Evaluation report on the PhD dissertation of Hongyu Su

« Multilabel Classification through Structured Output Learning - Methods and Applications »

PhD dissertation submitted by Hongyu Su
to the school of Science of the University of Aalto

Advisor : Juho Rousu

This thesis tackles multilabel classification using structured output approaches. Within supervised machine learning, multilabel classification is a very practically relevant framework that finds applications in many domains including text analysis, computer vision, or bioinformatics. Among existing methods, the thesis focuses on structured output learning techniques based on margin maximization. The main advantage of these methods for multilabel classification is that they can expressly take into account dependencies between the labels, typically assumed to be encoded in the form of a graph. The main limitations of these methods are both the fact that the output graph needs to be known in advance and also their computational inefficiency, in particular when the output graph is highly connected. The present thesis advances the state of the art in this domain along two main directions. First, the candidate formulates a new relevant learning problem, called network response prediction, and provides an efficient technique to solve this problem based on the adaptation of existing structured output learning approaches. Second, he proposes several computationally efficient algorithms for addressing structured output problems where the output graph is unknown (and therefore needs to be learned from the data).

The thesis manuscript is divided into two main parts. The first part (79 pages) is a general overview of the work carried out by the candidate, while the second part (99 pages) reproduces the text of five peer-reviewed publications co-authored by the candidate (from 2010 to 2014). I examine below both parts in turn before concluding with a general assessment of the thesis.

General overview

The overview is divided into 7 chapters. After a good general introduction to the thesis in Chapter 1, Chapters 2 and 3 provide the necessary background respectively on supervised (single label) classification and (flat) multilabel learning. The selected references and the level of the details adopted in the descriptions are appropriate. One minor remark I have is that the distinction between flat and structured output techniques is not totally natural to me and could have been better emphasized. Some methods categorized as flat indeed exploits a dependency structure on the labels (e.g., classifier chains) and structured output techniques could also be categorized into the algorithm adaptation subfamily (as they are multilabel adaptations of single label margin maximization techniques such as SVM). Chapter 4 and 5 give an overview of the contributions of the authors. Chapter 4 first surveys the most prominent structured output prediction models and then introduces the problem of response mode prediction and gives a sketch of the original method proposed by the candidate to solve this problem (Publication I, see below). Chapter 5 then introduces successively the different methods developed in publications II to V to handle structured output prediction problems with an unknown output graph. The descriptions in these two chapters are clear and provide a good summary of the main contributions of the thesis. Chapter 6 gives implementation details for all proposed methods and provides links to public repositories collecting the source code. Chapter 7 concludes the overview with a discussion of the main results and an enumeration of several, very relevant, future work directions. Overall, going through this overview was pleasant and gave me useful keys to understand the scientific approach adopted by the candidate before I started reading the actual publications. It is appropriately organized and contains enough details. One regret however is that this overview remains rather factual in its description of the different contributions. The author does not really try to put his contributions into a broader perspective with respect to the individual publications.

The writing and the language of this overview are not as polished as in the peer-reviewed publications that follow but they are good enough. There remain several minor typos or errors but I will transmit them to the candidate and they can be easily fixed before publication.

Publications

The contributions are collected into five peer-reviewed publications.

Publication I introduces the network response prediction problem, whose goal is to predict the subnetwork of a complex network that is activated in response to a given action. This problem finds many applications, e.g. for predicting message spreading in social networks. The authors approach this problem through supervised learning (mapping actions to subnetworks) with a specifically designed max-margin structured output model. The resulting max-margin optimization problem requires to solve an inference problem, which is shown in the paper to be NP-hard. Two approximate inference methods are therefore proposed: one based on a semidefinite programming relaxation (with an approximation guarantee but not scalable) and one greedy approach (with no guarantee but scalable). Through extensive experiments on several real-world benchmark problems, the resulting method, called SPIN, is shown to outperform competing approaches from the literature. Overall, this publication represents a very consistent and significant contribution to the field.

