Decomposition and structuring of the output space in multi-label classification

Gjorgji Madjarov
What is a decomposition of the output space?

- The output space in multi-label learning

A global model predicts all labels at once

A decomposition of the above multi-label problem

- A set of local models predict one label each
  - Multi-label problem decomposed into several single-label problems
Binary relevance methods

- Binary relevance ($BR$)
Binary relevance methods

• Binary relevance ($BR$)

\[ X_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D} \} \]
Binary relevance methods

- Classifier chains (**CC**)

\[ x_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D} \} \]

\[ x_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D}, BP_{M_{10}} \} \]
Binary relevance methods

• Classifier chains (CC)

\[ X_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D} \} \]

\[ X_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D}, BP_{M_{10}} \} \]
Binary relevance methods

- Classifier chains (CC)

\[ X_i = \{x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D}\} \]

\[ X_i = \{x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D}, BP_{M_{10}}, BP_{M_{20}}\} \]
Binary relevance methods

- Classifier chains ($CC$)

\[ X_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D} \} \]

\[ X_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D}, BP_{M_{i_1}}, BP_{M_{i_2}} \} \]
Binary relevance methods

• Classifier chains (CC)

\[ x_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D} \} \]

\[ x_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D}, BP_{M_{10}}, BP_{M_{20}}, BP_{M_{30}}, BP_{M_{40}} \} \]
Pairwise methods

- Calibrated label ranking (CLR)
Pairwise methods

- Calibrated label ranking (CLR)

\[ X_i = \{x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D}\} \]
Pairwise methods

- Calibrated label ranking (CLR)

\[ X_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, ..., x_{i_D} \} \]
Pairwise methods

• Calibrated label ranking (CLR)

\[ X_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D} \} \]
Two stage architecture

$M_{10}$  $M_{20}$  $M_{30}$  $M_{40}$  $M_{12}$  $M_{13}$  $M_{14}$  $M_{23}$  $M_{24}$  $M_{34}$
Two stage architecture
Two stage architecture

\[ M_{10} \quad M_{20} \quad M_{30} \quad M_{40} \]

\[ M_{12} \quad M_{13} \quad M_{14} \quad M_{23} \quad M_{24} \quad M_{34} \]
Two stage architecture
Two stage architecture
Two stage architecture

\[ x_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, ..., x_{i_D} \} \]
Two stage architecture

\[ X_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D} \} \]
Two stage architecture

$$x_i = \{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{iD} \}$$

Labels

2 4 0 3 1
Two stage classifier chains architecture
Two stage classifier chains architecture

\[ x_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D} \} \]
Two stage classifier chains architecture

\[ x_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D} \} \]

\[ x_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D}, P_{M_{10}}, P_{M_{20}}, P_{M_{30}}, P_{M_{40}} \} \]
Two stage classifier chains architecture

\[ x_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D} \} \]

\[ x_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D}, P_{M_{10}}, P_{M_{20}}, P_{M_{30}}, P_{M_{40}} \} \]
Two stage classifier chains architecture

\[ X_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D} \} \]

\[ X_i = \{ x_{i_1}, x_{i_2}, x_{i_3}, \ldots, x_{i_D}, P_{M_{20}}, P_{M_{30}} \} \]
What is the structuring of the output space?

- Hierarchical multi-label classification
  - A hierarchical structure imposed on the label space
The importance of the label hierarchy in HMC

• The task of learning predictive models for hierarchical multi-label classification is addressed

• Investigation is made on
  • the differences in performance and interpretability of the local and global models
    • whether including information in the form of hierarchical relationships among the labels helps to improve the performance of the predictive models
  • inclusion of the information on the output structure also improves the performance of ensemble models.
The importance of the label hierarchy in HMC

- Two local and two global modeling tasks that exploit different amounts of the information provided by the label hierarchy were considered.

Local approaches

Single-label classification

Model 1
Label 1/1

Model 2
Label 1/2

Model 3
Label 2/1

Hierarchical single-label classification (HSC)

Model 4
Model 1
Label 1/1

Model 2
Label 1/2

Model 3
Label 2/1

Model 5

Global approaches

Multi-label classification

Model
Label 1/1
Label 1/2
Label 2/1

Hierarchical multi-label classification (HMC)
The importance of the label hierarchy in HMC - conclusions

- Label hierarchy improves the predictive performance of single trees
- HMC trees should be used on domains with well populated label hierarchy
- HSC tree architecture should be used if the number of labels per example is closer to one

- Label hierarchy brings less (or no) advantage in terms of predictive performance to ensembles
- However, there are considerable differences in the learning time between global and local ensemble methods
- HMC ensembles are much more efficient in terms of learning time than the single-label ensembles and should be used if time is an issue (especially random forests)
But what if we don’t have a structure?

• Derive a structure from the data
  • Input space
  • Output space
  • Combination of the input and output space (no experimental results)
An example of a ML dataset and its transformed HMC dataset.

<table>
<thead>
<tr>
<th>example</th>
<th>features</th>
<th>original labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( x_{1,1}, x_{1,2}, \ldots, x_{1,n} )</td>
<td>( { \lambda_1 } )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( x_{2,1}, x_{2,2}, \ldots, x_{2,n} )</td>
<td>( { \lambda_3, \lambda_5 } )</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>( x_{3,1}, x_{3,2}, \ldots, x_{3,n} )</td>
<td>( { \lambda_6 } )</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>( x_{4,1}, x_{4,2}, \ldots, x_{4,n} )</td>
<td>( { \lambda_1, \lambda_6 } )</td>
</tr>
<tr>
<td>( x_5 )</td>
<td>( x_{5,1}, x_{5,2}, \ldots, x_{5,n} )</td>
<td>( { \lambda_1, \lambda_2, \lambda_6 } )</td>
</tr>
</tbody>
</table>
Structuring of the output space – output data

• Label hierarchy based on the clustering of occurrence profiles of labels across instances
  • Identifying the relationships between labels by using expert provided information (maybe some features are not relevant for particular problem)
  • Not very relevant if the output space is sparse
Structuring of the output space – output data conclusions

- We have compared four different approaches to deriving label hierarchies
  - balanced k-means
  - hierarchical agglomerative clustering (single and complete linkage)
  - PCTs

- The hierarchies derived by using balanced k-means are clearly better to the ones derived by using the other approaches, yielding the highest improvements in predictive performance
Structuring of the output space — output space conclusions

• We have also compared data-derived hierarchies to expert-provided ones (where such hierarchies are available)

• The results reveal that they have approximately the same utility, i.e., both yield similar improvements in predictive performance
Structuring of the output space – input data

• We construct label hierarchies from the relevance scores of the features for every label
  • Each label from the output space is described (represented) by the relevance scores of the descriptive features for that particular label computed by using Relief
  • Balanced K-means (k=2,3,4,5)
Structuring of the output space – input data conclusions

• Great improvements as compared to the approach that does not use the structured output

• More general approach for structuring the output space (applicable even for multi-class classification problems)

• One extra step
  • Compute the relevance scores of the features for each label in the classification problem
Further work

• Combining the descriptions of the labels that come from (both) the input (relevance score) and the output (co-occurrence relationships) space

• Decomposing the data-derived output space

• Structuring with constraints