Research Statement of Chun-Nam Yu

As a machine learning researcher my main research interest is in the area of structured output learning, especially large-margin approaches based on support vector machines and kernel methods. For applications, I am interested in machine learning problems arising from biological and medical data.

Structured Output Learning

Unlike in classification and regression where the output variables are simple binary or real values, in *structured output learning* the prediction outputs are complex objects such as trees, sequences, alignments, graph cuts, and other combinatorial structures. As an example of structured prediction task, given a set of 'golden' protein structure alignments produced by biologists, how can we learn to predict accurate sequence to structure alignments on new unknown sequences? As another example, given a set of bilingual documents with possible side information such as dictionaries, how can we make a computer learn to produce accurate and useful translations between the two languages? These structured prediction tasks are prevalent in natural language processing, computer vision, and computational biology. Knowing how to perform these tasks well is central to solving the problem of artificial intelligence.

At its core, structured output learning can be considered as the 'inverse' problem of combinatorial optimization, with the goal of learning good parameters to reproduce a given set of target output structures. The main research challenge in this area is to learn prediction models that can generalize well to unseen examples, and to learn them in a *time-efficient* and *label-efficient* manner. It is important to be time-efficient because combinatorial optimization algorithms are involved in learning, and label-efficient because acquiring labeled examples (e.g., protein alignments, bilingual translations) is much more costly when compared to classification or regression problems.

My thesis research made contributions to the theory and practice of structured output learning using *structural support vector machines* (structural SVMs) [5], a popular discriminative structured output learning algorithm. Structural SVMs [3] allow flexible features to be incorporated into the prediction model. As opposed to joint distribution estimation in generative models, it directly minimizes the prediction loss, giving state-of-art performance on many structured output learning tasks. As a large-margin method it also does not require the computation of a partition function for normalization as required in conditional random fields, making it very suitable for learning tasks where approximations are used for NP-hard decoding problems. However, the basic structural SVM algorithm cannot handle latent variables. For example, the body part labels in the problem of recognizing humans in an image, the syntactic structures (parse trees, part-of-speech tags) in machine translation, are usually not observed as part of the training examples given and are therefore latent. For such applications it is important to model these latent variables in order to build accurate structured output predictors. In my thesis I introduced the first algorithm for training *structural SVMs with latent variables* [9]. I gave conditions on the type of loss function that can be minimized effectively, and proposed an algorithm based on the concave-convex procedure to solve the training problem. The proposed algorithm and the associated software package have been used by various research groups to solve structured output prediction problems involving latent variables in computer vision and natural language processing, improving upon the state of the art on many tasks. Recently, I introduced another extension to *structural SVM that allows prior knowledge* about the learning task to be incorporated as soft constraints [6]. For example, these constraints can state that each sentence in a part-of-speech tagging output has to contain at least one noun and one verb. The constraints can also state that in the segmentation of a piece of text or image into parts, the parts labels have to be close to some known prior frequencies. I showed that through the use of these prior knowledge constraints the proposed algorithm is able to improve prediction accuracies when few labeled examples are available, and achieves equal or improved accuracies when compared to other state-of-art semi-supervised learning algorithms on several tasks.

There are several future directions that I want to explore as extensions to the above works. In many structured output learning tasks, there are usually additional side information available apart from the training input-output pairs. These include lexical resources such as dictionaries or a large unlabeled corpus in natural language processing, weakly labeled images from other sources in computer vision, and manually curated knowledge bases on gene ontology and protein pathways in computational biology. One way to incorporate such side information is to train generative models with it, and then include the model outputs as feature functions for the discriminative training of structured prediction models. Examples of such an approach include language modeling, generative kernels, and representation learning via deep belief nets. Alternatively, we can include this side information via prior knowledge constraints in our transductive structural SVM formulation [6]. I am interested in understanding which approach is more effective in exploiting such side information, and if there is any way to combine the strengths of both. In our work on latent structural SVM [9], we focused on improving the modeling flexibility and prediction accuracies through the introduction of latent variables. In addition to providing a more accurate modeling of the specific structured output prediction task, in many applications the latent variables have interesting structures themselves and can serve as an intermediate layer of representation. These include the syntactic structures in machine translation, and the body parts model or object parts model in human/object recognition. I am interested in exploring the use of these latent structures for multi-task learning and transfer learning. This can be seen as an analog of multi-task feature learning in classification, where the goal is to discover a common set of sparse features or low-dimensional subspace for accurate classification. Understanding how to exploit such common latent structure, and how to employ weak supervision from side information, are keys to improving the label efficiency of structured output learning.

In addition to reducing the labeled data requirement, I am also interested in improving the training time of structured output prediction models. The training of discriminative structured output prediction models requires repeated application of decoding to tune the parameters. Training time can become long when the training set is large, or when the decoding involves a combinatorial optimization problem that does not scale well with instance size (e.g., NP-hard inference). In the past I have worked on the application of the *cutting plane algorithm to speed up the training of structural SVMs* [2], and also improving the training time when *non-linear kernels* are used [8, 4]. The work in [4] also received an ECML best paper award. These works improve the training time by reducing the number of cutting plane computations in training, thus reducing the total number of decoding required. As future work I am interested in exploring the use of multiple approximate inference algorithms at different stages of optimization to speed up training. Given an NP-hard decoding problem, there are usually multiple approximate inference algorithms available (e.g., loopy belief propagation, linear programming relaxation, local search), with different accuracies and runtime at different regions of the parameter space. Such an approach will be similar to the portfolio-based approach used in many satisfiability problem solver such as SATzilla. Other directions that I am interested in exploring include parallelization and stochastic optimization.

