On the Optimality of Multi-Label Classification under Subset Zero-One Loss for Distributions Satisfying the Composition Property

Maxime Gasse MAXIME.GASSE@LIRIS.CNRS.FR
Alex Aussem ALEXANDRE.AUSSEM@LIRIS.CNRS.FR
Haytham Elghazel HAYTHAM.ELGHAZEL@LIRIS.CNRS.FR
LIRIS, UMR 5205, University of Lyon 1, 69622 Lyon, France

Abstract

The benefit of exploiting label dependence in multi-label classification is known to be closely dependent on the type of loss to be minimized. In this paper, we show that the subsets of labels that appear as irreducible factors in the factorization of the conditional distribution of the label set given the input features play a pivotal role for multi-label classification in the context of $0/1$ loss minimization, as they divide the learning task into simpler independent multi-class problems. We establish theoretical results to characterize and identify these irreducible label factors for any given probability distribution satisfying the Composition property. The analysis lays the foundation for generic multi-label classification and optimal feature subset selection procedures under this subclass of distributions. Our conclusions are supported by carefully designed experiments on synthetic and benchmark data.

1. Introduction

Multi-label Classification (MLC) is a challenging problem in many real-world application domains, where each instance can be assigned simultaneously to multiple binary labels (Dembczynski et al., 2012; Read et al., 2009; Madjarov et al., 2012; Kocev et al., 2007; Tsoumakas et al., 2011). Formally, learning from multi-label examples amounts to finding a mapping from a space of features to a space of labels. Given a multilabel training set $\mathcal{D}$, the goal of MLC is to find a function which is able to map any unseen example to its proper set of labels. From a Bayesian point of view, this problem amounts to modeling the conditional joint distribution $p(\mathbf{Y}|\mathbf{X})$, where $\mathbf{X}$ is a random vector in $\mathbb{R}^d$ associated with the input space, $\mathbf{Y}$ a random vector in $\{0,1\}^n$ associated with the labels, and p the probability distribution defined over $(\mathbf{X}, \mathbf{Y})$.

The problem of modeling $p(\mathbf{Y}|\mathbf{X})$ may be tackled in various ways (Luaces et al., 2012; Cherman et al., 2012; Read et al., 2009; Blockeel et al., 1998; Kocev et al., 2007; Gasse et al., 2014). Each of these approaches is supposed to capture - to some extent - the relationships between labels. An important question remains: what shall we capture from the statistical relationships between labels exactly to solve the multi-label classification problem? The question received increasing attention in the last years. So far, there was a consensus among researchers that, to improve the performance of MLC algorithms, label dependencies have to be incorporated into the learning process (Tsoumakas & Vlahavas, 2007; Guo & Gu, 2011; Zhang & Zhang, 2010; Bielza et al., 2011). In a recent paper, however, Dembczynski et al. (2012) showed that the expected benefit of exploiting label dependence is tightly dependent on the type of loss to be minimized and, most importantly, one cannot expect the same MLC method to be optimal for different types of losses at the same time (the reader is directed to Dembczynski et al. (2012) and references therein for further details about closed-form solution for risk-minimizing prediction).

In this study we are mainly concerned with risk-minimizing prediction for the subset $0/1$ loss which is commonly applied as performance metric in MLC experimental studies. More specifically, we establish several theorems to characterize: 1) the irreducible label factors (denoted as ILFs) in the factorization of the conditional distribution of the label set given the input features (i.e., minimal subsets $\mathbf{Y}_{LF} \subseteq \mathbf{Y}$ such that $\mathbf{Y}_{LF} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_{LF} \mid \mathbf{X}$) under the assumption that the probability distribution satisfies the Composition property; and 2) the ILFs' Markov boundaries (that are not necessarily unique under this subclass of distributions). The latter problem is closely related to the feature subset selection problem in the MLC context (Gharroudi et al., 2014; Lee & Kim, 2013; Spolaôr et al.,

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

2013) which has not yet been underpinned by theoretical results to the best of our knowledge. We emphasize that the present analysis conducted in this paper prepares the ground for a generic class of correct procedures for solving the MLC problem under the subset $0/1$ loss, for any given distribution satisfying the Composition property.

The rest of the paper is organized as follows: Section 2 discusses some key concepts used along the paper and state some results that will support our analysis. Section 3 addresses the ILF decomposition and the feature selection problem in the MLC context. Finally, we propose in section 4 a straightforward instantiation of our ILF decomposition procedure, called ILF-Compo, and present a number of experimental studies, using both synthetic and benchmark data.

2. Preliminaries

We define next some key concepts used along the paper and state some results that will support our analysis. In this paper, upper-case letters in italics denote random variables (e.g., X, Y) and lower-case letters in italics denote their values (e.g., x, y). Upper-case bold letters denote random variable sets (e.g., $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$) and lower-case bold letters denote their values (e.g., $\mathbf{x}, \mathbf{y}, \mathbf{z}$). We denote by $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ the conditional independence between $\mathbf{X}$ and $\mathbf{Y}$ given the set of variables $\mathbf{Z}$. To keep the notation uncluttered, we use $p(\mathbf{y} \mid \mathbf{x})$ to denote $p(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x})$. For consistency, we assume that $p(\emptyset) = 1$ and thus $\mathbf{X} \perp\!\!\!\perp \emptyset \mid \mathbf{Z}$ is always true, where $\emptyset$ denotes the empty set of random variables. We will assume the reader is familiar with the concepts of d-separation in Bayesian networks (BNs) (Pearl, 1989).

2.1. Conditional independence properties

We now present some properties of conditional independence. Let $\mathbf{U}$ denote a set of random variables, and p the probability distribution defined over $\mathbf{U}$. Let $\mathbf{X} \subseteq \mathbf{U}$, any $\mathbf{M} \subseteq (\mathbf{U} \setminus \mathbf{X})$ such that $\mathbf{X} \perp\!\!\!\perp \mathbf{U} \setminus (\mathbf{X} \cup \mathbf{M}) \mid \mathbf{M}$ is called a Markov blanket of $\mathbf{X}$ in $\mathbf{U}$. By extension, let $\mathbf{V} \subseteq \mathbf{U}$. Any $\mathbf{M} \subset (\mathbf{V} \setminus \mathbf{X})$ such that $\mathbf{X} \perp\!\!\!\perp \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{M}) \mid \mathbf{M}$ is called a Markov blanket of $\mathbf{X}$ in $\mathbf{V}$. It is worth noting that if $\mathbf{M}$ is a Markov blanket of $\mathbf{X}$ in $\mathbf{U}$, $\mathbf{M} \cap \mathbf{V}$ is not necessarily a Markov blanket of $\mathbf{X}$ in $\mathbf{V}$. Any minimal Markov blanket of $\mathbf{X}$ (i.e. none of its proper subsets is a Markov blanket of $\mathbf{X}$) is called a Markov boundary (MB) of $\mathbf{X}$. The following two theorems are proven in Pearl (1989):

