Information theoretic feature selection in multi-label data through composite likelihood

Konstantinos Sechidis, Nikolaos Nikolaou, and Gavin Brown

School of Computer Science
University of Manchester
Motivation: Multi-label Learning

- Multi-label: Each datapoint can be associated to $>1$ labels
Motivation: Multi-label Learning

- Multi-label: Each datapoint can be associated to $> 1$ labels
- Applications

- Bioinformatics: 1 gene/protein, many functions
- Text Mining: 1 webpage/document, many categories
- Image Retrieval: 1 image, many semantic concepts

Common characteristic of these domains: Large number of features
Motivation: Multi-label Learning

- Multi-label: Each datapoint can be associated to $>1$ labels
- Applications
  - Bioinformatics: 1 gene/protein, many functions
Motivation: Multi-label Learning

- Multi-label: Each datapoint can be associated to $>1$ labels
- Applications
  - Bioinformatics: 1 gene/protein, many functions
  - Text Mining: 1 webpage/document, many categories
Motivation: Multi-label Learning

- Multi-label: Each datapoint can be associated to $>1$ labels
- Applications
  - Bioinformatics: 1 gene/protein, many functions
  - Text Mining: 1 webpage/document, many categories
  - Image Retrieval: 1 image, many semantic concepts

Male, Person, Motorbike, Vehicle Building
Female, Person, Building
Male, Person
Rabbit, Animal Car, Vehicle
Motivation: Multi-label Learning

- **Multi-label:** Each datapoint can be associated to > 1 labels

- **Applications**
  - Bioinformatics: 1 gene/protein, many functions
  - Text Mining: 1 webpage/document, many categories
  - Image Retrieval: 1 image, many semantic concepts

- **Common characteristic of these domains:** Large number of features
Feature Selection: Find minimal subset of features with maximal useful information
Feature Selection

- Feature Selection: Find minimal subset of features with maximal useful information

- Filters: Functions that assign a “utility” score to each feature
Feature Selection

- Feature Selection: Find minimal subset of features with maximal useful information

- Filters: Functions that assign a “utility” score to each feature

- In this work we discuss information-theoretic filters
Feature Selection

- Feature Selection: Find minimal subset of features with maximal useful information

- Filters: Functions that assign a “utility” score to each feature

- In this work we discuss information-theoretic filters

- Filter Assumption: model and feature selection are independent
Brown et al. (JMLR 2012) unified many heuristic information-theoretic filter criteria for feature selection.

Conditional Likelihood Maximization under model

\[
\begin{align*}
X & \rightarrow Y \\
\tau, \theta & \\
\end{align*}
\]
Brown et al. (JMLR 2012) unified many heuristic information-theoretic filter criteria for feature selection.

Conditional Likelihood Maximization under model

\[ \begin{align*}
    X &\quad \tau, \theta \\
    \downarrow & \\
    Y &
\end{align*} \]

Negative log-likelihood asymptotically decomposes into 3 terms:

\[ \lim_{N \to \infty} -\ell = \text{model term} + \text{feature selection term} + \text{Bayes error} \]

Feature selection is mutual info \( I(X_\theta; Y) \)
Single-label Feature Selection Criteria

Feature space independence assumptions:
Full:

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td></td>
</tr>
<tr>
<td>$X_2$</td>
<td></td>
</tr>
<tr>
<td>$X_3$</td>
<td>$Y$</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>$X_d$</td>
<td></td>
</tr>
</tbody>
</table>
Single-label Feature Selection Criteria

Feature space independence assumptions:
Full:

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td></td>
</tr>
<tr>
<td>$X_2$</td>
<td></td>
</tr>
<tr>
<td>$X_3$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
</tr>
<tr>
<td>$X_d$</td>
<td>$Y$</td>
</tr>
</tbody>
</table>
Single-label Feature Selection Criteria

Feature space independence assumptions:
Full:

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td></td>
</tr>
<tr>
<td>$X_2$</td>
<td>$Y$</td>
</tr>
<tr>
<td>$X_3$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
</tr>
<tr>
<td>$X_d$</td>
<td></td>
</tr>
</tbody>
</table>
Single-label Feature Selection Criteria

Feature space independence assumptions:
Full:

