International Journal of Approximate Reasoning 150 (2022) 258–272

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

# On the relative value of weak information of supervision for learning generative models: An empirical study

Jerónimo Hernández-González [a,*], Aritz Pérez [b]

[a] Departament de Matemàtiques i Informàtica, Universitat de Barcelona (UB), Gran Via de les Corts Catalanes 585, Barcelona, Spain
[b] Basque Center for Applied Mathematics, Al. Mazarredo 14, Bilbao, Spain

ARTICLE INFO

*Article history:*
Received 2 July 2022
Received in revised form 8 August 2022
Accepted 22 August 2022
Available online 31 August 2022

*Keywords:*
Weak supervision
Model learning
Generative models
Empirical study

ABSTRACT

Weakly supervised learning is aimed to learn predictive models from partially supervised data, an easy-to-collect alternative to the costly standard full supervision. During the last decade, the research community has striven to show that learning reliable models in specific weakly supervised problems is possible. We present an empirical study that analyzes the value of weak information of supervision throughout its entire spectrum, from none to full supervision. Its contribution is assessed under the realistic assumption that a small subset of fully supervised data is available. Particularized in the problem of learning with candidate sets, we adapt Cozman and Cohen [1] key study to learning from weakly supervised data. Standard learning techniques are used to infer generative models from this type of supervision with both synthetic and real data. Empirical results suggest that weakly labeled data is helpful in realistic scenarios, where fully labeled data is scarce, and its contribution is directly related to both the amount of information of supervision and how meaningful this information is.



## 1. Introduction

In supervised classification, one of the essential issues is the requirement that training data is fully labeled. Labeling training data requires expert knowledge, and this is usually costly. In the last few decades, different alternative learning frameworks which relax this requirement have been proposed. Semi-supervised learning [2], for example, aims to learn with only a subset of the training examples labeled. Positive-unlabeled learning [3] aims to learn models with only a subset of positive-labeled samples. In these two frameworks, training examples are labeled or unlabeled. However, there exist many novel frameworks where data instances provide a (small) portion of information of supervision, that is, the labeling of the training data is only partially missing. All these frameworks form the subarea of weak supervision [4,5].

Most works in the literature focus on a single (realization of) learning problem with weak supervision. In this work, we propose to observe the whole class of weakly supervised problems. From this perspective, it is easy to see that different realizations of the weakly supervised learning framework, even using the same supervision schema (i.e., how the information of supervision is provided), might involve very different amounts of information of supervision [4,5]. For example, it is not the same if 4 or 10 workers provide annotations in the popular crowdsourced labeling problem [6]. Thus, weakly supervised learning is considered to cover the whole spectrum from fully-labeled to unsupervised learning; from complete to none class information, including intermediate scenarios which may carry misleading information.

* Corresponding author.
*E-mail addresses:* jeronimo.hernandez@ub.edu (J. Hernández-González), aperez@bcamath.org (A. Pérez).

https://doi.org/10.1016/j.ijar.2022.08.012

In this paper, we explore the whole spectrum of labeling scenarios throughout a comprehensive empirical study on synthetic and real data. We aim to assess the contribution of the weakly labeled examples to the learning process of generative models, based on the key study of Cozman and Cohen [1] on semi-supervised learning. Following their setting, we consider Bayesian networks as generative models, which allows us to create probabilistic models of different complexity to test the assumptions of correct and incorrect model. We use standard learning techniques to infer them from data: maximum likelihood estimation and an expectation-maximization (EM) method to deal with the missing label information. We focus on the problem of learning from candidate labels [7], of which both fully supervised and unsupervised are particular cases. This framework is selected because it is flexible enough to cover the whole spectrum of weakly supervised scenarios, and it remains simple to analyze. As a reasonable weakly supervised learning scenario, we assume that a subset of fully labeled examples is available and, thus, our analysis is midway between purely weakly supervised and semi-supervised learning. We show that weakly labeled data is helpful mainly when fully labeled data is scarce, although it heavily depends on the meaningfulness of the provided weak supervision: (i) meaningful class information contributes more or less depending on the amount of supervision, but it usually does not harm; (ii) class information which is not meaningful only contributes in very limited scenarios, and it could harm the performance of a model learned only with the fully labeled subset. The general character of the results is constrained to the use of maximum likelihood estimation to learn probabilistic models with a standard EM algorithm. Considering that practitioners usually resort to weakly labeled data when there are not enough fully labeled examples, the scenario where weak supervision helps might be considered realistic, although the meaningfulness of the information of supervision might be unmanageable.

The rest of the paper is organized as follows. Firstly, we summarize the relevant related work. In Section 3, we present our framework and, consecutively, the type of models and learning techniques used. From Section 5 on, we present the empirical study, its results, and discussion. The manuscript finishes with conclusions and several ideas for future work.

## 2. Related work

Weakly supervised learning (WSL) is a supervised machine learning paradigm where samples provide only partial information about their true class. This information is not conclusive but might help to elucidate the true class label of the samples. Weak supervision may appear in many different forms, which have motivated diverse approaches to learning from the specific type of supervision. The information of supervision may involve a certain degree of uncertainty and/or noise, and, overall, it may range from almost complete information to no information at all. In addition, this weak information can be assigned to individual instances or to sets of instances. In the latter case, instances in the same set share a single piece of information of supervision (e.g., [8,9]). In the former, each instance receives its own piece of information: e.g., candidate labels [7], or semi-supervised learning (SSL) [2], which can be seen as a particularly extreme case of WSL where the class information of the samples is complete or empty. Different surveys which cover more or less comprehensively the subfield of WSL problems have been presented [4,5]. However, it is arguably a recent topic and general, formal theoretical analyses are still scarce. Available theoretical studies focus on specific problems.

There exists extensive theoretical literature on the relative value of unlabeled data for SSL. When assuming the correct model, works such as [10] support that unlabeled data is helpful to improve the performance of the learned model. [11] and [12] estimated the minimum amount of labeled data required in binary SSL for a specific family of mixture models, and [13] extended it to multiclass classification. Cohen and Cozman [1], in a study about the risks of using unlabeled data with generative models, conclude that unlabeled data helps mainly in two situations: (i) when the learning model is correct, and (ii) when the model is incorrect and the amount of labeled data is limited relative to the number of parameters –complexity– of the model to learn. In the former scenario, the unlabeled data reduces the variance of the model and maintains its bias. In the latter, the decrease of the variance term compensates for the increase of the bias. Several decisions limit the extent of the study, such as the use of simply binary classes. They point out that, in other situations, unlabeled data might degrade the performance of the learned classifier. In a finite-sample theoretical work, Singh et al. [14] show that, in certain situations, SSL outperforms supervised learning in terms of error bounds. They emphasize the importance of studying the benefits of SSL using finite sample regimes, as these are the scenarios where SSL might appear superior. Whereas SSL studies usually assume that labeled and unlabeled data comes from the same distribution and labels are missing completely at random (labels are missing independently of the problem variables), Chawla and Karakoulas [15] carried out an empirical study that goes beyond and analyses SSL under other missing data mechanisms: Missing at random (labels are missing according to some function on the descriptive variables of the problem) and missing not at random (when the actual value of the label can determine that it is not given, i.e., sample-selection bias). van Engelen and Hoss [16] provide an up-to-date revision of the SSL state-of-the-art for the interested reader.

The problem of learning from candidate labels receives diverse names in the literature: partial or multiple labels [7, 17], candidate labeling [18] or superset learning [19,20]. This problem, which we selected for this study as it allows us to cover the whole spectrum from supervised to unsupervised data, has also been theoretically studied. Cour et al. [7] provide a theoretical framework for effective learning using training sets with partial labels and unsupervised data. Their framework relies on some assumptions about the distribution underlying the partial labeling. They establish a relationship between classification and superset error, pivoting on their definition of ambiguity degree. Liu and Dietterich [19] use the same definition of ambiguity degree to provide a sample complexity analysis based on Natarajan dimension. Many works address learning with candidate labels as the disambiguation of the sets of labels. Cour et al. [7] present a method

for disambiguating partially labeled data by assuming some properties about their underlying distribution, and establish conditions for its asymptotic consistency. It minimizes a convex loss function adapted to partial labeling. Based on the same loss function, Cabannes et al. [21] present a general framework that allows for the development of statistically consistent learning algorithms for different classification problems. Recently, they reviewed the disambiguation principle [22] and proposed a disambiguation algorithm under the assumptions of [21], establishing its learning rate. Wang et al. [23] propose a learning algorithm for datasets that are only partially labeled with candidate sets (the rest is unlabeled). The algorithm is a propagation procedure that iteratively disambiguates the weakly labeled and unlabeled data. Learning from candidate sets can be seen as a particular case of the versatile soft labeling scheme provided by Dempster-Shafer mass functions [24–27], and thus it has also received considerable attention from the field of fuzzy systems. For example, in the problem of learning from fuzzy labels, where instances are labeled using fuzzy sets, Campagner [28] uses statistical learning theory to theoretically study the problem and analyze sample complexity and risk bounds for it. Couso and Dubois [29] describe several likelihood functions that can be considered for this problem, depending on the purpose of the study, for a single latent variable.

In this paper, we focus on the contribution of weak information of supervision depending on its different traits. We assume the existence of a subset of samples fully labeled among a large set of weakly labeled samples, a realistic scenario as reflected in the available real-world datasets [7,30]. We aim to shed light on the circumstances in which the weakly labeled data is really useful depending on the relative size of the fully/weakly labeled subsets, and the quantity and quality of the supervision. To do so, we will modify the conditions under which the labeling is performed, moving beyond the assumptions of previous works. Up to our knowledge, no previous theoretical or empirical work has addressed these questions in the case of candidate labeling or generally in weak supervision.

