Turun yliopisto
University of Turku

Pre-examination statement of Hongyu Su's scientific dissertation "Multilabel Classification through Structured Output Learning - Methods and Applications"

As the pre-examiner appointed by the Aalto University School of Science, I would like to state the following:

The scientific dissertation authored by MSc. Hongyu Su concerns structured output learning approaches for multi-label classification problems. The dissertation is composed of about 55 page introductory part and five scientific articles of which one is published in a journal and four in conference proceedings. The introduction consists of a review of the multi-label classification methods in the existing literature and a description of two branches of structured output learning for multi-label classification, namely the case in which the graph of label dependencies is known in advance and the case in which the dependency graph is not known and is either inferred from an auxiliary data or substituted with an ensemble of models trained with randomly drawn graphs. The review about the related literature is very coherent and comprehensive, and lays the ground for the actual contributions of the thesis, the structured output approaches.

The first paper concerns the structured output prediction approach with the presence of a known dependency graph between the class labels. The learning task is first rigorously defined and the rest of the paper concentrates on new learning algorithms for solving it. The article is very mature and the presented algorithms are elegant. Hongyu Su made a major contribution for this paper, whose first author he also is. The paper is published in ICML, one of the most prestigious conferences in the field of machine learning.

The second publication uses maximum margin conditional random fields for the multilabel molecular classification problem. The problem of the label dependency graph being unknown is remedied via taking advantage of auxiliary datasets in which such information is available. The paper is first authored by Hongyu Su, who also designed and implemented the learning approach used it to do the experiments while inventing the idea and writing the article jointly with other authors, and it is published in an established conference in bioinformatics.

The above work with multilabel molecular classification is continued in the third article, in which the authors use a set of random dependency graphs instead of those extracted from auxiliary data sets. MMCRF models are trained with each of the generated random graphs and the models are used together to form a majority voting ensemble. Hongyu Su first authored the article published in an established conference in bioinformatics.

The fourth article continues the study of the ensemble multi-label learning methods and introduces two more sophisticated aggregation approaches for learning with a set of random output dependency graphs. The first method, instead of using the previously considered majority voting for each microlabel, sets each microlabel to the one with the highest average local max-marginal score among the set of models. The second method first constructs a global average dependency graph from the edge potentials of the set of random graphs and infers the multilabel vector from that graph. The methods are also theoretically analysed with tools that have previously been used for analysing only single-label ensembles. Hongyu Su is the first author of the paper that is published in one of the top journals of the field of machine learning.

The fifth publication presents a result about being able to accurately approxiate the mult-label learning with a complete output dependency graph with a learning approach using a set of

Informaatioteknologian laitos
Department of Information Technology

ICT-talo, Joukahaisenkatu 3-5, FI-20014 Turun yliopisto,
Finland
Puhelin/Telephone +358 2 333 8627
Faksi/Fax +358 2 333 8600
www.it.utu.fi

uniformly drawn spanning trees of the graph. The result is beneficial firstly because solving the argmax problem involved in learning with the complete graph is computationally infeasible, while the corresponding problem involving only the spanning trees is considerably easier to solve. The approximation ability of the complete graph with the set of random spanning trees and the learning performance of the proposed approach are backed by a rigorous theoretical analysis. Hongyu Su is the second author of this article and his contribution on the learning framework and the implementations was major. The article is published in one of the most appreciated conferences in the field of machine learning.

The learning algorithms are well formalized and their explanations are written with clear language. Altogether, the thesis is written in line with good scientific criteria, relevant literature is properly cited and discussed and the proposed methods are always contrasted with the state-of-the-art in the field. The new approaches, algorithms and theoretical results introduced in this thesis without a doubt represent a significant body of new scientific findings, and the contributions of Hongyu Su in the jointly written articles are sufficient for a doctoral thesis. The rigorous mathematical formalizations of the methods and the depth of the theoretical analysis demonstrates his ability of performing high quality research.

