Factorization of the Label Conditional Distribution for Multi-Label Classification

ECML PKDD 2015

International Workshop on Big Multi-Target Prediction

Maxime Gasse Alex Aussem Haytham Elghazel

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

September 11, 2015

Outline

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

This work was recently presented at ICML (Gasse, Aussem, and Elghazel 2015).

Unified probabilistic framework

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

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

Unified probabilistic framework

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

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

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

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

Unified probabilistic framework

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

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

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 $p(\mathbf{Y} \mid \mathbf{X})$

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

Unified probabilistic framework

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

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

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... under which loss function?

Hamming loss vs Subset 0/1 loss

Hamming loss

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

Subset 0/1 loss

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

Hamming loss vs Subset 0/1 loss

Hamming loss

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

BR (Binary Relevance) is optimal, with c parameters

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

Subset 0/1 loss

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

LP (Label Powerset) is optimal, with 2^c parameters

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

Hamming loss vs Subset 0/1 loss

Hamming loss

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

BR (Binary Relevance) is optimal, with c parameters

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

Subset 0/1 loss

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

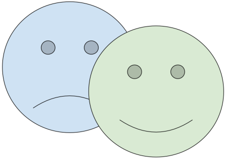
LP (Label Powerset) is optimal, with 2^c parameters

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

$p(\mathbf{Y} \mid \mathbf{x})$ much harder to estimate than $p(Y_i \mid \mathbf{x})$... can we use the label dependencies to better model $p(\mathbf{Y} \mid \mathbf{x})$?

Hamming loss vs Subset 0/1 loss

A quick example: who is in the picture?



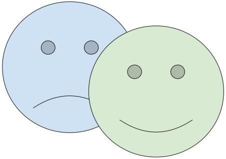
Jean	René	$p(J, R \mid \mathbf{x})$
0	0	0.02
0	1	0.10
1	0	0.13
1	1	0.75

HLoss optimal : $J = 1,\ R = 1\ (88\%, 85\%)$

SLoss optimal : $J = 1,\ R = 1\ (75\%)$

Hamming loss vs Subset 0/1 loss

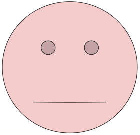
A quick example: who is in the picture?



Jean	René	$p(J, R \mid \mathbf{x})$
0	0	0.02
0	1	0.10
1	0	0.13
1	1	0.75

HLoss optimal : $J = 1,\ R = 1\ (88\%, 85\%)$

SLoss optimal : $J = 1,\ R = 1\ (75\%)$



Jean	René	$p(J, R \mid \mathbf{x})$
0	0	0.02
0	1	0.46
1	0	0.44
1	1	0.08

HLoss optimal : $J = 1,\ R = 1\ (52\%, 54\%)$

SLoss optimal : $J = 0,\ R = 1\ (46\%)$

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 (unique) 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 $\mathcal{G}$.

$\mathcal{O}(c^2 2^c)$ 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 $\mathcal{G}$.

$\mathcal{O}(c^2 2^c)$ 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).

Typical counter-example

The exclusive OR relationship,

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

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 $\mathcal{G}$.

$\mathcal{O}(c^2)$ 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 $\mathcal{G}$.

$\mathcal{O}(c^2)$ 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

Generic procedure

- 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 $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid \mathbf{M}_i$ to build $\mathcal{G}$.
- Extract the partition $\mathbf{ILF} = \{\mathbf{Y}_{LF_1}, \ldots, \mathbf{Y}_{LF_m}\}$ from $\mathcal{G}$.
- Decompose the multi-label problem into a series of independent multi-class problems.

ILF-Compo algorithm

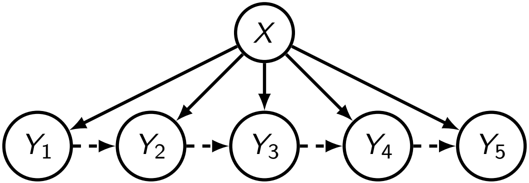
Generic procedure

- 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 $\{Y_i\} \not\perp\!\!\!\perp \{Y_j\} \mid \mathbf{M}_i$ to build $\mathcal{G}$.
- Extract the partition $\mathbf{ILF} = \{\mathbf{Y}_{LF_1}, \dots, \mathbf{Y}_{LF_m}\}$ from $\mathcal{G}$.
- Decompose the multi-label problem into a series of independent multi-class problems.