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td></td>
</tr>
<tr>
<td>$X_2$</td>
<td></td>
</tr>
<tr>
<td>$X_3$</td>
<td>$Y$</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>$X_d$</td>
<td></td>
</tr>
</tbody>
</table>
Feature space independence assumptions:
Full:

\[
\begin{array}{cc}
\text{Features} & \text{Label} \\
X_1 & \text{X} \\
X_2 & \text{X} \\
X_3 & \text{X} \\
\vdots & \vdots \\
X_d & \text{X} \\
\end{array}
\]

\[J(M) = I(X; Y)\]
Single-label Feature Selection Criteria

Feature space independence assumptions:
Full:

\[
\begin{array}{c|c}
\text{Features} & \text{Label} \\
X_1 & Y \\
X_2 & \\
X_3 & \\
\vdots & \\
X_d & \\
\end{array}
\]

\[J_{MIM}(X_k) = I(X_k; Y)\]
Single-label Feature Selection Criteria

Feature space independence assumptions:

Full:

\[ J_{MIM}(X_k) = I(X_k; Y) \]

Partial (i.e. pairwise dependencies):

\[ J_{MIM}(X_k) = I(X_k; Y) \]
Feature space independence assumptions:

**Full:**

\[ J_{MIM}(X_k) = I(X_k; Y) \]

**Partial** (i.e. pairwise dependencies):

\[ J_{MIM}(X_k) = I(X_k; Y) \]
Single-label Feature Selection Criteria

Feature space independence assumptions:

**Full:**

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td></td>
</tr>
<tr>
<td>$X_2$</td>
<td></td>
</tr>
<tr>
<td>$X_3$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
</tr>
<tr>
<td>$X_d$</td>
<td>$Y$</td>
</tr>
</tbody>
</table>

$J_{MIM}(X_k) = I(X_k; Y)$

**Partial** (i.e. pairwise dependencies):

<table>
<thead>
<tr>
<th>Features</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td></td>
</tr>
<tr>
<td>$X_2$</td>
<td></td>
</tr>
<tr>
<td>$X_3$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td></td>
</tr>
<tr>
<td>$X_d$</td>
<td>$Y$</td>
</tr>
</tbody>
</table>

Information theoretic feature selection in multi-label data through composite likelihood
Single-label Feature Selection Criteria

Feature space independence assumptions:

Full:

\[ J_{MIM}(X_k) = I(X_k; Y) \]

Partial (i.e. pairwise dependencies):

\[ J_{MIM}(X_k) = |X_\theta| \sum_{j=1}^{d} I(X_{\theta j}; X_k; Y) \]
Feature space independence assumptions:

**Full:**

\[ J_{MIM}(X_k) = I(X_k; Y) \]

**Partial** (i.e. pairwise dependencies):

\[ J_{JMI}(X_k) = \sum_{j=1}^{\mid \mathbf{X}_\theta \mid} I(X_{\theta_j}X_k; Y) \]
Single-label Feature Selection Criteria

Feature space independence assumptions:

**Full:**

\[ J_{\text{MIM}}(X_k) = I(X_k; Y) \]

**Partial** (i.e. pairwise dependencies):

\[ J_{\text{JMI}}(X_k) = \sum_{j=1}^{|X_\theta|} I(X_{\theta_j}X_k; Y) \]

**None:**

\[ J_{\text{MIM}}(X_k) = I(X_k; Y) \]

\[ J_{\text{JMI}}(X_k) = \sum_{j=1}^{|X_\theta|} I(X_{\theta_j}X_k; Y) \]
Single-label Feature Selection Criteria

Feature space independence assumptions:

Full:

\[
\begin{align*}
J_{\text{MIM}}(X_k) &= I(X_k; Y) \\
J_{\text{MIM}}(X_k) &= I(X_k; Y)
\end{align*}
\]

Partial (i.e. pairwise dependencies):

\[
J_{\text{JMI}}(X_k) = \left| X_\theta \right| \sum_{j=1}^{X_\theta} I(X_{\theta j}; X_k; Y)
\]

None:

\[
J_{\text{CMI}}(X_k) = I(X_k; Y| X_\theta)
\]
Single-label Feature Selection Criteria

Feature space independence assumptions:

Full:

\[ J_{\text{MIM}}(X_k) = I(X_k; Y) \]

Partial (i.e. pairwise dependencies):

\[ J_{\text{JMI}}(X_k) = \sum_{j=1}^{X_\theta} I(X_{\theta j}; X_k; Y) \]

None:

\[ J_{\text{CMI}}(X_k) = I(X_k; Y|X_\theta) \]
Next, extend to multi-label where $Y$ is $q$-dimensional.
Next, extend to multi-label where $Y$ is $q$-dimensional.

