



Aalto University
School of Science
and Technology

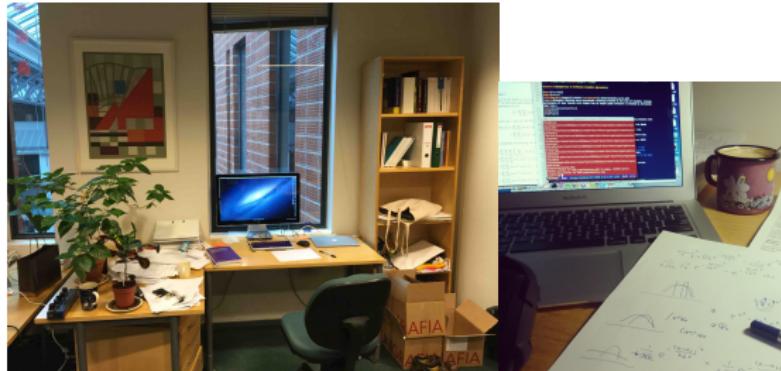
About me

Hongyu Su

Helsinki Institute for Information Technology HIIT
Department of Computer Science
Aalto University

December 1, 2015

- ▶ Hongyu Su
- ▶ Born 1984 (integrity, hard-work, open, creative)
- ▶ I am a postdoc in Helsinki Institute for Information Technology and Aalto University since 2015.5.
- ▶ Research: build advance machine learning models to solve large scale data analysis problem (research project).
- ▶ Equivalent to: continuously learning, thinking, implementing, reporting.



Educations



- ▶ Bachelor in Computer Science and Engineering, Xidian University, 2007
- ▶ Master in Bioinformatics, University of Helsinki, 2010
- ▶ Phd in Information and Computer Science, Aalto University, 2015
- ▶ '*Everything comes with a price. Everything. Some things just cost more than others.*' -Brom
- ▶ Some things don't have a price tag!

Phd, 2011.01-2015.04, Helsinki, Finland

- ▶ The topic is **machine learning and optimization research on structured data**.
- ▶ Machine learning is to estimate outcomes (unknown) from data (known).
- ▶ Ask non-trivial machine learning questions and provide solutions.
 - ▶ Computer vision, identify object in the image.


$$(+1, +1, -1, -1, -1, +1, +1)$$

boat sea sun beach people ice land

- ▶ News articles can be assigned to multiple categories.


$$(+1, +1, -1, -1, -1, -1, -1)$$

news economics sports politics movie science art

Results

Methods and technologies that are published in TOP machine learning journal and conference.

Multilabel Structured Output Learning with Random Spanning Trees of Max-Margin Markov Networks [Implementation contributed by Marie Marbach, Sven Sra, Bertrand Michel, John Shawe-Taylor]

Let $\mathcal{V} = \{1, \dots, n\}$, $\mathcal{A}_v = \{1, \dots, |A_v|\}$. Let \mathcal{C}_v denote the sets of labels associated with node v .

$$P(w, v) := \frac{e^{-d_w(v)}}{\sum_{w' \in \mathcal{V}} e^{-d_w(v)}}, \text{ where } d_w(v)$$

Structured Output Prediction of Anti-cancer Drug Activity

Hongyu Su, Marko Dolinek, and John Shawe-Taylor
*Department of Computer Science,
UCL, London, UK*

Abstract: We present a novel and simple structured prediction framework for drug activity classification of a molecule. Our model predicts a set of labels, which is annotated with a node of a graph. The graph is encoded by a Matrix Tree transform that measures the probability of an active set of labels given a molecule. A typical loss function based on submodularity is applied to the set of predicted labels. In addition, we also propose the multilabel structured classification framework that is able to predict the likelihood of drug activity against 10 cancer test sets independently. The test results are summarized in the end of the paper.

Multilabel Structured Output Learning with Random Spanning Trees of Max-Margin Markov Networks [Implementation contributed by Marie Marbach, Sven Sra, Bertrand Michel, John Shawe-Taylor]

Keywords: Multilabel classification, Structured output, Random spanning trees, Max-Margin Markov Networks

Authors:
Hongyu Su
Universität Regensburg, Institut für Informationstechnologie (IT),
Department of Mathematics and Computer Science, Aalto University, Finland
John Shawe-Taylor
London, UK
Bertrand Michel
Institute for Information Technology (MIT),
Department of Mathematics and Computer Science, Aalto University, Finland
Marie Marbach
Universität Regensburg, Institut für Informationstechnologie (IT),
Department of Mathematics and Computer Science, Aalto University, Finland

Abstract: Structured prediction, which handles multiple dependent predictions simultaneously, has been extensively applied to various tasks in machine learning and computer vision, such as action recognition, scene classification, relation extraction, multi-label classification, and semantic role labeling. The main idea of structured prediction is to model the dependencies between the multiple outputs. In this work, we propose a new framework for structured prediction, named multilabel structured output learning with random spanning trees of max-margin markov networks, which is able to predict the likelihood of drug activity against 10 cancer test sets independently. The test results are summarized in the end of the paper.

