On the Optimality of Multi-Label Classification under Subset 0/1 Loss (assuming the Composition property)

International Conference on Machine Learning

Maxime Gasse Alex Aussem Haytham Elghazel

LIRIS Laboratory, UMR 5205 CNRS
University of Lyon 1, France

July 9, 2015

Outline

- Unified probabilistic framework
  - Multi-label classification and loss functions
  - Subset 0/1 loss minimization
- Factorization of the joint conditional distribution of the labels
  - Irreducible label factors
  - The ILF-Compo algorithm
- Experimental results
  - Toy problem
  - Benchmark data sets

Multi-label classification

Find a mapping $\mathbf{h}$ from a space of features $\mathbf{X}$ to a space of labels $\mathbf{Y}$

$$\mathbf{x} \in \mathbb{R}^d, \ \mathbf{y} \in \{0,1\}^n, \ \mathbf{h}\colon \mathbf{X} \to \mathbf{Y}$$

Multi-label classification

Find a mapping $\mathbf{h}$ from a space of features $\mathbf{X}$ to a space of labels $\mathbf{Y}$

$$\mathbf{x} \in \mathbb{R}^d, \ \mathbf{y} \in \{0,1\}^n, \ \mathbf{h}\colon \mathbf{X} \to \mathbf{Y}$$

Example

- To which categories does a movie belong? (war, fiction...)
- Which objects are present on a picture? (cat, human, tree...)
- Which emotions does a music trigger? (love, anger...)

Multi-label classification

Find a mapping $\mathbf{h}$ from a space of features $\mathbf{X}$ to a space of labels $\mathbf{Y}$

$$\mathbf{x} \in \mathbb{R}^d, \ \mathbf{y} \in \{0,1\}^n, \ \mathbf{h}\colon \mathbf{X} \to \mathbf{Y}$$

Example

- To which categories does a movie belong? (war, fiction...)
- Which objects are present on a picture? (cat, human, tree...)
- Which emotions does a music trigger? (love, anger...)

Any classification problem with non-mutually exclusive classes can be formulated as a multi-label learning problem.

Probabilistic framework

The risk-minimizing model $\mathbf{h}^\star$ with respect to a loss function L is defined over $p(\mathbf{X}, \mathbf{Y})$ as

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

Probabilistic framework

The risk-minimizing model $\mathbf{h}^\star$ with respect to a loss function L is defined over $p(\mathbf{X}, \mathbf{Y})$ as

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

The point-wise best prediction requires only $p(\mathbf{Y} \mid \mathbf{X})$

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

Probabilistic framework

The risk-minimizing model $\mathbf{h}^\star$ with respect to a loss function L is defined over $p(\mathbf{X}, \mathbf{Y})$ as

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

The point-wise best prediction requires only $p(\mathbf{Y} \mid \mathbf{X})$

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

The current trend is to exploit label dependence to improve MLC. However, Dembczynski et al. (2012) observe that "a concrete connection between the type of multi-label classifier used and the loss to be minimized is rarely established, implicitly giving the misleading impression that the same method can be optimal for different loss functions."

Hamming loss: a decomposable loss function over $\mathbf{Y}$

Some loss function do not necessarily require the estimation of $p(\mathbf{Y} \mid \mathbf{X})$. One such function is the **Hamming loss**

$$L_H(\mathbf{y}, \mathbf{h}(\mathbf{x})) = 1/n \sum_{i=1}^{n} \mathbf{1}(y_i \neq h_i(\mathbf{x}))$$

Hamming loss: a decomposable loss function over $\mathbf{Y}$

Some loss function do not necessarily require the estimation of $p(\mathbf{Y} \mid \mathbf{X})$. One such function is the **Hamming loss**

$$L_H(\mathbf{y}, \mathbf{h}(\mathbf{x})) = 1/n \sum_{i=1}^{n} \mathbf{1}(y_i \neq h_i(\mathbf{x}))$$

Here the risk-minimizing prediction is given by the mode of each marginal distribution $p(Y_i \mid \mathbf{X})$

$$\mathbf{h}_H^\star(\mathbf{x}) = \bigcup_{i=1}^{n} \arg\max_{y_i} p(y_i \mid \mathbf{x})$$

This approach is called **Binary Relevance** (BR).