## 3. Framework

This paper considers the problem of supervised classification, where the objective is to learn a classifier from data. Formally, let $\boldsymbol{X} = (X_1, \ldots, X_v)$ be a set of descriptive random variables, and $C$ a special random variable $C$, the class variable. Let $\Omega_{X_j}$ be the set of possible values that $X_j$ can take, i.e., its domain. In this work, we assume all the variables to be discrete without loss of generality. Specifically, each of the possible values of the special class variable, $C$ is called class label, $c \in \Omega_C$. Let us define an instance $\boldsymbol{x}$ as a realization of $\boldsymbol{X}$. In classical supervised learning, a training dataset $\{(\boldsymbol{x}_i, c_i)\}_{i=1}^N$ is available for model learning, where each instance $\boldsymbol{x}_i$ is provided together with its associated class label $c_i$ and is usually assumed to be independent and identically distributed (iid) according to the real distribution, $p(\boldsymbol{X}, C)$. A learning algorithm infers from data a classifier, that is, a mapping function between the descriptive variables and the class, $h : \boldsymbol{X} \to C$. Given a loss function (e.g., the standard 0-1 loss), the objective of the learning process is usually to find the classifier $h$ that minimizes the expectation of the loss function or, in practice, the empirical expect loss on the training set.

In the general framework of weakly supervised learning [5], the information of supervision of the training data is not fully available. That is, training data is not provided as a set of instance-label pairs, but as a set of instances $\{\boldsymbol{x}_i\}_{i=1}^N$ that is accompanied by imprecise and/or noisy information about the respective classes. This lack of supervised information is different depending on the specific problem. Specifically, in this work we assume the specific weak supervision schema of *candidate labels* [7,5], where information of supervision is provided on a per-instance basis in the form of a set of labels $S_i$ accompanying each instance $\boldsymbol{x}_i$. The set of labels $S_i \subseteq \Omega_C$ contains the true class label for $\boldsymbol{x}_i$, $c_i \in S_i$ [19]. Thus, the dataset is $D = \{(\boldsymbol{x}_i, S_i)\}_{i=1}^N$. In this case, the information of supervision provided for each instance does not specify the actual label, but it may allow for discarding a few labels. The objective remains exactly the same.

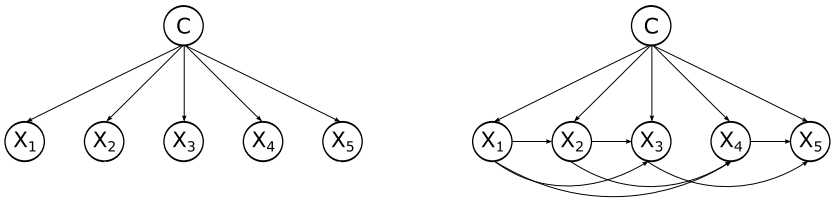
### 3.1. Weak supervision as a spectrum

When discussing the particularities of weakly supervised problems, we usually note that training data is partially labeled. However, little attention is paid to the specific amount or level of supervision. In fact, the same weak supervision schema can involve extremely different levels of supervision [4,5].

Let us define the *amount of supervision* provided by a candidate set $S$ as $A(S) = 1 - \log|S|/\log|\Omega_C|$, which can be understood as the inverse of the normalized entropy of $S$ assuming constant probability among classes $c \in S$. When $|S_i| = 1$ the corresponding example $\boldsymbol{x}_i$ is actually *completely labeled* (maximum information, $A(S) = 1$). Similarly, when the set of candidate labels contains all possible class labels in the problem, $|S| = |\Omega_C|$, it can be considered as *completely unlabeled* (minimum information, $A(S) = 0$). In this way, one may realize that the size of the candidate sets alone determines the amount of information of supervision in this problem, ranging from full supervision to complete unsupervision. Note that the average amount of supervision is easily measurable for any given dataset.

Not only amount of information of supervision, but the information provided also needs to be useful to learn to discriminate. The composition of the label sets can determine the amount of discriminative information that is provided, from which classifiers are learned. Candidate sets can be seen as the corruption of the original information, the real class $c_i$, a process known as coarsening in the statistics literature. Many related works[1] assume *coarsening at random* (CAR) [31,32],

[1] E.g., superset assumption [20] is CAR, with the specific condition of $p(S|c) = p(S|c') = 1/2^{|\Omega|-1}, \forall c, c' \in \Omega_C, S \in \mathcal{P}(\Omega_C),\ c, c' \in S$ [29].




(a) Naive Bayes (b) A $K$-DB with $K = 2$

**Fig. 1.** Structure of naive Bayes (NB) and of a $K$-dependence Bayesian classifier ($K$-DB).

that is, that the probability of observing a subset $S$ factorizes as the product of the sum of the probabilities of the elements of $S$ times the probability of $S$ according to certain distribution $\pi$ over all non-empty subsets of labels $\mathcal{P}(\Omega_C)$, $p(S|\boldsymbol{x}) = p(C \in S|\boldsymbol{x})\pi(S|C \in S)$ with $p(C \in S|\boldsymbol{x}) = \sum_{c \in S} p(c|\boldsymbol{x})$. The second factor $\pi$ can be considered as a procedure that corrupts the information ($S$) about the real –unknown– class $c$ independently from $c$. Adopting $p(c|S, \boldsymbol{x}) = p(c|\boldsymbol{x})/p(c \in S|\boldsymbol{x})$ if $c \in S$ (and 0 otherwise), Gill et al. [31] show that $p(S|\boldsymbol{x}, c)$ is constant in $c \in S, \forall S \in \mathcal{P}(\Omega_C)$. But this assumption might not be completely realistic. Adapting the ideas of Chawla and Karakoulas [15] for SSL, missing not at random involves different coarsening mechanisms that lead to diverse distributions $p(c|\boldsymbol{x}, S) \neq p(c|\boldsymbol{x}, S')$ for $S, S' \in \mathcal{P}, c \in S, S'$. This involves that not all the sets are equally probable $p(S|\boldsymbol{x}, c) \neq p(S'|\boldsymbol{x}, c)$ (non-CAR can be understood in similar terms). Thus, co-occurrence is a non-CAR event that has been previously described in this context [19]: there exists $c' \in \Omega_C$ such that when an example of class $c$ ($c \neq c'$) is observed, the provided candidate set $S$ includes both $c, c' \in S$ with probability $s$. As the probability of co-occurrence increases ($s \to 1$), discrimination between $c$ and $c'$ classes becomes more difficult (in line with consequently lying teachers, described by Lugosi [33] for noisy labeling). Similarly, adversarial labelers [34] can be seen as non-CAR events where noisy information is provided purposely to hamper the learning process. A labeling might be considered adversarial when the candidate set $S$ contains, apart from real class $c$, other labels $c' \in S, c \neq c'$ which are improbable according to the real –unknown– model. We use the term *data meaningfulness* to refer to this conditioning factor which, to be measured, access to the real model is required.

## 4. Learning tool-set

The main goal of this work is to study the value of weakly supervised data throughout the spectrum of scenarios in which it might be present, as described above. The use of standard models and learning techniques, which are presented in this section, allows us to focus on our main goal.

As in [1], our empirical study uses generative classification models from the family of Bayesian network classifiers (BNCs) [35]. These are based on Bayesian networks (BN), one of the main types of probabilistic graphical models [36]. This family of classifiers has a directed acyclic graph structure specially designed to deal with supervised classification. Specifically, BNs provide an appropriate framework for working with simulated data –specifically, they allow for randomly generating distributions with a controlled number of parameters and sampling samples is efficient–, and learning the MLE parameters can be efficiently performed. Accounting for the partially missing information, we use an expectation-maximization (EM) algorithm [37] to learn from weakly supervised data.

### 4.1. Bayesian network classifiers

Bayesian networks are probabilistic graphical models which use directed acyclic graphs (DAG) to encode a set of conditional dependencies between random variables. Following the DAG, the join distribution factorizes as the product of conditional probability distributions for each variable given its parents. Thus, a BN is described as a pair $(G, \theta)$, where $G$ is the DAG and $\theta$ are the parameters that determine the conditional probability distribution of each variable. Both the DAG and the model parameters could be learned from data. In this work, we first analyze the effect of the amount of supervision under the correct model assumption, i.e., the true distribution underlying the data can be represented by $(G, \theta)$. To do so, we fix the true structure underlying the data, $G$, and we focus on the parametric learning of the BNCs. We also study other scenarios where this assumption does not hold and the learned model type is less expressive than the generative one.

We use BN models as probabilistic classifiers [35]. Specifically, we consider the family of $K$-dependence Bayesian classifiers ($K$-DB) [38]. The DAG of a $K$-DB classifier is biased towards supervised classification: all the descriptive variables have the class variable $C$ as a parent and, moreover, at most $K$ other parents among the descriptive variables (Fig. 1). The classification rule is defined as:

$$\hat{c} = \operatorname*{argmax}_c p(C = c) \prod_{j=1}^{v} p(X_j = x_j | \boldsymbol{\Pi}_j = \boldsymbol{\pi}_j, C = c)$$

where $\boldsymbol{\Pi}_i$ is a set of $K$ descriptive-variable parents of $X_i$, and $\boldsymbol{\pi}_j$ is a realization of them. The popular naive Bayes classifier [39] corresponds to a 0-dependence Bayesian (0-DB) classifier, and the tree-augmented naive Bayes [40] is a 1-DB.