What independence assumptions can we make in label space?
Next, extend to multi-label where $Y$ is $q$-dimensional

What independence assumptions can we make in label space?

In this work we examined:

- Binary Relevance (BR) vs Label Powerset (LP)
Multi-label Extension: LP Transformation

- Label Powerset (LP): No independence among labels
Multi-label Extension: LP Transformation

- Label Powerset (LP): No independence among labels
- Binary $q$-label problem $\Rightarrow$ 1 single-label, $2^q$-class problem

Pros: dependencies among labels are accounted for
Cons: probability estimates unreliable (curse of dimensionality)
Multi-label Extension: LP Transformation

- Label Powerset (LP): No independence among labels
- Binary $q$-label problem $\Rightarrow$ 1 single-label, $2^q$-class problem

<table>
<thead>
<tr>
<th>Animal</th>
<th>Building</th>
<th>Vehicle</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1 0 1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0 1 1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Pros: dependencies among labels are accounted for
Cons: probability estimates unreliable (curse of dimensionality)
Multi-label Extension: LP Transformation

- Label Powerset (LP): No independence among labels
- Binary $q$-label problem $\Rightarrow$ 1 single-label, $2^q$-class problem

<table>
<thead>
<tr>
<th>Animal</th>
<th>Building</th>
<th>Vehicle</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1 0 1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0 1 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pros: dependencies among labels are accounted for

Pros: dependencies among labels are accounted for
Multi-label Extension: LP Transformation

- **Label Powerset (LP):** No independence among labels
- **Binary $q$-label problem $\Rightarrow$ 1 single-label, $2^q$-class problem**

<table>
<thead>
<tr>
<th>Animal</th>
<th>Building</th>
<th>Vehicle</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

- **Pros:** dependencies among labels are accounted for
- **Cons:** probability estimates unreliable (curse of dimensionality)
Feature space independence assumptions:
Full:

Features

\[ X_1 \]
\[ X_2 \]
\[ X_3 \]
\[ \vdots \]
\[ X_d \]

Labels

\[ Y_1 \]
\[ Y_2 \]
\[ \vdots \]
\[ Y_q \]
Multi-label Extension: LP Transformation

Feature space independence assumptions:
Full:

Features

\[ X_1 \]
\[ X_2 \]
\[ X_3 \]
\[ \vdots \]
\[ X_d \]

Labels

\[ Y_1 \]
\[ Y_2 \]
\[ \vdots \]
\[ Y_q \]
Feature space independence assumptions:
Full:

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$X_d$</td>
<td>$Y_q$</td>
</tr>
</tbody>
</table>
Feature space independence assumptions:
Full:

\[
\begin{align*}
&\text{Features} & \quad & \text{Labels} \\
X_1 & & Y_1 \\
X_2 & & Y_2 \\
X_3 & & \vdots \\
\vdots & & Y_q \\
X_d & & 
\end{align*}
\]
Multi-label Extension: LP Transformation

Feature space independence assumptions:
Full:

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>...</td>
</tr>
<tr>
<td>$X_d$</td>
<td>$Y_q$</td>
</tr>
</tbody>
</table>
Multi-label Extension: LP Transformation

Feature space independence assumptions:
Full:

\[
J_{MIM}^{LP}(X_k) = I(X_k; Y_{1:q})
\]
Feature space independence assumptions:

Full:

\[ J_{MIM}^{LP}(X_k) = I(X_k; Y_{1:q}) \]

Partial (i.e. pairwise dependencies):

\[ J_{MIM}^{LP}(X_k) = I(X_k; Y_1) \]
Multi-label Extension: LP Transformation

Feature space independence assumptions:

**Full:**

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>...</td>
</tr>
<tr>
<td>$X_d$</td>
<td>$Y_q$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Partial (i.e. pairwise dependencies):</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>$X_1$</td>
</tr>
<tr>
<td>$X_2$</td>
</tr>
<tr>
<td>$X_3$</td>
</tr>
<tr>
<td>$X_d$</td>
</tr>
</tbody>
</table>

$$J_{MIM}^{LP}(X_k) = I(X_k; Y_{1:q})$$
### Multi-label Extension: LP Transformation