Theorem 2.1. *Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and $\mathbf{W}$ denote four mutually disjoint subsets of $\mathbf{U}$. Any probability distribution p satisfies the following four properties:*

Symmetry $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \Rightarrow \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$
Decomposition $\mathbf{X} \perp\!\!\!\perp (\mathbf{Y} \cup \mathbf{W}) \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$
Weak Union $\mathbf{X} \perp\!\!\!\perp (\mathbf{Y} \cup \mathbf{W}) \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid (\mathbf{Z} \cup \mathbf{W})$
Contraction $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid (\mathbf{Z} \cup \mathbf{W}) \wedge \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp (\mathbf{Y} \cup \mathbf{W}) \mid \mathbf{Z}$

If p is strictly positive, then p satisfies the previous four properties plus the following property:

Intersection $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid (\mathbf{Z} \cup \mathbf{W}) \wedge \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid (\mathbf{Z} \cup \mathbf{Y}) \Rightarrow \mathbf{X} \perp\!\!\!\perp (\mathbf{Y} \cup \mathbf{W}) \mid \mathbf{Z}$

If p is faithful to a DAG $\mathcal{G}$, then p satisfies the previous five properties plus the following property:

Composition $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \wedge \mathbf{X} \perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp (\mathbf{Y} \cup \mathbf{W}) \mid \mathbf{Z}$

Theorem 2.2. *If p satisfies the Intersection property then each $\mathbf{X} \subseteq \mathbf{U}$ has a unique Markov boundary $\mathbf{MB}_{\mathbf{X}}$.*

It follows from Theorem 2.2 that the Markov boundaries are unique when p is faithful to a DAG $\mathcal{G}$. However, the Theorem says nothing about distributions that do not satisfy the Intersection property. In fact, many real-life distributions violate the Intersection property and contain non-unique Markov boundaries as discussed for instance in Statnikov et al. (2013); Peña et al. (2007).

2.2. Minimization of multi-label loss functions

In the framework of MLC, one can consider a multitude of loss functions. The risk of a classifier $\mathbf{h}$ is defined formally as the expected loss over the joint distribution,

$$R_L(\mathbf{h}) = \mathbb{E}_{\mathbf{X},\mathbf{Y}}[L(\mathbf{Y}, \mathbf{h}(\mathbf{X}))]$$

where $L(\cdot)$ is a loss function on multi-label predictions. The pointwise risk-minimizing model $\mathbf{h}^\star(\mathbf{x})$ is given by

$$\mathbf{h}^\star(\mathbf{x}) = arg \min_{y} \mathbb{E}_{\mathbf{Y}|\mathbf{x}}[L(\mathbf{Y}, \mathbf{y})]$$

In this study, we focus on a class of loss functions which explicitly requires the estimation of the joint conditional probability distribution $p(\mathbf{Y} \mid \mathbf{X})$. This class includes most non label-wise decomposable loss functions as discussed in Dembczynski et al. (2012) for instance. Modeling the entire joint distribution has also the desirable advantage that one can easily sample from the estimated joint distribution to deliver an optimal prediction under any loss function. In this paper, we will focus on the subset $0/1$ loss, which generalizes the well-known $0/1$ loss from the conventional to the multi-label setting,

$$L_S(\mathbf{Y}, \mathbf{h}(\mathbf{X})) = \theta(\mathbf{Y} - \mathbf{h}(\mathbf{X}))$$

where $\theta(\mathbf{x}) = 1$ if $\mathbf{x} \neq \mathbf{0}$ and $\theta(\mathbf{x}) = 0$ otherwise. The risk-minimizing prediction for subset $0/1$ loss is given by the mode of the distribution (Dembczynski et al., 2012),

$$\mathbf{h}^\star(\mathbf{x}) = arg \max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x})$$

Admittedly, this loss function may appear overly stringent, especially in the case of many labels. Still, it is nowadays routinely used as a performance metric in almost all MLC studies. While the focus will be on the subset $0/1$ loss in this study, the estimation of the joint conditional probability distribution is also required for other losses. Another example of loss function for which the optimal prediction was explicitly established is the instance-wise F-measure loss, which essentially corresponds to the harmonic mean of precision and recall. Dembczynski et al. showed in (2011) that the F-measure loss can be minimized in an efficient manner using $n^2 + 1$ parameters of the conditional joint distribution over labels: $p(\mathbf{Y} = \mathbf{0} \mid \mathbf{x})$ and the n^2 values of $p_{ik} = p(Y_i = 1, \sum_{i=1}^{n} Y_i = k \mid \mathbf{x})$ for $i, k = 1 \ldots, n$. Other losses, like the *Jaccard distance*, are suspected to require the estimation of the joint probability but are more difficult to analyze. Note that risk-minimizing predictions with label-wise decomposable losses (e.g. Hamming loss), and also specific non label-wise decomposable losses like the rank loss, can be solved on the basis of the marginal distributions $p(Y_i|\mathbf{x})$ alone. Hence, for such loss functions there is in principle no need for modeling conditional dependence between the labels. This does not exclude the possibility of first modeling the joint distribution and then perform a proper marginalization procedure. Although not pursued here, readers interested in exploring the connections between loss functions and optimal MLC predictions are encouraged to consult Dembczynski et al. (2011) as well as references therein.

2.3. Label factor decomposition

We shall now introduce the concept of label factor that will play a pivotal role in the factorization of the conditional distribution $p(\mathbf{Y} \mid \mathbf{X})$.

Definition 2.1. We say that $\mathbf{Y}_{LF} \subseteq \mathbf{Y}$ is a label factor *iff* $\mathbf{Y}_{LF} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_{LF} \mid \mathbf{X}$. Additionally, $\mathbf{Y}_{LF}$ is said irreducible if it is non-empty and has no other non-empty label factor as proper subset.

The key idea behind label factors is the decomposition of the conditional distribution of the labels into a product of factors

$$p(\mathbf{Y} \mid \mathbf{X}) = \prod_{j=1}^{L} p(\mathbf{Y}_{LF_j} \mid \mathbf{X}) = \prod_{j=1}^{L} p(\mathbf{Y}_{LF_j} \mid \mathbf{M}_{LF_j})$$

where $\{\mathbf{Y}_{LF_1}, \ldots, \mathbf{Y}_{LF_L}\}$ is a partition of label factors and $\mathbf{M}_{LF_j}$ is a Markov blanket of $\mathbf{Y}_{LF_j}$. From the above definition, we have $\mathbf{Y}_{LF_i} \perp\!\!\!\perp \mathbf{Y}_{LF_j} \mid \mathbf{X}, \forall i \neq j$. Under subset $0/1$ loss, we seek a factorization into a product of minimal factors in order to facilitate the estimation of the mode of the conditional distribution, also called the most probable explanation (MPE),

$$\begin{aligned}\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}) &= \prod_{j=1}^{L} \max_{\mathbf{y}_{LF_j}} p(\mathbf{y}_{LF_j} \mid \mathbf{x}) \\ &= \prod_{j=1}^{L} \max_{\mathbf{y}_{LF_j}} p(\mathbf{y}_{LF_j} \mid \mathbf{m}_{LF_j})\end{aligned}$$

This paper aims to obtain theoretical results for the characterization of the irreducible label factors $\mathbf{Y}_{LF_j}$ in order to be able to estimate the MPE more effectively.

