Structured Output Prediction of Anti-Cancer Drug Activity

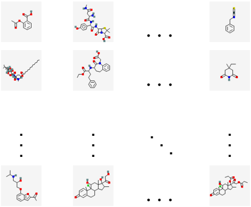
Hongyu Su, Markus Heinonen, Juho Rousu

Department of Computer Science
University of Helsinki

Pattern Recognition in Bioinformatics, Nijmegen, Netherlands
September 23, 2010

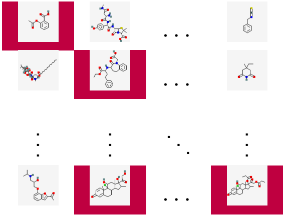
Drug bioactivity classification

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



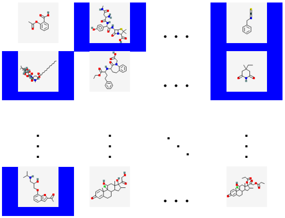
Drug bioactivity classification

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



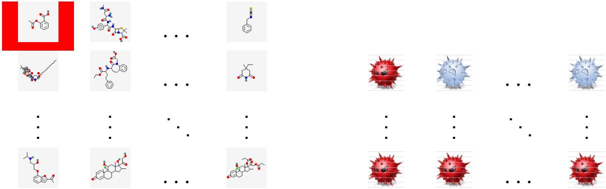
Drug bioactivity classification

- 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 :

$$x_i \xrightarrow{predict} y_i, y_i \in \{0,1\}$$

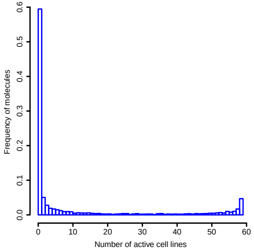
- Multilabel classification: Multiple labels (targets) associate with each example.

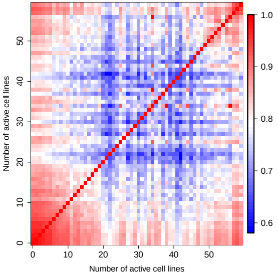
$$x_i \xrightarrow{predict} \mathbf{y_i} = y_1 \times y_2 \times \cdots \times y_k,\ y_i \in \{0,1\}$$

- 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

NCI-cancer Dataset

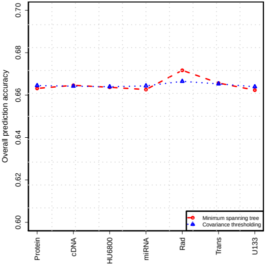
- NCI-cancer dataset contains > 4000 molecules with anti-cancer activity against $\sim$60 cancer celllines (cancer types).
- Histogram shows the distribution of molecules according to the activity.
  - Each bar contains molecules active against given number of targets
  - Skewed multilabel distribution
- 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

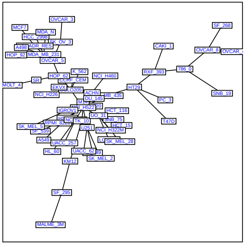




Output representation: embedding of a labeled network

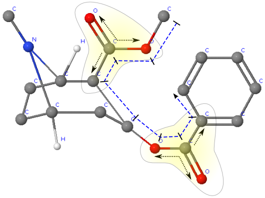
- No pre-existing structure between the drug targets, but lots of microarray data on the cell lines them selves: Reverse-phase lysate, cDNA, Affymetric HU6800, miRNA, ABC transporter Radiation RNA array
- Each gives a correlation matrix between the cell lines (how similarly the cell lines respond)
- 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: $\psi_{e,u}(\mathbf{y}) = 1$ iff edge e is labeled u in **y**

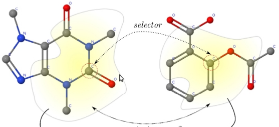




Input representation: Kernels over molecular graphs

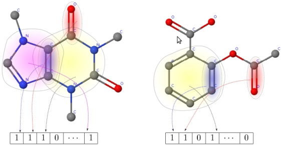
- Various kernels applicable for molecular graphs, and have previously been used in single-label molecular classification tasks
  - Walk kernels (top picture): count matching walks (e.g. C-O-C-C-C-O-C-C-C) in two molecular graphs
  - Weighted decomposition kernel (middle): matches neighbourhoods of same-labeled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





Carbon	selector
Carbon	5
Nitrogen	2
Oxygen	1

Carbon	selector
Carbon	7
Nitrogen	0
Oxygen	1



1	1	1	0	···	1

1	1	0	1	···	0

Method: Max-margin Conditional Random Field (MMCRF)

