NSF-ITR/IM PROJECT: 2002 Abstracts

From Bits to Information:
Statistical Learning Technologies for Digital Information Management Search
(Brown University, M.I.T, Oregon State University, and University of Illinois at Urbana-Champaign)

Brown University:

Project Title: Support Vector Machines for Multiple Instance Learning
Participants: Stuart Andrews and Thomas Hofmann
Abstract: Multiple Instance Learning (MIL) is an important generalization of standard supervised binary classification. In MIL labels are not available for individual training patterns, but are associated with sets of patterns, which introduces additional uncertainty and challenges for machine learning algorithms. In contrast to much of the previous work that has focused on special purpose learning architectures, we have generalized competitive classification algorithms, namely Support Vector Machines and the kernel perceptron, to deal with the MIL problem. We are currently investigating applications in text categorization and multimedia retrieval, by representing documents as sets of passages and images as sets of regions/objects, respectively.

Project Title: Polycategorical Categorization for Personalized Information Filtering
Participants: Ioannis Tsochantaridis and Thomas Hofmann
Abstract: Polycategorical categorization is an extension of standard classification in which items are labeled by multiple binary labels. We are particularly interested in cases with large numbers of overlapping categories and a priori unknown dependencies between labels. The main application of this approach is in personalized information filtering, where our approach addresses the fundamental problem of combining collaborative and content-based filtering. To this extend, we have developed a generalization of Transductive Support Vector Machines that is able to incorporate probabilistic labeling information derived from a statistical inter-label dependency model. This technique has shown to result in improvements on standard collaborative filtering benchmarks.

Project Title: Learning Stochastic Models of the Web
Participants: David Gondek and Thomas Hofmann
Abstract: Stochastic models of hypertext repositories are important for gaining insight in the way content is organized and linked on the Web, which in turn is highly relevant for various types of information access (search, categorization, filtering, etc.) and information gathering. We are investigating stochastic graph models which center around the notion of Web communities and which combine hyperlink analysis with semantic analysis of textual document content. We are currently developing an intelligent Web crawler that uses statistical Web models to predict the content of an unseen Web page with partially known in-links. The expected content is predicted and evaluated with a text categorizer to derive a score, which is in turn used to organize a priority cue for focused crawling. Current experiments use crawls of the Internet Archive as well as the TREC Web WT10g data for evaluation.

Project Title: Semantic Features for Document Categorization & Retrieval
Participants: Lijuan Cai and Thomas Hofmann
Abstract: This project investigates the use of semantic document representations, derived from unsupervised learning techniques as well as from available semantic knowledge bases like WordNet to improve document categorization and information retrieval. In particular, we will focus on the use of boosting techniques to combine different types of features based on terms, phrases and semantic concepts. We are currently comparing and benchmarking different techniques on the TREC9 filtering task.

Massachusetts Institute of Technology:

Project Title: Improvements in Multi-class Document Classification
Participants: Jason Rennie and Ryan Rifkin
Abstract: The problem of multi-class document classification arises in many applications, including the development of automatic email response systems, and the automated development of structured document hierarchies. We hope to improve the state of the art in multi-class document classification by combining extremely powerful binary classifiers, such as Support Vector Machines, in novel ways; specifically, we seek to extend and improve known methods for combining classifiers that are derived from the theory of error correcting codes.
Project Title: Fast Multiclass Document Classification Using RLSC  
**PI:** T. Poggio  
**Participants:** Ryan Rifkin  
**Abstract:** It has already been shown that multiclass SVMs in a one-vs-all or ECOC framework represent a highly robust approach to multiclass document classification. In this project, we show that similar results can be obtained using the much simpler RLSC (Regularized Least-Squares Classification) in place of the SVM. Whereas training an SVM requires the solution of a quadratic programming problem, training RLSC requires only the solution of a single system of linear equations. By taking advantage of the linear kernel and the sparse structure of the data, we obtain an algorithm that is more than an order of magnitude faster than the SVM on large datasets, while providing essentially identical accuracy.

Project Title: Component-based Face Detection  
**PI:** T. Poggio  
**Participants:** Bernd Heisele and Thomas Serre  
**Abstract:** We present a component-based, trainable system for detecting frontal and near-frontal views of faces in still gray images. The system consists of a two-level hierarchy of Support Vector Machine (SVM) classifiers. On the first level, component classifiers independently detect components of a face. On the second level, a single classifier checks if the geometrical configuration of the detected components in the image matches a geometrical model of a face. We propose a method for automatically learning components by using 3-D head models. This approach has the advantage that no manual interaction is required for choosing and extracting components. Experiments show that the component-based system is significantly more robust against rotations in depth than a comparable system trained on whole face patterns.

