NSF-ITR/IM PROJECT

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: **Polycategorical Categorization for Personalized Information Filtering**  
Pl: T. Hofmann  
Participants: Ioannis Tsochandaritis 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. We assume that a population of users rate subsets of documents from a shared repository according to their interests and preferences. Each user may label/rate documents in a different manner, but users will typically have interests in common. Currently there are two methods available: (i) collaborative filtering, which ignores the content and models the dependency structure between user ratings to make predictions, (ii) information filtering which solves the classification for each user independently by ignoring dependencies between user ratings. We propose to combine both methods by learning a prior dependency model for ratings and then use this model to generate probabilistic labels for missing ratings. The labeled data as well as the probabilistically labeled data are then used as input for a kernel-based classifier.

Project Title: **Learning Stochastic Models of the Web**  
Pl: T. Hofmann  
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.). 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. Current experiments use crawls of the Internet Archive as well as TREC Web data for evaluation. We are also looking into ways of learning semantic kernels for documents.

Project Title: **Hierarchical Document Categorization**  
Pl: T. Hofmann  
Participant: Thomas Hofmann  
Abstract: Most document categorization problems involve large numbers of categories that are organized in a taxonomy or category lattice. We investigate ways to take advantage of a hierarchy or lattice in order to improve classification accuracy. Methods considered include Support Vector Machines as well as naïve Bayes classifiers.

Massachusetts Institute of Technology:

Project Title: **Classification of Audio and Video Events using Text Sources as Supervision**  
Pl: E. Grimson  
Participant: Chris Stauffer  
Abstract: We have previously shown the efficacy of unsupervised video monitoring in video surveillance. This project seeks to extend those capabilities in three ways. First, we intend to add audio event classification and anomaly detection. Second, we intend to include many different sources of audio/visual data, including indoor surveillance, outdoor surveillance, and broadcast television. The final aspect of this project is to add supervision. Since it is impractical to consider manually labeling these data streams, we are investigating other sources of supervision, including web schedules for classifying general conference room use, program descriptions from TV Guide for classification of program type, and closed caption text for classification of scene types or scene elements. These sparse, noisy, incomplete, and sometimes incorrect sources of supervision raise multiple difficulties in classification.

Project Title: **Audio, Visual and Closed-Captioning Analysis**  
Pl: E. Grimson  
Participants: Chris Stauffer and Huizhen Yu  
Abstract: Closed-captioning is a relatively new and interesting source of information that relates to the audio and video content of movies, television programs, and newscasts. Closed captioning is free and abundant and it contains information about what is said, who is saying it, and even indirect information about other content in the A/V stream. As a natural link to conversational language, CC can augment the temporal analysis of audio and video sequences in a semi-supervised fashion that may be more powerful than manual labeling. We will investigate deriving a rich representation from A/V/CC sources and applying it to a variety of tasks such as modeling, clustering and data mining. We will also explore transferring to situations where closed captioning is not available.
Abstract: We present a component-based, trainable system for detecting frontal and near-frontal views of faces in 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.

Project Title: A Component-based Approach to Face Detection
Participant: Thomas Serre
Abstract: We present an approach for image retrieval using a very large number of highly selective features and efficient online learning. Our approach is predicated on the assumption that each image is generated by a sparse set of visual "causes" and that images which are visually similar share causes. We propose a mechanism for computing a very large number of highly selective features, which capture some aspects of this causal structure (in our implementation there are over 46,000 highly selective features). At query time a user selects a few example images, and a technique known as "boosting" is used to learn a classification function in this feature space. By construction, the boosting procedure learns a simple classifier, which only relies on 20 of the features. As a result a very large database of images can be scanned rapidly, perhaps a million images per second. Finally, we will describe a set of experiments performed using our retrieval system on a database of 3000 images.

Project Title: Boosting Image Retrieval
Participant: Kinh Tieu
Abstract: We present an approach for image retrieval using a very large number of highly selective features and efficient online learning. Our approach is predicated on the assumption that each image is generated by a sparse set of visual "causes" and that images which are visually similar share causes. We propose a mechanism for computing a very large number of highly selective features, which capture some aspects of this causal structure (in our implementation there are over 46,000 highly selective features). At query time a user selects a few example images, and a technique known as "boosting" is used to learn a classification function in this feature space. By construction, the boosting procedure learns a simple classifier, which only relies on 20 of the features. As a result a very large database of images can be scanned rapidly, perhaps a million images per second. Finally, we will describe a set of experiments performed using our retrieval system on a database of 3000 images.