#### Feature space independence assumptions:

**Full:**

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$Y_q$</td>
</tr>
<tr>
<td>$X_d$</td>
<td></td>
</tr>
</tbody>
</table>

\[
J_{MIM}^{LP}(X_k) = I(X_k; Y_{1:q})
\]

**Partial (i.e. pairwise dependencies):**

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td></td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$Y_q$</td>
</tr>
<tr>
<td>$X_d$</td>
<td></td>
</tr>
</tbody>
</table>
Multi-label Extension: LP Transformation

Feature space independence assumptions:

Full:

\[ J_{MIM}^{LP}(X_k) = I(X_k; Y_{1:q}) \]

Partial (i.e. pairwise dependencies):

\[ J_{MIM}^{LP}(X_k) = |X_\theta| \sum_{j=1}^{q} I(X_k X_\theta^j; Y_{1:q}) \]
Multi-label Extension: LP Transformation

Feature space independence assumptions:

Full:

\[
J_{LP_{MIM}}(X_k) = I(X_k; Y_{1:q})
\]

Partial (i.e. pairwise dependencies):

\[
J_{LP_{JMI}}(X_k) = \sum_{j=1}^{\theta} I(X_k X_{\theta,j}; Y_{1:q})
\]
Multi-label Extension: LP Transformation

Feature space independence assumptions:

Full:

\[ J_{MIM}^{LP}(X_k) = I(X_k; Y_{1:q}) \]

Partial (i.e. pairwise dependencies):

\[ J_{JMI}^{LP}(X_k) = \sum_{j=1}^{|X_0|} I(X_k X_{\theta_j}; Y_{1:q}) \]

None:

Information theoretic feature selection in multi-label data through composite likelihood
Multi-label Extension: LP Transformation

Feature space independence assumptions:

- **Full:**
  \[ J_{LP}^{MIM}(X_k) = I(X_k; Y_{1:q}) \]

- **Partial** (i.e. pairwise dependencies):
  \[ J_{LP}^{JMI}(X_k) = \sum_{j=1}^{\lfloor X\phi \rfloor} I(X_k X_{\phi_j}; Y_{1:q}) \]

- **None:**
  \[ J_{LP}^{CMI}(X_k) = I(X_k; Y_{1:q}) \]
Multi-label Extension: LP Transformation

Feature space independence assumptions:

**Full:**

\[ J_{\text{LP}}^{\text{MIM}}(X_k) = I(X_k; Y_{1:q}) \]

**Partial** (i.e. pairwise dependencies):

\[ J_{\text{LP}}^{\text{JMI}}(X_k) = \sum_{j=1}^{\left| X_\theta \right|} I(X_kX_{\theta j}; Y_{1:q}) \]

**None:**

\[ J_{\text{LP}}^{\text{CMI}}(X_k) = I(X_k; Y_{1:q}\mid X_\theta) \]
Binary Relevance (BR): Full Independence among labels
Multi-label Extension: BR Transformation

- Binary Relevance (BR): Full Independence among labels
- Binary $q$-label problem $\Rightarrow q$ independent single-label, binary problems
Multi-label Extension: BR Transformation

- Binary Relevance (BR): Full Independence among labels
- Binary $q$-label problem $\Rightarrow$ $q$ independent single-label, binary problems

<table>
<thead>
<tr>
<th>Animal</th>
<th>Building</th>
<th>Vehicle</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

Pros: more reliable probability estimates

Cons: dependencies among labels are not accounted for

Information theoretic feature selection in multi-label data through composite likelihood
Multi-label Extension: BR Transformation

- Binary Relevance (BR): Full Independence among labels
- Binary $q$-label problem $\Rightarrow$ $q$ independent single-label, binary problems

<table>
<thead>
<tr>
<th>Animal</th>
<th>Building</th>
<th>Vehicle</th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Pros: more reliable probability estimates
Multi-label Extension: BR Transformation

- Binary Relevance (BR): Full Independence among labels
- Binary $q$-label problem $\Rightarrow q$ independent single-label, binary problems

Pros: more reliable probability estimates
Cons: dependencies among labels are not accounted for
Multi-label Extension: BR Transformation

Feature space independence assumptions:

Full:

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>...</td>
</tr>
<tr>
<td>$X_d$</td>
<td>$Y_q$</td>
</tr>
</tbody>
</table>

Information theoretic feature selection in multi-label data through composite likelihood
Feature space independence assumptions:
Full:

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>…</td>
</tr>
<tr>
<td>$X_d$</td>
<td>$Y_q$</td>
</tr>
</tbody>
</table>
Feature space independence assumptions:
Full:

