RESEARCH STATEMENT

Romer E. Rosales

RESEARCH OBJECTIVES

My primary research interest is applied machine learning and data mining, with special focus on the analysis and modeling of mid to large-scale, real-world data/information sources as a means to understand and, in particular, exploit this information in more valuable ways. I am interested in the analysis of systems and phenomena through data and ultimately in enhancing them.

We are in the midst of a dramatic growth in the diversity and amount of data available electronically at a low cost. As a consequence, we are now able to address and pose numerous research questions (some completely novel) that can lead to more meaningful and innovative uses of this data. Such applications have fundamental implications in areas as varied as business, society, healthcare, and science. For example, modeling of web search/information already generates revenue of billions of dollars for a few innovative companies; analysis of on-line product ratings is redefining the marketing and sales landscape; automated analysis of health records is making strides (with remarkable government support) towards changing the way medicine is fundamentally practiced; modeling patterns of economic/human/social activity is at the core of new business models (in addition to being of special interest for national security); business intelligence increasingly relies on automated forms of large-scale data analysis.

My research is based on the development of computational and mathematical models and algorithms capable of representing and efficiently extracting valuable information from the available data, which in turn permit the successful utilization of this data in productive, interesting, novel ways. Based on my academic and industrial research experience, this approach is suitable for addressing the problems of interest because it allows for conceiving general, abstract models that are amenable to formal analyses and can be more systematically linked to algorithmic solutions or even new problems. Highly-available processing power can then be used for the automation of information/knowledge analysis and extraction. This approach focuses on the use of data as a means of objectively evaluating competing solutions (allowing for continuous, automated experimentation), it is grounded on solid theoretical (e.g.; mathematical and computational) principles, and it has shown ample success in addressing challenging problems as described in this statement.

In order to make progress towards these goals I focus on: (1) analyzing/developing mathematical models and algorithms aimed at large-scale experimentation with data, (2) maintaining a correct understanding of end-user, business, and engineering needs or limitations to help identify, validate, and support our work on legitimate problems, and (3) innovating by tackling new possibilities unlocked by new forms of data and operational processes (new tools, applications, business models, services) based on our research findings.

PRESENT CONTRIBUTIONS

1. MODELS OF FREE TEXT

Natural language and semi-structured text as information source is ubiquitous. I lead R&D work for exploiting the text/narrative information in health records for various information retrieval/extraction (IR/IE) products: (1) patient/document identification [P1,P5,P10], or how to identify patients/documents based on the free text content in the health records; (2) concept search [P2], or how to search for concepts rather than keywords in text documents; and (3) language model domain adaptation, or how to use text data in one domain to aid in problems involving data in a different but related domain (with the aid of active learning) [12]. Each of them is a piece in the puzzle to address a fundamental question: How can we efficiently identify and automatically learn the occurrence of events/concepts in text? For example, what type of information is the user seeking based on his typed query (air travel, shopping, news, facts)? Which user review/article gave positive quality/design/usability feedback for a product/company? Did the patient stop smoking more than one year ago?

2. LEARNING FROM CROWDS

Data can now be easily shared and collectively processed by a large number of entities (Wikipedia contributions; product ratings; user geo-location, actions, and tweets); an effect coined *Crowdsourcing* (Howe, 2009) and exploited by the Amazon Mechanical Turk platform. In the supervised machine learning context, this translates to having not one labeler/ground-truth but many

labelers (potentially unreliable). This new scenario renders traditional supervised learning sub-optimal but also creates interesting problems. In [1] we showed how to efficiently utilize the labels provided by several annotators and, in particular, when annotator effectiveness or accuracy varies depending on the data instance presented (e.g.; running shoes are more accurately rated by runners). To the best of our knowledge this was the first work to address this setting. This also addresses related problems such as the case when ground-truth is by nature not available (e.g., what are the best results for this search query) or expensive to obtain (e.g.; a biopsy can provide ground-truth about cancer lesions but at a high cost). We also showed how the model is suitable for handling missing annotators, estimating ground-truth, and evaluating labelers.

3. ACTIVE SENSING What if we have annotator(s), but we want to use their time more efficiently by advising them which data to focus on. This is the usual active learning scenario. Now envision that we could also decide what new observations (not just annotations) need to be made in order to learn more efficiently. For example, in diagnosing disease, what medical test is optimal in order to learn how to diagnose a particular disease/condition? From a sensor network perspective this is equivalent to choosing what sensor to observe to decrease an uncertainty function. In [3,10] we presented an approach to achieve this by modeling the problem as a Gaussian process.

4. APPROXIMATE INFERENCE IN GRAPHICAL MODELS An effective way to represent the structure of probability distributions is by means of a graph. The (exponential) computational complexity of inference (on the tree-width of the triangulated graph) motivates approximations. We introduced a general approximate inference algorithm based on an iterative process that sequentially eliminates variables and obtains an approximation of the marginal distribution of the remaining variables by means of marginal tree-structured distributions of subset of variables. The method's ability to include dependencies induced by marginalization but not represented in the original graph may underlie the excellent empirical results obtained. Interestingly, we were able to derive a related distributed strategy in the flavor of the sum-product algorithm. I believe this work is merely a first step in a new and exciting research direction [21].