Subset 0/1 loss: a non-decomposable loss function over $\mathbf{Y}$

Other loss functions require the full modeling of $p(\mathbf{Y} \mid \mathbf{X})$. One such function is the **Subset** $0/1$ **loss**

$$L_S(\mathbf{y}, \mathbf{h}(\mathbf{x})) = \mathbf{1}(\mathbf{y} \neq \mathbf{h}(\mathbf{x}))$$

Subset 0/1 loss: a non-decomposable loss function over $\mathbf{Y}$

Other loss functions require the full modeling of $p(\mathbf{Y} \mid \mathbf{X})$. One such function is the **Subset** $0/1$ **loss**

$$L_S(\mathbf{y}, \mathbf{h}(\mathbf{x})) = \mathbf{1}(\mathbf{y} \neq \mathbf{h}(\mathbf{x}))$$

Here the risk-minimizing prediction is given by the mode of the joint conditional distribution $p(\mathbf{Y} \mid \mathbf{X})$

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

This approach is called **Label Powerset** (LP).

Factorization of the joint conditional distribution

Depending on the dependency structure between the labels and the features, the problem of modeling the joint conditional distribution may actually be decomposed into a product of label factors

$$p(\mathbf{Y} \mid \mathbf{X}) = \prod_{\mathbf{Y}_{LF} \in \mathcal{P}_{\mathbf{Y}}} p(\mathbf{Y}_{LF} \mid \mathbf{X}),$$

$$\arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}) = \bigcup_{\mathbf{Y}_{LF} \in \mathcal{P}_{\mathbf{Y}}} \arg\max_{\mathbf{y}} p(\mathbf{y}_{LF} \mid \mathbf{x}),$$

with $\mathcal{P}_{\mathbf{Y}}$ a partition of $\mathbf{Y}$.

Factorization of the joint conditional distribution

Depending on the dependency structure between the labels and the features, the problem of modeling the joint conditional distribution may actually be decomposed into a product of label factors

$$p(\mathbf{Y} \mid \mathbf{X}) = \prod_{\mathbf{Y}_{LF} \in \mathcal{P}_{\mathbf{Y}}} p(\mathbf{Y}_{LF} \mid \mathbf{X}),$$

$$\arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}) = \bigcup_{\mathbf{Y}_{LF} \in \mathcal{P}_{\mathbf{Y}}} \arg\max_{\mathbf{y}} p(\mathbf{y}_{LF} \mid \mathbf{x}),$$

with $\mathcal{P}_{\mathbf{Y}}$ a partition of $\mathbf{Y}$.

Definition

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 *iff* none of its non-empty proper subsets is a label factor.

Factorization of the joint conditional distribution

Depending on the dependency structure between the labels and the features, the problem of modeling the joint conditional distribution may actually be decomposed into a product of label factors

$$p(\mathbf{Y} \mid \mathbf{X}) = \prod_{\mathbf{Y}_{LF} \in \mathcal{P}_{\mathbf{Y}}} p(\mathbf{Y}_{LF} \mid \mathbf{X}),$$

$$\arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}) = \bigcup_{\mathbf{Y}_{LF} \in \mathcal{P}_{\mathbf{Y}}} \arg\max_{\mathbf{y}} p(\mathbf{y}_{LF} \mid \mathbf{x}),$$

with $\mathcal{P}_{\mathbf{Y}}$ a partition of $\mathbf{Y}$.

Definition

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 *iff* none of its non-empty proper subsets is a label factor.

We seek a factorization into irreducible label factors **ILF**.

Graphical characterization

Theorem

Let $\mathcal{G}$ 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 $\exists \mathbf{Z} \subseteq \mathbf{Y} \setminus \{Y_i, Y_j\}$ 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 $\mathcal{G}$.

Graphical characterization

Theorem

Let $\mathcal{G}$ 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 $\exists \mathbf{Z} \subseteq \mathbf{Y} \setminus \{Y_i, Y_j\}$ 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 $\mathcal{G}$.

$\mathcal{O}(n^2 2^n)$ pairwise tests of conditional independence to characterize the irreducible label factors.

Graphical characterization

Theorem

Let $\mathcal{G}$ 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 $\exists \mathbf{Z} \subseteq \mathbf{Y} \setminus \{Y_i, Y_j\}$ 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 $\mathcal{G}$.

$\mathcal{O}(n^2 2^n)$ pairwise tests of conditional independence to characterize the irreducible label factors.

Much easier if we assume the Composition property.

The Composition property

The dependency of a whole implies the dependency of some part

$$\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \;\Rightarrow\; \mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \;\vee\; \mathbf{X} \not\perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z}$$

The Composition property

The dependency of a whole implies the dependency of some part

$$\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \;\Rightarrow\; \mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \;\vee\; \mathbf{X} \not\perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z}$$

Weak assumption: several existing methods and algorithms assume the Composition property (e.g. forward feature selection).