I warmly recommend the thesis for acceptance. In addition, this thesis consisting of a thorough introductory chapter and articles in top machine learning forums can, on my behalf, also be considered as being among the top 10% of the dissertations in the field of computer science.

Attached is a list of corrections and suggestions regarding the introductory part. The most of these concern only notation or typos, whose implementation can be carried out under the supervision of the thesis advisors.

Yours sincerely,

7. 1. 2015

Tapio Pahikkala
Phd, Adjunct Professor
Department of Information Technology
University of Turku
Finland

Attachment: List of corrections and suggestions to the dissertation manuscript

Page 24: In Eq. (2.14), P(y_i|phi(x_i)) should be P(y_i|x_i)?

Pages 27-28: The kernel function is first defined via the symmetry and positive semidefiniteness properties (Def. 5) and then the feature map is subsequently defined via the RKHS (Def. 6), which is quite unusual in the kernel methods literature. In Def. 5, the sum goes over all pairs of inputs in the whole input space, but this is not well-defined for infinite or continuous input spaces. Instead, please rewrite the PSD property so that x_i and x_j in the sum would go through every element in a particular finite subset of the input space, and state that the condition has to hold for all finite subsets of the input space. Def. 6 is not the standard definition of a reproducing kernel Hilbert space. However, since the concept is not used anywhere else in the thesis, I suggest replacing the current definitions 5 and 6 with a definition where a function is defined to be a kernel if it can be expressed as an inner product between some feature mappings of the inputs, as it is often defined in the literature. Then, you can continue by stating that every kernel is positive semidefinite and give the definition of the PSD property (e.g. the current Def. 5).

Pages 30-31: Please check that the distribution indices are correct on every line of the Algorithm 1. D^t should probably be D^{t-1} on line 4? Also, please be consistent in the notation you use in the text when referring to the algorithm. For example, on page 30, the passage "... current distribution D^t (line 3) ..." uses a different index for D than what is actually used on line 3 of Algorithm 1.

Page 40: Again, please check the distribution indices from the Algorithm 2, especially D^t on line 5. Also, should y_i on line 4 be hat{y}_i?

Page 43 and throughout the rest of the thesis: The symbol of an edge e is sometimes bolded and sometimes not. Please be consistent with this as the bolding makes a difference with other label symbols.

Page 45: To stress the connection between (4.3) and (4.4), you could substitute the log sum in (4.4) with log z_{x_i,w}. Please also check the boldings of the symbols as sometimes the calligraphic Y seems to be in bold, like in (4.4) and sometimes not, like in (4.3).

Page 51: The right hand side of the first nonnumbered equation should probably be <w, phi(a,G_a)>.

Page 51-52: The notation in equations (4.9.) and (4.10) is confusing. The argument of H* has an index i (probably should not have) and the index i is also used in the definition of s_{y_e} but it only indexes w. In article I, it also indexes phi indicating the ith feature of action a. Please clarify the notation. It would also pay to define the shape of w and the feature map for the graph, since the triple index consisting of i, e, and y_e is challenging to picture without reading through the article first.

Page 52: The marginal gain for the nodes F_m is undefined in Lemma 1. The set V_p^H of activated vertices under the sum is also undefined.

Page 56: "... e denotes a matrix of ones" -> "... e denotes a vector of ones".

Informaatioteknologian laitos
Department of Information Technology

ICT-talo, Joukahaisenkatu 3-5, FI-20014 Turun yliopisto, Finland
Puhelin/Telephone +358 2 333 8627
Faksi/Fax +358 2 333 8600
www.it.utu.fi

Page 58: In the first equation, the second occurrence of the vector w is missing an index e. Please be consistent with the index of the microlabel which is i in the text of the last paragraph and j in the last equation.

Page 59: Please open the concept of the consensus graph to some extent, as the description of the MAM method is otherwise too challenging to picture.

Informaatioteknologian laitos
Department of Information Technology

ICT-talo, Joukahaisenkatu 3-5, FI-20014 Turun yliopisto,
Finland
Puhelin/Telephone +358 2 333 8627
Faksi/Fax +358 2 333 8600
www.it.utu.fi