3. Problem analysis

We shall assume throughout that $\mathbf{X}$ is the feature set, $\mathbf{Y}$ the label set, $\mathbf{U} = \mathbf{X} \cup \mathbf{Y}$ and p a probability distribution defined over $\mathbf{U}$.

3.1. Label factor algebraic structure

We first show that label factors can be characterized as an algebraic structure satisfying certain axioms. Let $\mathbf{LF}$ denote the set of all label factors defined over $\mathbf{U}$, and $\mathbf{ILF}$ the set of all irreducible label factors. It is easily shown that $\{\mathbf{Y}, \emptyset\} \subseteq \mathbf{LF}$. More specifically, the collection of all label factors in $\mathbf{U}$ can be ordered via subset inclusion to obtain a lattice bounded by $\mathbf{Y}$ itself and the null set.

Theorem 3.1. $\forall \mathbf{Y}_{LF_i}, \mathbf{Y}_{LF_j} \in \mathbf{LF}$, *then* $\mathbf{Y}_{LF_i} \cup \mathbf{Y}_{LF_j} \in \mathbf{LF}$ *and* $\mathbf{Y}_{LF_i} \cap \mathbf{Y}_{LF_j} \in \mathbf{LF}$. *Moreover, the decomposition of* $\mathbf{Y}$ *into irreducible label factors is unique.*

Proof of Theorem 3.1. First, we prove that $\mathbf{Y}_{LF_i} \cup \mathbf{Y}_{LF_j} \in \mathbf{LF}$. From the label factor assumption for $\mathbf{Y}_{LF_i}$ and $\mathbf{Y}_{LF_j}$ we have $\mathbf{Y}_{LF_i} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_{LF_i} \mid \mathbf{X}$ and $\mathbf{Y}_{LF_j} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_{LF_j} \mid \mathbf{X}$. Using the Weak Union property we obtain that $\mathbf{Y}_{LF_i} \perp\!\!\!\perp \mathbf{Y} \setminus (\mathbf{Y}_{LF_i} \cup \mathbf{Y}_{LF_j}) \mid \mathbf{X} \cup (\mathbf{Y}_{LF_j} \setminus \mathbf{Y}_{LF_i})$, and similarly with the Decomposition property we get $\mathbf{Y}_{LF_j} \setminus \mathbf{Y}_{LF_i} \perp\!\!\!\perp \mathbf{Y} \setminus (\mathbf{Y}_{LF_i} \cup \mathbf{Y}_{LF_j}) \mid \mathbf{X}$. We may now apply the Contraction property to show that $\mathbf{Y}_{LF_i} \cup \mathbf{Y}_{LF_j} \perp\!\!\!\perp \mathbf{Y} \setminus (\mathbf{Y}_{LF_i} \cup \mathbf{Y}_{LF_j}) \mid \mathbf{X}$. Therefore, $\mathbf{Y}_{LF_i} \cup \mathbf{Y}_{LF_j}$ is a label factor by definition. Second, we prove that $\mathbf{Y}_{LF_i} \cap \mathbf{Y}_{LF_j} \in \mathbf{LF}$. From the label factor assumption for $\mathbf{Y}_{LF_i}$ and $\mathbf{Y}_{LF_j}$ we have $\mathbf{Y}_{LF_i} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_{LF_i} \mid \mathbf{X}$ and $\mathbf{Y}_{LF_j} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_{LF_j} \mid \mathbf{X}$. Using the Weak Union property we obtain $\mathbf{Y}_{LF_i} \cap \mathbf{Y}_{LF_j} \perp\!\!\!\perp (\mathbf{Y} \setminus (\mathbf{Y}_{LF_i} \cup \mathbf{Y}_{LF_j})) \cup (\mathbf{Y}_{LF_j} \setminus \mathbf{Y}_{LF_i}) \mid \mathbf{X} \cup (\mathbf{Y}_{LF_i} \setminus \mathbf{Y}_{LF_j})$, and similarly with the Decomposition property we get $\mathbf{Y}_{LF_i} \cap \mathbf{Y}_{LF_j} \perp\!\!\!\perp \mathbf{Y}_{LF_i} \setminus \mathbf{Y}_{LF_j} \mid \mathbf{X}$. We may now apply the Contraction property to show that $\mathbf{Y}_{LF_i} \cap \mathbf{Y}_{LF_j} \perp\!\!\!\perp \mathbf{Y} \setminus (\mathbf{Y}_{LF_i} \cap \mathbf{Y}_{LF_j}) \mid \mathbf{X}$. Therefore, $\mathbf{Y}_{LF_i} \cap \mathbf{Y}_{LF_j}$ is a label factor by definition. Third, we prove by contradiction that the decomposition of $\mathbf{Y}$ into irreducible label factors is unique. Suppose it is not the case, then there exists two distinct and overlapping irreducible label factors $\mathbf{Y}_{LF_i}$ and $\mathbf{Y}_{LF_j}$, i.e. $\mathbf{Y}_{LF_i} \neq \mathbf{Y}_{LF_j}$

and $\mathbf{Y}_{LF_i} \cap \mathbf{Y}_{LF_j} \neq \emptyset$. As $\mathbf{Y}_{LF_i} \cap \mathbf{Y}_{LF_j}$ is also a label factor, then due to the irreducible label factor assumption we have either $\mathbf{Y}_{LF_i} = \mathbf{Y}_{LF_j}$ or $\mathbf{Y}_{LF_i} \cap \mathbf{Y}_{LF_j} = \emptyset$ which shows the desired result. □

It follows from Definition 2.1 and Theorem 3.1 that the set of all irreducible label factors **ILF** is a partition of $\mathbf{Y}$.

3.2. Irreducible label factor characterization

We shall now characterize the ILFs and their Markov boundaries for probability distributions satisfying the Composition property.

Lemma 3.2. *Suppose p supports the Composition property. Let Y_i and Y_j denote two distinct labels in $\mathbf{Y}$ and define by $\mathbf{Y}_{LF_i}$ and $\mathbf{Y}_{LF_j}$ their respective irreducible label factor, then we have $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid \mathbf{X} \Rightarrow \mathbf{Y}_{LF_i} = \mathbf{Y}_{LF_j}$.*

Proof of Theorem 3.2. By contradiction, suppose $\mathbf{Y}_{LF_i} \neq \mathbf{Y}_{LF_j}$. Then, owing to Theorem 3.1 we have $\mathbf{Y}_{LF_i} \cap \mathbf{Y}_{LF_j} = \emptyset$ and thus $Y_j \in \mathbf{Y} \setminus \mathbf{Y}_{LF_i}$. From the label factor assumption of $\mathbf{Y}_{LF_i}$, we also have that $\mathbf{Y}_{LF_i} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_{LF_i} \mid \mathbf{X}$, which yields $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \mid \mathbf{X}$ due to the Decomposition property. This concludes the proof. □

The inverse here is clearly not true.