Project Title: Algorithms for Learning from Partially Labeled Data  
**PI:** T. Poggio and T. Jaakkola  
**Participant:** Martin Szummer  
**Abstract:** The partially labeled learning problem is studied from both a theoretical and practical perspective. We examine principled means of exploiting partially labeled data, and formulate a set of algorithms that follow these principles. Applications include text classification, object detection and recognition and classification of bioinformatics data.

Project Title: Machine Learning for Text Summarization  
**PI:** T. Poggio  
**Participants:** Osamu Yoshimi and Luis Pérez-Breva  
**Abstract:** Text summarization is often considered as the twofold problem of sentence extraction and summary evaluation. We study text summarization within the formal framework of machine learning that maps these problems into feature selection and model selection respectively. An initial implementation of this framework working with Reuters news stories uses HMM to infer the meaning of words, and extracts the sentences closest to the headline. A multiple non-exclusive hierarchical classification task determines the best model. Preliminary results suggest already that summaries may be a powerful alternative to full text for low level of detail classification task, which confirms the human intuition about the use of summaries.

Project Title: Semantic Inference for Text Representation  
**PI:** T. Poggio  
**Participant:** Luis Pérez-Breva  
**Abstract:** One of the main problems in information retrieval is Representation. The lack of a quantitative counterpart to textual information prevents researchers from applying machine learning techniques already being used in a variety of other fields. We study a probabilistic model of semantic information in a document that shall allow us to combine the power of spectral clustering (aka Latent Semantic Analysis) and Bayesian Inference. Thus incorporate measures of word similarity beyond simple word match, and allow enhancements to textual measures: TF-IDF, Bag of Words, ... by incorporating semantic information. Increase the amount of information made available to existing machine learning techniques.

Project Title: Applying Partially Labeled Data towards Object Recognition  
**PI:** T. Poggio  
**Participant:** Andy Crane  
**Abstract:** We are exploring various algorithms and their performance on partially labeled data sets. From the unlabeled samples, we attempt to create generative features and use these in discriminative classifiers. The primary application is object recognition and detection.

Project Title: Classification of Yahoo News from Images and Captions  
**PI:** T. Poggio  
**Participant:** Alexandros Kyriakides and Giorgos Zacharia  
**Abstract:** Perform classification of images obtained from news articles. The images are assigned to one or more topic categories or classes. We will be performing multi-class, multi-label classification. In multi-class classification, as distinguished from binary classification, there are more than two classes. In multi-label classification each image may have more than one class label. For example, a single picture of the U.S. President on a visit to Greece, could fall into each of the following categories: *U.S. President, Greek politics, U.S. politics.*
Abstract: Many current text classification algorithms operate on a bag-of-words representation of documents. However, such a form of input representation discards a great deal of information that is potentially useful in document classification (such as layout information). We tested whether people use such information in document classification and whether they show differences in classification performance as a function of text presentation (fully formatted text vs. bag-of-words). Results indicate that performance on the full text presentation is only slightly better than on the bag-of-words presentation. This suggests that people very rapidly construct semantic representations of text, even if no structural information about the text is available. One plan for further research is to investigate the sensitivity of people to statistical patterns in language. It is possible, for instance, that people only pay attention to less frequent words when characterizing a text and distinguishing it from others (i.e. classifying it).

Project Title: Informaton Extraction from Financial News  
Participant: Giorgos Zacharia  
Abstract: Extracting automatically relevant information from wire news is of increasing importance in a variety of areas. We are developing state of the art learning techniques to extract information from on-line CNN financial news.

Project Title: Auxilary Information and View-based Classification  
Participant: Alex Rakhlin  
Abstract: At the training time we often have more information about the training set than we use. For example, for face detection we might have an extra label of whether the training face is a frontal or a profile view. This extra information is not available at the testing time. In this project we study how this auxiliary label can be used to achieve better classification performance. In the case of learning faces vs. non-faces, weighted sum of two SVMs (frontal and profile) seems to perform better than other methods (e.g. one-vs.-all SVM).

