Aalto University

306 APPLICATION FOR PRE-EXAMINATION AND PERMISSION TO PUBLISH THE DISSERTATION

Doctoral Programme
Doctoral Programme in Science

PERSONAL DATA

Name and academic degree	Student number	Gender
Hongyu Su, M.Sc.	829305	Female X Male
Street address: Konemiehentie 2	Telephone number: +358 442793169	
Postal code and city: FI-00076 Aalto	E-mail address: Hongyu.su@aalto.fi	
Research field: Information and Computer Science		

DETAILS OF DISSERTATION MANUSCRIPT

Type of manuscript	Language of dissertation
Article dissertation X Monograph Other	English
Name of manuscript (in the language the dissertation is written): Multilabel Classification through Structured Output Learning - Methods and Applications	

Supervising professor (name)	**Thesis advisor** (name and academic degree)
Juho Rousu	Juho Rousu, Ph.D.
Department (abbreviation): ICS	Work place: Aalto University
E-mail address: Juho.rousu@aalto.fi	E-mail address: Juho.rousu@aalto.fi

SIGNATURE OF APPLICANT which signifies the acceptance of proposed pre-examiners.

Date	Signature
11/06/14	Hongyu Su

PRELIMINARY EXAMINERS PROPOSED BY SUPERVISOR

Preliminary examiner (name and academic degree/position)	Preliminary examiner (name and academic degree/position)
Pierre Geurts, Ph.D./Associate Professor	Tapio Pahikkala, Ph.D./Senior lecturer, Docent
Work address: University of Liege, Department of Electrical Engineering and Computer Science	Work address: University of Turku, Department of Information Technology
Telephone: +32 43664815	Telephone: +358 405572077
E-mail address: p.geurts@ulg.ac.be	E-mail address: Tapio.pahikkala@utu.fi

Date	Signature of supervising professor
11/06/14	Juho Rousu

DECISION OF DOCTORAL PROGRAMME COMMITTEE

Preliminary examiners assigned (date)	Deadline for pre-examination statements

Date	Signature

Permission to publish granted	Date	Signature

APPENDICES

For detailed information on the required appendices and for contact information, please visit into.aalto.fi.


Aalto University

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Hongyu Su

Name of the doctoral dissertation
Multilabel Classification through Structured Output Learning - Methods and Applications

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 0/2014

Field of research Information and Computer Science

Manuscript submitted XX.XX.XXXX **Date of the defence** XX.XX.XXXX

Permission to publish granted (date) XX.XX.XXXX **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Multilabel classification as an important research field in machine learning arises naturally from many real world applications. For example, in document classification, a research article can be categorized into "science", "drug" and "genomics" at the same time. The goal of multilabel classification is to make reliable prediction on multiple output labels for a given input example. As multiple interdependent output labels can be "on" and "off" simultaneously, the central problem in multilabel classification is to explore the label correlation in order to make accurate prediction. Compared to previous flat multilabel classification approach that treats the multiple labels as a flat vector, structured output learning builds an output graph connecting multiple labels in order to explore the label correlation in a comprehensive manner. The main question studied in the thesis is how to tackle multilabel classification through structured output learning. Within this scope we discuss several subproblems.

The thesis starts with extensive review on classification learning covering both single-label and multilabel classification settings. The first problem that we address is how to solve the multilabel classification problem when output graph is observed as a-priori. We revisit several well established structured output learning algorithms and study the network response prediction problem within the context of social network analysis. We realize that the current structured output learning algorithms rely on the output graph to gain representation power of label dependency. Therefore, the second problem that we address is how to use structured output learning when there is no pre-established output graph. More specifically, we examine the potential of learning on a set of random output graphs when the "real" one is hidden. This problem is relevant as in general multilabel classification problems there does not exist any output graph that reveals the label dependency. It is also difficult to extract the dependency structure from data. The third problem that we address is how to analyze the proposed learning algorithms in a theoretical manner. Especially, we want to explain the behavior of the proposed models and to study the generalization error.

The main contributions of the thesis are the new learning algorithms that widen the applicability of structured output learning. For the problem with observed structure, the proposed algorithm "SPIN" is able to predict a directed acyclic graph from an observed underlying network. For general multilabel classification problems without pre-established output graph, we proposed several learning algorithms that combine many structured output learners built on random output graphs. In addition, we develop a joint learning and inference framework that is based on the Max-Margin learning over a random sample of spanning trees. The theoretic analysis also guarantees the generalization error of the proposed method.