**Algorithm 1** Pseudo-code for the implemented EM algorithm.

**procedure** EM($D = \{(\boldsymbol{x}_i, S_i)\}_{i=1}^N$)
  $\theta \leftarrow \theta^{(0)}$
  **while** not converged **do**
    $q_i(c) \propto p(c|\boldsymbol{x}_i;\theta)\mathbb{I}[c \in S_i], \forall i, c$ ▷ Update $q$: E-step
    $\theta \leftarrow \text{argmax}_\theta \, \mathbb{E}_{\boldsymbol{c}\sim q} \log p(\boldsymbol{c}|\boldsymbol{x};\theta)$ ▷ Update $\theta$: M-step
  **end while**
  **return** $\theta$
**end procedure**

In this work, we aim to account for the complexity of the model when studying the contribution of weakly supervised data, in a similar way to what Cozman and Cohen [1] did for SSL. Inspired by their experimental design, we also consider the family of $K$-DB classifiers to easily generate models of increasing complexity. Note that the number of parameters of $K$-DB models is exponential on $K$. Assuming that the cardinality of all the descriptive variables is the same ($|\Omega_X|$), the number of free parameters is $|\Omega_C| - 1 + v \cdot (|\Omega_X| - 1) \cdot |\Omega_C| \cdot |\Omega_X|^K$.

### 4.2. An expectation-maximization algorithm

The maximum likelihood estimation (MLE) of the parameters of a discrete BN can be computed in closed form using complete data [41]. However, as in the weakly supervised problems the information of supervision is incomplete, we resort to the expectation-maximization strategy [37]. EM is a theoretically-founded iterative procedure that allows dealing with the maximum likelihood estimation of parameters in the presence of missing data. In each iteration, two steps are interleaved. Firstly, in the expectation (E) step, the missing values are replaced by their expected value given the current model, $q$. Secondly, in the maximization (M) step, the maximum likelihood parameters are estimated using the complete data. In general, this procedure converges to a local maximum.

In this case, the likelihood can be expressed as,

$$\mathcal{L}(\theta; D) = p(D;\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i, S_i;\theta) = \prod_{i=1}^{N} \sum_{c\in\Omega_C} p(\boldsymbol{x}_i, c, S_i;\theta)$$

Considering conditional independence of $\boldsymbol{X}$ and $S$ given $C$ and superset [20] assumptions, we obtain

$$\mathcal{L}(\theta; D) = \prod_{i=1}^{N} \sum_{c\in\Omega_C} p(\boldsymbol{x}_i, c;\theta) p(S_i|c) = \prod_{i=1}^{N} \sum_{c\in\Omega_C} p(\boldsymbol{x}_i, c;\theta) \frac{1}{2^{|\Omega_C|-1}} \mathbb{I}[c \in S_i]$$

and, as our objective is to find the parameters that maximize this expression, it can be reduced to

$$\mathcal{L}(\theta; D) \propto \prod_{i=1}^{N} \sum_{c\in\Omega_C} p(\boldsymbol{x}_i, c;\theta)\mathbb{I}[c \in S_i] = \prod_{i=1}^{N} \sum_{c\in S_i} p(\boldsymbol{x}_i, c;\theta)$$

This likelihood expression is in line with Couso and Dubois [29]'s *visible likelihood*, but including descriptive variables too. Similarly, it is equivalent to the likelihood for *imprecise data* formulated by Denoeux [25]. The log-likelihood is,

$$\log \mathcal{L}(\theta; D) \propto \sum_{i=1}^{N} \log \sum_{c\in S_i} p(\boldsymbol{x}_i, c;\theta) \tag{1}$$

Considering that several candidate sets might have a single element ($i \in \{1, \ldots, N_f\}$ s.t. $|S_i| = 1$), which by definition ($c_i \in S_i, \forall i$) can be considered fully labeled samples, we can rewrite the previous expression as,

$$\log \mathcal{L}(\theta; D) \propto \sum_{i=1}^{N_f} \log p(\boldsymbol{x}_i, c_i;\theta) + \sum_{i=N_f+1}^{N} \log \sum_{c\in S_i} p(\boldsymbol{x}_i, c;\theta)$$

where $N = N_f + N_w$. The first term is the log-likelihood of the labeled subset. The second term can be understood as a generalized expectation criterion [42] with the log-likelihood of the observed candidate sets as score function, and a constraint function returns 1 only if $c \in S$ and 0 otherwise.

Algorithm 1 shows the pseudocode of our EM-based method, a generative version of the proposal by Jin and Ghahramani [17] and equivalent to the EM algorithm presented by Denoeux [25, Sect. 4.1]. In the E-step, the expected value of the missing values (class label of weakly supervised examples) is calculated as $q_i(c) \propto p(c|\boldsymbol{x}_i;\theta)$ only for candidate labels, $c \in S_i$. In the M-step, using the data completed probabilistically with the $q$ estimates, we compute the MLE parameters. These two steps iterate until convergence. This algorithm can be interpreted as a self-labeling approach for weakly supervised data, in which the estimate of the class distribution provided by the model, $q_i$, is used to refine the parameters of the model, $\theta$, in an iterative way [43].

## 5. Design of the empirical study

We aim to study the contribution of weakly supervised data as an addition to a (small) subset of fully labeled data. We explore a vast spectrum of labeling scenarios, ranging from (near) full supervision to (almost) unsupervised data, including data with non-meaningful information. We conduct our experiments based on a series of intuitions that lead to the proposed experimental setup.

### 5.1. Intuitions

The information of supervision is necessary for supervised machine learning. A *robust* learning technique should be able to learn from enough (weakly) labeled data, even if a few labels are misleading. With this idea in mind, our intuitions on the usefulness of weakly labeled data include that:
(*i*) Weakly supervised data might not be useful if there is enough labeled data, although it should not harm either.
(*ii*) The enhancement due to the use of weakly labeled data is correlated with the amount of information provided. More information might involve more weakly labeled instances, and/or smaller candidate sets.
(*iii*) Weak supervision helps as long as it is meaningful. In other words, the ranks over $c \in \Omega_C$ obtained with $p(c \in S|\boldsymbol{x})$ or with $p(c \mid \boldsymbol{x})$ tend to be the same.

Through this section, we present different experiments carried out to test the contribution of weakly supervised data and the validity of our intuitions.

### 5.2. Experimental setting

Two different sets of experiments have been carried out. To be able to explore the whole spectrum of weak supervision, the first set of experiments was carried out with synthetic data. We synthetically create a generative model, sample a fully labeled dataset from it, and finally transform this dataset into weakly labeled. The second set of experiments uses real data labeled with candidate sets.

*Synthetic data generation.* As generative models, we use $K$-DB classifiers. The structure is randomly generated. Given an ordering of descriptive variables, the $K$ parents of each variable are selected randomly among the previous variables in the ordering. Then, the class variable is added as a parent of every descriptive variable. The parameters of the model are randomly generated by sampling a Dirichlet distribution with all hyper-parameters $\alpha = 1$. For the convenience of Bayesian networks, discrete variables are considered. Specifically, we use 16 binary descriptive variables plus a class variable with 6 possible values. We use generative models of different complexity ($K = 1$ or $K = 4$). We sample the generative $K$-DB model to obtain a dataset of the desired size and perform our train/test split. Only the training partition is transformed into a weakly labeled dataset. Thus, the test partition remains fully labeled.

We simulate different experimental scenarios that try to cover as much as possible the spectrum of weak supervision. The next parameters are consistently used in all the experimental scenarios for synthetic generation of weakly labeled data: the size of a basic subset which is left fully labeled ($N_f$), the relative size of the weakly labeled subset ($r = N_w/N_f$), and the size of the candidate sets ($l = |S|$). First of all, $N$ instances are sampled from the generative model where $N = N_f + N_f \cdot r$. The real labels of $N_f \cdot r$ instances are replaced by their corresponding candidate set. The candidate set $S$ for an instance $\boldsymbol{x}$ is formed using a distribution $\tilde{p}(c|\boldsymbol{x})$ which is defined based on the distribution $p_M(c|\boldsymbol{x})$ given by the generative model, where the real class $c$ has $\tilde{p}(c|\boldsymbol{x}) = 0$. $S$ includes $c$ by default, plus $l - 1$ false positive labels which are simply the labels with largest probability in distribution $\tilde{p}(c|\boldsymbol{x})$. We handle $\tilde{p}(c|\boldsymbol{x})$ to simulate different experimental conditions:
**Experimental condition I: honest label sets**. Label sets are assumed to be produced by a reasonable doubt of an honest labeler, and thus the distribution

$$\tilde{p}(c'|\boldsymbol{x}) \propto p_M(c'|\boldsymbol{x}) \text{ over } c' \in \Omega_C, c' \neq c$$

is proportional to the distribution over the class labels given the instance $\boldsymbol{x}$ according to the generative model, $p_M$, with the probability of the real label $c$ equal to 0, $\tilde{p}(c|\boldsymbol{x}) = 0$. This is in line with the CAR assumption.
**Experimental condition II: misleading label sets**. Label sets are assumed to be produced by a malicious labeler who tries to induce a confusion, and thus the distribution

$$\tilde{p}(c'|\boldsymbol{x}) \propto 1 - p_M(c'|\boldsymbol{x}) \text{ over } c' \in \Omega_C, c' \neq c$$

is *inversely* proportional to the distribution over the classes $p_M(c|\boldsymbol{x})$, with the probability of the real label $c$ equal to 0, $\tilde{p}(c|\boldsymbol{x}) = 0$. In this case, the generation of candidate sets is non-CAR.
**Experimental condition III: co-occurrence**. Label sets are assumed to be produced by a reasonable labeler who confuses pairs of labels. Let us formally define the probability of co-occurrence, $s \in [0, 1]$, as:

$$\forall c \in \Omega_C, \exists \check{c} \neq c \text{ such that } \forall(\boldsymbol{x}, c, S), p(\check{c} \in S) = s$$