In summary, I believe structured output learning is an important research area within machine learning because of two main reasons. First, from a practical point of view, advances in structured output learning have improved the application performance of many tasks in natural language processing, computer vision, and computational biology. While it is possible to study specific application problems in isolation in these fields, there are substantial benefits in studying learning problems from multiple domains together as general structured prediction tasks. Many ideas invented in one field can be quite readily transferred to another. Also, by considering multiple problems it becomes much easier to distinguish domain-specific optimizations from the underlying general learning principles. My preferred approach to studying structured output learning involves abstracting and generalizing from specific applications and observations in multiple domains, to see if there is a common algorithmic idea behind them. This is important because by abstracting away the essentials we can provide a framework that allows new applications to be modeled more rapidly. It also allows the limits of the learning framework to be revealed when there are new applications that cannot be modeled. Second, many important and recurrent questions in machine learning, including semi-supervised/multi-task/transfer/representation learning, can be viewed in a new light when considered in structured prediction settings. Many algorithms designed for these problems depend on theoretical models that treat data as point clouds in a vector space, and do not generalize to the structured prediction case. I believe we can obtain more complete answers to the above important questions through research in structured output learning.

Biomedical Informatics

Apart from designing algorithms for structured output prediction, I also work on specific machine learning applications, especially on biomedical applications of machine learning. Previously I have applied *structural SVMs to build accurate alignment models* between new protein sequences and known protein structures [10], an important intermediate step in template-based homology modeling of proteins. The application of structural SVMs significantly reduced the alignment error compared to several state-of-art models. During my postdoctoral training at the University of Alberta, I had the chance to meet and collaborate with faculty members from the biology department and the medical school, who had many interesting biomedical problems that can be solved with machine learning. In a collaboration with the Cross Cancer Institute at the University of Alberta, I have worked on the problem of improving the accuracy of prognostic models using information stored in electronic healthcare records. We developed a new method called *multi-task logistic regression* (MTLR) for *predicting personalized survival distributions for cancer patients* [7]. Unlike traditional regression models in survival analysis which focus on prognostic factor discovery, MTLR is designed to give accurate survival rate prediction for individual patients. It gives much more accurate survival rate prediction for a large cohort of patients drawn from the Alberta Cancer Registry, when compared to traditional survival regression models such as Cox or Aalen regressions. This improved survival time prediction can help doctors and cancer patients to make better decisions on treatment planning.

There are several ongoing projects that I am collaborating with PhD students in my postdoc supervisor's group. In one project we are designing supervised learning approaches to eliminate technical noise in DNA microarray data. This will allow the combination of data from multiple studies to obtain more accurate disease classifiers based on gene expression. Recently we submitted a related paper on a theoretical model on the expected overlap between gene expression signatures from different studies [1]. In another project we are working on transfer learning algorithms that takes feature value shifts into account, so that a prognostic model learned from a particular training patient population can be adapted more robustly to a test population with different patient composition. In addition to these projects, there are many other biomedical datasets that can benefit from improved analysis with machine learning, compared to the standard processing techniques currently used by biologists and medical practitioners. These biomedical data, including DNA microarray, metabolomic profiles and MRI scans, are usually high-dimensional and noisy, with many of them also structured. My research goal in this area is to develop machine learning algorithms to help analyze such data to assist the diagnosis of diseases and prognostic modeling for patients, and structured output learning will play an essential part in the analysis and integration of these data.

References

[1] B. Damavandi, C.-N. Yu, S. Damaraju, and R. Greiner. Explaining the gene signature anomaly: Estimating the overlap of two ranked lists. *submitted to ISMB 2012*.

[2] T. Joachims, T. Finley, and C.-N. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 76(4), 2009.

[3] T. Joachims, T. Hofmann, Y. Yue, and C.-N. Yu. Predicting structured objects with support vector machines. *Communications of the ACM, Research Highlight*, 52(11):97–104, 2009.

[4] T. Joachims and C.-N. Yu. Sparse kernel SVMs via cutting-plane training. *Machine Learning (ECML PKDD 2009 Special Issue)*, 76(2-3):179–193, 2009.

[5] C.-N. Yu. *Improved Learning of Structural Support Vector Machines: Training with Latent Variables and Non-linear Kernels*. PhD thesis, Cornell University, 2010.

[6] C.-N. Yu. Transductive learning of structural SVMs via prior knowledge constraints. In *AISTATS*, 2012.

[7] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *NIPS*, 2011.

[8] C.-N. Yu and T. Joachims. Training structural SVMs with kernels using sampled cuts. In *KDD*, 2008.

[9] C.-N. Yu and T. Joachims. Learning structural SVMs with latent variables. In *ICML*, 2009.

[10] C.-N. Yu, T. Joachims, R. Elber, and J. Pillardy. Support vector training of protein alignment models. In *RECOMB*, 2007.