$X_1$ $X_2$ $X_3$ $\ldots$ $X_d$ $Y_1$ $Y_2$ $\ldots$ $Y_q$
Multi-label Extension: BR Transformation

Feature space independence assumptions:
Full:

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td></td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>$X_d$</td>
<td>$Y_q$</td>
</tr>
</tbody>
</table>
Multi-label Extension: BR Transformation

Feature space independence assumptions:
Full:

Features | Labels
--- | ---
$X_1$ | $Y_1$
$X_2$ | $Y_2$
$X_3$ | ... 
... | ...
$X_d$ | $Y_q$

Information theoretic feature selection in multi-label data through composite likelihood
Feature space independence assumptions:

Full:

\[
J_{MIM}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)
\]
Multi-label Extension: BR Transformation

Feature space independence assumptions:

Full:

\[ J_{MIM}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l) \]

Partial (i.e. pairwise dependencies):

\[ X_1 \quad Y_1 \]
\[ X_2 \quad Y_2 \]
\[ \vdots \quad \vdots \]
\[ X_d \quad Y_q \]
**Multi-label Extension: BR Transformation**

Feature space independence assumptions:

**Full:**

- Features: $X_1, X_2, X_3, \ldots, X_d$
- Labels: $Y_1, Y_2, Y_3, \ldots, Y_q$

**Test statistic:**

$$J_{BR\text{MIM}}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$$

**Partial** (i.e. pairwise dependencies):

- Features: $X_1, X_2, X_3, \ldots, X_d$
- Labels: $Y_1, Y_2, Y_3, \ldots, Y_q$

Information theoretic feature selection in multi-label data through composite likelihood
**Multi-label Extension: BR Transformation**

Feature space independence assumptions:

**Full:**

\[
\begin{align*}
    &X_1 & \quad & Y_1 \\
    &X_2 & \quad & Y_2 \\
    &X_3 & \quad & \vdots \\
    &X_d & \quad & Y_q
\end{align*}
\]

**Partial** (i.e. pairwise dependencies):

\[
\begin{align*}
    &X_1 & \quad & Y_1 \\
    &X_2 & \quad & Y_2 \\
    &X_3 & \quad & \vdots \\
    &X_d & \quad & Y_q
\end{align*}
\]

\[
J_{MIM}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)
\]
**Multi-label Extension: BR Transformation**

Feature space independence assumptions:

**Full:**

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>$Y$</td>
</tr>
<tr>
<td>$X_d$</td>
<td>$Y_q$</td>
</tr>
</tbody>
</table>

$J_{BR}^{MIM}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$

**Partial** (i.e. pairwise dependencies):

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$Y_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$Y_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>$Y$</td>
</tr>
<tr>
<td>$X_d$</td>
<td>$Y_q$</td>
</tr>
</tbody>
</table>

Information theoretic feature selection in multi-label data through composite likelihood
Feature space independence assumptions:

**Full:**

\[
\begin{align*}
J_{BR_{MIM}}^{\text{MIM}}(X_k) &= \sum_{l=1}^{q} I(X_k; Y_l) \\
J_{BR_{JMI}}^{\text{JMI}}(X_k) &= \sum_{j=1}^{X_\theta} \sum_{l=1}^{q} I(X_kX_{\theta_j}; Y_l)
\end{align*}
\]

**Partial** (i.e. pairwise dependencies):

\[
\begin{align*}
J_{BR_{MIM}}^{\text{MIM}}(X_k) &= \sum_{l=1}^{q} I(X_k; Y_l) \\
J_{BR_{JMI}}^{\text{JMI}}(X_k) &= \sum_{j=1}^{X_\theta} \sum_{l=1}^{q} I(X_kX_{\theta_j}; Y_l)
\end{align*}
\]
Multi-label Extension: BR Transformation

Feature space independence assumptions:

Full:

Partial (i.e. pairwise dependencies):

None:

$J_{MIM}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$

$J_{JMI}^{BR}(X_k) = \sum_{j=1}^{q} \sum_{l=1}^{q} I(X_{kX_{\theta_{j}}}; Y_l)$

Information theoretic feature selection in multi-label data through composite likelihood
Multi-label Extension: BR Transformation

Feature space independence assumptions:

**Full:**

- Features: \(X_1, X_2, X_3, \ldots, X_d\)
- Labels: \(Y_1, Y_2, Y_3, \ldots, Y_q\)