**Table 1**
Description of the real datasets [30] used in this paper, before and after preprocessing. Mean frequency of co-occurrence† is measured as the maximum proportion of instances of real class $c$ that have label $c' \neq c$ in the candidate set too. Mean values per real label $c$ and standard deviations are given. Meaningfulness‡ is estimated as one minus the average normalized mean rank of the false positive labels $c'$ in the sets according to the probability given by a complex model ($K = 4$) learned with the ground-truth labels.

| Characteristic | Birdac | Lost | MSRCv2 |
|---|---|---|---|
| No. instances | 4.998 | 1.122 | 1.758 |
| No. labels | 13 | 14 | 23 |
| No. fully labeled insts. ($\lvert S_i\rvert = 1$) | 1.632 | 67 | 140 |
| No. weakly labeled insts. ($\lvert S_i\rvert > 1$) | 3.366 | 1.055 | 1.618 |
| Candidate set size | $2.74 \pm 0.72$ | $2.31 \pm 0.46$ | $3.34 \pm 1.17$ |
| Mean frequency of co-occurrence† | $0.72 \pm 0.17$ | $0.53 \pm 0.12$ | $0.76 \pm 0.21$ |
| | Original datasets | | |
| No. instances | 3.029 | 838 | 1.046 |
| No. labels | 6 | 6 | 6 |
| No. fully labeled insts. ($\lvert S_i\rvert = 1$) | 526 | 159 | 229 |
| No. weakly labeled insts. ($\lvert S_i\rvert > 1$) | 2.503 | 679 | 817 |
| Candidate set size | $2.59 \pm 0.66$ | $2.24 \pm 0.43$ | $3.05 \pm 0.84$ |
| Mean frequency of co-occurrence† | $0.70 \pm 0.13$ | $0.56 \pm 0.12$ | $0.63 \pm 0.04$ |
| Estimated meaningfulness‡ | 0.51 | 0.47 | 0.49 |
| | Adapted datasets | | |

i.e., $s$ represents the probability that a certain label $\check{c}$ appears in the label sets of instances with real class $c$. In practice, we fix $\check{c}$ to the label that, on average, has the largest probability given instances that have $c$ as the real class s.t. $\check{c} \neq c$:

$$\check{c} = \arg_{\check{c} \neq c} \max \sum_{(\boldsymbol{x}, c)} p_M(\check{c}|\boldsymbol{x}).$$

Let us define $\mathcal{S}$ as the subset of labels that is consistently included in the candidate set, $\mathcal{S} \subseteq S$. With probability $s$, $\mathcal{S} = \{c, \check{c}\}$, and with probability $1 - s$, $\mathcal{S} = \{c\}$. Thus, we can define the distribution as

$$\tilde{p}(c'|\boldsymbol{x}) \propto p_M(c'|\boldsymbol{x}) \text{ over } c' \in \Omega_C, c' \notin \mathcal{S}$$

proportional to $p_M(c|\boldsymbol{x})$, with the probability of the labels in $\mathcal{S}$ equal to 0. Exactly, $|S| - |\mathcal{S}|$ labels are obtained from it. In this case, the generation of candidate sets is non-CAR. Note that the probability of co-occurrence imposes a harder learning condition than the ambiguity degree considered by [7,19]. Whereas ambiguity degree is defined as an upper bound on the maximum probability, over all instances $(\boldsymbol{x}, c)$, of an extra label $\check{c} \neq c$ appearing together with $c$, the probability of co-occurrence sets the probability of an extra label $\check{c} \neq c$ appearing together with $c$, for all true labels $c$ and all instances $(x, c)$.

We have generated experimental scenarios resulting from the combination of the following parametric values: 10 different fully labeled subset sizes ($N_f \in \{33, 66, \ldots, 333\}$), 4 different values for the relative size of the weakly labeled subset ($r \in \{0.5, 1, 2, 5\}$), 4 candidate set sizes ($l \in \{2, \ldots, 5\}$), all three experimental conditions (honest, misleading, and co-occurrence with 4 probabilities of co-occurrence, $s \in \{0.25, 0.5, 0.75, 1.0\}$). We use generative models of limited ($K = 1$) and large ($K = 4$) complexity, whereas all learned models have $K = 1$. When both generative and learned models have limited complexity ($K = 1$), we allow the learning process to learn the parameters over the original $K$-DB structure. This allows us to inspect the role of weakly labeled data, the main goal of the paper, in the favorable scenario of a correct model [1] where the contribution of weak supervision is expected to be observed in a cleaner way. When the generative model ($K = 4$) is more complex than the learned one ($K = 1$), the correct model assumption does not hold. This arguably is a reasonable simulation of real-world scenarios, where the generative model is likely to be more complex than the learned one. To account for the variance coming from the synthetic data generation, every single experiment with a specific combination of generative hyper-parameters is replicated up to 900 times (30 generative models $\times$ 30 sampled datasets). The displayed results show averaged metrics over the repetitions.

*Real data preprocessing.* We use three real datasets with candidate label sets [30]: Birdac, Lost and MSRCv2, described in Table 1. These were adapted to our experimental setup. Firstly, we binarize all the descriptive variables and remove constant and completely correlated features. We reduce the set of labels to 6, as used in the synthetic data generation. The majority classes are maintained (with the exception of the most common label of Birdac, which is always and only present in label sets of size 1). Only instances the real class of which is maintained are kept. Removed labels are also dropped off the candidate sets. Table 1 describes the adapted data too.

In these experiments, candidate set size, meaningfulness, and co-occurrence level are fixed (see the specific values in Table 1). We simulate different experimental scenarios by generating subsamples from the combination, whenever it is

possible, of 10 different fully labeled subset sizes ($N_f \in \{33, 66, \ldots, 333\}$) and 4 different values for the relative size of the weakly labeled subset ($r \in \{0.5, 1, 2, 5\}$). Note for example that $N_f = 333$ is not possible with *Lost* and *MSRCv2* datasets, as their number of fully labeled instances is smaller. Similarly, in these cases $N_f = 200$ together with $r = 5$ is not possible as the number of weakly labeled instances is smaller than $200 \cdot 5$. The generative model is completely unknown in these real datasets, and a generative complex model is arguably a fair assumption. Thus, these experiments do not follow the favorable scenario of the correct model [1]. To account for the variance coming from subsampling, every single experiment with a specific combination of generative hyper-parameters is replicated 30 times (different subsamples). The displayed results show averaged metrics over the repetitions.

### 5.3. Validation

All the results have been estimated by means of a $5 \times 3$-fold cross-validation [44]. Real labels are available for validation for both real and synthetic data, which eases model evaluation. In the former case, the actual labels are also provided in the real datasets. In the latter, transformation into weakly labeled is only performed on the training subset, whereas the validation subset is left fully labeled.

We consider two different metrics. On the one hand, the Macro-F1 metric calculates the harmonic mean of the average precision and recall over all the class labels,

$$MacroF1 = \frac{2 \cdot \overline{pr} \cdot \overline{re}}{\overline{pr} + \overline{re}}$$

with

$$\overline{pr} = \frac{1}{|\Omega_C|} \sum_{c \in \Omega_C} \frac{TP^c}{TP^c + FP^c} \quad \text{, and} \quad \overline{re} = \frac{1}{|\Omega_C|} \sum_{c \in \Omega_C} \frac{TP^c}{TP^c + FN^c}$$

where $TP^c$, $FP^c$, and $FN^c$ are the counts for true positives, false positives, and false negatives, respectively, regarding class $c$. The larger the value a classifier obtains, the better it performs. On the other hand, the Brier score measures the difference among the probability distributions over the classes and the real class label (one-hot encoding),

$$BrierScore = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{|\Omega_C|} (\hat{p}_{ic} - \mathbb{I}[c = c_i])^2$$

In this case, the smaller the value a classifier obtains, the better it adjusts the probabilities (the better it performs).

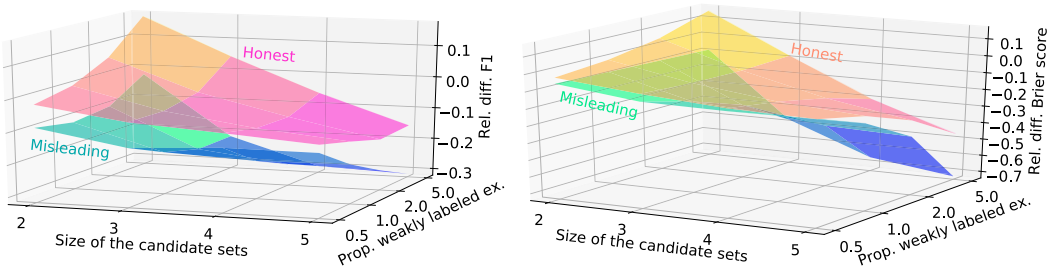
## 6. Views of the empirical study

From the extensive experimental setting explained above, we have extracted several snapshots that aim to analyze the contribution of weak supervision through the interaction among the different key factors of the weakly supervised learning framework: the size of the fully labeled subset, the relative size of the weakly labeled subset, the size and the meaningfulness of the candidate sets, including the co-occurrence of labels with a certain probability. Each following subsection displays a snapshot of the study where up to three key factors are confronted while the rest of the experimental settings are fixed.

