Multilabel Prediction of Drug Activity

Hongyu Su, Markus Heinonen, Juho Rousu

Department of Computer Science
University of Helsinki

Machine Learning in Systems Biology, Edinburgh
October 16, 2010

Drug bioactivity classification

- Given molecule, predict active/not active
- State of the art method: SVM with graph kernels over the molecules
Drug bioactivity classification

- Given molecule, predict active/not active
- State of the art method: SVM with graph kernels over the molecules
Drug bioactivity classification

- Given molecule, predict active/not active
- State of the art method: SVM with graph kernels over the molecules
Predicting activity against multiple targets

- There are numerous targets (different viruses, cancer types, ...) that share characteristics.
- Can we predict the activity better by learning against all available targets at the same time?
Multilabel classification

- **Single label classification**:
  \[ x_i \xrightarrow{\text{predict}} y_i, y_i \in \{0, 1\} \]

- **Multilabel classification**: Multiple labels (targets) associate with each example.
  \[ x_i \xrightarrow{\text{predict}} y_i = y_1 \times y_2 \times \cdots \times y_k, \ y_i \in \{0, 1\} \]

- **Basic approach**: Build a single-label classifier for each individual label, compose the multilabels from their output
  - Does not benefit from possible statistical dependencies between labels

- **Structured output prediction**: utilize structure (graph, tree, sequence) of the output to predict the multilabel in a single shot
  - Leverage on the correlation of neighboring labels
Method: Max-Margin Conditional Random Field (MMCRF)


- Relative of M³N (Taskar et al. 2003) - but assumes fixed output structure, different optimization algorithm

- Generalization of the hierarchical multilabel classifier HM³ (Rousu et al. 2005;2006) to fixed general graphs.

- Based on Conditional Random Field model over a network of outputs:

\[
P(y|x) \propto \prod_{e \in \mathcal{E}} \exp(w_e^T \varphi_e(x, y_e)),
\]
NCI-cancer Dataset

- NCI-cancer dataset contains $> 4000$ molecules with anti-cancer activity against $\sim 60$ cancer cell lines (cancer types).
- **Histogram** shows the distribution of molecules according to the activity.
  - Each bar contains molecules active against given number of targets
  - Skewed multilabel distribution
- **Heatmap** shows the similarity between pair of activity groups.
  - Inactive molecules are mutually similar
  - So are molecules that are active against all targets
  - And the extremes are similar to each other
Output representation: embedding of a labeled network

- No pre-existing structure between the drug targets, but lots of microarray data on the cell lines themselves:
  Reverse-phase lysate, cDNA, Affymetric HU6800, miRNA, ABC transporter
  Radiation RNA array

- Each gives a correlation matrix between the cell lines (how similarly the cell lines respond)

- Extract network from the correlation matrix: Maximum weighted spanning tree, Correlation thresholding, ...

- Multilabel $y$ induces a labeling of the network

- Embed the (labelled) network to a feature space: $\psi_{e,u}(y) = 1$ iff in multilabel $y$ edge $e$ is labeled $u$, $u \in \{00,01,10,11\}$
Input representation: Kernels over molecular graphs

- Various kernels applicable for molecular graphs, and have previously been used in single-label molecular classification tasks
  - Walk kernels (top picture): count matching walks (e.g. C-O-C-C-C-O-C-C-C) in two molecular graphs
  - Weighted decomposition kernel (middle): matches neighbourhoods of same-labeled nodes in two molecular graphs
  - Tanimoto kernel (bottom): kernel over user-defined salient substructures (molecular fingerprints)
- Tanimoto works the best
Joint feature map $\varphi(x, y)$

- Learning happens in feature space joint for inputs and outputs
- The feature map contains products of all input (molecule graph) and output feature (edge-labeling) pairs via the tensor (outer) product:
  \[ \varphi(x, y) = \phi(x) \otimes \psi(y) \]
- The formulation lets us learn context (edge-labeling) specific feature weights for a global set of input features
  - No assumption of alignment between input and output features
Learning MMCRF: overview

The MMCRF framework consists of the following components

- **Max-margin learning**: Maximize the margin between real example \( \varphi(x_i, y_i) \) and all the incorrect pseudo-examples \( \varphi(x_i, y) \), whilst controlling the norm of the weight vector
- **Use of kernels** \( K(x, x') \) to tackle high-dimensionality of input feature maps
- **Use of graphical model techniques** for tackle the exponential size of the multilabel space

\[
\min_{\mathbf{w}, \xi \geq 0} \left( \frac{1}{2} \left\| \mathbf{w} \right\|^2 + C \sum_{i=1}^{n} \xi_i \right)
\]

s.t. \( \mathbf{w}^T \varphi(x_i, y_i) - \mathbf{w}^T \varphi(x_i, y) \geq \ell \Delta(y_i, y) - \xi_i, \forall x_i, y. \)
Data preprocessing

- Three versions of the dataset prepared
  - Full data.
  - With no zero active molecules (group 0 removed).
  - With middle-active molecules (groups 0-10 and 50-59 removed)
- 5-fold stratified cross-validation used:
  - divide each activity group into 5-folds
  - merge across groups to create global folds
  - ensures that each group is represented in each fold
The scatter plots show prediction accuracy (left) and F1 (right) of MMCRF (y-axis) against SVM (x-axis) for each cell line (blue dots).

In terms of accuracy the two methods work equally well.

In terms of F1, MMCRF better than SVM.
Prediction Accuracy/F1: Zero-actives removed

- The scatter plots show prediction accuracy (left) and F1 (right) of MMCRF (y-axis) against SVM (x-axis) for each cell line (blue dots).
- MMCRF significantly better in terms of accuracy and F1.

![Accuracy Scatter Plot](image1)
![F1 Score Scatter Plot](image2)
Prediction Accuracy/F1: Middle-actives only

- The scatter plots show prediction accuracy (left) and F1 (right) of MMCRF (y-axis) against SVM (x-axis) for each cell line (blue dots)
- MMCRF significantly better in terms of accuracy and F1

\[\text{Accuracy (middle active molecules)}\]
\[\text{F1 score (middle active molecules)}\]

\[\text{p-value from sign test is 1.245e-6}\]
\[\text{p-value from sign test is 0.018}\]
\[\text{p-value from sign test is 0.04}\]
\[\text{p-value from sign test is 3.0e-5}\]
The plot shows the running time required for training MMCRF (1 multilabel model) and SVM (libsvm) (59 single label models).

MMCRF (native Matlab code) scales better than libsvm (C++) on large datasets.
Conclusions

• We proposed a structured output prediction approach for the classification of drug-like molecules.

• It is, to our knowledge, the first multilabel classification approach for the problem.

• The method is able to utilize the the statistical dependencies between multiple labels by means of a network constructed from auxiliary data available for the targets.

• In our experiments, the MMCRF outperforms the state-of-the-art SVM

• Future work includes
  • studying the effect of the output structure to predictive accuracy (learning algorithms, tree vs. general graph, other graph-theoretic properties)
  • better tackling of the skewness of the multilabel distribution
  • deeper look at cell line and drug molecule properties that explain good/bad performance