Example. Consider $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$ and $\mathbf{X} = \emptyset$, with Y_1, Y_2, Y_3 three binary variables such that $Y_1 = Y_2 + Y_3$ ($+$ denotes the OR operator). Then we have $\{Y_1\} \not\perp\!\!\!\perp \{Y_2\} \mid \mathbf{X}$ and $\{Y_1\} \not\perp\!\!\!\perp \{Y_3\} \mid \mathbf{X}$, which implies that $\mathbf{ILF} = \{\{Y_1, Y_2, Y_3\}\}$. Here Y_2 and Y_3 are in the same irreducible label factor, yet we have $\{Y_2\} \perp\!\!\!\perp \{Y_3\} \mid \mathbf{X}$.

Lemma 3.3. *Suppose p supports the Composition property. Consider $\mathbf{Y}_{LF} \in \mathbf{ILF}$, then for all nonempty proper subset $\mathbf{Z}$ of $\mathbf{Y}_{LF}$, we have $\mathbf{Z} \not\perp\!\!\!\perp \mathbf{Y}_{LF} \setminus \mathbf{Z} \mid \mathbf{X}$.*

Proof of Lemma 3.3. By contradiction, suppose such a $\mathbf{Z}$ exists. Then we have $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y}_{LF} \setminus \mathbf{Z} \mid \mathbf{X}$. From the label factor assumption of $\mathbf{Y}_{LF}$, we also have that $\mathbf{Y}_{LF} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_{LF} \mid \mathbf{X}$, and therefore $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_{LF} \mid \mathbf{X}$ due to the Decomposition property. We may now apply the Composition property on these two statements to obtain $\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Z} \mid \mathbf{X}$ which contradicts the irreducible label factor assumption of $\mathbf{Y}_{LF}$. This concludes the proof. □

Theorem 3.4. *Suppose p supports the Composition property. Let $\mathcal{G}$ be an undirected graph whose nodes correspond to the random variables in $\mathbf{Y}$ such that Y_i and Y_j are connected in $\mathcal{G}$ if and only if $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid \mathbf{X}$. Then Y_i and Y_j are in same irreducible label factor if and only if a path exists between Y_i and Y_j in $\mathcal{G}$.*

Proof of Theorem 3.4. If a path exists between Y_i and Y_j in $\mathcal{G}$ then either Y_i and Y_j are directly connected, and thus $\mathbf{Y}_{LF_i} = \mathbf{Y}_{LF_j}$ due to Lemma 3.2, or there exists a sequence of intermediate nodes $Y_k, Y_{k+1}, \ldots$ such that $\mathbf{Y}_{LF_k} = \mathbf{Y}_{LF_{k+1}}$, and by induction $\mathbf{Y}_{LF_i} = \mathbf{Y}_{LF_j}$. We may now prove the converse. Suppose Y_i and Y_j belong to the same irreducible label factor $\mathbf{Y}_{LF}$, and define $\{\mathbf{W}_i, \mathbf{W}_j\}$ a partition of $\mathbf{Y}_{LF}$ such that $Y_i \in \mathbf{W}_i$ and $Y_j \in \mathbf{W}_j$. Consider W_k^i a label in $\mathbf{W}_i$. Using the Composition property, we have that either $\{W_1^i\} \not\perp\!\!\!\perp \mathbf{W}_j \mid \mathbf{X}$ or $(\mathbf{W}_i \setminus \{W_1^i\}) \not\perp\!\!\!\perp \mathbf{W}_j \mid \mathbf{X}$. Let us apply the Composition property again on the second expression, we obtain that $\{W_2^i\} \not\perp\!\!\!\perp \mathbf{W}_j \mid \mathbf{X}$ or $(\mathbf{W}_i \setminus \{W_1^i, W_2^i\}) \not\perp\!\!\!\perp \mathbf{W}_j \mid \mathbf{X}$. If we proceed recursively, we will necessarily find a variable $W_k^i \in \mathbf{W}_i$ such that $\{W_k^i\} \not\perp\!\!\!\perp \mathbf{W}_j \mid \mathbf{X}$. In other words, there exists at least one variable W_k^i in $\mathbf{W}_i$, such that $\{W_k^i\} \not\perp\!\!\!\perp \mathbf{W}_j \mid \mathbf{X}$. Likewise, we can proceed along the same line to exhibit a variable W_l^j in $\mathbf{W}_j$ such that $\{W_k^i\} \not\perp\!\!\!\perp \{W_l^j\} \mid \mathbf{X}$. In other words, for every partition $\{\mathbf{W}_i, \mathbf{W}_j\}$ of $\mathbf{Y}_{LF}$, there exists at least one label $W_k^i \in \mathbf{W}_i$ and one label $W_l^j \in \mathbf{W}_j$ such that $\{W_k^i\} \not\perp\!\!\!\perp \{W_l^j\} \mid (\mathbf{X} \cup \mathbf{Z})$. We proved that the irreducible label factor $\mathbf{Y}_{LF}$ containing $\{Y_i, Y_j\}$ is a connected component in $\mathcal{G}$, and therefore Y_i and Y_j are connected in $\mathcal{G}$. □

Theorem 3.5. *Suppose p supports the Composition property. Then, the statement $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \mid \mathbf{X}$ is strictly equivalent to $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \mid \mathbf{M}_i$ for every Markov blanket $\mathbf{M}_i$ of Y_i in $\mathbf{X}$.*

Proof of Theorem 3.5. We may rewrite $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \mid \mathbf{X}$ as $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \mid (\mathbf{X} \setminus \mathbf{M}_i) \cup \mathbf{M}_i$ for any $\mathbf{M}_i$. From the Markov blanket assumption for $\mathbf{M}_i$, we also have $\{Y_i\} \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{M}_i \mid \mathbf{M}_i$. Using the Contraction property, we obtain $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \cup (\mathbf{X} \setminus \mathbf{M}_i) \mid \mathbf{M}_i$. Using the Decomposition property, we obtain $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \mid \mathbf{M}_i$. Conversely, suppose there exists a Markov blanket $\mathbf{M}_i$ such that $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \mid \mathbf{M}_i$. From the Markov blanket assumption, we also have $\{Y_i\} \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{M}_i \mid \mathbf{M}_i$. Using the Composition property, we obtain $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \cup (\mathbf{X} \setminus \mathbf{M}_i) \mid \mathbf{M}_i$. Using the Weak Union, we obtain $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \mid \mathbf{X}$ which proves the equivalence. □

Theorem 3.4 suggests a practical way to construct the graph $\mathcal{G}$ provided that the Composition property holds. While the Composition property assumption is less stringent than the Faithfulness assumption, we provide an example where the Composition property does not hold.