**Partial** (i.e., pairwise dependencies):

- Features: \(X_1, X_2, X_3, \ldots, X_d\)
- Labels: \(Y_1, Y_2, Y_3, \ldots, Y_q\)

**None:**

- Features: \(X_1, X_2, X_3, \ldots, X_d\)
- Labels: \(Y_1, Y_2, Y_3, \ldots, Y_q\)

Mathematical expressions:

- For Full: \(J_{BR}^{MIM}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)\)
- For Partial: \(J_{BR}^{JMI}(X_k) = \sum_{j=1}^{X_q} \sum_{l=1}^{q} I(X_k X_{\theta_j}; Y_l)\)
- For None: \(J_{BR}^{CMI}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l| X_{\theta})\)
Feature space independence assumptions:

**Full:**

$$J_{BR_{MIM}}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$$

**Partial** (i.e. pairwise dependencies):

$$J_{BR_{JMI}}^{BR}(X_k) = \sum_{j=1}^{X_\theta} \sum_{l=1}^{q} I(X_k X_{\theta_j}; Y_l)$$

**None:**

$$J_{BR_{CMI}}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l|X_\theta)$$

Information theoretic feature selection in multi-label data through composite likelihood
Summarizing, the criteria based on feature space $X$ and label space $Y$ independence assumptions:

<table>
<thead>
<tr>
<th>Label space independence assumptions</th>
<th>Feature space independence assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label Powerset (none)</td>
<td>CMI (none)</td>
</tr>
<tr>
<td></td>
<td>$J_Y:\text{none}$</td>
</tr>
<tr>
<td></td>
<td>$J_X:\text{none}$</td>
</tr>
<tr>
<td></td>
<td>$J_Y:\text{none}$</td>
</tr>
<tr>
<td></td>
<td>$J_X:\text{partial}$</td>
</tr>
<tr>
<td></td>
<td>$J_Y:\text{none}$</td>
</tr>
<tr>
<td></td>
<td>$J_X:\text{full}$</td>
</tr>
<tr>
<td>Binary Relevance (full)</td>
<td>CMI (none)</td>
</tr>
<tr>
<td></td>
<td>$J_Y:\text{full}$</td>
</tr>
<tr>
<td></td>
<td>$J_X:\text{none}$</td>
</tr>
<tr>
<td></td>
<td>$J_Y:\text{full}$</td>
</tr>
<tr>
<td></td>
<td>$J_X:\text{partial}$</td>
</tr>
<tr>
<td></td>
<td>$J_Y:\text{full}$</td>
</tr>
<tr>
<td></td>
<td>$J_X:\text{full}$</td>
</tr>
</tbody>
</table>
Summarizing, the criteria based on feature space $X$ and label space $Y$ independence assumptions:

<table>
<thead>
<tr>
<th>Label space independence assumptions</th>
<th>Feature space independence assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label Powerset (none)</td>
<td>$CMI$ (none)</td>
</tr>
<tr>
<td></td>
<td>$JMI$ (partial)</td>
</tr>
<tr>
<td></td>
<td>$MIM$ (full)</td>
</tr>
<tr>
<td></td>
<td>$Doquire &amp; Verleysen (2013)$</td>
</tr>
<tr>
<td>Binary Relevance (full)</td>
<td>$J_Y:none$</td>
</tr>
<tr>
<td></td>
<td>$X:partial$</td>
</tr>
<tr>
<td></td>
<td>$J_Y:none$</td>
</tr>
<tr>
<td></td>
<td>$X:full$</td>
</tr>
<tr>
<td></td>
<td>$J_Y:full$</td>
</tr>
<tr>
<td></td>
<td>$X:full$</td>
</tr>
<tr>
<td></td>
<td>$J_Y:full$</td>
</tr>
<tr>
<td></td>
<td>$X:full$</td>
</tr>
</tbody>
</table>
Summarizing, the criteria based on feature space $X$ and label space $Y$ independence assumptions:

