1. Analyzing Brain Tumors (http://www.cs.ualberta.ca/~btgp)

Gliomas are diffuse, invasive brain tumors, that are often treated by irradiating the cancerous regions. Fortunately, portions of such tumors can often be detected in Magnetic Resonance (MR) images of the brain of a cancer patient. Unfortunately, there can be other “radiographically occult” tumor cells, that need to be treated as well. This project (in collaboration with Prof. J. Sander (CSD) and Dr A. Murtha (MD, Radiation Oncologist) addresses the task of using prior knowledge (eg, brain templates), and various imaging techniques, to find the complete tumor volume (both visible and invisible), given a set of MR images.

The first step is to accurately locate the visible tumor region. This involved a long pipeline of pre-processing steps. to find relevant features for each voxel. [C9] addressed the major challenge of determining exactly which features (allow a learner to) produce an effective classifier. Another issue is combining the labels for the individual voxels, to find a “consistent” labeling for the entire image; here we extended the now-standard Condition Random Fields by using (in essence) a support vector machine; this produced better results [C11], fairly efficiently [C2]. The overall result is one of the best tumor segmentation systems (from T1, T2 and T1c MR images). [C6] uses these segmented images to address the challenge of learning a system that can predict how the glioma will grow. (This uses the assumption that “where the tumor is visible tomorrow, it is invisible today”.)

This work has resulted in two MSc theses (M Morris and and M Schmidt), the latter nominated for the “Best MSc thesis” prize. We have also applied for a patent for this technology, and discussed it in an Alberta-wide radio show1. This work is also mentioned as one of the four research “success stories” in the “Alberta Surplus” newsletter2 that was sent to every Albertan resident.

2. WEBIC: An All-WWW Recommendation System (http://www.web-ic.com)

There are currently a large number of recommendation systems, each designed to give useful advice to the user. Essentially all such Web recommendation systems are specific to single web site; e.g., Amazon.com’s recommendation system is designed to suggest Amazon.com pages to users currently visiting the Amazon.com website. Such systems can base their recommendation on where other “similar” users have gone, using notions like the “support” and “confidence” of various pages and trajectories, based on the dozens to thousands of previous user visits to each page.

Our goal, however, is a system that can locate and recommend “information content (IC) pages” — pages the current user must see to complete his/her task — from essentially anywhere on the web. As most of the billions of pages have essentially no visits, or at least none that we know, support and confidence are not meaningful here. We therefore need to use a very different technology for this class of tasks.

Our WEBIC system first extracts the “browsing properties” of each word encountered in the user’s current click-stream — eg, how often each word appears in the title of a page in this sequence, or in the “anchor” of a link that was followed, etc. It then uses a user- and site-independent model, learned from a set of annotated web logs acquired in a user study, to determine which of these words is likely to appear in an IC page [C18],[C25],[C26]. We then show how to use these words to find IC-pages themselves, and demonstrate empirically that this browsing-based approach works effectively [C15],[C17]. This work, in collaboration with Prof G Häubl (UofAlberta Business Faculty) and postdoc B Price, is the basis of T Zhu’s PhD thesis. It was also described in both an Alberta-wide radio show3, and in several articles that have been repeated in dozens of publications around the world. This research is also the foundation of a current start-up company.

1http://innovationalberta.com/article.php?articleid=624
3. Proteome Analyst  (http://www.cs.ualberta.ca/~bioinfo/PA)

Proteome Analyst (PA) is a publicly available, high-throughput, web-based system for predicting various properties of each protein in an entire proteome. Using machine-learned classifiers, PA can predict, for example, the GeneQuiz general function and Gene Ontology (GO) molecular function of a protein [J9],[J7]. In addition, PA is one of the most accurate and most comprehensive systems for predicting subcellular localization, the location within a cell where a protein performs its main function [J6]. These functions are organized in a hierarchy; [C10] investigates how a learner should exploit such hierarchical information. PA produces a Support Vector Machine classifier, which is amenable to a graphical and interactive approach to explain its predictions; transparent predictions increase the user’s confidence in, and understanding of, PA [C5]. [J1] describes an extension that predicts which of an organism’s proteins participate in each of a set of pathways This work is in collaboration with Profs P. Lu, D. Szafron and D. Wishart, as well as many MSc and undergrad students.

4. Budgeted Learning

Researchers often use clinical trials to collect the data needed to evaluate some hypothesis, or produce a classifier. During this “training” process, they have to pay the cost of performing each test. Many studies will run a comprehensive battery of tests on each subject, for as many subjects as their budget will allow, in a “round robin” (RR) fashion. We consider a more general model, where the researcher can sequentially decide which single test to perform on which specific individual, then use the result of this test (together with earlier information) to make the next decision; again subject to spending only the available funds. Our goal here is to use these funds most effectively, to collect the data that leads to the most accurate classifier.

We first explore the simplified “coins version” of this task. After observing that this is NP-hard, we consider a range of heuristic algorithms, both standard and novel, and observe that our “biased robin” approach is both efficient and much more effective than most other approaches, including the standard RR approach [C19]. We then apply these ideas to learning a naïve-bayes classifier and observed similar behavior [C22]. Finally, we consider the most realistic model, where both the researcher gathering data to build the classifier, and the user (eg, physician) applying this classifier to an instance (patient) must pay for the features used — eg, the researcher has $10,000 to acquire the feature values needed to produce an optimal $30/patient classifier. Again, we see that our novel approaches are almost always much more effective than the standard RR model [C12].

This work, with postdoc O Madani, became the MSc theses of A Kapoor and D Lizotte.

5. Belief Net Algorithms

Many standard tasks inherently involve reasoning probabilistically — e.g., about correlations between patient data and his disease state. (Bayesian) belief networks have become the representation of choice for many AI researchers and practitioners, as they provide a succinct way to encode such probabilistic information, which allows them to reason about these situations effectively. This has led to an explosion of algorithms for both learning and reasoning with these systems.

A general Belief Net contains both a structure, which specifies what depends on what, and a set of parameters, “Conditional Probability Table values”, which indicate the strength of these connections. We have investigated ways to learn the best generative structure [C33],[J11], as well as the best discriminative structure [C31],[C16] for producing a good classifier. We have also provided a new, and effective, algorithm for learning the best discriminative parameters (for a fixed structure) [C27],[C21],[J5].

We have also provided an effective way to compute the variance around a belief net response [C29] (with Statistics Professor P Hooper and students), and used this as part of a tool for combining
different belief-net based classifiers [C4] (with PhD student Chi-Hoon Lee and postdoc Shaojun Wang). With postdoc Shaojun Wang and others, we have also used undirected probabilistic models for modeling language [C3],[C13].

Refereed journal papers


Refereed Conference Articles (Full paper refereed, under 1-in-3 acceptance rate)


“Best Student Paper Prize”


“RunnerUp, “Best Paper Prize”


Other Publications I have also co-edited 1 conference proceedings (for the “International Conference on Machine Learning” ICML), published 28 other “lightly” refereed papers, and 20 posters and invited (but not refereed) publications.