Project Title: Object Detection with Contextual Information  
Participant: Alex Rakhlin  
Abstract: Contextual information is necessary for robust detection of objects in scenes. Most of the time detection of one object is not independent from detection of another object, especially if they are parts of some more complex object. If we build separate detectors for each of the simpler objects, this dependence information needs to be integrated in a probabilistic model. Graphical models seem an appropriate framework for this.

Project Title: Experimental Characterization of Human Text Classification Performance  
Participants: Pawan Sinha and Florian Wolf  
Abstract: Many current text classification algorithms operate on a bag-of-words representation of documents. However, such a form of input representation discards a great deal of information that is potentially useful in document classification (such as layout information). We tested whether people use such information in document classification and whether they show differences in classification performance as a function of text presentation (fully formatted text vs. bag-of-words). Results indicate that performance on the full text presentation is only slightly better than on the bag-of-words presentation. This suggests that people very rapidly construct semantic representations of text, even if no structural information about the text is available. One plan for further research is to investigate the sensitivity of people to statistical patterns in language. It is possible, for instance, that people only pay attention to less frequent words when characterizing a text and distinguishing it from others (i.e. classifying it).

Project Title: Representation and Recognition of Shapes  
Participant: Pedro F. Felzenszwalb  
Abstract: We are developing a new representation for deformable shapes. Besides describing shapes in a natural and precise manner, our representation can characterize non-rigid shapes at multiple levels of specificity. The representation yields fast algorithms for various computational tasks in image analysis. For example, we have developed an algorithm that can efficiently localize a specific deformable shape in an image. A second algorithm can be used to detect potential objects in images without using a specific shape model. We believe these algorithms will be useful for content-based image classification and retrieval.

Project Title: Continuation Methods for Mixing Heterogeneous Sources  
Participant: Adrian Corduneanu  
Abstract: A number of modern learning tasks involve estimation from heterogeneous information sources. This includes classification with labeled and unlabeled data as well as other problems with analogous structure such as competitive (game theoretic) problems. The associated estimation problems can be typically reduced to solving a set of fixed point equations (consistency conditions). We introduce a general method for combining a preferred information source with another in this setting by evolving continuous paths of fixed points at intermediate allocations. We explicitly identify critical points along the unique paths to either increase the stability of estimation or to ensure a significant departure from the initial source. The homotopy continuation approach is guaranteed to terminate at the second source, and involves no combinatorial effort. We illustrate the power of these ideas in text classification with labeled and unlabeled data.

Project Title: The Bundled-SVM: Bridging the Gap Between the SVM and Naive Bayes  
Participants: Kai Shih, Yu-Han Chang, Jason Rennie and David Karger  
Abstract: The Support Vector Machine (SVM) typically outperforms other algorithms on text classification problems, but requires super-linear training time (in the number of training documents). In contrast, linear time algorithms like Naive Bayes have lower performance, but can easily handle huge training sets. In this project, we have developed a technique that creates a continuum of classifiers between the SVM and a Naive Bayes-like algorithm. This allows the user to trade-off accuracy for faster training. Along this continuum is a classifier that scales linearly, but, in
experiments, performs better than both Naive Bayes and using the SVM with a subsampled data set. Through theoretical arguments and experimental results, we show why our algorithm should be preferred to data-reduction techniques such as subsampling.

Project title: A Better Model for Text Classification
Participants: Kai Shih, Jaime Teevan, Jason Rennie and David Karger
Abstract: Discriminative classification techniques, such as the Support Vector Machine (SVM) have received much attention recently for text and other classification problems. Probabilistic models, such as multinomial Naive Bayes, perform poorly in comparison, especially on data sets with an uneven distribution of training examples. In this project, we show that a major reason for the poor performance is the poor match of the multinomial to text feature distributions. We find that the feature distribution is better matched to a power-law distribution. This motivates taking logs of term frequencies; experiments show great improvements over the regular multinomial model. Furthermore, the multinomial performs poorly when some classes have small training sets compared to other classes. We find that treating classification for such classes as an anomaly detection problem can greatly improve performance. A combination of these two techniques gives us performance on standard text data sets nearing that of the SVM.

