

Structured Project Master 2016-2017

Structured Output Prediction of Anti-Cancer Drug Activity

1^{er} avril 2017

Introduction

Machine learning has become increasingly important in the domain of drug discovery. In medical cancer treatment, people usually work on a collected dataset that contain numerous of information about different targets, for example different kind of viruses, cancer types. . . that share the same characteristics. Several methods like inductive logic programming and artificial neural networks are used for example in the task of molecule classification which consists to predict the presence or absence of the bioactivity of interest. However, classification methods focusing on a single target variable are probably not optimally suited to drug screening applications where large number of target cell lines are to be handled. A very useful question was posed : Can we predict the activity better by learning against all available targets at the same time ? The paper “Structured Output Prediction of Anti-Cancer Drug Activity” will give a point of view about solving multilabel classification in biological molecule. The main goal of the publication is predicting the active or not active anti-cancer treatments with given molecules which are represented via kernels based on molecular graphs. Among the 59 cancer cell lines, 2305 molecules were experimented by the graph kernels before applying multilabel classification learning algorithms. In case of this research, structured Support Vector Machine (SVM) and Max Margin Conditional Random Field (MMCRF) will be used and compared the results at the end.

1 Problem

The goal of this article is to apply multilabel learning approach for molecular classification. It consists to predict the presence or absence of the bioactivity of interest (given molecule, predict active/not active). For do that, the basis approach consists to build a single-label classifier for each individual label, compose the multilabels from their output. Indeed, in the case of single label classification, for given molecule x_i predicts y_i with $y_i \in \{0, 1\}$. For the multilabel classification, multiple labels (targets) associate with each example, for x_i predicts

$y_i = y_1 \times y_2 \times \dots \times y_k$ with $y_i \in \{0, 1\}$. This approach doesn’t benefit from possible statistical dependencies between labels.

To overcome this difficulty, the authors propose one method which belongs to the structured output prediction family. In the next section, we present briefly this approach.

2 Description of method

The authors proposed the Max-Margin Conditional Random Field (MMCRF). This method married graphical methods and kernels. It uses structure (graph, tree, sequence) of the output to predict the multilabel in a single shot. Then, the drug targets (cancer cell lines) are organized in a Markov network, drug molecules are represented by kernel.

The MMCRF algorithm takes like input representation kernels over molecular graphs. So, it takes as input a matrix $K = (k(x_i, x_j))_{i,j=1}^m$ (m is the number of molecules) of kernel values $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ between the training patterns where $\phi(x)$ denotes a feature description of an input pattern (drug molecule in our case) and a label matrix $Y = (y_i)_{i=1}^m$ containing the multilabels $y_i = (y_{i1}, \dots, y_{ik})$ of the training patterns. The components $y_j \in \{-1, +1\}$ of the multilabel correspond to different cancer cell lines.

A major challenge for any statistical learning model is to define a measure of similarity. There exists various kernels applicable for molecular graphs : Walk kernels, Weighted decomposition kernel and Tanimoto kernel. In our case, we use Tanimoto kernel.

Definition of Tanimoto kernel

Let denote \mathbf{u}, \mathbf{v} two molecules and d be an integer. Consider the feature map ϕ_d and the corresponding kernel k_d . The Tanimoto kernel k_d^t is defined by :

$$k_d^t(u, v) = \frac{k_d(u, v)}{k_d(u, u) + k_d(v, v) - k_d(u, v)}$$

Then, we have a square matrix of shape (m, m).

In order to use MMCRF to classify drug molecules, the authors build a Markov network for the cell lines used as the output, with nodes corresponding to cell lines and edges to potential statistical dependencies. The algorithm assumes also an associative network $\mathcal{G} = (V, E)$ where node $j \in V$ corresponds to the j’th component of the multilabel and the edges $e = (j, j') \in E$ correspond to a microlabel dependency structure.

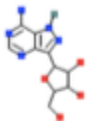
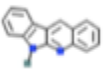
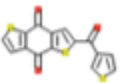
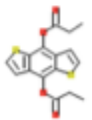
In conclusion, the MMCRF framework consists of the following components :

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

3 Presentation of data

For this project, we use the National Cancer Institute (NCI) dataset. Our study focuses on **2305** molecules (No-zero active dataset). For each molecule tested against a certain cell

line, the dataset provide a bioactivity outcome that we use as the classes (active, inactive) For making our study, we are two different datasets : Gram matrix of molecules and label dataset of molecules.

Tested Substance			Outcome	Score	loggi50 [M]
Structure	CID	SID			
	265953	405110	Active	72	-7.177
	67484	521809	Active	64	-6.44
	388302	519572	Active	67	-6.749
	388303	519573	Active	64	-6.437