Example. Consider $\mathbf{Y} = \{Y_1, Y_2, Y_3\}$ and $\mathbf{X} = \emptyset$, with Y_1, Y_2, Y_3 three binary variables such that $Y_1 = Y_2 \oplus Y_3$ ($\oplus$ denotes the exclusive OR operator). Here p does not satisfy the Composition property, and Theorem 3.4 does not apply any more. We have $\{Y_1\} \perp\!\!\!\perp \{Y_2\} \mid \mathbf{X}$, $\{Y_1\} \perp\!\!\!\perp \{Y_3\} \mid \mathbf{X}$ and $\{Y_2\} \perp\!\!\!\perp \{Y_3\} \mid \mathbf{X}$, and yet it is easily shown that $\mathbf{ILF} = \{\{Y_1, Y_2, Y_3\}\}$.

Notice that the search for ILFs relies on our ability to infer Markov boundaries for individual labels, which by the way may not be unique. This can in principle be achieved with any off the shelf Markov boundary discovery algorithm that is correct for distributions satisfying the Composition property.

The second fundamental problem that we wish to address involves finding a Markov boundary (or blanket) for a given ILF that consists of several labels. The problem is closely related to the optimal feature subset selection problem in the multi-label context. Notice that multi-label feature subset selection has recently received some attention (Gharroudi et al., 2014; Lee & Kim, 2013; Spolaôr et al., 2013). However, the empirical work has not yet been underpinned by theoretical results to our knowledge. We address the following question: Can we form the joint Markov boundary of a set of labels from the Markov boundaries of the single labels $\{Y_i\}$? Answering this question is not completely trivial as we shall see.

Theorem 3.6. *Suppose p satisfies the Composition property. Let $\mathbf{Y}_1$, $\mathbf{Y}_2$, ... and $\mathbf{Y}_n$ denote non-empty disjoint subsets of the label set $\mathbf{Y}$ and let $\mathbf{M}_i$ be a Markov boundary of $\mathbf{Y}_i$ in $\mathbf{X}$. Then, $\mathbf{M} = \mathbf{M}_1 \cup \ldots \cup \mathbf{M}_n$ is a Markov blanket for $\mathbf{Y}_1 \cup \ldots \cup \mathbf{Y}_n$ in $\mathbf{X}$. Moreover, if we remove recursively from $\mathbf{M}$ any X such that $\mathbf{Y}_1 \cup \ldots \cup \mathbf{Y}_n \perp\!\!\!\perp \{X\} | (\mathbf{M} \setminus \{X\})$, we obtain a Markov boundary.*

Proof of Theorem 3.6. We show first that the statement holds for $n = 2$ and then conclude that it holds for all n by induction. From the Markov blanket assumption for $\mathbf{Y}_1$ in $\mathbf{X}$ we have $\mathbf{Y}_1 \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{M}_1 \mid \mathbf{M}_1$. Using the Weak Union property, we obtain that $\mathbf{Y}_1 \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{M} \mid \mathbf{M}$. Similarly we can derive $\mathbf{Y}_2 \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{M} \mid \mathbf{M}$. Combining these two statements yields $\mathbf{Y}_1 \cup \mathbf{Y}_2 \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{M} \mid \mathbf{M}$ due to the Composition property. This, along with the fact that $\mathbf{M}_1 \subseteq \mathbf{X}$, $\mathbf{M}_2 \subseteq \mathbf{X}$, and hence $\mathbf{M} \subseteq \mathbf{X}$ defines $\mathbf{M}$ as a Markov blanket for $\mathbf{Y}_1 \cup \mathbf{Y}_2$ in $\mathbf{X}$. We shall now prove that removing recursively from $\mathbf{M}$ any X such that $\mathbf{Y}_i \perp\!\!\!\perp \{X\} | (\mathbf{M} \setminus \{X\})$ yields a Markov boundary. From the Markov blanket assumption we may write $\mathbf{Y}_i \perp\!\!\!\perp \mathbf{U} \setminus (\mathbf{Y}_i \cup \mathbf{M}) \mid (\mathbf{M} \setminus \{X\}) \cup \{X\}$. Using the Contraction property on these two statements then yields $\mathbf{Y}_i \perp\!\!\!\perp \{X\} \cup (\mathbf{U} \setminus (\mathbf{Y}_i \cup \mathbf{M})) \mid \mathbf{M} \setminus \{X\}$, which defines a Markov blanket. We may now prove that the resulting Markov blanket $\mathbf{M}$ is minimal. Let us suppose that it is not the case, i.e. there exists a non-empty proper subset $\mathbf{Z} \subset \mathbf{M}$ such that $\mathbf{Y}_i \perp\!\!\!\perp \mathbf{Z} \cup (\mathbf{U} \setminus (\mathbf{Y}_i \cup \mathbf{M})) \mid \mathbf{M} \setminus \mathbf{Z}$. From the Decomposition property we may write $\mathbf{Y}_i \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{M} \setminus \mathbf{Z}$, and then from the Weak Union property $\mathbf{Y}_i \perp\!\!\!\perp \{X\} \mid \mathbf{M} \setminus \{X\}$, with $\{X\} \subset \mathbf{M}$. Since we ensured that such an $\{X\}$ variable does not exist, we may conclude that $\mathbf{Z} = \emptyset$, and hence that $\mathbf{M}$ is minimal. This concludes the proof. □

Algorithm 1 ILF-Compo

Input: $\mathcal{D}$ a data set, $\mathbf{X}$ the set of features, $\mathbf{Y}$ the set of labels, $(\cdot \perp\!\!\!\perp \cdot \mid \cdot)$ a statistical test of conditional independence, MB_{alg}, a Markov boundary learning algorithm, MC_{alg}, a multi-class classification algorithm.
Initialize $\mathbf{ILF} \leftarrow \emptyset$, $\mathbf{Y}_{done} \leftarrow \emptyset$.
for all $Y_i \in \mathbf{Y}$ **do**
  Compute $\mathbf{M}_i$ of Y_i in $\mathbf{X}$ using MB_{alg}
while $\mathbf{Y} \setminus \mathbf{Y}_{done} \neq \emptyset$ **do**
  Select arbitrarily one label Y_i from $\mathbf{Y} \setminus \mathbf{Y}_{done}$
  Initialize $\mathbf{Y}_{LF} \leftarrow \{Y_i\}$
  while $\mathbf{Y}_{LF} \setminus \mathbf{Y}_{done} \neq \emptyset$ **do**
    Select arbitrarily one label Y_j from $\mathbf{Y}_{LF} \setminus \mathbf{Y}_{done}$
    Add Y_j to $\mathbf{Y}_{done}$
    for all $Y_k \in \mathbf{Y} \setminus (\mathbf{Y}_{done} \cup \mathbf{Y}_{LF})$ **do**
      if $\{Y_j\} \not\perp\!\!\!\perp_{\mathcal{D}} \{Y_k\} \mid \mathbf{M}_j$ **or** $\{Y_j\} \not\perp\!\!\!\perp_{\mathcal{D}} \{Y_k\} \mid \mathbf{M}_k$ **then**
        Add Y_k to $\mathbf{Y}_{LF}$
  Add $\mathbf{Y}_{LF}$ to **ILF**
Solve the MLC problem using **ILF** and MC_{alg}.