The Composition property

The dependency of a whole implies the dependency of some part

$$\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \;\Rightarrow\; \mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z} \;\vee\; \mathbf{X} \not\perp\!\!\!\perp \mathbf{W} \mid \mathbf{Z}$$

Weak assumption: several existing methods and algorithms assume the Composition property (e.g. forward feature selection).

Counter-example

Exclusive OR relationships break the Composition property,

$$A = B \oplus C \;\Rightarrow\; \{A\} \not\perp\!\!\!\perp \{B, C\} \;\wedge\; \{A\} \perp\!\!\!\perp \{B\} \;\wedge\; \{A\} \perp\!\!\!\perp \{C\}$$

Graphical characterization - assuming Composition

Theorem

Suppose p supports the Composition property. Let $\mathcal{G}$ 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 $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid \mathbf{X}$. 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 $\mathcal{G}$.

Graphical characterization - assuming Composition

Theorem

Suppose p supports the Composition property. Let $\mathcal{G}$ 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 $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid \mathbf{X}$. 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 $\mathcal{G}$.

$\mathcal{O}(n^2)$ pairwise tests only. Moreover,

Graphical characterization - assuming Composition

Theorem

Suppose p supports the Composition property. Let $\mathcal{G}$ 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 $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid \mathbf{X}$. 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 $\mathcal{G}$.

$\mathcal{O}(n^2)$ pairwise tests only. Moreover,

Theorem

Suppose p supports the Composition property and consider $\mathbf{M}_i$ an arbitrary Markov blanket of Y_i in $\mathbf{X}$. Then, $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid \mathbf{X}$ is true iff $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid \mathbf{M}_i$.

ILF-Compo algorithm

The generic procedure goes as follows

- For each label Y_i compute $\mathbf{M}_i$ a Markov boundary in $\mathbf{X}$.
- For each pair of labels (Y_i, Y_j) check $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid \mathbf{M}_i$.
- Build the decomposition graph $\mathcal{G}$.
- Extract the partition $\mathbf{ILF} = \{\mathbf{Y}_{LF_1}, \ldots, \mathbf{Y}_{LF_L}\}$ from $\mathcal{G}$.
- Decompose the multi-label problem into a series of independent multi-class problems.

ILF-Compo algorithm

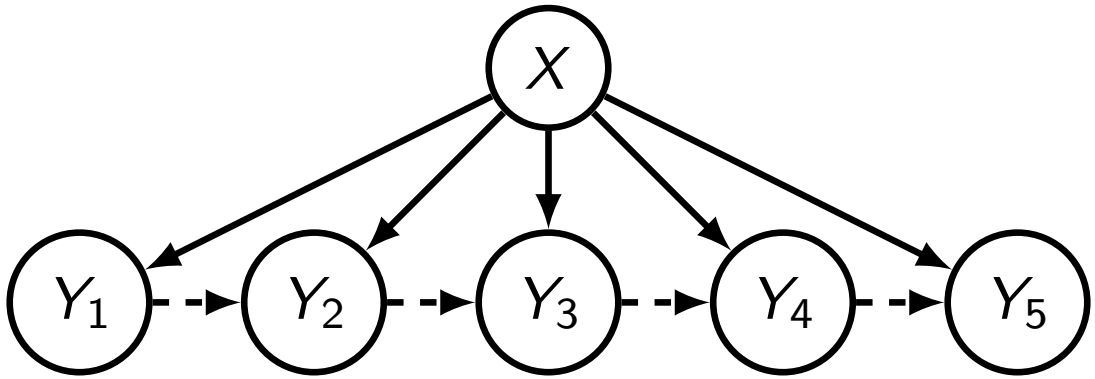
The generic procedure goes as follows

- For each label Y_i compute $\mathbf{M}_i$ a Markov boundary in $\mathbf{X}$.
- For each pair of labels (Y_i, Y_j) check $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid \mathbf{M}_i$.
- Build the decomposition graph $\mathcal{G}$.
- Extract the partition $\mathbf{ILF} = \{\mathbf{Y}_{LF_1}, \ldots, \mathbf{Y}_{LF_L}\}$ from $\mathcal{G}$.
- Decompose the multi-label problem into a series of independent multi-class problems.

In the experiments we implement ILF-Compo with

- IAMB a constraint-based Markov boundary learning algorithm (Tsamardinos, Aliferis, and Statnikov 2003);
- Mutual Information (MI)-based test of independence;
- Random Forest classifier.