Oregon State University:

Project Title: Sensitivity of Spatial Metrics to Pixel Misclassification Rate
Participants: William Langford and Tom Dietterich
Abstract: In many image-processing applications, individual pixels are classified and then the classified pixels are merged to yield regions or segments. We are studying the relationship between errors at the pixel level and errors in the resulting regions. Our application problem is drawn from landscape ecology, where regions are characterized by metrics such as the patch compaction index, which measures the extent to which an spatial region is approximately square. A common assumption has been that the per-pixel misclassification rate is proportional to the error rate in the patch compaction. We showed that this is not true in general, because pixel errors can lead to the fusing of regions in the image, with resulting catastrophic changes in the patch compaction index. We also showed that smoothing of the pixel classifications, while it can reduce per pixel error rates, can increase the error in spatial landscape metrics.

Project Title: Supervised Reinforcement Learning
Participants: Xin Wang and Tom Dietterich
Abstract: In earlier work, Wei Zhang showed that reinforcement learning could be applied to learn application-specific evaluation functions for combinatorial optimization problems in industry, specifically, resource-constrained scheduling in NASA's space shuttle program. We are exploring ways to generalize his results to the problem setting that we call Supervised Reinforcement Learning. In this setting, a series of training examples is provided for a sequential decision-making task (such as scheduling). The goal is to learn an evaluation function that performs well on these examples and that also generalizes well to new instances. We have developed new kernel-based reinforcement learning algorithms for this setting and applied them to problems of scheduling and deterministic control. However, these methods are not fully satisfactory, so we are exploring a new model-based policy gradient method. In this method, a model of the reinforcement learning problem is acquired through experimentation, the behavior of a parameterized policy on this model is then employed to compute the gradient of the policy performance with respect to its parameters.

Project Title: Content-Based Image Retrieval of Herbarium Samples
Participants: Tom Dietterich, Dan Forrest and Ashit Gandhi
Abstract: Imagine you are hiking in the forest and you see an unusual plant. You want to know the genus and species of this plant, so you clip off a leaf, take it home, and scan it into your computer. Now you want to retrieve images of similar plants from online plant databases such as those being constructed by the Missouri Botanical Garden, the NY Botanical Garden, and the Oregon State Herbarium. We are developing an initial test bed for this problem based on 6 species of maples and oaks native to Oregon. We have developed a shape-based matching algorithm that borrows dynamic programming methods from DNA sequence matching and applies them to matching the sequence of local curvatures of the shape of a leaf. This allows us to match partial, overlapping, and occluded leaves successfully.

Project Title: Machine Learning for Sequential and Spatial Data
Participants: Tom Dietterich, Adam Ashenfelter, Saket Joshi
Abstract: The goal of this project is to build off-the-shelf algorithms for classifying sequences and spatial objects (e.g., pixels Markov Random Fields, require substantial hand-tweaking to work well in new applications. We seek discriminative methods that work well across a wide range of applications without manual tuning. Our current focus is on ensembles of recurrent classifiers. These make minimal modeling assumptions and can be trained and evaluated efficiently. We are incorporating this approach into the R public domain statistical package.
Project Title: **Sub-pixel Classification of Remote Sensed Images**  
**PI:** T. Dietterich  
**Participants:** Tom Dietterich and Diane Damon  
**Abstract:** Many satellite-based instruments collect earth-surface images at coarse resolution (e.g., one pixel = 1km square or 0.5km square). Training data is available at higher resolutions, and sub-pixel classification attempts to predict the fraction of each pixel that belongs to various land-cover classes (forest, grassland, are soil, crops, city, etc.). We are exploring new regression tree and kernel based methods for sub-pixel classification. We are also exploring ways that spatial information can be applied to improve the accuracy of these classifications.

Project Title: **Multiclass Classification with Support Vector Machines**  
**PI:** T. Dietterich  
**Participants:** G. Valentini (Univ. of Genoa) and Tom Dietterich  
**Abstract:** Support vector machines are binary classifiers. To handle multiple classes, some method is required to decompose the multiclass problem into a set of two-class problems that can be solved individually or collectively. We are exploring methods based on our earlier work on error-correcting output coding (ECOC). ECOC is an ensemble method, and it works primarily by learning multiple versions of each segment of the multiclass decision boundary. Each version of the segment is learned in the context of different other segments, which leads to diversity and, hence, to variance reduction through voting. The fundamental question is whether support vector machines can benefit from this (or any other) form of ensemble learning. We have conducted a bias/variance analysis of SVMs which shows that SVMs trained on small samples can exhibit substantial variance, which suggests that variance-reduction methods, such as bagging and ECOC, will be able to improve SVM performance. The studies also show that for Gaussian kernels, the most important tuning parameter is the scale of the Gaussians and that good performance is obtained over a wide range of values.