Notice that Theorem 3.6 holds for any set of labels and is not restricted in particular to ILFs. It shows that an ILF's Markov boundary can easily be obtained by combining the Markov boundaries of its individual labels, followed by a backward step in order to remove recursively the redundant features. Finding such a Markov boundary is useful to reduce the dimension of the input feature vector, thereby reducing the computation burden for training the multi-class classifiers.

The theoretical analysis conducted in this section lays the foundation for generic procedures to solve the MLC problem under the subset $0/1$ loss for any probability distribution that satisfies the Composition property. Consider the straightforward implementation called ILF-Compo described in Algorithm 1. The procedure goes as follows: i) learn the Markov boundary of every label in $\mathbf{X}$, and test for each pair of labels whether $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \mid \mathbf{M}_i$ or $\{Y_i\} \perp\!\!\!\perp \{Y_j\} \mid \mathbf{M}_j$ using any conditional independence test. While the two expressions are mathematically equivalent when the Composition holds, the tests may end up with distinct decisions for numerical reasons because the larger the size of the conditioning set, the less accurate are the independence tests. Notice that many other heuristics, more or less conservative, could be envisaged here. Then, ii) build up an undirected graph $\mathcal{G}$ owing to Theorem 3.5, and extract the ILFs owing to Theorem 3.4. Finally iii) decompose the MLC problem into a series of independent multi-class problems. So, ILF-Compo relies on a conditional independence test, a Markov boundary learning algorithm and a multi-class classification model.

The correctness of ILF-Compo relies on Theorem 3.5 and

on the fact that the mode of the conditional distribution is optimal for the subset $0/1$ loss as discussed in the preliminaries. While the procedure is mathematically sound, it may not necessarily translate - of course - into a reduction of $0/1$ loss in practical MLC scenarios for at least one reasons. Indeed, the judgments on conditional independence $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ are made by performing fallible statistical tests. Typically, they are based on either a G^2 or a χ^2 independence test when the data set is discrete and a Fisher's Z test when it is continuous in order to decide on dependence or independence, that is, upon the rejection or acceptance of the null hypothesis of conditional independence. The main limitation of these tests is the rate of convergence to their limiting distributions, which is particularly problematic when dealing with small sample sizes or sparse contingency tables. The decision of accepting or rejecting the null hypothesis depends on the freedom degree which grows exponentially with the size of the conditional set. For a practical comparison of conditional independence tests (i.e. parametric, permutation and Shrinkage tests), we defer to Scutari & Brogini (2012); Tsamardinos & Borboudakis (2010).

4. Experiments

This section presents a number of experimental studies, using both synthetic and benchmark data. Our aim is not to perform a through comparison of ILF-Compo against state-of-the-art MLC algorithms but instead to corroborate our theoretical findings by means of empirical evidence. We first investigate on a toy problem the extent to which ILF-Compo can solve the MLC problem when the degree of label dependencies is varied. Then, we assess the ability of ILF-Compo to reduce the subset $0/1$ loss on real-world multi-label data.

ILF-Compo was implemented in the R language, upon the *bnlearn* package from (Scutari, 2010). There exists in the literature a wealth of Markov boundary discovery algorithms, whose correctness is usually demonstrated for probability distribution satisfying the faithfulness assumption which is stronger than the Composition property. In our experiments, we use the Incremental Association Markov Boundary discovery algorithm (IAMB) proposed by Peña et al. (2007) which was proved to be correct for distributions satisfying the Composition property. Within both IAMB and ILF-Compo, a semi-parametric Mutual Information conditional independence tests was employed with $\alpha = 10^{-3}$ and 100 permutations as discussed in Tsamardinos & Borboudakis (2010). We used the Random Forest classifier from Breiman (2001), implemented in the *randomForest* R package from Liaw & Wiener (2002), as our multi-class classification model.

4.1. Toy problem

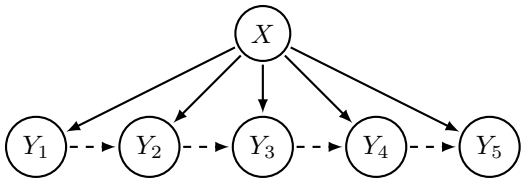


Figure 1. BN structure of our toy problem. Dashed lines indicate possibly missing edges.

Consider the toy problem depicted in Figure 4.1. The aim is to illustrate the importance of the ILF decomposition to minimize of the *subset* $0/1$ *loss*. As ILF-Compo includes BR and LP as special cases, we examine whether the extraction of ILFs translates to improved *subset* $0/1$ *loss* with respect to these two standard algorithms. Consider a (single) discrete variable X with 16 modalities, and 5 labels $Y_1, Y_2, \ldots, Y_5$. As may be seen, each label Y_i is dependent on X, and possibly on the preceding label Y_{i-1} according to the presence/absence of edges between labels as indicated by short-dashed arrows in Figure 4.1. By removing intentionally certain edges, we shall consider the following distinct ILF decompositions:

- DAG 1: $\mathbf{ILF} = \{\{Y_1\}, \{Y_2\}, \{Y_3\}, \{Y_4\}, \{Y_5\}\}$;
- DAG 2: $\mathbf{ILF} = \{\{Y_1, Y_2\}, \{Y_3, Y_4\}, \{Y_5\}\}$;
- DAG 3: $\mathbf{ILF} = \{\{Y_1, Y_2, Y_3\}, \{Y_4, Y_5\}\}$;
- DAG 4: $\mathbf{ILF} = \{\{Y_1, Y_2, Y_3, Y_4\}, \{Y_5\}\}$;
- DAG 5: $\mathbf{ILF} = \{\{Y_1, Y_2, Y_3, Y_4, Y_5\}\}$.

For each of these BN structures, we generate a random probability distribution by sampling uniformly the conditional probability table of each node in the DAG from a unit simplex as discussed in (Smith & Tromble, 2004). The process is repeated 1000 times for each DAG, to obtain 5*1000 random probability distributions p. From each distribution, we draw 7 training samples with respectively 50, 100, 200, 500, 1000, 2000 and 5000 instances and one testing sample with 5000 instances. ILF-Compo, LP and BR are then run on each training set using the same multi-class base learner, and the *subset* $0/1$ *loss* is assessed on the test set.

There are two things we want to evaluate: the quality of the decomposition and the quality of the MLC. As we know the ground truth of the ILF decomposition, it may be used as a gold standard to assess the efficiency of the ILF decomposition returned by ILF-Compo. To do so, we compute the Rand index of each decomposition. The idea behind the Rand index is to view the ILF decomposition as a series of decisions, one for each of the $N(N-1)/2$ pairs of labels. Ideally, one would like to assign two labels to the same ILF if and only if they are in the true ILF. The Rand index measures the percentage of decisions that are correct. The

result of this experiment are reported in Table 1 in terms of Rand index and in Figure 2 in terms of mean global accuracy (defined as 1-$0/1$ loss) with respect to the size of the training set.