Keywords

ISBN (printed) 000-000-00-0000-0 **ISBN (pdf)** 000-000-00-0000-0

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages **urn** http://urn.fi/URN:ISBN:000-000-00-0000-0

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Hongyu Su, Aristides Gionis, Juho Rousu. Structured Prediction of Network Response. In *Proceedings of the* 31^{th} *International Conference on Machine Learning (ICML 2014)*, Beijing, China, 2014. JMLR W&CP volume 32:442-450, June 2014.

II Hongyu Su, Markus Heinonen, Juho Rousu. Multilabel Classification of Drug-like Molecules via Max-margin Conditional Random Fields. In *Proceedings of the* 5^{th} *International Conference on Pattern Recognition in Bioinformatics (PRIB 2010)*, Nijmegen, The Netherlands, 2010. Springer LNBI volume 6282:265-273, September 2010.

III Hongyu Su, Juho Rousu. Multi-task Drug Bioactivity Classification with Graph Labeling Ensembles. In *Proceedings of the* 6^{th} *International Conference on Pattern Recognition in Bioinformatics (PRIB 2011)*, Delft, The Netherlands, 2011. Springer LNBI volume 7035:157-167, November 2011.

IV Hongyu Su, Juho Rousu. Multilabel Classification through Random Graph Ensembles. *Machine Learning*, DOI:10.1007/s10994-014-5465-9, Published Online 26 Pages, September 2014.

V Mario Marchand, Hongyu Su, Emilie Morvant, Juho Rousu, John Shawe-Taylor. Multilabel Structured Output Learning with Random Spanning Trees of Max-Margin Markov Networks. In *Proceedings of the* 28^{th} *Advances in Neural Information Processing Systems (NIPS 2014)*, Accepted 9 Pages, December 2014.

Hongyu Su
07. Nov. 2014

Author's Contribution

Publication I: "Structured Prediction of Network Response"

Publication I presents a novel formalism of the network response prediction problem, and proposes a structured output prediction algorithm for the problem. The proposed algorithm SPIN captures the contextual information and improves the state-of-the-art models in terms of the prediction performance.

The definition of the problem and the initial modeling idea were developed jointly by the authors. The author made major contribution to designing the learning framework and the optimization algorithm. The author implemented the whole learning system including SDP and the GREEDY inference algorithms. The author designed and performed the experiments and analyzed the results. The author worked jointly in the designing and writing the research article.

Publication II: "Multilabel Classification of Drug-like Molecules via Max-margin Conditional Random Fields"

Publication II presents a structured output prediction model for multilabel molecular activity classification problem. The new model incorporates the correlation of the output variables into the learning process and surpasses the previous single-label classification approach in terms of the prediction performance.

The original modeling idea was jointly proposed. The author implemented the learning system that applying the structured output learning algorithm for the task. The author designed and performed the experiments and analyzed the results. The author participated in designing

and writing the research articles.

Publication III: "Multi-task Drug Bioactivity Classification with Graph Labeling Ensembles"

Publication III extends Publication II by learning a majority voting ensemble of a set of structured output classifiers built from random output graphs.

The idea of the ensemble learning strategy was jointly developed. The author made major contribution to the design of the algorithm. The author implemented the algorithm, designed and performed the experiments, and analyzed the results. The author participated in designing and writing the research articles.

Publication IV: "Multilabel Classification through Random Graph Ensembles"

Publication IV extends Publication III by introducing two aggregation framework (AMM and MAM) which perform the inference before or after combining the multiple structure output learners. The theoretical study in Publication IV explains the performance of the proposed MAM model. The performance of the proposed models are examined on a set of heterogeneous multilabel prediction problems.

The modeling idea was developed jointly by authors. The author designed and implemented the learning frameworks and the optimization algorithms. The author developed the theorem explaining the performance of MAM. The author designed and conducted the experiments and analyzed the results. The author worked jointly in the designing and writing the research article.

Publication V: "Multilabel Structured Output Learning with Random Spanning Trees of Max-Margin Markov Networks"

Publication V is a major step forward of Publication IV by introducing the joint learning and inference framework and developing rigorous learning theory to backup the algorithm.

The idea was initialized jointly by the authors. The author worked jointly on the development of the theories and the proofs. The author made major contribution to the learning framework and optimization algorithms. The author implemented the whole learning system. The author designed and performed the experiments and analyzed the results. The author participated in designing and writing the research articles.

Hanghyn Su
07. Nov. 2014

I approve the author's contribution as sufficient for a doctoral dissertation.

Espoo, 7. November 2014

Juha Rousu
Associate Professor
Supervisor