Experiment on toy problem



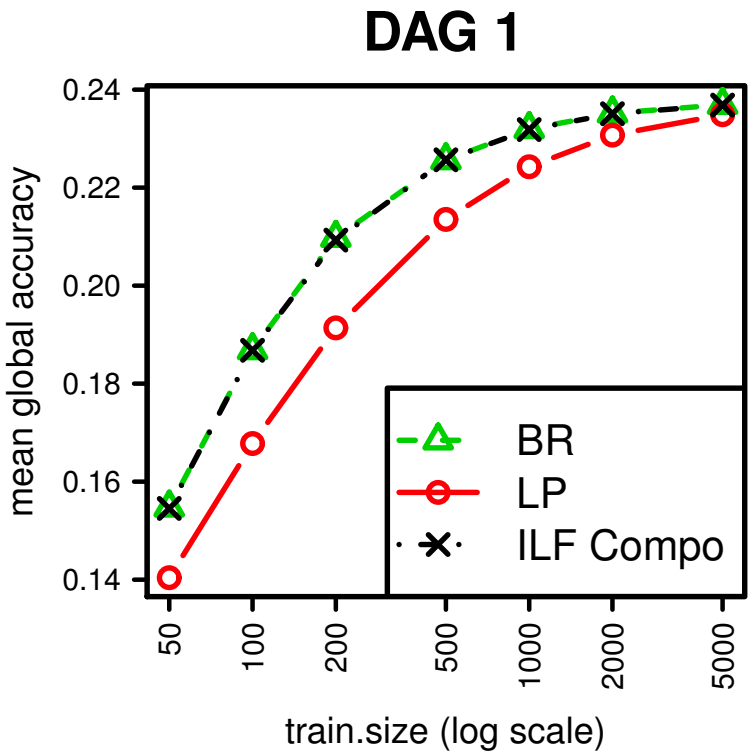
Generic toy DAG (Bayesian network).

We build 5 distinct irreducible factorizations:

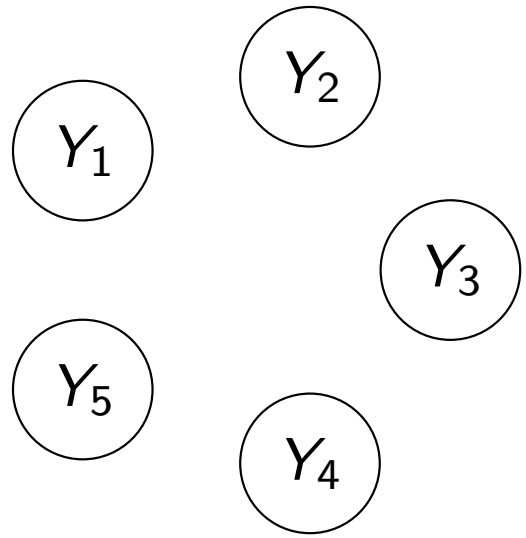
- DAG 1: **ILF** $= \{\{Y_1\}, \quad \{Y_2\}, \quad \{Y_3\}, \quad \{Y_4\}, \quad \{Y_5\}\}$;
- DAG 2: **ILF** $= \{\{Y_1, Y_2\}, \quad \{Y_3, Y_4\}, \quad \{Y_5\}\}$;
- DAG 3: **ILF** $= \{\{Y_1, Y_2, Y_3\}, \quad \{Y_4, Y_5\}\}$;
- DAG 4: **ILF** $= \{\{Y_1, Y_2, Y_3, Y_4\}, \quad \{Y_5\}\}$;
- DAG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

Experiment on toy problem

$$\textbf{ILF} = \{\{Y_1\}, \quad \{Y_2\}, \quad \{Y_3\}, \quad \{Y_4\}, \quad \{Y_5\}\}$$



Mean global accuracy[1] over 1000 random distributions.