In view of these curves, several conclusions may be drawn. First, ILF-Compo (black curve) compares favorably to BR (green curve) and LP (red curve) in all cases. Second, LP is asymptotically optimal with the size of the training set as expected. The asymptotic difference between LP and BR is more pronounced as we move from DAG 1 to DAG 5. In fact, the larger the conditional dependence between the labels, the more BR and LP diverge asymptotically. Nonetheless, it happens that BR outperforms LP on small sample sizes (< 500), despite not being optimal, as shown in Fig. 2b. The reason is that LP needs more observations to safely estimate $p(\mathbf{Y} \mid \mathbf{X})$ than BR to estimate each $p(Y_i \mid \mathbf{X})$. Third, when the labels are all independent to each other conditionally on X (see DAG 1), the ILFs are reduced to singletons and ILF-Compo boils down to BR. This not a surprise as BR is optimal in that case. At the opposite extreme, when the successive labels are pairwise dependent and the unique ILF consists of all the labels, ILF-Compo boils down to LP (DAG 5). Between these two extreme cases (DAGs 2, 3, 4), ILF-Compo always outperforms BR and LP.

Overall, the results reported in Table 1 are in nice agreement with our theoretical findings. It is worth noting, though, that in the extreme case where almost all the labels are conditionally dependent (e.g., DAG 4 and 5), ILF-Compo can fail spectacularly to identify the correct ILFs with small sample sizes (< 200). This is clearly due to the lack of robustness of the statistical test employed with a limited amount of samples. However, this has little impact in terms of subset $0/1$ loss for small sample sizes since BR, LP and ILF-Compo perform poorly as well. As the sample size increases, the quality of the decomposition becomes significantly better.

Table 1. Rand index, averaged over 1000 runs, of the decomposition output by ILF-Compo versus the optimal decomposition on the toy problem.

samp. size	DAG 1	DAG 2	DAG 3	DAG 4	DAG 5
50	99.9 ± 0.7	80.4 ± 2.2	60.6 ± 2.7	40.7 ± 2.8	1.4 ± 3.8
100	99.6 ± 2.0	85.0 ± 6.5	69.0 ± 7.8	49.9 ± 8.0	13.7 ± 10.1
200	99.3 ± 2.7	94.5 ± 6.4	83.5 ± 8.1	65.3 ± 8.3	34.4 ± 9.8
500	99.0 ± 3.0	99.0 ± 3.2	92.1 ± 5.3	75.6 ± 7.7	49.8 ± 9.4
1000	99.0 ± 3.0	99.3 ± 2.6	95.4 ± 5.3	82.4 ± 8.4	60.1 ± 10.8
2000	99.1 ± 2.9	99.0 ± 3.0	98.0 ± 4.1	88.8 ± 7.2	70.0 ± 9.5
5000	99.2 ± 2.7	98.9 ± 3.2	99.2 ± 2.8	94.0 ± 5.7	79.0 ± 9.1

4.2. Benchmark on real-world data

We may now report on the experiments performed on 10 real-world multi-label data sets. These data sets come from different problem domains including text, biology, and music. They can be found on the Mulan[1] repository, except for *image* which comes from Zhou[2] (Maron & Ratan, 1998). Of course, we have no idea whether the Composition axiom holds in these distributions, nor do we know the true ILFs decomposition. To increase the difficulty of the task, the data sets were also duplicated. The duplication was performed in a way that maintains the probabilistic structure of each set of variables while imposing their mutual independency, by permutating the rows on the duplicated variables. So by design, these augmented distributions have - at least - two irreducible label factors. We only compared here ILF-Compo and LP - as they are both intended to minimize the $0/1$ loss - and no feature selection was performed to avoid biasing the experimental results. When necessary, continuous variables were binarized upon median value in order to run our discrete independence test. A 5x2-fold cross-validation was performed on each dataset, and predictions were aggregated over the 10 test folds to estimate the subset $0/1$ loss.

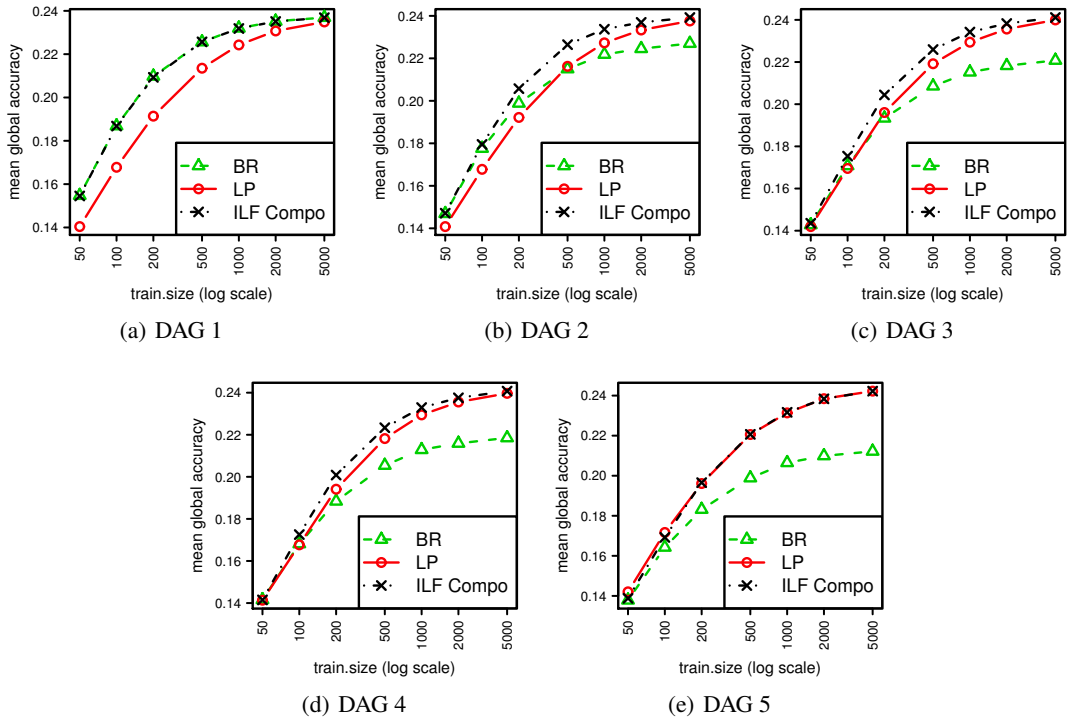
Table 2. Global accuracy on the original and the duplicated benchmarks.