Experimental setup

- IAMB a constraint-based Markov boundary learning algorithm (Tsamardinos, Aliferis, and Statnikov 2003);
- Mutual Information-based test of independence ($\alpha = 10^{-3}$) (Tsamardinos and Borboudakis 2010);
- Random Forest classifier.

Experiment on toy problem



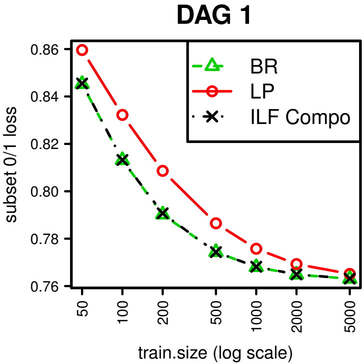
Generic toy DAG (Bayesian network).

We build 5 distinct irreducible factorizations:

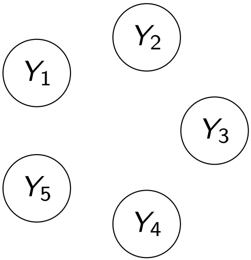
- DAG 1: $\mathbf{ILF} = \{\{Y_1\}, \quad \{Y_2\}, \quad \{Y_3\}, \quad \{Y_4\}, \quad \{Y_5\}\}$;
- DAG 2: $\mathbf{ILF} = \{\{Y_1, Y_2\}, \quad \{Y_3, Y_4\}, \quad \{Y_5\}\}$;
- DAG 3: $\mathbf{ILF} = \{\{Y_1, Y_2, Y_3\}, \quad \{Y_4, Y_5\}\}$;
- DAG 4: $\mathbf{ILF} = \{\{Y_1, Y_2, Y_3, Y_4\}, \quad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

Experiment on toy problem

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



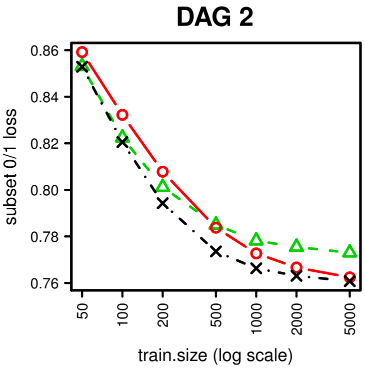
Subset 0/1 loss over 1000 random distributions.



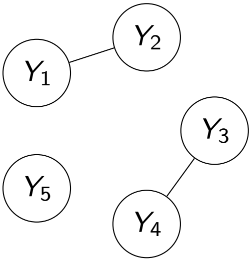
Decomposition graph.

Experiment on toy problem

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



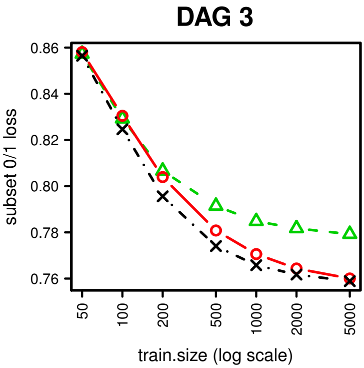
Subset 0/1 loss over 1000 random distributions.



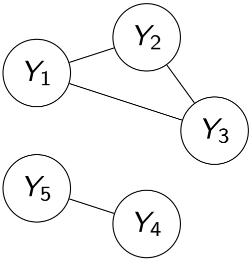
Decomposition graph.