Publications II and III focus on the problem of predicting drug activities in different cancer cell lines. The main contribution of these papers is to address this problem as a multilabel classification task, with each label corresponding to a given cell line, while it was addressed so far only with single label methods (mostly SVM). Both publications exploit the max-margin conditional random fields (MMCRF) method. To be applicable, this method however requires an output graph on labels (ie., cell lines). Publication II evaluates output graphs derived from a variety of auxiliary datasets collecting information about the cell lines. Publication III proposes an original ensemble technique that trains separate MMCRF models on random output graphs and then aggregate their predictions through a simple majority vote (per label). The two approaches are compared against single label SVM on the NCI-Cancer dataset. Both multilabel approaches outperform significantly single label SVM and the ensemble method shows better performance than MMCRF trained with a graph derived from auxiliary data. Results in these papers clearly show the potential of multilabel approaches for drug activity prediction. As future work and in the sake of getting the best possible results for this application, an interesting next step would be to compare the MMCRF approaches with other (flat) multilabel approaches.

Publication IV extends the work in publication III by proposing two alternative aggregation operators for the random output graph ensemble methods, called AMM and MAM. Extensive experiments are carried out on 10 standard multilabel benchmark problems to compare the two aggregation operators as well as different ways to generate random output graphs and to assess the ensemble approach with respect to single label SVM and some other competing multilabel approaches. The comparison shows the superiority of the MAM approach with respect to other methods, although unfortunately the difference with respect to single label SVM does not appear to be statistically significant. An interesting theoretical analysis is also carried out that shows that the reconstruction error of the score function with MAM is guaranteed to be lower than the average reconstruction error of the individual base models. I think it would have been interesting to include in the comparison some related flat multilabel methods also based on the ensemble idea, such as for example the ensemble of classifier chains method (which generates random chains of labels, Read et al., 2009) or the random k-labelsets method (Tsoumakas and Vlahavas, 2007). The original ensemble framework proposed in publications III and IV nevertheless already appears as a plausible approach to handle multilabel problems with unknown output graph with structured output prediction methods.

Publication V addresses the problem of training a structured output prediction model with a complete output graph. This is a very appealing way to handle problems where the output graph is unknown, by letting the method learn from the data which output dependencies are useful through the optimization of the edge-associated weights. The paper addresses the intractability of the inference by first reformulating the score function for the complete graph as the expectation of the scores of random spanning trees and then by replacing this expectation with a finite sample estimate over a small number of random spanning trees. The paper gives theoretical guarantees on the performance obtained when jointly optimizing over a random sample of trees, in terms of both achievable margin and generalization error. An efficient algorithm with probabilistic guarantees is then derived to perform the inference jointly on an ensemble of trees. The resulting new multilabel method, called RTA, is shown to outperform MAM and other competitors on the same ten datasets as in publication IV. Overall, this paper is a very impressive piece of work and to me, it represents one of the most significant contributions of the thesis.

Overall assessment

From a scientific point of view, the work presented in the thesis is of very high quality. The main scientific outputs of the thesis are three original structured output prediction algorithms that are all motivated from real applicative needs and very well thought out from an algorithmic point of view. These algorithms are studied both theoretically and empirically in an extensive and fair way and their implementations have been made available to the community. The thesis work furthermore opens several very interesting directions for further investigation, both from an applicative and from a theoretical point of view.

The candidate's research has led to five peer-reviewed publications, four as first author and one as co-author. The contribution of the candidate in each publication is clearly explained in the manuscript and he seems to have had an instrumental role in all of them. These papers have been published in top machine learning venues (two of the most selective conferences, ICML and NIPS, and one of the best journals, Machine Learning) and in one good bioinformatics conference (PRIB). Overall, the publication record of the candidate is very good.

In conclusion, the thesis undoubtedly demonstrates the capacity of Hongyu Su to propose and validate novel ideas and to tackle both theoretical and practical questions in his research domain. In consideration of the overall quality of his work, I therefore recommend that **permission to publish the dissertation should be granted**.

Liège, January 7th, 2015

Pierre Geurts
Associate Professor,
University of Liège