### 6.1. Contribution of weak supervision with increasingly larger weakly-labeled subset and candidate sets

The first factor helps us understand how many samples we need to take advantage of weak supervision under fair conditions. For this purpose, the real labels of a subset of $N_f = 33$ examples are known. The second factor, the size of the candidate sets, is directly related to the amount of supervision, as explained in Section 3.1. Firstly, in Fig. 2 we study the interaction of these two dimensions on two meaningfulness scenarios (honest vs. misleading labels) with synthetic data under the correct model assumption. As we aim to analyze the behavior of weakly supervised learning as a complement to fully labeled data, we show the results as the *difference* of the performance of the classifier learned with weak supervision minus that of an equivalent classifier learned only with the fully labeled subset, relative to that of the *real* generative model.[2] In general, performance improves as the size of the candidate sets decreases, and as the proportion of weakly labeled samples increases. When the size of the candidate set is the largest ($|S| = 5$), the performance stops decreasing and recovers only when the proportion of weakly labeled samples is really large. That is, large amounts of weakly supervised samples are required to recover when the information within a single candidate set is small (large $|S|$). The largest gain is

[2] Any value **larger than** 0 points out a scenario where the use of weakly supervised data helps to outperform the classifier learned with just fully labeled data. These values are relative to the performance of the real generative model on the same dataset. A value of 1 would mean that the learned model performs as well as the real model.




(a) Results in terms of Macro-F1 (b) Results in terms of Brier score

**Fig. 2.** Graphical description of the value of weak supervision as the size of the candidate sets and the proportion of weakly labeled examples increase on synthetic data. In both figures, two surfaces display results with misleading (labels in $S$ are really improbable) and honest (labels in $S$ are probable) labels, respectively. All the surfaces show the difference of performance of a classifier learned with weak supervision minus that of a similar classifier learned only with the labeled subset; in terms of Macro-F1 in a), of Brier score in b), relative to the performance of the real model. Other parameters are fixed to a small fully labeled subset ($N_f = 33$), simple generative and learned models ($K = 1$), and no induced co-occurrence ($s = 0$).

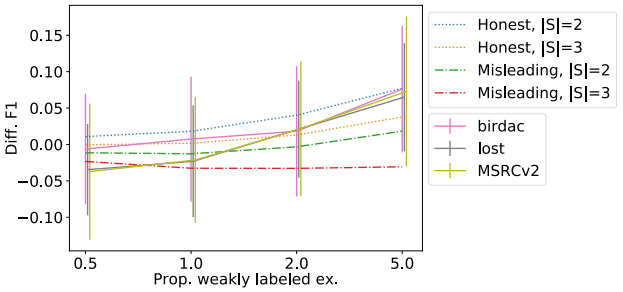


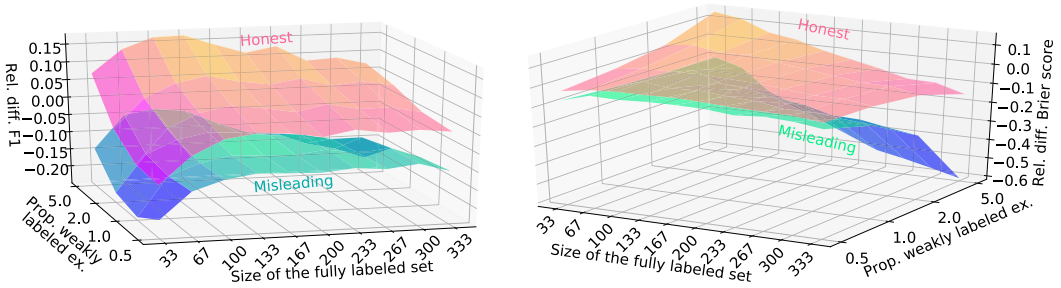

**Fig. 3.** Display of different scenarios of candidate set sizes and proportions of weak supervision on real and synthetic data. Dashed lines depict results on synthetic data in the experimental conditions I (Honest) and II (Misleading) with candidate set sizes $|S| = \{2, 3\}$. Solid lines depict results on real data: Birdac, Lost and MSRCv2 (see in Table 1 their mean candidate set size and estimated label meaningfulness). All the lines show the difference of performance in terms of Macro-F1 of a classifier learned with weak supervision minus that of a similar classifier learned only with the labeled subset. Other parameters are fixed to a small fully labeled subset ($N_f = 33$), complex generative ($K = 4$) and simple learned ($K = 1$) models, and no induced co-occurrence ($s = 0$). (Figures are colored in the web version to make interpretation easier.)

observed when the candidate sets are small and the proportion of weakly labeled samples is large, that is, when there is more information of supervision. Results are consistently worse within the misleading labeling scenario. Whereas misleading labels are not that harmful when candidate sets are small, the advantage of having candidate sets filled in with honest labels stands out as more unfavorable experimental setups are explored. Similar behaviors are observed in terms of both F1 and Brier score.

Fig. 3 shows the equivalent snapshot of the empirical study with real data. As candidate set size is fixed in real data, we only display one of the dimensions (proportion of weakly labeled examples) by simulating increasingly large datasets with subsamples of the real data. We include results with synthetic data (but using the incorrect model assumption, i.e., the generative model is more complex than the learned one) as a reference to interpret the results with real data. We show the results as the *difference* of the performance of the classifier learned with weak supervision minus that of an equivalent classifier learned only with the fully labeled subset. This time, the performance is not relative to that of the *real* generative model, which is unknown. Only results in terms of Macro-F1 are shown since, due to the lack of the real model, normalized performance cannot be obtained and Brier score, unscaled, would display results hardly interpretable. Results with *Lost* and *MSRCv2* datasets show similar behaviors, whereas that of *Birdac* dataset is slightly better. As the proportion of weakly labeled samples increases, the experiments with real data stabilize among the results with synthetic data of $|S| = 2$ and $|S| = 3$ in the honest scenario. Note that the mean candidate size in all three datasets is in the interval $[2.2, 3.05]$ and their estimated meaningfulness show intermediate values (see Table 1).

### 6.2. Contribution of weak supervision with increasingly larger fully-labeled and weakly-labeled subsets

The interaction between the relative sizes of fully-labeled and weakly-labeled subsets helps us understand when weakly supervised data can effectively provide information of supervision. Using the same display as before, Fig. 4 shows the interaction of the relative sizes of these subsets on two meaningfulness scenarios (honest vs. misleading labels) with synthetic data under the correct model assumption. In all the cases, when the fully labeled subset is small, proportionally increasing the size of the weakly supervised subset enhances the most the performance of the models. The gain obtained with weak supervision is more limited as the size of the fully labeled subset increases. When a large fully labeled subset is available, many weakly labeled samples can harm the performance in terms of probability calibration, as observed when measuring




(a) Results in terms of Macro-F1 (b) Results in terms of Brier score

**Fig. 4.** Graphical description of the value of weak supervision as the amount of fully labeled data and the proportion of weakly labeled examples increase on synthetic data. In both figures, two surfaces display results with misleading (labels in $S$ are really improbable) and honest (labels in $S$ are probable) labels, respectively. All the surfaces show the difference of performance of a classifier learned with weak supervision minus that of a similar classifier learned only with the labeled subset; in terms of Macro-F1 in a), of Brier score in b), relative to the performance of the real model. Other parameters are fixed to small candidate sets ($|S| = 2$), simple generative and learned models ($K = 1$), and no induced co-occurrence ($s = 0$).

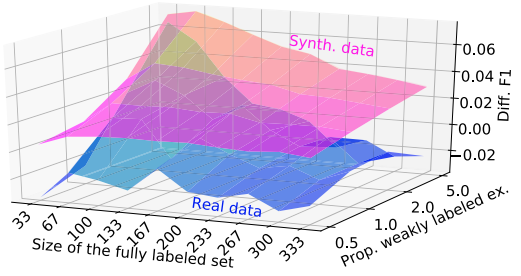



**Fig. 5.** Graphical description of the value of weak supervision as the amount of fully labeled data and the proportion of weakly labeled examples increase on real and synthetic data. One surface summarizes results on three real datasets: Birdac, Lost and MSRCv2 (see Table 1). The surface with results on synthetic data involves honest labeling and incorrect model assumption (complex generative, $K = 4$, and simple learned, $K = 1$, models). Both surfaces show the difference of performance in terms of Macro-F1 of a classifier learned with weak supervision minus that of a similar classifier learned only with the labeled subset. Other parameters are fixed to small label sets ($|S| = 2$) and no induced co-occurrence ($s = 0$).