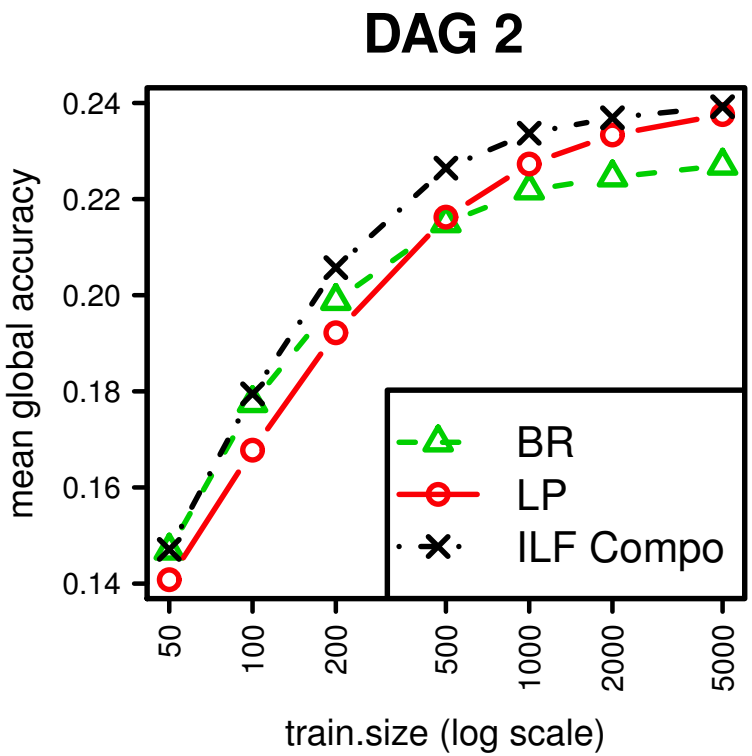


Decomposition graph.

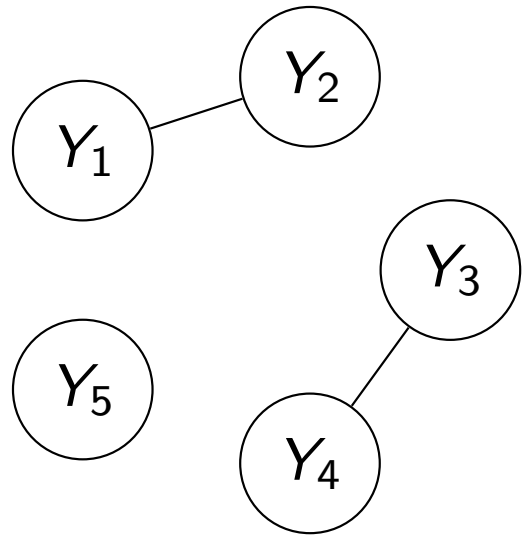
[1] 1 - subset 0/1 loss (higher is better)

Experiment on toy problem

$$\textbf{ILF} = \{\{Y_1, Y_2\}, \quad \{Y_3, Y_4\}, \quad \{Y_5\}\}$$



Mean global accuracy[1] over 1000 random distributions.

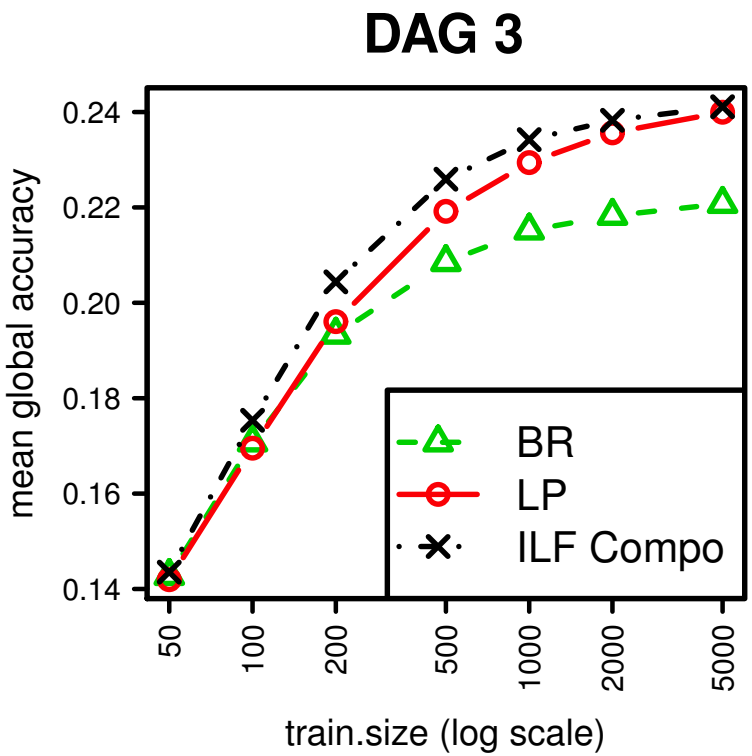


Decomposition graph.