**University of Illinois at Urbana-Champaign:**

Project Title: **Intermediate Knowledge Representations that Facilitate Learning**  
**PI:** D. Roth  
**Participants:** Dav Zimak, Chad Cumby and Shivani Agarwal  
**Abstract:** Learning becomes easy once the correct input representation has been chosen, for example, one that produces linearly separable point sets. We have several projects in the direction of: (1) automatically generating intermediate representations to aid supervised learning algorithms, (2) developing methods that allow the use of relational representations and of learning relational definitions, and (3) developing a flexible knowledge representation language that can be used along with feature efficient learning algorithms. We study applications of this general knowledge representation paradigm in the context of learning in the natural language domain (e.g., information extraction) and visual recognition. (4) Developing kernels for Boolean functions and relational functions and studying computational complexity of algorithms that use kernels.

Project Title: **Inference with Classifiers**  
**PI:** D. Roth  
**Participants:** Dan Roth and Vasin Punyakanok  
**Abstract:** In many situations it is necessary to make decisions that depend on the outcomes of several different classifiers in a way that provides a coherent inference that satisfies some constraint. These constraints might arise from the sequential nature of the data or other domain specific constraints. We study several general approaches to this problem and are evaluating those in the context of inference problems in natural language -- identifying phrase structure and question-answering. The approaches studied are: (1) A Markovian approach that extends standard HMMs to allow the use of a rich observation structure and of general classifiers to model state-observation dependencies. We study both generative and conditional models. (2) Extensions of constraint satisfaction formalisms. Currently the focus is on developing hierarchical models. (3) Markov Random fields. We study a more general model in which constraints of more general structures can be developed.

Project Title: **Learning Coherent Concepts**  
**PI:** D. Roth  
**Participants:** Ashutosh Garg and Vasin Punyakanok  
**Abstract:** This research seeks to develop an integrated view - theoretical understanding, algorithms development and experimental evaluation - for learning coherent concepts. These are learning scenarios that are common in cognitive learning - where multiple learners co-exist and may learn different functions on the same input, but there are mutual compatibility constraints on their outcomes. Our effort will consist of developing a learning theory for these situations and of studying algorithmic ways to exploit them in natural language inferences. The theoretical study concentrates on developing a semantics for the coherency conditions and study it from a learning theory point of view. The goal is to understand in what ways does learning become easier and more robust in these situations. The algorithmic study concentrates on developing ways to exploit coherency and makes use of several important problems in natural language processing as a testbed for investigating chaining of coherent classifiers and inferences that rely on the outcomes of several classifiers.
Project Title: **Constraint Classification: A New Approach to Multiclass Classification**  
**PI:** D. Roth  
Participants: Dav Zimak and Yair Even-Zohar  
Abstract: We develop a new view of multiclass classification and introduce the constraint classification problem, a generalization that captures many flavors of multiclass classification. In particular, our framework captures multiclass classification, ranking problems and multilable classification and winner-take-all (WTA) algorithms. We study both algorithmic issues and theoretical issues such as sample bounds.Algorithmically, based on our view, we develop a learning algorithm that learns via a single linear classifier in high dimension and can also be viewed as a network of properly trained linear classifiers in a low dimension. We also study distribution independent bounds for many multiclass-learning algorithms, including winner-take-all (WTA), as well as margin-based generalization bounds.

Project Title: **Learning Sparse Representations for Object Detection**  
**PI:** D. Roth  
Participants: Shivani Agarwal and Ashutosh Garg  
Abstract: We study an approach for learning to detect objects in still gray images that is based on a sparse, part-based representation of objects. A vocabulary of information-rich object parts is automatically constructed from a set of sample images of the object class of interest. Images are then represented using parts from this vocabulary, along with spatial relations observed among them. Based on this representation, a feature-efficient learning algorithm is used to learn to detect instances of the object class. The framework developed can be applied to any object with distinguishable parts in a relatively fixed spatial configuration. So far we have experimented on images of side views of cars. Our experiments show that the method achieves a high-detection accuracy on a difficult test set of real-world images, and is highly robust to partial occlusion and background variation.