Abstract: Machine learning has become increasingly widely used in bioinformatics and biostatistics. In particular, Quantitative Structure-Activity Modeling (QSAR), that is, the prediction of drug activity (e.g., mutagenicity, cytotoxicity, etc.) is routinely built using machine learning models. In particular, these QSAR models can be learned from the most recent crowds available [1].

However, due to the lack of drug activity data, the task of drug activity prediction – one of the main applications in QSAR – has been tackled as a multi-label classification problem. Most of the state-of-the-art methods [2–5] have tried to solve this problem using linear regression-based methods. In our previous work [1, 6], we have proposed a novel learning framework based on kernel-based methods for multilabel prediction, providing new insights and showing its potentialities. Now we propose to move on to another task related to drug screening: pharmacophore identification.

Keywords: QSAR, machine learning, multilabel classification

Authors:
Hongyu Su
London, UK
John Shawe-Taylor
London, UK
Marko Dolinek
London, UK

Abstract: We show that the usual naive baseline for conditional Markov networks can be beaten as far as the separation rate of the training samples. We also show that a linear model with squared loss does not always perform best. We compare the performance of the various methods for the complete graph and provide conditions under which the proposed method is competitive. This work is related to our previous work on pharmacophore learning, in which we studied learning using the sigmoid.

Keywords: QSAR, machine learning, multilabel classification

Authors:
Hongyu Su
London, UK
John Shawe-Taylor
London, UK
Marko Dolinek
London, UK

Multitask Drug Bioactivity Classification with Graph Labeling Ensembles

Hongyu Su, John Shawe-Taylor

Keywords: multitask learning, structured output, graph labeling, ensemble learning

Authors:
Hongyu Su
London, UK
John Shawe-Taylor
London, UK

Abstract: Drug bioactivity prediction is an important task in computational chemistry and pharmacology. One way to learn drug bioactivities is to learn a set of classifiers for different drug targets and then to aggregate their predictions. Another way is to learn a single classifier that takes into account the dependencies between different drug targets. In this work, we propose a new framework for multitask drug bioactivity prediction, named multitask drug bioactivity classification with graph labeling ensembles. The key difference between this approach and the single target approach is that each target in the multitask setting is associated with a graph, representing the dependencies between the different targets. These graphs can be learned from the data or provided by domain experts. Our proposed approach is able to learn a single classifier that takes into account the dependencies between the different drug targets.

Keywords: multitask learning, drug bioactivity prediction, graph labeling

Authors:
Hongyu Su
London, UK
John Shawe-Taylor
London, UK

Multilabel Classification through Random Graph Ensembles

Hongyu Su, John Shawe-Taylor

Keywords: Multilabel classification, Random graph ensembles, Ensemble learning

Authors:
Hongyu Su
London, UK
John Shawe-Taylor
London, UK

Abstract: Given an instance of size $n \times p$, the goal is to find an optimal configuration of nodes (\mathcal{V}) and edges (\mathcal{E}) for a random graph ensemble to represent the underlying data. This problem is NP-hard. To overcome this difficulty, we propose two efficient approximation algorithms: one for the case of small nodes and another for the case of large nodes. The first algorithm is based on a local search strategy that iterates over the nodes and edges of the graph. The second algorithm is based on a global search strategy that explores all possible configurations of nodes and edges. Our experimental results show that the proposed algorithms are able to find good configurations of nodes and edges in a reasonable amount of time.

Keywords: Multilabel classification, Random graph ensemble, Graph learning

Authors:
Hongyu Su
London, UK
John Shawe-Taylor
London, UK

Dissertation

- ▶ My dissertation **Multilabel classification through structured output learning - methods and applications**.
 - ▶ **Advanced** machine learning methods to push the boundary of multilabel classification.
 - ▶ Solving many real-world **nontrivial** machine learning problems: document classification, image annotations, molecular classification, bioinformatics, social network analysis.