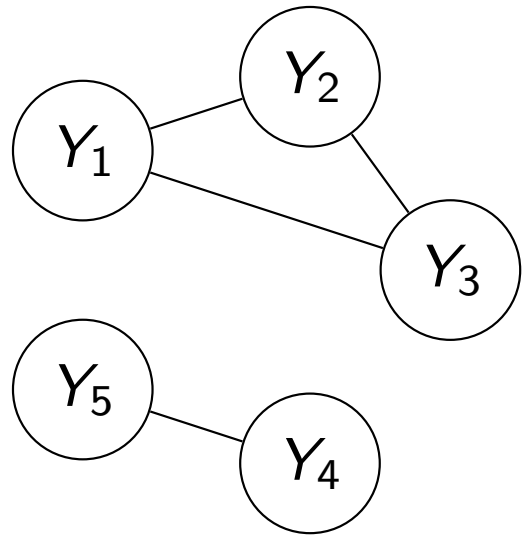
[1] 1 - subset 0/1 loss (higher is better)

Experiment on toy problem

$$\textbf{ILF} = \{\{Y_1, Y_2, Y_3\}, \quad \{Y_4, Y_5\}\}$$



Mean global accuracy[1] over 1000 random distributions.

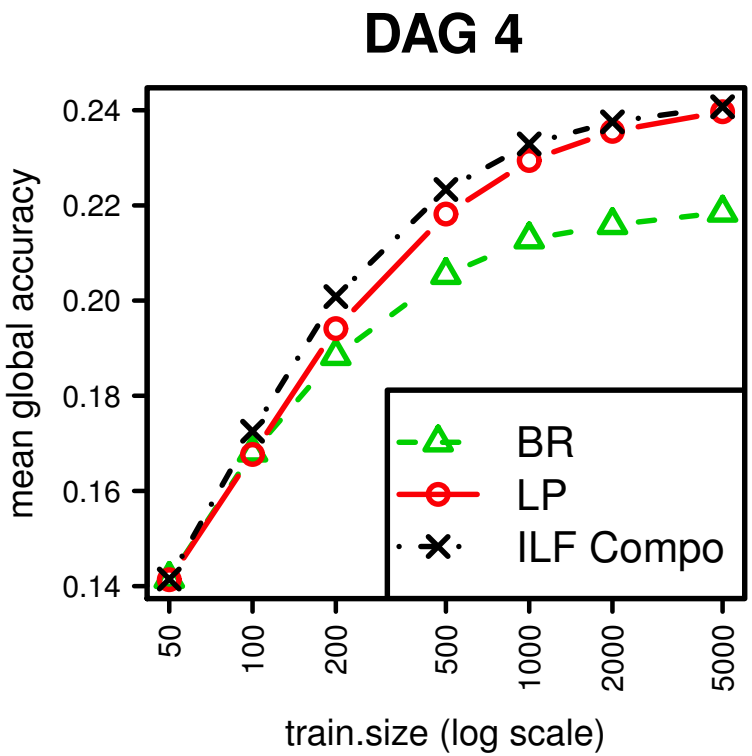


Decomposition graph.

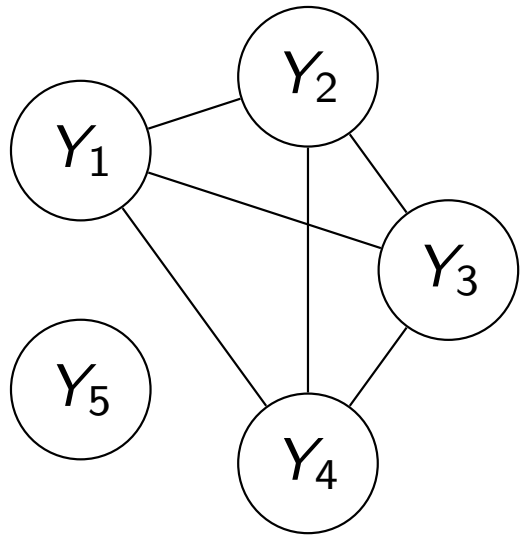
[1] 1 - subset 0/1 loss (higher is better)

Experiment on toy problem

$$\textbf{ILF} = \{\{Y_1, Y_2, Y_3, Y_4\}, \quad \{Y_5\}\}$$



Mean global accuracy[1] over 1000 random distributions.