Project Title: Maximum Likelihood with Partially Labeled Data for Classification
Participants: Adrian Corduneanu and Tommi Jaakkola
Abstract: When we impose suitable constraints on classification models, unlabeled data provide, in addition to labeled data alone, information about class distributions, which could be used to improve the learning of the model. The benefit of using unlabeled data along with labeled data to train classification models may be significant, especially for tasks in which labeling is time consuming and requires expensive expert knowledge. The simplest way to make use of unlabeled data is to incorporate it into the likelihood criterion, which can be maximized by a suitable EM algorithm. Previous experiments show that maximum likelihood with partially labeled data does improve classification in some situations, but can also degrade it significantly in others. The purpose of this project is to identify in a principled way model families for which maximum likelihood with partially labeled data improves classification, and to quantify the importance of unlabeled data in such situations. In addition, we want to point out the problems of maximum likelihood with partially labeled data that can arise under other model constraints, and to exhibit improved criteria and algorithms that address them. Models considered include mixture of normals, Naive Bayes for web document classification, leading to general graphical models with missing nodes.

Project Title: Active Information Retrieval
Participant: Tommi Jaakkola
Abstract: We propose a new approach to effective information retrieval. The user is successively queried for distinctions at varying levels of abstraction and is permitted to respond with multiple selections or may choose not to respond. The information is in each case unambiguously interpreted and incorporated by the system. The next query is chosen optimally to minimize the need for any further exchange. The system is also capable of determining whether or not the document of interest is in the (portion of the) database being consulted. In more technical terms, we make use of a stochastic substitution matrix that is derived, e.g., from local relations among the documents in the database. The notion of substitution permits us to define what is meant by the optimal interaction as well as construct algorithms to achieve it. All the queries are carried out under resource constraints, which pertain either to the amount of information presented to the user per iteration or to the maximum/average number of iterations that are allowed. The effect of the resource constraints can be assessed prior to the interaction by using the properties of the substitution matrix as well as bounds on the successive reduction in the uncertainty about the information being sought.

Project Title: Information Extraction from Financial News
Participants: Luis Perez-Breva and 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: 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. Preliminary results indicate that this approach has promise.

Project Title: Maximum Likelihood with Partially Labeled Data for Classification
Participants: Adrian Corduneanu and Tommi Jaakkola
Abstract: When we impose suitable constraints on classification models, unlabeled data provide, in addition to labeled data alone, information about class distributions, which could be used to improve the learning of the model. The benefit of using unlabeled data along with labeled data to train classification models may be significant, especially for tasks in which labeling is time consuming and requires expensive expert knowledge. The simplest way to make use of unlabeled data is to incorporate it into the likelihood criterion, which can be maximized by a suitable EM algorithm. Previous experiments show that maximum likelihood with partially labeled data does improve classification in some situations, but can also degrade it significantly in others. The purpose of this project is to identify in a principled way model families for which maximum likelihood with partially labeled data improves classification, and to quantify the importance of unlabeled data in such situations. In addition, we want to point out the problems of maximum likelihood with partially labeled data that can arise under other model constraints, and to exhibit improved criteria and algorithms that address them. Models considered include mixture of normals, Naive Bayes for web document classification, leading to general graphical models with missing nodes.

Project Title: Active Information Retrieval
Participant: Tommi Jaakkola
Abstract: We propose a new approach to effective information retrieval. The user is successively queried for distinctions at varying levels of abstraction and is permitted to respond with multiple selections or may choose not to respond. The information is in each case unambiguously interpreted and incorporated by the system. The next query is chosen optimally to minimize the need for any further exchange. The system is also capable of determining whether or not the document of interest is in the (portion of the) database being consulted. In more technical terms, we make use of a stochastic substitution matrix that is derived, e.g., from local relations among the documents in the database. The notion of substitution permits us to define what is meant by the optimal interaction as well as construct algorithms to achieve it. All the queries are carried out under resource constraints, which pertain either to the amount of information presented to the user per iteration or to the maximum/average number of iterations that are allowed. The effect of the resource constraints can be assessed prior to the interaction by using the properties of the substitution matrix as well as bounds on the successive reduction in the uncertainty about the information being sought.

