIGMOD symposium on principles of database systems, pp 155–166
30. Fürnkranz J, Gamberger D, Lavrač N (2012) Foundations of rule learning. Springer, Heidelberg
31. Gollapudi S, Sharma A (2009) An axiomatic approach for result diversification. In: Proceedings of the 18th international conference on world wide web, pp 381–390
32. Kosub S (2019) A note on the triangle inequality for the Jaccard distance. Pattern Recogn Lett 120:36–38
33. Morewedge CK, Kahneman D (2010) Associative processes in intuitive judgment. Trends Cogn Sci 14(10):435–440

34. Kahneman D (2011) Thinking, fast and slow. Macmillan, New York
35. Zou Z, Li J, Gao H, Zhang S (2010) Mining frequent subgraph patterns from uncertain graph data. IEEE Trans Knowl Data Eng 22(9):1203–1218
36. Mukherjee AP, Xu P, Tirthapura S (2015) Mining maximal cliques from an uncertain graph. In: 2015 IEEE 31st international conference on data engineering, pp 243–254. IEEE
37. Jampani R, Pudi V (2005) Using prefix-trees for efficiently computing set joins. In: international conference on database systems for advanced applications, pp 761–772
38. Snoek C, Worring M, Gemert J, Geusebroek J-M, Smeulders A (2006) The challenge problem for automated detection of 101 semantic concepts in multimedia, pp 421–430. https://doi.org/10.1145/1180639.1180727
39. Sajnani H, Saini V, Kumar K, Gabrielova E, Choudary P, Lopes C (2012) Classifying yelp reviews into relevant categories. Univ. California Press, Berkeley, CA USA, Tech. Rep, Mondego Group
40. Duygulu P, Barnard K, de Freitas JF, Forsyth DA (2002) Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: European conference on computer vision, pp 97–112
41. Katakis I, Tsoumakas G, Vlahavas I (2008) Multilabel text classification for automated tag suggestion. In: Proceedings of the ECML/PKDD, vol 18
42. Klimt B, Yang Y (2004) The enron corpus: a new dataset for email classification research. In: European conference on machine learning, pp 217–226
43. Pestian JP, Brew C, Matykiewicz P, Hovermale D, Johnson N, Cohen KB, Duch W (2007) A shared task involving multi-label classification of clinical free text. In: Processing of biological, translational, and clinical language, pp 97–104. Association for Computational Linguistics, Prague, Czech Republic. https://aclanthology.org/W07-1013
44. Briggs F, Huang Y, Raich R, Eftaxias K, Lei Z, Cukierski W, Hadley SF, Hadley A, Betts M, Fern XZ, et al (2013) The 9th annual mlsp competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment. In: 2013 IEEE international workshop on machine learning for signal processing (MLSP), pp 1–8
45. Tsoumakas G, Katakis I, Vlahavas I (2008) Effective and efficient multilabel classification in domains with large number of labels. In: Proceedings of ECML/PKDD 2008 workshop on mining multidimensional data (MMD'08), vol 21, pp 53–59
46. Turnbull D, Barrington L, Torres D, Lanckriet G (2008) Semantic annotation and retrieval of music and sound effects. IEEE Trans Audio Speech Lang Process 16(2):467–476
47. Klein Y, Rapp M, Loza Mencía E (2019) Efficient discovery of expressive multi-label rules using relaxed pruning. In: International conference on discovery science, pp 367–382. Springer
48. Zhang M-L, Li Y-K, Liu X-Y, Geng X (2018) Binary relevance for multi-label learning: an overview. Front Comp Sci 12(2):191–202
49. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
50. Hintze JL, Nelson RD (1998) Violin plots: a box plot-density trace synergism. Am Stat 52(2):181–184
51. Hofmann H, Kafadar K, Wickham H (2011) Letter-value plots: Boxplots for large data. Technical report, had.co.nz
52. Organization WH, et al (1978) International classification of diseases:[9th] ninth revision. Basic tabulation list with alphabetic index. World Health Organization, Geneva

Martino Ciaperoni is a PhD student in Computer Science at Aalto University under the supervision of Prof. Aristides Gionis. His main research interests include probabilistic methods, machine learning, and databases.

Han Xiao was a PhD student in Computer Science at Aalto University under the supervision of Prof. Aristides Gionis. His main research interests include graph mining and machine learning.

Aristides Gionis received the PhD degree from Stanford University, in 2003. He is a WASP professor with KTH Royal Institute of Technology, an adjunct professor in Aalto University, and a research fellow in ISI Foundation. Previously he was a senior research scientist in Yahoo! Research. He is currently serving as an associate editor in Data Mining and Knowledge Discovery, Transactions on Knowledge Discovery from Data, and Transactions on the Web. His research interests include data mining, web mining, and social-network analysis.