Work hard for 4 years and then

- Defense



- Karonkka



- Phd



Awards (name vs money)

- ▶ Chinese government awards for outstanding Phd candidate



- ▶ Some awards from Aalto university.

What did I learn from Phd?

- ▶ Strong expertise in
 - ▶ Machine learning and mathematical modelings
 - ▶ Optimization research: linear/nonlinear optimizations
 - ▶ Algorithm and data analysis / recommender system
 - ▶ Large scale data analysis
- ▶ Solid programming skills in
 - ▶ Python, Matlab, C
 - ▶ Hadoop, Spark, SQL
 - ▶ SVN, Git, Jekyll, JavaScript
 - ▶ Website and blog at www.hongyusu.com
 - ▶ GitHub at www.github.com/hongyusu
- ▶ A creative brain to solve challenging problems with modern technologies.
- ▶ An open mind that always wants to learn.
- ▶ **Be able to work hard for the long-term goal.**

A big fan of modern information technologies

- ▶ I am curious, open, and want to learn new technologies.
- ▶ I am innovative, want to use new technologies to change daily life.
- ▶ Know new technologies very well, e.g., deep learning, big data, Kafka, Spark, IoT.
- ▶ Maintain a technical blog at www.hongyusu.com on technology innovations.
- ▶ Double blade: new are not always good, e.g., I like old-fashion mechanical keyboard.



A big fan of sports

- ▶ I enjoy competitions and aggressive sport for example basketball.



- ▶ Now I try to discover unknown part of myself
 - ▶ Downhill snowboarding.
 - ▶ Bouldering (license)
 - ▶ Paragliding
 - ▶ Open water diving (license)
- ▶ A lot of gyms.

A cat person

- ▶ Pabulo



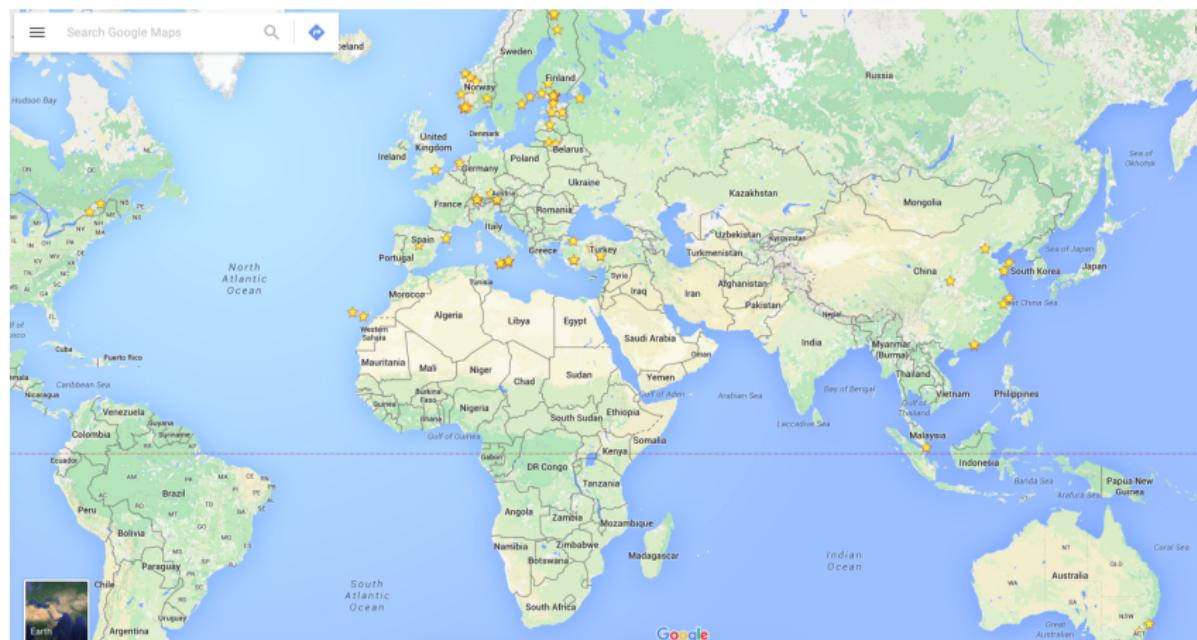
- ▶ Miu



- ▶ Cats are independent and make my life not very technical.

A hiker

I like to discover new places.



A photographer with a Flickr account

I like to memorize great moments.



A bottle collector

I like to taste new things.



'Stay hungry, stay foolish.' - Steve Jobs