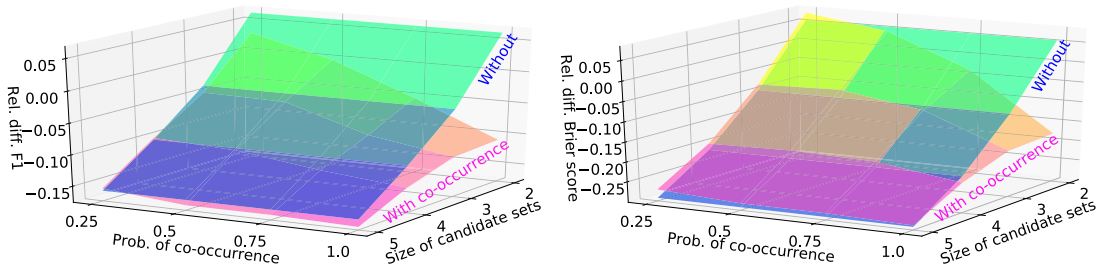
Brier score. When the false positive labels are not meaningful (misleading scenario), the performance gain is more limited and the harm of large amounts of weakly-labeled samples is more clearly observed. In all these experiments, small candidate sets ($|S| = 2$) are considered. Although not shown in these plots, our results show that as $|S|$ increases (i.e., the amount of information decreases) the performance gain declines, and differences between honest and misleading scenarios reduce.

Fig. 5 shows the equivalent snapshot of the empirical study with real data. We can display both dimensions (size of the fully labeled subset, and proportion of weakly labeled examples) by subsampling the real data. Results with all three real datasets are aggregated (some experimental setups cannot be simulated with all the datasets). Results with synthetic data (honest scenario with $|S| = 2$, but assuming the incorrect model) are also included as a reference to interpret the results with real data. The same performance behaviors depicted by synthetic data are observed with real data, although the performance drop is relatively much more prominent. This effect might be partially explained by the fact that the mean candidate size in all three real datasets is larger than 2 and their estimated meaningfulness shows intermediate values (see Table 1).

### 6.3. Contribution of weak supervision with increasingly larger candidate sets and probability of co-occurrence

The interaction between these two factors helps us understand how the consistent provision of a false label affects the performance in scenarios with different amounts of supervision, determined by the size of the candidate sets. In Fig. 6 we study the interaction of these two dimensions on synthetic data under the correct model assumption. One surface illustrates how the use of data with a certain co-occurrence degrades the performance, whereas a second surface shows how the performance remains if no co-occurrence is induced. The results are similar for both metrics, the performance drop increases with the probability of co-occurrence, as expected, and it is more relevant when the candidate sets are small. As the candidate sets are larger (the amount of information of supervision is reduced), the performance drop is smaller and even negligible.

Fig. 7 shows the equivalent snapshot of the empirical study with real data. Candidate set size and co-occurrence level are fixed in real data, but we can still display one of the dimensions as different repetitions (subsamples) show slightly different co-occurrence levels. Results with synthetic data (honest scenario with $|S| = \{2, 3\}$ with and without induced co-occurrence, with the incorrect model assumption) are also included as a reference to interpret the results with real data. The averaged results with all three real datasets are comparable to those of synthetic data with induced co-occurrence and $|S| = 2$ (and/or




(a) Results in terms of Macro-F1 (b) Results in terms of Brier score

**Fig. 6.** Graphical description of the value of weak supervision as the probability of co-occurrence in the candidate sets and the size of the candidate sets increase on synthetic data. In all figures, a surface shows results of experiments without inducing co-occurrence, whereas another surface shows results when inducing co-occurrence with increasing probability. All the surfaces show the difference of performance of a classifier learned with weak supervision minus that of a similar classifier learned only with the labeled subset; in terms of Macro-F1 in a), of Brier score in b), relative to the performance of the real model. Other parameters are fixed to a small fully labeled subset ($N_f = 33$), twice as weakly labeled samples ($r = 2$), and simple generative and learned models ($K = 1$).

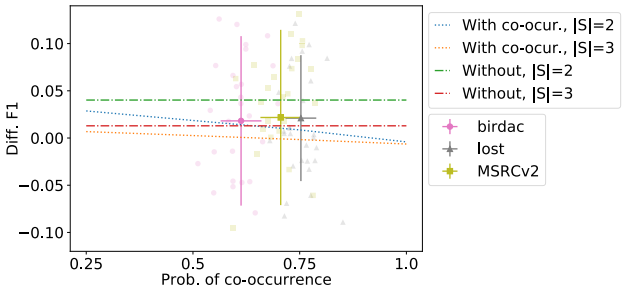
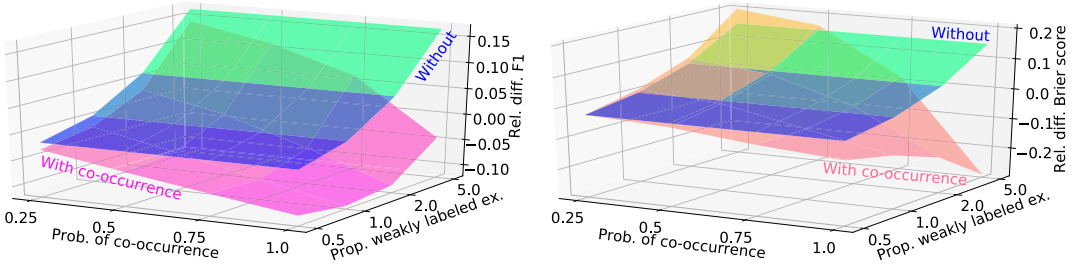



**Fig. 7.** Display of different scenarios of candidate set size as the probability of co-occurrence increases on real and synthetic data. Dashed lines depict results on synthetic data without and with increasing induced co-occurrence, for candidate set sizes $|S| = \{2, 3\}$. Each point depicts the result of a repetition (30) on a real dataset (Birdac, Lost or MSRCv2; see in Table 1 their co-occurrence level): performance vs. co-occurrence level of the specific training subsample (solid lines show mean and std. deviation of both axes). In all the cases, the difference of performance in terms of Macro-F1 of a classifier learned with weak supervision minus that of a similar classifier learned only with the labeled subset is shown. Other parameters are fixed to the honest scenario, a small fully labeled subset ($N_f = 33$), twice as weakly labeled samples ($r = 2$), and complex generative ($K = 4$) and simple learned ($K = 1$) models.




(a) Results in terms of Macro-F1 (b) Results in terms of Brier score

**Fig. 8.** Graphical description of the value of weak supervision as the probability of co-occurrence in the candidate sets and the proportion of weakly labeled examples increase on synthetic data. In all figures, a surface shows results of experiments without inducing co-occurrence, whereas another surface shows results when inducing consistent labels with increasing probability of co-occurrence. All the surfaces show the difference of performance of a classifier learned with weak supervision minus that of a similar classifier learned only with the labeled subset; in terms of Macro-F1 in a), of Brier score in b), relative to the performance of the real model. Other parameters are fixed to a small fully labeled subset ($N_f = 33$), candidate set size $|S| = 2$, and simple generative and learned models ($K = 1$).

without induced co-occurrence with $|S| = 3$). Remember that the real datasets have mean candidate sizes in the interval [2.2, 3.05] (see Table 1).

### *6.4. Contribution of weak supervision with increasingly larger weakly-labeled subset and probability of co-occurrence*

The interaction between these two factors helps us understand how the consistent provision of a false label affects the performance in scenarios with different numbers of weakly-labeled samples. In Fig. 8 we study the interaction of these two dimensions on synthetic data under the correct model assumption. As before, two surfaces display results with and without

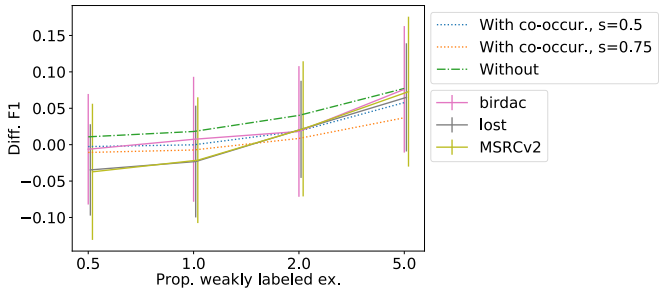



**Fig. 9.** Display of different scenarios of co-occurrence level as the proportion of weakly labeled examples increases on real and synthetic data. Dashed lines depict results on synthetic data without and with induced co-occurrence ($s = \{0, 0.5, 0.75\}$). Solid lines depict results on real data: Birdac, Lost and MSRCv2 (see in Table 1 their co-occurrence level). All the lines show the difference of performance in terms of Macro-F1 of a classifier learned with weak supervision minus that of a similar classifier learned only with the labeled subset. Other parameters are fixed to the honest scenario, with a small fully labeled subset ($N_f = 33$), and complex generative ($K = 4$) and simple learned ($K = 1$) models.

induced co-occurrence. Classification performance (F1) drops with a larger probability of co-occurrence, although it seems to rise when a large weakly-labeled subset is provided. Although this might seem in contradiction with the theoretical insights by [19], note that these experiments include a fully-labeled subset, which might be providing enough evidence for class disambiguation. In terms of probabilistic calibration, the performance drop is monotonically decreasing as both the probability of co-occurrence and the proportion of weakly-labeled samples increase.

Fig. 9 shows the equivalent snapshot of the empirical study with real data. Note that the results with real data of this figure are already presented in Fig. 3. This extra figure shows performance on real data in the context of different synthetic scenarios of co-occurrence $|s| = \{0, 0.5, 0.75\}$ (honest labels, under the incorrect model assumption). Results with all three real datasets end up resembling those of synthetic data with a probability of co-occurrence of $s = 0.5$ (and/or without induced co-occurrence). Seen together, Figs. 3 and 9 show the complexity of analyzing on real data how different factors affect performance, as several ones (co-occurrence level and candidate set size, in this comparison) might be used to explain parts of the results.