Project Title: Information Extraction from Financial News
Participants: Luis Perez-Breva and 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: 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. Preliminary results indicate that this approach has promise.

Project Title: Maximum Likelihood with Partially Labeled Data for Classification
Participants: Adrian Corduneanu and Tommi Jaakkola
Abstract: When we impose suitable constraints on classification models, unlabeled data provide, in addition to labeled data alone, information about class distributions, which could be used to improve the learning of the model. The benefit of using unlabeled data along with labeled data to train classification models may be significant, especially for tasks in which labeling is time consuming and requires expensive expert knowledge. The simplest way to make use of unlabeled data is to incorporate it into the likelihood criterion, which can be maximized by a suitable EM algorithm. Previous experiments show that maximum likelihood with partially labeled data does improve classification in some situations, but can also degrade it significantly in others. The purpose of this project is to identify in a principled way model families for which maximum likelihood with partially labeled data improves classification, and to quantify the importance of unlabeled data in such situations. In addition, we want to point out the problems of maximum likelihood with partially labeled data that can arise under other model constraints, and to exhibit improved criteria and algorithms that address them. Models considered include mixture of normals, Naive Bayes for web document classification, leading to general graphical models with missing nodes.

Project Title: Active Information Retrieval
Participant: Tommi Jaakkola
Abstract: We propose a new approach to effective information retrieval. The user is successively queried for distinctions at varying levels of abstraction and is permitted to respond with multiple selections or may choose not to respond. The information is in each case unambiguously interpreted and incorporated by the system. The next query is chosen optimally to minimize the need for any further exchange. The system is also capable of determining whether or not the document of interest is in the (portion of the) database being consulted. In more technical terms, we make use of a stochastic substitution matrix that is derived, e.g., from local relations among the documents in the database. The notion of substitution permits us to define what is meant by the optimal interaction as well as construct algorithms to achieve it. All the queries are carried out under resource constraints, which pertain either to the amount of information presented to the user per iteration or to the maximum/average number of iterations that are allowed. The effect of the resource constraints can be assessed prior to the interaction by using the properties of the substitution matrix as well as bounds on the successive reduction in the uncertainty about the information being sought.

Project Title: Information Extraction from Financial News
Participants: Luis Perez-Breva and 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: 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. Preliminary results indicate that this approach has promise.

Project Title: Maximum Likelihood with Partially Labeled Data for Classification
Participants: Adrian Corduneanu and Tommi Jaakkola
Abstract: When we impose suitable constraints on classification models, unlabeled data provide, in addition to labeled data alone, information about class distributions, which could be used to improve the learning of the model. The benefit of using unlabeled data along with labeled data to train classification models may be significant, especially for tasks in which labeling is time consuming and requires expensive expert knowledge. The simplest way to make use of unlabeled data is to incorporate it into the likelihood criterion, which can be maximized by a suitable EM algorithm. Previous experiments show that maximum likelihood with partially labeled data does improve classification in some situations, but can also degrade it significantly in others. The purpose of this project is to identify in a principled way model families for which maximum likelihood with partially labeled data improves classification, and to quantify the importance of unlabeled data in such situations. In addition, we want to point out the problems of maximum likelihood with partially labeled data that can arise under other model constraints, and to exhibit improved criteria and algorithms that address them. Models considered include mixture of normals, Naive Bayes for web document classification, leading to general graphical models with missing nodes.
**Project Title:** Reinforcement Learning for Active Learning  
**PI:** T. Poggio  
**Participant:** Christian Shelton  
**Abstract:** The problem of active learning (picking which examples to have labelled) can be thought of as a reinforcement learning problem. The goal is to pick a sequence of examples (using information gained after each choice) that results in the best reward (correct classification with the fewest examples needed). The success of active learning depends on two things. The first is the behavior of the classifier. The second is the domain of the problem. The former is best modeled explicitly since we have control over the classifier and understand its performance. The second can be learned from examples. We assume that we will be solving similar problems (in the same domain) repeatedly. In such situations, it is beneficial to learn about the interaction of the classifier and active learning with the learning setting. The key to using reinforcement learning is not ignoring our knowledge about the behavior of the classifier. We have already demonstrated how the incorporation of environment knowledge can be used in the case of memory (AI-Memo 2001-002). We will extend that result here to the case of active learning.