<table>
<thead>
<tr>
<th>Label space independence assumptions</th>
<th>Feature space independence assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label Powerset (none)</td>
<td>CMI (none)</td>
</tr>
<tr>
<td></td>
<td>Doquire &amp; Verleysen (2013)</td>
</tr>
<tr>
<td></td>
<td>JMI (partial)</td>
</tr>
<tr>
<td></td>
<td>$J_Y: \text{none}$</td>
</tr>
<tr>
<td></td>
<td>$X: \text{partial}$</td>
</tr>
<tr>
<td></td>
<td>MIM (full)</td>
</tr>
<tr>
<td></td>
<td>$J_Y: \text{full}$</td>
</tr>
<tr>
<td></td>
<td>$X: \text{none}$</td>
</tr>
<tr>
<td></td>
<td>$J_Y: \text{full}$</td>
</tr>
<tr>
<td>Binary Relevance (full)</td>
<td>$X: \text{partial}$</td>
</tr>
<tr>
<td></td>
<td>$X: \text{full}$</td>
</tr>
</tbody>
</table>
Summarizing, the criteria based on feature space $X$ and label space $Y$ independence assumptions:

<table>
<thead>
<tr>
<th>Label space independence assumptions</th>
<th>Feature space independence assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label Powerset (none)</td>
<td>CMI (none)</td>
</tr>
<tr>
<td></td>
<td>$\text{Doquire \ &amp; \ Verleysen (2013)}$</td>
</tr>
<tr>
<td>Binary Relevance (full)</td>
<td>JMI (partial)</td>
</tr>
<tr>
<td></td>
<td>$J^Y_{\text{none}}$ \text{X:partial}$</td>
</tr>
<tr>
<td></td>
<td>$\text{Spola\virg{o}r et al. (2013)}$</td>
</tr>
<tr>
<td></td>
<td>MIM (full)</td>
</tr>
<tr>
<td></td>
<td>$J^Y_{\text{full}}$ \text{X:full}$</td>
</tr>
<tr>
<td></td>
<td>$J^Y_{\text{full}}$ \text{X:partial}$</td>
</tr>
<tr>
<td></td>
<td>$\text{Young \ &amp; \ Pedersen (1997), Trohidis et al. (2008), ...}$</td>
</tr>
</tbody>
</table>
Experiments

- Compare

Procedure: Select $M$ top features under each criterion, classify, evaluate; vary $M$

Datasets: scene and yeast

Classification: ML-$k$NN, $k=7$

Evaluation: Hamming Loss (shown here) and Ranking Loss (similar)
Experiments

- Compare
  - effect of label space assumptions
Experiments

- Compare
  - effect of label space assumptions
  - effect of feature space assumptions

Procedure: Select $M$ top features under each criterion, classify, evaluate; vary $M$

Datasets: scene and yeast

Classification: ML-$k$NN, $k = 7$

Evaluation: Hamming Loss (shown here) and Ranking Loss (similar)
Experiments

- Compare
  - effect of label space assumptions
  - effect of feature space assumptions
  - our best criterion vs. state-of-the-art

Procedure: Select $M$ top features under each criterion, classify, evaluate; vary $M$

Datasets: scene and yeast

Classification: ML-$k$NN, $k = 7$

Evaluation: Hamming Loss (shown here) and Ranking Loss (similar)
Experiments

Compare

- effect of label space assumptions
- effect of feature space assumptions
- our best criterion vs. state-of-the-art

Procedure: Select $M$ top features under each criterion, classify, evaluate; vary $M$
Experiments

- Compare
  - effect of label space assumptions
  - effect of feature space assumptions
  - our best criterion vs. state-of-the-art

- Procedure: Select $M$ top features under each criterion, classify, evaluate; vary $M$

- Datasets: scene and yeast
Experiments

- Compare
  - effect of label space assumptions
  - effect of feature space assumptions
  - our best criterion vs. state-of-the-art

- Procedure: Select $M$ top features under each criterion, classify, evaluate; vary $M$

- Datasets: scene and yeast

- Classification: ML-$k$NN, $k = 7$
Experiments

- Compare
  - effect of label space assumptions
  - effect of feature space assumptions
  - our best criterion vs. state-of-the-art

- Procedure: Select $M$ top features under each criterion, classify, evaluate; vary $M$

- Datasets: scene and yeast

- Classification: ML-$k$NN, $k = 7$

- Evaluation: Hamming Loss (shown here) and Ranking Loss (similar)
Effect of Label Space Assumptions

Hamming loss

Yeast

Number of features selected

Information theoretic feature selection in multi-label data through composite likelihood
Effect of Feature Space Assumptions