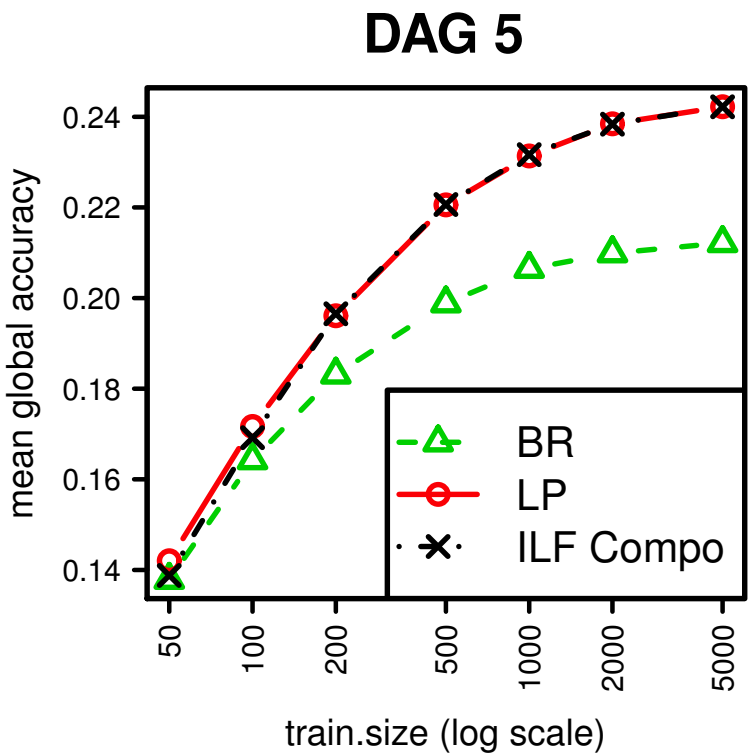


Decomposition graph.

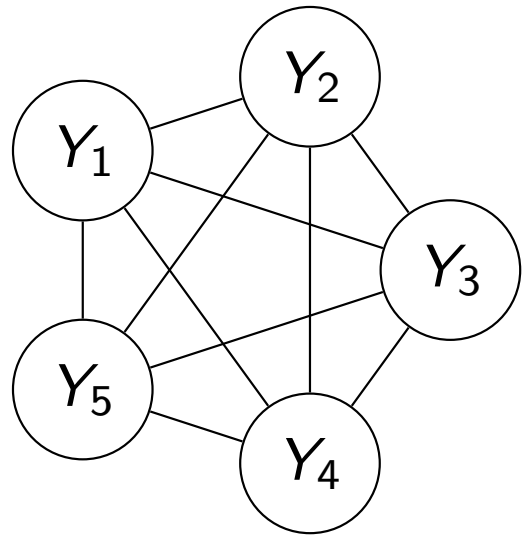
[1] 1 - subset 0/1 loss (higher is better)

Experiment on toy problem

$$\textbf{ILF} = \{\{Y_1, Y_2, Y_3, Y_4, Y_5\}\}$$



Mean global accuracy[1] over 1000 random distributions.



Decomposition graph.

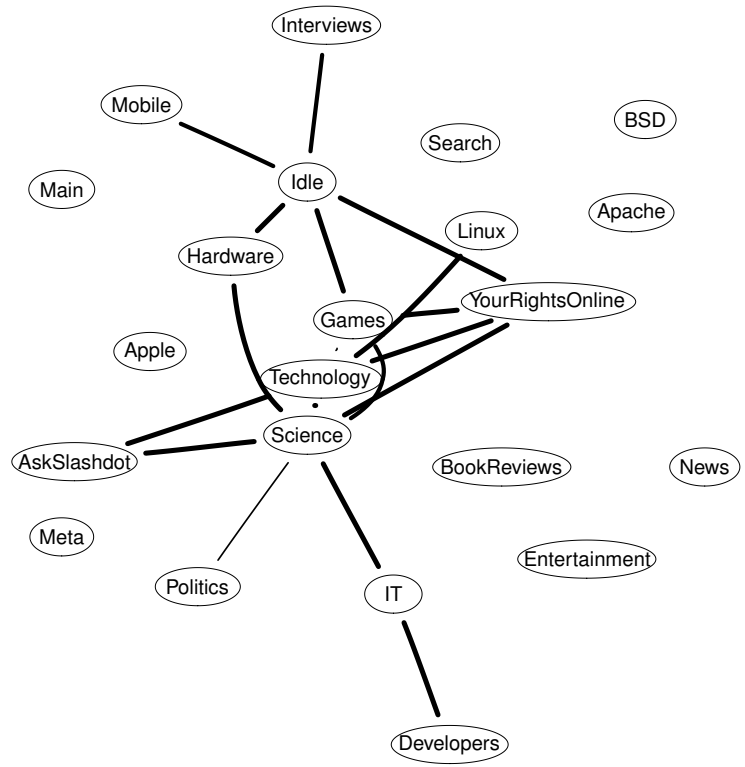
[1] 1 - subset 0/1 loss (higher is better)

Experiment on benchmark data sets

Mean global accuracy on the original benchmark, over 5x2-fold cross-validation.

