International Workshop on Big Multi-Target Prediction manuscript No.
(will be inserted by the editor)

On the Factorization of the Label Conditional Distribution in the context of Multi-Label Classification

Maxime Gasse · Alex Aussem · Haytham Elghazel

the date of receipt and acceptance should be inserted later

Abstract There is a consensus among researchers that, to improve the performance of multi-label learning algorithms, label dependencies have to be incorporated into the learning process. However, the benefit of exploiting label dependence in multi-label classification (MLC) is known to be closely dependent on the type of loss to be minimized. In this study, we show that identifying the irreducible factors in the factorization of the conditional distribution of the label set given the input features can play a pivotal role for MLC in the context of zero-one loss minimization, as it divides the learning task into simpler independent problems. We establish theoretical results to characterize and identify the irreducible label factors under various assumptions about the underlying probability distribution (i.e., Composition, Intersection, DAG-Faithfulness), which lays the foundation for practical irreducible label factor decomposition procedures for these subclasses of distributions. This discussion extends prior works published in ESWA[1] and recently presented at ICML[2].

Keywords multi-label learning · probabilistic graphical models · Markov boundary discovery · subset zero-one loss

Contribution summary

From a Bayesian point of view, multi-label learning amounts to modeling the conditional joint distribution $p(\mathbf{Y} \mid \mathbf{X})$. The key question is: what shall we capture from $p(\mathbf{Y} \mid \mathbf{X})$ exactly to solve the MLC problem? In a recent paper, Dembczynsk et al. showed that the expected benefit of exploiting label dependence depends 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. In particular, minimizing the *subset* $0/1$ *loss*, the *F-measure loss*

Maxime Gasse, Alex Aussem and Haytham Elghazel
Université Lyon 1, LIRIS, UMR5205, F-69622, France
E-mail: {maxime.gasse,alexandre.aussem,haytham.elghazel}@liris.cnrs.fr

[1] M. Gasse et al. (2014). A hybrid algorithm for Bayesian network structure learning with application to multi-label learning. *Expert Systems with Applications* 41.15, pp. 6755–6772.

[2] M. Gasse et al. (2015). On the Optimality of Multi-Label Classification under Subset Zero-One Loss for Distributions Satisfying the Composition Property. *ICML*. Ed. by F. R. Bach and D. M. Blei. Vol. 37. JMLR Proceedings. JMLR.org, pp. 2531–2539.

or the *Jaccard index* requires the modeling of the joint distribution (at least to some extent). In this study, we are mainly concerned with risk-minimizing prediction for the subset $0/1$ loss. We establish several theorems, under the assumption that the probability distribution satisfies the Composition, the Intersection or the DAG-Faithfulness properties, to characterize the so-called irreducible label factors (ILFs), that appear as (unique) irreducible factors 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 $\mathbf{Y}_{LF} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_{LF} \mid \mathbf{X}$). The ILF characterization for any distribution p follows,

Theorem 0.1. *Let $\mathscr{G}$ be an undirected graph whose nodes correspond to the random variables in $\mathbf{Y}$ and in which two nodes Y_i and Y_j are adjacent iff there exists $\mathbf{Z} \subseteq \mathbf{Y} \setminus \{Y_i, Y_j\}$ such that $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid (\mathbf{X} \cup \mathbf{Z})$. Then, two labels Y_i and Y_j belong to the same irreducible label factor iff a path exists between Y_i and Y_j in $\mathscr{G}$.*

In view of this result, the process of deciding upon whether $\exists \mathbf{Z} \subseteq \mathbf{Y} \setminus \{Y_i, Y_j\}$ such that $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid (\mathbf{X} \cup \mathbf{Z})$ is a combinatorial problem and may be challenging as the number of possible combinations for $\mathbf{Z}$ grows very quickly with the number of labels. In addition, performing a statistical test of independence conditioned on $\mathbf{X} \cup \mathbf{Z}$ may become problematic in discrete data, where the sample size required for high-confidence grows exponentially in the size of the conditioning set. Fortunately, we may derive more convenient characterizations if we assume the Intersection, the Composition or the DAG-Faithfulness property (see paper for further details). Examples of decomposition graphs learned from multi-label data sets are shown in Figure 1.

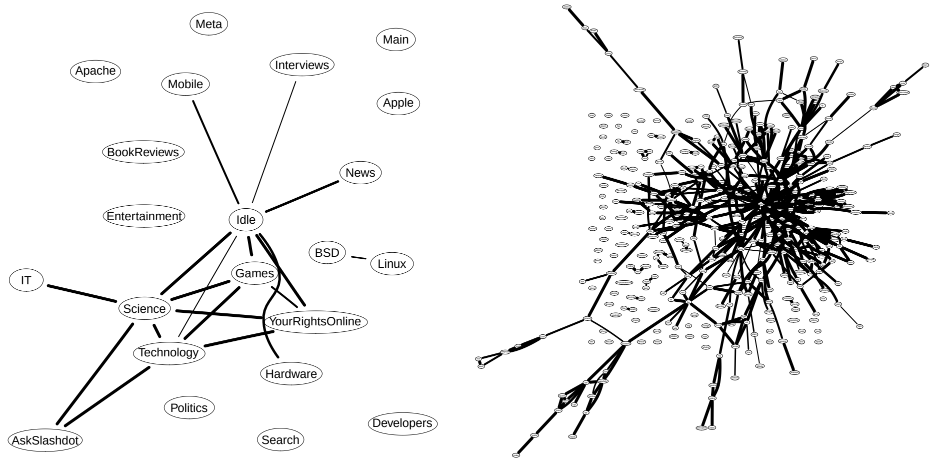

Fig. 1 Label decomposition graphs $\mathscr{G}$ obtained on Slashdot and Corel5 for illustration purposes.

In conclusion, the present analysis prepares the ground for a generic class of MLC decomposition procedures - that may be implemented in different manners - which are correct under a given set of assumptions underlying the joint distribution (e.g. Composition, Intersection, DAG-Faithfulness properties). Our experimental results demonstrate the usefulness of the ILF decomposition for the MLC problem under subset $0/1$ loss.