data set	LP	ILF-Compo
emotions	35.7 ± 2.5	35.5 ± 1.9
image	47.4 ± 0.7	47.7 ± 0.9
scene	73.8 ± 1.4	73.3 ± 1.1
yeast	26.4 ± 1.1	26.1 ± 1.6
slashdot	45.3 ± 1.3	42.4 ± 1.4
genbase	96.2 ± 1.2	96.6 ± 1.1
medical	68.9 ± 1.8	65.5 ± 1.4
enron	15.5 ± 0.5	16.0 ± 0.5
bibtex	22.0 ± 0.5	13.8 ± 0.8
corel5k	3.0 ± 0.3	2.9 ± 0.2
emotions2	4.8 ± 1.1	10.7 ± 1.7
image2	12.0 ± 1.0	21.0 ± 0.8
scene2	35.2 ± 1.3	50.3 ± 1.5
yeast2	2.3 ± 0.4	5.8 ± 0.5
slashdot2	8.9 ± 1.0	18.2 ± 0.7
genbase2	69.1 ± 3.8	93.1 ± 1.7
medical2	20.6 ± 2.2	27.8 ± 2.3
enron2	0.6 ± 0.3	2.5 ± 0.4
bibtex2	0.8 ± 0.1	0.5 ± 0.1
corel5k2	0.0 ± 0.0	0.0 ± 0.0

Table 2 reports the outputs of our algorithm in terms of global accuracy of each method over the 10 data sets. The difference between ILF-Compo and LP on the origi-

[1] http://mulan.sourceforge.net/datasets.html

[2] http://lamda.nju.edu.cn/data_MIMLimage.ashx



(a) DAG 1 (b) DAG 2 (c) DAG 3

(d) DAG 4 (e) DAG 5

Figure 2. Mean global accuracy of BR, LP and ILF-Compo on each DAG w.r.t. the training size in logarithmic scale. Results are averaged over 1000 random distributions.

nal benchmark datasets was not significant according to a Wilcoxon's signed rank test at a confidence level of 0.01: this is pretty much a dead heat between the 2 algorithms. However, ILF-Compo did significantly better on the duplicated benchmarks at $p < 0.01$.

Not shown here due to space restrictions, the graph structures obtained with ILF-Compo lend themselves naturally to interpretation. Several graphs $\mathcal{G}$ are densely connected, like Bibtex or Corel5K, while others are surprisingly sparse, like Genebase and Medical. Finally, graphical models have a number of advantages over alternative methods. They clearly lay bare useful information about the label dependencies which is crucial if one is interested in gaining an understanding of underlying domain. This is however well beyond the scope of this paper to delve deeper into the graph interpretation.

5. Discussion & Conclusion

In this paper, the multi-label classification and optimal feature subset selection problems under the subset $0/1$ loss were formulated within a unified probabilistic framework for a subclass of distributions satisfying the Composition property. This framework paves the way for the development of a broad class of correct MLC procedures optimal under the subset $0/1$ loss. A straightforward instantiation was proposed and evaluated on synthetic and real-world data. Significant improvements over LP were obtained with label conditional distributions exhibiting several irreducible factors.

Admittedly, the subset $0/1$ loss may appear overly stringent in practical applications. Finding theoretically correct algorithms for other non label-wise decomposable loss functions is still a great challenge. Nevertheless, we hope that our paper will convince others about the importance of label factor decomposition and the possibility to work with other loss functions.

Finally, as we usually have no idea whether the Composition axiom holds in the probability distribution underlying the data at hand (it is violated for instance when complex interactions exists between variables, such as the noisy parity problem), future work should aim to relax this assumption.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work is funded by the European Community through the European Nanoelectronics Initiative Advisory Council (ENIAC Joint Undertaking), under grant agreement no 324271 (ENI.237.1.B2013).

References

Bielza, C., Li, G., and Larrañaga, P. Multi-dimensional classification with bayesian networks. *International Journal of Approximate Reasoning*, 52(6):705–727, 2011.

Blockeel, H., Raedt, L. De, and Ramon, J. Top-down induction of clustering trees. In *ICML*, pp. 55–63, 1998.

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Cherman, E. Alvares, Metz, J., and Monard, M.C. Incorporating label dependency into the binary relevance framework for multi-label classification. *Expert Systems With Applications*, 39(2):1647–1655, 2012.

Dembczynski, K., Waegeman, W., Cheng, W., and Hüllermeier, E. An exact algorithm for f-measure maximization. In *NIPS*, pp. 1404–1412, 2011.

Dembczynski, K., Waegeman, W., Cheng, W., and Hüllermeier, E. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2): 5–45, 2012.

Gasse, M., Aussem, A., and Elghazel, H. A hybrid algorithm for bayesian network structure learning with application to multi-label learning. *Expert Systems with Applications*, 41(15):6755–6772, 2014.

Gharroudi, O., Elghazel, H., and Aussem, A. A comparison of multi-label feature selection methods using the random forest paradigm. In *AI*, pp. 95–106, 2014.

Guo, Y. and Gu, S. Multi-label classification using conditional dependency networks. In *IJCAI*, pp. 1300–1305, 2011.

Kocev, D., Vens, C., Struyf, J., and Dzeroski, S. Ensembles of multi-objective decision trees. In *ECML*, pp. 624–631, 2007.

Lee, J.S. and Kim, D.W. Feature selection for multi-label classification using multivariate mutual information. *Pattern Recognition Letters*, 34(3):349–357, 2013.

Liaw, A. and Wiener, M. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

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

Madjarov, G., Kocev, D., Gjorgjevikj, D., and Dzeroski, S. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104, 2012.

Maron, O. and Ratan, A.L. Multiple-instance learning for natural scene classification. In *ICML*, pp. 341–349, 1998.

Pearl, J. *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1989.

Peña, J. M., Nilsson, R., Björkegren, J., and Tegnér, J. Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning*, 45(2):211–232, 2007.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. Classifier chains for multi-label classification. In *ECML/PKDD*, volume 5782, pp. 254–269, 2009.

Scutari, M. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

Scutari, M. and Brogini, A. Bayesian Network Structure Learning with Permutation Tests. *Communications in Statistics - Theory and Methods*, 41(16-17), 2012.

Smith, N.A. and Tromble, R.W. Sampling Uniformly from the Unit Simplex. Technical report, Johns Hopkins University, 2004.

Spolaôr, N., Cherman, E. Alvares, Monard, M.C., and Lee, H.D. A comparison of multi-label feature selection methods using the problem transformation approach. *Electronic Notes in Theoretical Computer Science*, 292:135–151, 2013.

Statnikov, A.R., Lemeire, J., and Aliferis, C.F. Algorithms for discovery of multiple markov boundaries. *Journal of Machine Learning Research*, 14(1):499–566, 2013.

Tsamardinos, I. and Borboudakis, G. Permutation testing improves bayesian network learning. In *ECML/PKDD (3)*, pp. 322–337, 2010.

Tsoumakas, G. and Vlahavas, I.P. Random k -labelsets: An ensemble method for multilabel classification. In *ECML*, pp. 406–417, 2007.

Tsoumakas, G., Katakis, I., and Vlahavas, I.P. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079–1089, 2011.

Zhang, M.L. and Zhang, K. Multi-label learning by exploiting label dependency. In *KDD*, pp. 999–1008, 2010.