Hamming loss

Scene

Number of features selected

Yeast

Scene

Number of features selected
Hamming loss

Yeast

Scene

Number of features selected

Information theoretic feature selection in multi-label data through composite likelihood
Doquire & Verleysen (2013) : $J^Y_{X:\text{none}}$ with pruning of rare cases
**Doquire & Verleysen (2013)**: $J_{X:none}$ with pruning of rare cases

**Lee & Kim (2013)**: Multivariate Mutual Information
Doquire & Verleysen (2013) : $J_{X:none}^{Y:none}$ with pruning of rare cases

Lee & Kim (2013) : Multivariate Mutual Information

$J_{X:partial}^{Y:full}$ tends to outperform state-of-the-art criteria
**Doquire & Verleysen (2013)**: $J_{X: \text{none}}^{Y: \text{none}}$ with pruning of rare cases

**Lee & Kim (2013)**: Multivariate Mutual Information
Doquire & Verleysen (2013) : \( J_{X: \text{none}}^{Y: \text{none}} \) with pruning of rare cases

Lee & Kim (2013) : Multivariate Mutual Information

\( J_{X: \text{partial}}^{Y: \text{full}} \) dominates state-of-the-art criteria
Empirical Observations

- Caution: Only 2 datasets! But based on them it appears that...

- Agrees with Gharroudi et al. (2014) for multilabel-label wrappers

- Examining pairwise interactions seems a good compromise between capturing interdependencies vs obtaining reliable estimates...

- Agrees with Brown et al. (JMLR 2012) findings in single-label filters
Empirical Observations

- Caution: Only 2 datasets! But based on them it appears that...
- ...independence assumptions in label space matter less than in feature space
Empirical Observations

- Caution: Only 2 datasets! But based on them it appears that...
- ...independence assumptions in label space matter less than in feature space
  - Agrees with Gharroudi et al. (2014) for multilabel-label wrappers
Empirical Observations

- Caution: Only 2 datasets! But based on them it appears that...
- ...independence assumptions in label space matter less than in feature space
  - Agrees with Gharroudi et al. (2014) for multilabel-label wrappers
- ...in feature space, JMI gives best results
Empirical Observations

- Caution: Only 2 datasets! But based on them it appears that...
- ...independence assumptions in label space matter less than in feature space
  - Agrees with Gharroudi et al. (2014) for multilabel-label wrappers
- ...in feature space, JMI gives best results
  - Examining pairwise interactions seems a good compromise between capturing interdependencies vs obtaining reliable estimates...
Empirical Observations

- Caution: Only 2 datasets! But based on them it appears that...
  - independence assumptions in label space matter less than in feature space
    - Agrees with Gharroudi et al. (2014) for multilabel-label wrappers
  - ...in feature space, JMI gives best results
    - Examining pairwise interactions seems a good compromise between capturing interdependencies vs obtaining reliable estimates...
    - Agrees with Brown et al. (JMLR 2012) findings in single-label filters
Future Work: Incorporating Domain Knowledge

- Probabilistic framework allows explicit incorporation of domain knowledge...
Future Work: Incorporating Domain Knowledge

- Probabilistic framework allows explicit incorporation of domain knowledge...

- ...as informative priors $P(X)$ or $P(Y)$
Future Work: Incorporating Domain Knowledge

- Probabilistic framework allows explicit incorporation of domain knowledge...

- ...as informative priors $P(X)$ or $P(Y)$

- ...as to how the distribution $P(Y|X)$ is factored

Information theoretic feature selection in multi-label data through composite likelihood
Future Work: Incorporating Domain Knowledge

- Probabilistic framework allows explicit incorporation of domain knowledge...
- ...as informative priors $P(X)$ or $P(Y)$
- ...as to how the distribution $P(Y|X)$ is factored
- Thus constructing more problem specific filters
Future Work: The Bigger Picture

A typical machine learning pipeline

Data $\xrightarrow{x}$ Feature Selection (Filter) $\xrightarrow{x'}$ Classification (Model) $\xrightarrow{y}$ Evaluation (Loss Function)

Assumptions in every step, often conflicting...

...should investigate interplay between model, filter & loss function
Future Work: The Bigger Picture

- A typical machine learning pipeline

```
Data \rightarrow ^x \text{Feature Selection (Filter)} \rightarrow ^{x'} \text{Classification (Model)} \rightarrow y \text{Evaluation (Loss Function)}
```

- Assumptions in every step, often conflicting...
A typical machine learning pipeline

Assumptions in every step, often conflicting...
...should investigate interplay between model, filter & loss function
Thank you!
Kiitos!