**Project Title:** Principles for Learning from Partially Labeled Data  
**PI:** T. Poggio and T. Jaakkola  
**Participant:** Martin Szummer  
**Abstract:** The partially labeled data problem is studied from both a theoretical and practical perspective. We try to uncover the fundamental properties of unlabeled data that can be exploited, and develop and test algorithms using those properties. Applications include text classification and diagnosis from bioinformatics data.

**Project Title:** Classification of Yahoo News from Images and Captions  
**PI:** T. Poggio  
**Participant:** Giorgos Zacharia  
**Abstract:** Classifying images of news articles allows us to automatically select the right images for future articles. We are developing learning techniques that combine classifiers trained the images themselves, on the text of the articles, the meta information of the html tags of the articles, and the image captions.

**Project Title:** Experimental Characterization of Human Text Classification Performance  
**PI:** P. Sinha  
**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). It is thus reasonable to assume that the performance limits of current document classification algorithms are at least partly due to the fact that they use a very incomplete form of representation of documents. This project aims at determining what information humans rely on in document classification, given different classification tasks, different types of documents, etc. One goal, for instance, is to determine the usefulness of layout information to humans, compared to content information. With respect to developing document classification algorithms, using layout information could be particularly interesting, since it can relatively easily be automatically extracted from documents (using text formatting tags, for instance). The methods used in this project will include behavioral tests, such as document classification under various time constraints, as well as eye movement studies. Studying document classification under time constraints could allow conclusions as to what kinds of information humans rely on in “shallow” vs. “deep” document processing. Eye movement studies could help determine more exactly what portions of a document humans take into account in classification tasks. Such a paradigm would also allow controlling document display depending on a subject’s eye movements to certain regions in a document. Thus, it would allow for more precise manipulations of the amount of information that is available to accomplish the classification tasks.

**Oregon State University:**

**Project Title:** A Reinforcement Learning Approach to Image Segmentation  
**PI:** T. Dietterich  
**Participants:** William Langford and Tom Dietterich  
**Abstract:** Algorithms for image segmentation require extensive tuning for each new application. We are studying methods for automating this tuning via reinforcement learning. We are applying bottom-up region merging methods to aerial images of forests, deserts, and coastlines to segment them for applications in landscape ecology. Given a small number of images that have been hand-segmented according to an application-specific criterion, we attempt to learn a value function that can predict which region-merging decisions will lead to the best final outcome. This work is also leading to new methods for reinforcement learning in search spaces with large branching factors.

**Project Title:** Supervised Reinforcement Learning  
**PI:** T. Dietterich  
**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 are developing new kernel-based reinforcement learning algorithms for this setting and applying them to problems of scheduling and of determining the structure of proteins from NMR spectroscopy.

Project Title: Experimental Test Bed for Multimedia Document Classification and Retrieval  
Participants: Tom Dietterich and Dan Forrest  
Abstract: The goal of this work is to identify an application domain for testing document classification and retrieval methods that combine image and textual information. We are exploring several scientific domains where the data are based primarily on images with associated textual information. We will then build a corpus of labeled (and unlabeled) documents to be used by other members of the ITR team.

Project Title: Multi-class Classification with Support Vector Machines  
Participants: G. Valentini (University of Genoa), Tom Dietterich and Amit Goel  
Abstract: Support vector machines are binary classifiers. To handle multiple classes, some method is required to decompose the multi-class 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 multi-class 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. Experiments conducted thus far have shown 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.

University of Illinois at Urbana-Champaign:

Project Title: Intermediate Knowledge Representations that Facilitate Learning  
Participant: 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.

Project Title: Inference with Classifiers  
Participant: 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 are studying two general approaches for this problem and are evaluating those in the context of inference problems in natural language -- identifying phrase structure and question-answering. The first approach studied is 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. The second is an extension of constraint satisfaction formalisms.

Project Title: Learning Coherent Concepts  
Participant: 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: **A Sequential Model for Multi-class Classification**

**Participant:** Yair Even-Zohar

**Abstract:** Many classification problems require decisions among a large number of competing classes. These tasks, however, are not handled well by general purpose learning methods and are usually addressed in an ad-hoc fashion. We develop a general approach -- a sequential learning model that utilizes classifiers to sequentially restrict the number of competing classes while maintaining, with high probability, the presence of the true outcome in the candidates set. In particular, this model is used to develop an approach for hierarchical classification. We study theoretical and computational properties of the model and evaluate it in the context of several NLP learning tasks such as syntactic and semantic annotation of text and query classification.