5. ℓ_1 REGULARIZATION FOR LEARNING SPARSE MODELS It is often the case that a small fraction of covariates suffices to build a model comparable in performance with others employing a larger number. Obtaining an optimal subset of features is in general NP-hard. However, the use of the ℓ_1 penalty provides an implicit covariate selection strategy. In [13,38], we presented an exhaustive comparison of optimization techniques to solve the problem of minimizing $f(x) = L(x) + \lambda\|x\|_1$, a twice-differentiable loss function L subject to ℓ_1 regularization. We also introduced generalized versions of several existing methods and new strategies based on a smooth convex approximation for the regularizer and on a constrained optimization formulation (exploiting gradient-projection ideas). Our experiments across 13 optimization strategies in 7 different scenarios showed in what cases one method should be used over others. We believe this is among the most exhaustive studies on the use of the ℓ_1 norm for building sparse models in machine learning.

6. STRUCTURE LEARNING IN GRAPHICAL MODELS The above inference problem assumes that we know the *structure* of the distribution, given by a graph G, often an invalid assumption. A few approaches address the problem of estimating G for Markov random fields (MRF). In [8] we concentrated on learning the structure of discriminative models (Conditional Random Fields) and non-binary MRFs; to the best of our knowledge, the first work to do so. We developed a general method for learning (sparse) graph structures for this kind of models via block-ℓ_1 regularization. This involved casting the task as a convex optimization problem and introducing a ℓ_1-ℓ_∞ regularization while paying special attention to efficiency given the practical difficulties involved in computing second order information.

7. OTHER Knowledge transfer or domain adaptation via extensions to Query-by-Committee [12], controlled incorporation of expert domain knowledge into learning problems [2], computer vision, modeling body motion/pose [18,33], tracking people in video, image rendering by transferring statistical properties of one image onto another [24]; clustering data using neighborhood structure (rather than the usual similarity measures) [22]; gene expression analysis [22]; learning to rank [19], learning to detect heart motion abnormalities from ultrasound images[15] (obtaining a best paper award), and web-based interfaces for searching/annotating and managing electronic records [P2].

FUTURE DIRECTIONS

In the long-term, the goal of my research is to make a clear impact on how information is analyzed and utilized at a large scale by creating suitable machine learning/data mining models and algorithms. In the short-term, I believe that a successful approach for achieving this must include working on real, significant problems and potentially replacing/extending them with new ones; establishing purposeful collaboration in appropriate areas; and making continuous improvements while exploring novel directions of research. In pursuing these goals, the key machine learning areas to explore involve models and algorithms directed towards the full utilization of the *knowledge of the masses* (e.g.; learning from the interaction of multiple users with a system, including user annotations), the efficient use of limited expert knowledge (e.g.; actively seeking for the important information: labels or data), the appropriate incorporation of widely available data (*e.g.*; unlabeled data); and the better exploitation of (currently sub-utilized) unstructured data sources (*e.g.*; free text). The ability to efficiently work with large datasets (in terms of storage and processing) is of especial importance to achieving these goals.

1. UNDERSTANDING FREE TEXT The understanding of free text (not limited to natural language) is considered one of the critical technologies in the future of information services because free text is a natural communication interface, large amounts of information in many aspects of life are and continue to be stored as free text, it is at the intersection of other areas such as speech recognition and language translation, and numerous applications will benefit from better text understanding. Not surprisingly, leading companies in the information industry are spending billions in R&D to build the next generation automated text processing products and tools. I believe that my research direction based on learning to recognize text concepts is very promising because while focused, it also addresses core elements applicable to more general problems.

2. LEVERAGING THE POWER OF THE CROWDS Data generated by multiple entities about the same general subject (product review, news article, search result, opinion, medical diagnosis, and question in general) is becoming common place. As a focus area in machine learning, we are just at the beginning. The data availability will be so large that this will be a critical area to allow us to process information more efficiently and usefully. Some essential problems that still need to be tackled include choosing annotators efficiently, modeling annotator biases, evaluation of annotators, and estimating unavailable ground-truth.

3. MODELING AND ORGANIZING INFORMATION STREAMS Information is generated at astonishing rates. Thus, analyzing and organizing this on-line information stream can give us a glimpse into the world today. I want to understand to what extent it is possible to accurately use this information to predict possible user information needs/preferences, epidemics, economic collapse, security treats.

4. APPROXIMATE INFERENCE IN GRAPHICAL MODELS Our previous work in Focused Inference [21] represents a rather unique avenue for approaching the general inference problem in statistics. Because of its practical implications and potential, I wish to study and more carefully understand the properties and potential uses of the distributed algorithm we introduced.

These areas encapsulate my current view about this evolving field. I am guided by my experience in both industry and academia where I have had the fortune to learn and contribute. I strongly believe in their complementarity for the advancement of science and technology. These experiences have given me a rather unique vantage point for choosing research directions, for collaboration, for building support, and for developing new ideas. I am very confident and at the same time excited about the effect that machine learning and data mining technology will have on the future of business and society. We need to speak the language of data to understand and navigate today's world.