## 7. Discussion

In the previous section, we have displayed different snapshots of the empirical study. The main conclusion is that weakly labeled data contributes for machine learning in specific cases: when a relatively small fully labeled sample is available, and weakly labeled examples provide a considerable amount of supervision which should also be preferably meaningful. Nevertheless, there also exist scenarios where weak supervision might harm the learning process even under the correct model assumption (i.e., when the real model is learnable). In the rest of the section, we discuss the contribution factor by factor.

***Is using more weakly supervised examples always better?*** The (relative) amount of weakly supervised examples seems to have a positive effect, but it can be harmful in case of unfavorable configuration of other factors.

In the favorable scenario of honest candidate labels and assuming the correct model, more data is usually positive. For the task of classification, according to Fig. 4a, a larger weakly supervised subset usually helps and, in the most unfavorable scenarios, harm is limited. A similar behavior is observed for the task of calibration (Fig. 4b): it is usually helpful, although in this case severe performance drops happen in the most unfavorable scenarios.

The amount of examples with weak supervision cannot be analyzed as an isolated factor. More weakly supervised data is clearly helpful when the amount of fully labeled data is short, but it might harm the learning process if enough fully labeled data is available. The gain is clearly observed when candidate sets are small. Fig. 2 suggests that as the amount of information of supervision is reduced ($|S|$ increases), the proportion of weakly supervised samples required to observe a performance gain needs to be increased. Note that this somehow contradicts our first intuition (Sect. 5.1). The meaningfulness of the label sets is also determining. In the case of misleading label sets, more weakly supervised data can hurt the behavior of the classifier. Finally, the gain obtained with larger weakly labeled data is reduced as the size of the label sets increases, or even reverted into a performance loss (Fig. 2). Under the correct model assumption, this last observation might seem in contraction with Cohen and Cozman [1] (and our first intuition), who argue that unlabeled data does not harm in this setting. The apparent discrepancy could be explained by the fact that ours is a more complex experimental setting where, for example, the cardinality of the class variable is large (in their case, it is fixed to 2).

The experiments with real data provide evidence for this question when the correct model assumption does not hold. These appear to be well aligned with synthetic experiments, and its performance gain as the proportion of the weakly labeled subset increases (Figs. 3 and 9) is well described by other data characteristics (Table 1). As might be expected, the same relationships between the amount of weakly labeled data and other factors seem to hold (Fig. 5), but the performance gain is restricted to a smaller set of favorable conditions (e.g., when the fully labeled subset is really small). Candidate

set size and the meaningfulness of their labels are fixed in real data, but additional experiments where the correct model assumption does not hold available as supplementary material confirm these insights.

The most interesting observation is that the largest gain consistently happens in a practical scenario: practitioners usually consider employing weakly labeled data when fully labeled data is not enough. As the difficulty of obtaining fully labeled examples is usually much larger than that of obtaining weakly labeled examples, one might expect the relative proportion of weakly labeled samples ($r$) to be large. This realistic scenario is in principle favorable for considering weakly labeled data, whenever other factors are not adverse.

***How does the size of the candidate set affect?*** This factor is directly related to the amount of information of supervision provided. In scenarios where label sets are large, the provided information approaches that of unsupervised learning (none), whereas it might resemble the supervision information of supervised learning (full) as label sets get tiny. This intuition (second one in Sect. 5.1) is clearly observed in the experimental results. For example, in Fig. 2 the performance in terms of both F1 and Brier score consistently drops as the size of the label sets is increased. However, it relates to other factors, such as the meaningfulness of its labels. The performance drop observed when larger candidate sets are used is more severe if the labels are misleading.

As candidate set size is fixed in real data, our empirical study with this type of data cannot capture directly the referred effect. It can be seen, however, that the observed performance in Fig. 3 seems to be reasonably well explained by the candidate set size, as it is surrounded by simplified synthetic scenarios with similar characteristics.

Depending on the labeling procedure, the size of the candidate set is a factor that is partially actionable. If the labeling process tolerates it, one can ask the labeler to provide as few labels as possible (while making sure that the real one is in the set). The practitioner could even decide to filter out annotations with too large candidate sets according to the evidence provided by this study.

***How does the composition of the candidate label sets affect?*** In line with our third intuition (Sect. 5.1), meaningful (honest) labels consistently provide more information, as the performance of the learned classifiers is better than that with misleading labels both in terms of F1 and Brier score (Figs. 2 and 4). This empirical study is based on generative models and, in this way, one assumes that the data observed has followed a sampling procedure from a certain –unknown– model. A composition of the label sets where labels are included based on their probability of being the real label follows the rationale of a reasonable sampling procedure, as it provides meaningful information about X-y relationship. In fact, our method has been derived from a likelihood function (Eq. (1)) which implicitly assumes this type of labeling. Thus, everything is implicitly designed to learn with honest label sets, which explains that this experimental setting outperforms tests that follow the opposite rationale (labels are included based on their probability of **not** being the real label; *misleading* label sets). In this sense, our learning methods cannot be expected to be robust against misleading information.

The underlying hypothesis of the supervision schema used in this study is that, if the labeler is doubting among a few classes, he/she would be asked to include them all. Doubt would denote similarity in this mindset. However, if the labeler (or labeling method) is not reliable and might include other classes, then the learning process will malfunction. Although meaningfulness is defined based on the generative model, which is not available for real data, we have estimated it using a complex 4-DB model for the three real datasets. In all the cases, the estimated meaningfulness is around 0.5: i.e., the additional labels are not always the most probable ones. This observation challenges the models for candidate generation that our methods usually assume. The bad news, in this case, is that this factor is usually hard (or even impossible) to control. Thus, there would be room for novel methods which are based on alternative label-set generation models.

***How does co-occurrence affect?*** Following the discussion about the composition of the candidate sets, one might expect that the learning method could struggle to learn to disambiguate labels in case of co-occurrence (i.e., a pair of labels is provided together with high frequency), as a particular case of our intuition 3 (Sect. 5.1). Our empirical study, when using synthetic data, provides clear evidence of the harmful effect of co-occurrence on the learning performance (Figs. 6 and 8): the larger the probability of induced co-occurrence, the larger the performance drop. We study the interaction of this factor with the size of the candidate sets and the amount of weakly labeled samples. Induced co-occurrence seems to have a smaller effect on the results as the size of the label sets increases (Fig. 6), but note that in this case, a large candidate size leads to large minimum co-occurrence probabilities. That is, one of the problems that might be associated with large candidate sets is the inherent co-occurrence. Likewise, classification performance (F1) seems to be affected by co-occurrence independently from the amount of weakly labeled samples (Fig. 8). In contrast, calibration performance (Brier score) drops as both the probability of co-occurrence and the size of the weakly labeled subset increase. A possible explanation for this divergence might be that the provided small set of fully labeled examples is enough to resolve the discrimination between the co-occurrent labels.

This factor is also fixed in real data, and we cannot manipulate it in our experimental setting. Taking a closer look at the experimental repetitions (with different subsamples), we have observed (Fig. 7) that the mean classification performance is in line with that of experiments with synthetic data and candidate set size $|S| = 2$, or even better.

Among other theoretical insights in line with our findings, these factors regarding the composition of the label sets can be interpreted in similar terms to concepts such as *ambiguity degree* [7,19]. These previous works are based on the so-called ambiguity degree to identify scenarios where learning is actually possible and establish theoretical learning conditions [7, 19]. They show that learning to discriminate becomes harder as co-occurrence increases ($s \to 1$, in our setup). Our results

would come to support the idea that a relatively small subset of fully labeled samples can allow us to learn a better classifier (note the different behavior of the performance in terms of F1 and Brier score in Fig. 8 as the probability of co-occurrence increases), i.e., they provide enough information for the model to learn to discriminate.

### 7.1. General ideas

This empirical study is based on the experimental setup of [1]. As they do, we assume the correct model to grasp the full potential of weak supervision in the set of experiments with synthetic data. In the experiments with real data, this assumption does not hold and, moreover, many of the experimental factors considered in this study cannot be disentangled. Thus, it is difficult to isolate and analyze the individual effect of each factor. We have replicated the study with generative models of large ($K = 4$) complexity to simulate an experimental setup where the correct model does not hold. Similar figures to the ones displayed in this manuscript are available in the supplementary material.[3] In line with the results reported in [1], in case of not using the correct model, the learning task becomes harder: the contribution of the weakly labeled data is consistently more limited and, consequently, the set of scenarios where weak supervision clearly helps is much more reduced than the ones presented in this paper. Results with real data, completely in line with behaviors observed with synthetic data, provide additional validation to this empirical study. To boost reproducibility, source code is also available as supplementary material.

Although we have employed a specific supervision schema (candidate labels), as our goal with this selection was to be able to cover a vast spectrum of the weak supervision space, we expect that these results might generalize to other weakly supervised learning problems as far as one can establish how much information of supervision the data involves, and/or how informative the labeling might be. These concepts are related to the factors used in this empirical study such as the size of the candidate sets or the meaningfulness of the labels in them.

## 8. Conclusions

This paper presents an empirical study of the contribution of weakly supervised data in the context of generative models. Specifically, we aim to show the contribution of weakly supervised data for enhancing the performance of generative models for classification under the realistic scenario that assumes that a small subset of fully labeled data is available. Specifically, we focus on the problem of learning from candidate labels [7,19], where each training instance is provided with a label set that always includes the real label. Inspired by the key work by Cozman and Cohen [1], we employ standard learning techniques for generative models and this type of data. To be able to explore the whole spectrum of weakly labeled scenarios, we use synthetic scenarios simulating different empirical conditions and study how the performance of the models is affected. We also include experiments with real data labeled with candidate sets to explore these ideas in even more realistic scenarios.