Experiment on toy problem

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



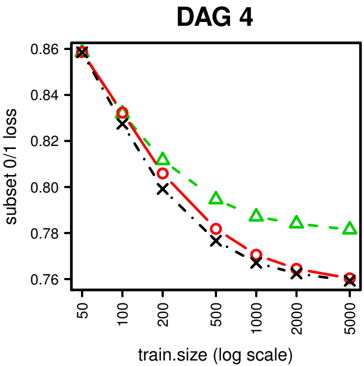
Subset 0/1 loss over 1000 random distributions.



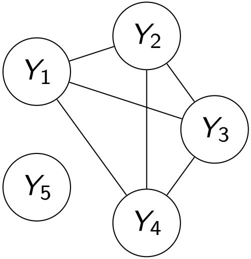
Decomposition graph.

Experiment on toy problem

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



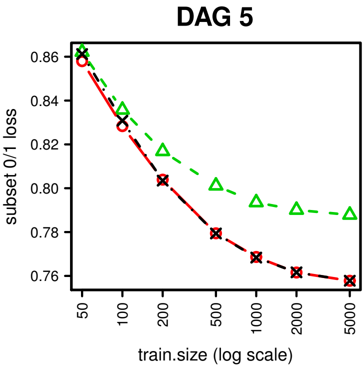
Subset 0/1 loss over 1000 random distributions.



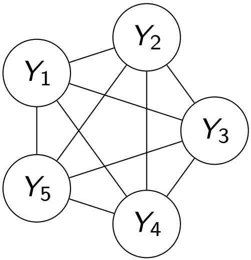
Decomposition graph.

Experiment on toy problem

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



Subset 0/1 loss over 1000 random distributions.



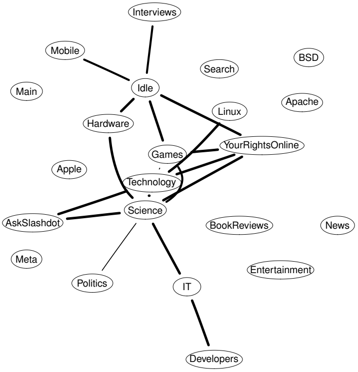
Decomposition graph.

Experiment on benchmark data sets

Mean Subset 0/1 loss on the original benchmark (5x2 CV).

Dataset	ILF-Compo	LP	BR
emotions	64.5	64.3	70.0
image	52.3	52.6	69.5
scene	36.7	36.2	45.9
yeast	73.9	73.6	84.5
slashdot	57.6	54.7	64.5
genbase	3.4	3.8	3.4
medical	34.5	**31.1**	37.5
enron	**84.0**	84.5	89.5
bibtex	86.2	**78.0**	88.4
corel5k	97.1	97.0	99.8

Not statistically different from LP.



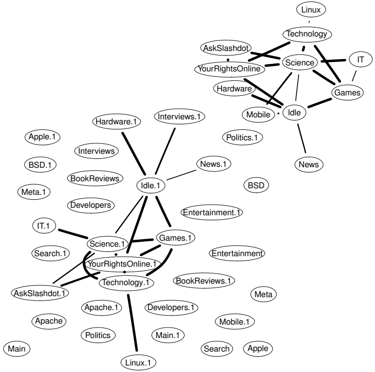
Decomposition obtained with ILF-Compo on slashdot.

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 Subset 0/1 loss on the duplicated benchmark (5x2 CV).

Dataset	ILF-Compo	LP	BR
emotions2	**89.3**	95.2	94.0
image2	**79.0**	88.0	94.6
scene2	**49.7**	64.8	78.9
yeast2	**94.2**	97.7	98.5
slashdot2	**81.8**	91.1	89.8
genbase2	**6.9**	30.9	6.7
medical2	**72.2**	79.4	79.4
enron2	**97.5**	99.4	99.2
bibtex2	99.5	**99.2**	99.4
corel5k2	99.9	99.9	99.9



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 conditional label dependence for other loss functions

Factorization of the Label Conditional Distribution for Multi-Label Classification

ECML PKDD 2015
International Workshop on Big Multi-Target Prediction

Maxime Gasse Alex Aussem Haytham Elghazel

Thank you!