4 Preprocessing

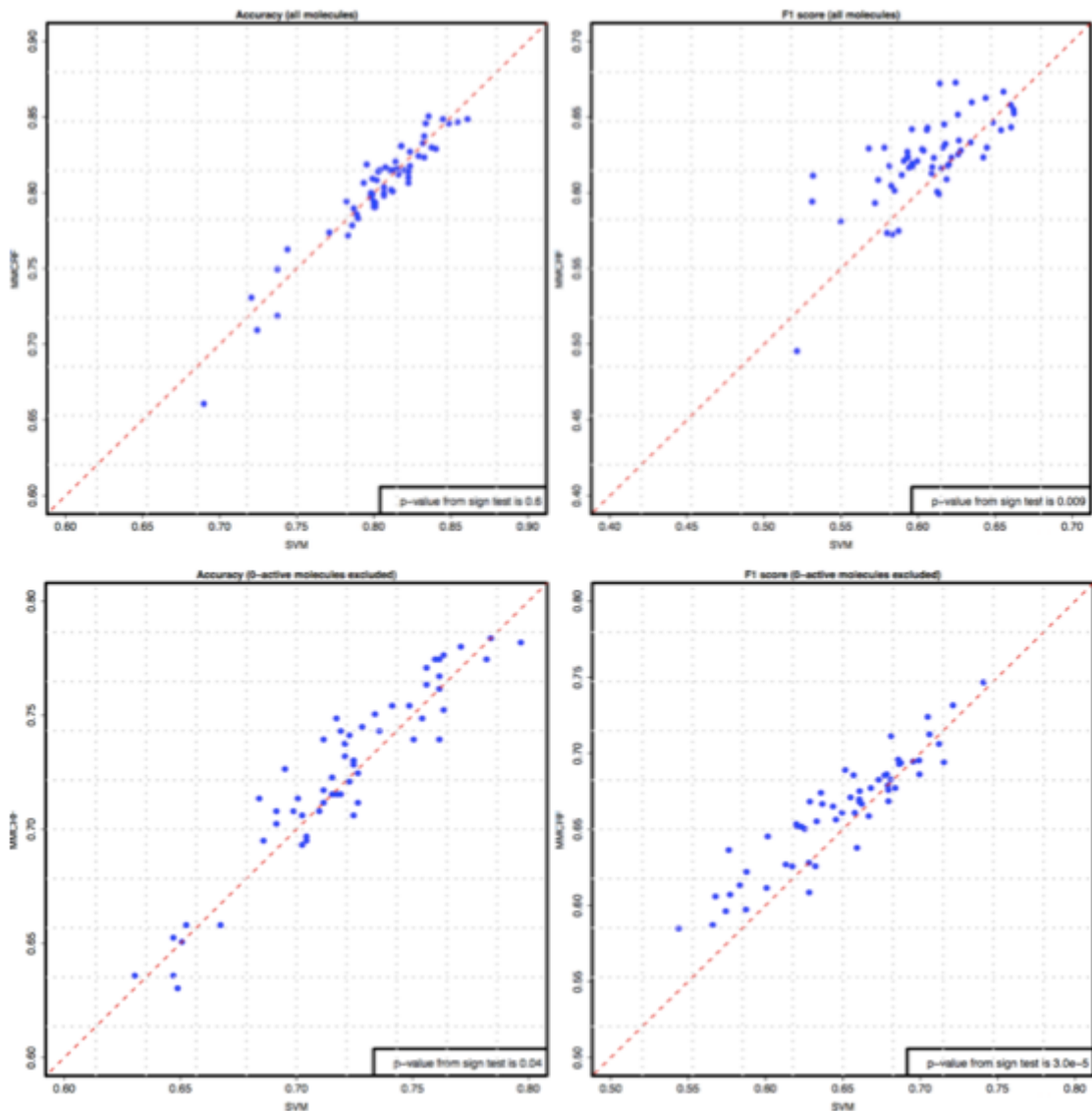
In the preprocessing part, we compute the Gram matrix and the graph network.

5 Results and discussions

Since the limitation of time, we used both Matlab (MMCRF) and Python (Strutured Multilabel SVM) for implementation of the code and final results. In particular, we benefited aslo the package Pystruct for running Structured SVM with different margin C parameters and ChainCFR which is might better be called Maximum Margin Random Fields to comparer the results with the publication. F1 scores with two tailed sign test has been used to observe more about the testing performances.

Because the paper have concluded that Tanimoto kernel is slightly better than others in microlabels F1 score thus we used Tanimoto to experience all the test. Overall, with Tanimoto kernel, the accuracies of SVM and MMCRF are respectively 61.2% and 66.3% which are less than their values in paper. Additionally, F1 scores are as well as slightly different from the obtained results in research (49.7%,56.4% in comparison with 52.7%,56.2%), we

expected the reasons driver from the different construction in algorithm between Pystruct package and LibSVM software written in C++. However, there is a same conclusion that MMCRF is markedly more accurate than SVM in both the test and MMCRF in additionally improves significantly the F1 scores over SVM. In the research, the authors supposed that SVM will be better than MMCRF in case of negative class (inactive anti-cancer) but we did not observe that result.



Due to the large amount of processing data (5000 complete molecules among 40000 initials), the training time of classifiers become seriously important. The training time of both methods increased exponential in term of expansion of training sets. Nevertheless, the training time of SVM become very slowly from training size of 2500. Overall, to run all the data set, the SVM requires around 845.88s-715.9s in compared with MMCRF 459.33s-421.2s that gives us a potential attention.

Conclusion

We observe clearly that in multi-label classification problem such that testing the drug activity of anti-cancer, MMCRF method could be a very interesting and potential approach to solve the problem in term of its accuracy and training time in comparison with other classic classification such that structured SVM especially in case of very large size of dataset.