Empirical results show that larger amounts of supervision, i.e., smaller candidate sets, usually determine the contribution of the weakly supervised data. However, not all the information of supervision is helpful: only meaningful information is. When the weakly supervised data includes, for example, adversarial or non-discriminable information, it might harm the performance of a learned model. It can also be observed that, when the objective is purely discrimination, even a small amount of certain class information is enough to overcome the learning difficulties anticipated by previous theoretical studies.

There are questions still open after this empirical study. A similar study could be carried out broadening the scope, such as using discriminative models or analyzing the impact of relevant characteristics of the underlying domain problem (e.g., if it is a class imbalance problem, if the weak supervision is closely related to difficult instances, etc.). Many weakly supervised problems assume that the information of supervision is never completely wrong. For example, learning from candidate labels assumes that the real label is always included in the label set. It would be of interest to extend our study to cover scenarios where this assumption does not hold.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

J. Hernández-González is a Serra Húnter fellow. A. Pérez is supported by the Spanish Ministry of Science, Innovation and Universities - AEI through the BCAM Severo Ochoa excellence accreditation SEV-2017-0718, and the Basque Government through the BERC 2022-2025 program.

[3] https://jhernandezgonzalez.github.io/supp_empweak.html.

## References


[1] F.G. Cozman, I. Cohen, Risks of semi-supervised learning: how unlabeled data can degrade performance of generative classifiers, in: Semi-Supervised Learning, MIT Press, 2006, pp. 57–72 (Ch. 4).
[2] O. Chapelle, B. Schökopf, A. Zien, Semi-Supervised Learning, MIT Press, 2006.
[3] F. Denis, R. Gilleron, F. Letouzey, Learning from positive and unlabeled examples, Theor. Comput. Sci. 348 (1) (2005) 70–83, https://doi.org/10.1016/j.tcs.2005.09.007.
[4] D. García-García, R.C. Williamson, Degrees of supervision, in: Proceedings of the 25th Conference on Neural Information Processing Systems Workshops, 2011.
[5] J. Hernández-González, I. Inza, J.A. Lozano, Weak supervision and other non-standard classification problems: a taxonomy, Pattern Recognit. Lett. 69 (2016) 49–55, https://doi.org/10.1016/j.patrec.2015.10.008.
[6] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, L. Moy, Learning from crowds, J. Mach. Learn. Res. 11 (2010) 1297–1322.
[7] T. Cour, B. Sapp, B. Taskar, Learning from partial labels, J. Mach. Learn. Res. 12 (2011) 1501–1536.
[8] S. Kumar, H.A. Rowley, Classification of weakly-labeled data with partial equivalence relations, in: Proceedings of the 11th IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[9] J. Hernández-González, I. Inza, J.A. Lozano, Learning Bayesian network classifiers from label proportions, Pattern Recognit. 46 (12) (2013) 3425–3440, https://doi.org/10.1016/j.patcog.2013.05.002.
[10] T. Zhang, F.J. Oles, A probability analysis on the value of unlabeled data for classification problems, in: Proceedings of the 17th International Conference on Machine Learning, 2000.
[11] V. Castelli, T.M. Cover, On the exponential value of labeled samples, Pattern Recognit. Lett. 16 (1) (1995) 105–111, https://doi.org/10.1016/0167-8655(94)00074-D.
[12] V. Castelli, T. Cover, The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter, IEEE Trans. Inf. Theory 42 (6) (1996) 2102–2117, https://doi.org/10.1109/18.556600.
[13] J. Ortigosa-Hernández, I. Inza, J.A. Lozano, Semisupervised multiclass classification problems with scarcity of labeled data: a theoretical study, IEEE Trans. Neural Netw. Learn. Syst. 27 (12) (2016) 2602–2614, https://doi.org/10.1109/TNNLS.2015.2498525.
[14] A. Singh, R.D. Nowak, X. Zhu, Unlabeled data: now it helps, now it doesn't, in: Advances in Neural Information Processing Systems (NeurIPS), vol. 21, 2009.
[15] N.V. Chawla, G. Karakoulas, Learning from labeled and unlabeled data: an empirical study across techniques and domains, J. Artif. Intell. Res. 23 (1) (2005) 331–366, https://doi.org/10.1613/jair.1509.
[16] J.E. van Engelen, H.H. Hoos, A survey on semi-supervised learning, Mach. Learn. 109 (2) (2020) 373–440, https://doi.org/10.1007/s10994-019-05855-6.
[17] R. Jin, Z. Ghahramani, Learning with multiple labels, in: Advances in Neural Information Processing Systems (NeurIPS), vol. 15, 2002.
[18] J. Luo, F. Orabona, Learning from candidate labeling sets, in: Advances in Neural Information Processing Systems (NeurIPS), vol. 23, 2010.
[19] L.-P. Liu, T.G. Dietterich, Learnability of the superset label learning problem, in: Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 3602–3610.
[20] E. Hüllermeier, W. Cheng, Superset learning based on generalized loss minimization, in: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2015, pp. 260–275.
[21] V. Cabannes, A. Rudi, F. Bach, Structured prediction with partial labelling through the infimum loss, in: Proceedings of the 37th International Conference on Machine Learning, 2020, pp. 1207–1216.
[22] V. Cabannes, F. Bach, A. Rudi, Disambiguation of weak supervision with exponential convergence rates, in: Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 1147–1157.
[23] Q.W. Wang, Y.F. Li, Z.H. Zhou, Partial label learning with unlabeled data, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019, pp. 3755–3761.
[24] B. Quost, T. Denoeux, S. Li, Parametric classification with soft labels using the evidential EM algorithm: linear discriminant analysis versus logistic regression, Adv. Data Anal. Classif. 11 (2017) 659–690, https://doi.org/10.1007/s11634-017-0301-2.
[25] T. Denoeux, Maximum likelihood estimation from uncertain data in the belief function framework, IEEE Trans. Knowl. Data Eng. 25 (1) (2013) 119–130, https://doi.org/10.1109/TKDE.2011.201.
[26] T. Denoeux, Maximum likelihood estimation from fuzzy data using the em algorithm, Fuzzy Sets Syst. 183 (1) (2011) 72–91, https://doi.org/10.1016/j.fss.2011.05.022.
[27] E. Côme, L. Oukhellou, T. Denoeux, P. Aknin, Learning from partially supervised data using mixture models and belief functions, Pattern Recognit. 42 (3) (2009) 334–348, https://doi.org/10.1016/j.patcog.2008.07.014.
[28] A. Campagner, Learnability in "learning from fuzzy labels", in: Proceedings of the 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2021, pp. 1–6.
[29] I. Couso, D. Dubois, A general framework for maximizing likelihood under incomplete data, Int. J. Approx. Reason. 93 (2018) 238–260, https://doi.org/10.1016/j.ijar.2017.10.030.
[30] L.-P. Liu, T.G. Dietterich, A conditional multinomial mixture model for superset label learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 557–565.
[31] R.D. Gill, M.J. van der Laan, J.M. Robins, Coarsening at random: characterizations, conjectures, counter-examples, in: Proceedings of the First Seattle Symposium in Biostatistics, 1997, pp. 255–294.
[32] M. Jaeger, Ignorability in statistical and probabilistic inference, J. Artif. Intell. Res. 24 (1) (2005) 889–917, https://doi.org/10.1613/jair.1657.
[33] G. Lugosi, Learning with an unreliable teacher, Pattern Recognit. 25 (1) (1992) 79–87, https://doi.org/10.1016/0031-3203(92)90008-7.
[34] P.G. Ipeirotis, F. Provost, J. Wang, Quality management on Amazon Mechanical Turk, in: Proceedings of the ACM SIGKDD Workshop on Human Computation, 2010, pp. 64–67.
[35] C. Bielza, P. Larrañaga, Discrete Bayesian network classifiers: a survey, ACM Comput. Surv. 47 (1) (2014) 1–43, https://doi.org/10.1145/2576868.
[36] S.L. Lauritzen, Graphical Models, vol. 17, Clarendon Press, 1996.
[37] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc., Ser. B, Methodol. 39 (1) (1977) 1–38.
[38] M. Sahami, Learning limited dependence Bayesian classifiers, in: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD 1996), 1996, pp. 335–338.
[39] D.J. Hand, K. Yu, Idiot's Bayes — not so stupid after all?, Int. Stat. Rev. 69 (3) (2001) 385–398, https://doi.org/10.2307/1403452.
[40] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (2–3) (1997) 131–163, https://doi.org/10.1023/A:1007465528199.
[41] D. Heckerman, A tutorial on learning with Bayesian networks, Tech. Rep. MSR-TR-95-06, Learning in Graphical Models, 1995.
[42] G.S. Mann, A. McCallum, Generalized expectation criteria for semi-supervised learning with weakly labeled data, J. Mach. Learn. Res. 11 (2010) 955–984.
[43] J. Lienen, E. Hüllermeier, Credal self-supervised learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2021.
[44] J.D. Rodríguez, A. Pérez, J.A. Lozano, A general framework for the statistical analysis of the sources of variance for classification error estimators, Pattern Recognit. 46 (3) (2013) 855–864, https://doi.org/10.1016/j.patcog.2012.09.007.