- Relative of M^3N (Taskar et al.) and HM^3 (Rousu et al.) but for fixed general graphs.
- Based on Conditional Random Field model over a network of outputs:

$$P(\mathbf{y}|x) = \frac{1}{Z(x,\mathbf{w})} \prod_{e \in \mathcal{E}} \mathbf{exp}(\mathbf{w}_e^T \varphi_e(x, \mathbf{y}_e)),$$

- Joint feature map contains products of all input (molecule graph) and output feature (edge-labeling) pairs via the tensor (outer) product:

$$\varphi(x,\mathbf{y}) = \phi(x) \otimes \psi(\mathbf{y})$$

- Lets us learn context (edge-labeling) specific feature weights

Method: Max-margin Conditional Random Field (MMCRF)

- Relative of M^3N (Taskar et al.) and HM^3 (Rousu et al.) but for fixed general graphs.
- Based on Conditional Random Field model over a network of outputs:

$$P(\mathbf{y}|x) = \frac{1}{Z(x, \mathbf{w})} \prod_{e \in \mathcal{E}} \mathbf{exp}(\mathbf{w}_e^T \varphi_e(x, \mathbf{y}_e)),$$

- Joint feature map contains products of all input (molecule graph) and output feature (edge-labeling) pairs via the tensor (outer) product:

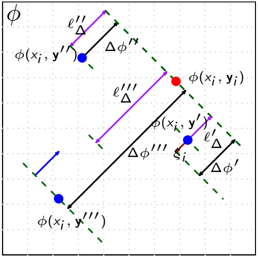
$$\varphi(x, \mathbf{y}) = \phi(x) \otimes \psi(\mathbf{y})$$

- Lets us learn context (edge-labeling) specific feature weights

Learning MMCRF: overview

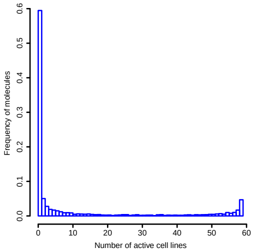
The MMCRF framework consists of the following components

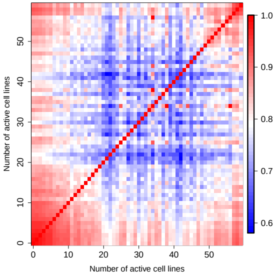
- Max-margin learning: Maximize the margin between real example $\varphi(x_i, \mathbf{y}_i)$ and all the incorrect pseudo-examples $\varphi(x_i, \mathbf{y})$, whilst controlling the norm of the weight vector
- Use of kernels $K(x, x')$ to tackle high-dimensionality of input feature maps
- Use of graphical model techniques for tackle the exponential size of the multilabel space
    - Marginal dual representation to obtain polynomial size (dual) variable set
    - Probabilistic inference (loopy belief propagation) over the marginal dual polytope to give fast updates during optimization



Data preprocessing

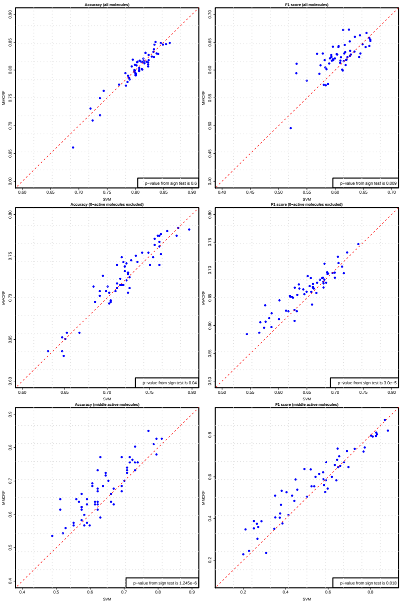
- Three versions of the dataset prepared
  - Full data.
  - With no zero active molecules (group 0 removed.
  - With middle-active molecules (groups 0-10 and 50-59 removed)
- 5-fold stratified cross-validation used:
  - divide each activity group into 5-folds
  - merge across groups to create global folds
  - ensures that each group is represented in each fold





Prediction Accuracy/F1

- The scatter plot shows prediction accuracy (left) and F1 (right) of MMCRF (y-axis) against SVM (x-axis).
- Three versions of the NCI-cancer dataset shown from top to bottom: Full, No-zero-actives, Middle-actives
- In terms of F1 (right-hand side plots), MMCRF always better than SVM
- In terms of accuracy (left-hand-side plots), MMCRF and SVM equally good on the full data, MMCRF better if zero-actives are removed



Accuracy (all molecules)

p−value from sign test is 0.6

F1 score (all molecules)

p−value from sign test is 0.009

Accuracy (0−active molecules excluded)

p−value from sign test is 0.04

F1 score (0−active molecules excluded)

p−value from sign test is 3.0e−5

Accuracy (middle active molecules)

p−value from sign test is 1.245e−6

F1 score (middle active molecules)

p−value from sign test is 0.018

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)
  - deeper look at cell line and drug molecule properties that explain good/bad performance