Dataset	ILF-Compo	LP	BR
emotions	35.5	35.7	30.0
image	47.7	47.4	30.5
scene	73.3	73.8	54.1
yeast	26.1	26.4	15.5
slashdot	42.4	45.3	35.5
genbase	96.6	96.2	96.6
medical	65.5	**68.9**	62.5
enron	**16.0**	15.5	10.5
bibtex	13.8	**22.0**	11.6
corel5k	2.9	3.0	0.2

Not statistically different from LP.



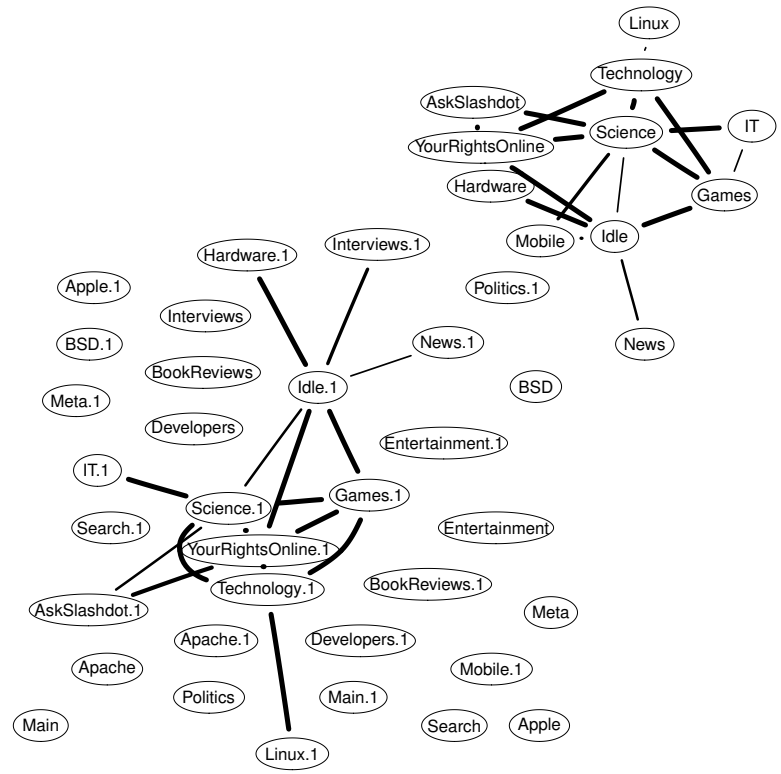
Decomposition obtained with ILF-Compo on slashdot.

Experiment on benchmark data sets - duplicated

We duplicate each data set and permute the rows on the duplicated variables. By design, the resulting data set contains at least two irreducible label factors.

Mean global accuracy on the duplicated benchmark, over 5x2-fold cross-validation.

Dataset	ILF-Compo	LP	BR
emotions2	**10.7**	4.8	6.0
image2	**21.0**	12.0	5.4
scene2	**50.3**	35.2	21.1
yeast2	**5.8**	2.3	1.5
slashdot2	**18.2**	8.9	10.2
genbase2	**93.1**	69.1	93.3
medical2	**27.8**	20.6	20.6
enron2	**2.5**	0.6	0.8
bibtex2	0.5	**0.8**	0.6
corel5k2	0.0	0.0	0.0



Decomposition obtained with ILF-Compo on slashdot2.

Conclusion

- The MLC problem under Subset 0/1 loss was formulated within a unified probabilistic framework.

Conclusion

- The MLC problem under Subset 0/1 loss was formulated within a unified probabilistic framework.
- An optimal factorization method was proposed for a subclass of distributions satisfying the Composition property.

Conclusion

- The MLC problem under Subset 0/1 loss was formulated within a unified probabilistic framework.
- An optimal factorization method was proposed for a subclass of distributions satisfying the Composition property.
- A straightforward instantiation showed that significant improvements can be obtained over LP when the conditional distribution of the labels exhibits several irreducible factors.

Conclusion

- The MLC problem under Subset 0/1 loss was formulated within a unified probabilistic framework.
- An optimal factorization method was proposed for a subclass of distributions satisfying the Composition property.
- A straightforward instantiation showed that significant improvements can be obtained over LP when the conditional distribution of the labels exhibits several irreducible factors.

Future work

- Relax the Composition property
- Exploit label dependence for other loss functions

On the Optimality of Multi-Label Classification under Subset 0/1 Loss (assuming the Composition property)

International Conference on Machine Learning

Maxime Gasse Alex Aussem Haytham Elghazel

Thank you!