VirginiaTech
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Computer Science Seminar Series, 2010
National Capital Region

Text mining in the biomedical domain

Speaker: Prof. Hongfang Liu
Department of Biostatistics, Bioinformatics, and Biomathematics
Georgetown University Medical Center

Friday, March 19, 2010
1:00PM- 2:00PM, NVC 325

Abstract

Machine learning has been used extensively in retrieving, extracting, and managing biomedical knowledge. This talk will focus on our ongoing efforts on the two challenges faced when applying machine learning to biomedical text classification tasks: learning from positives and unlabeled, and learning on imbalanced text classification. Text classification is the task of classifying a piece of free text (e.g., documents, paragraphs, or sentences) into two or more classes which can help people finding information in a huge collection of documents. Text classification research and practice have exploded in the past decade and been widely published via conferences and journals in information retrieval (IR), natural language processing (NLP), machine learning, and other fields. Retrieving relevant documents to specific topics can be defined as text classification task where relevant documents are positives and irrelevant documents are negatives. When defining text classification tasks in the biomedical domain, we face several challenges. One is the lack of negative instances. It is usually the case that positive documents can be obtained from existing biomedical databases with evidence articles cross-referenced. However, a set of negative instances is often hard to come by. The target machine learning task is learning from positives and unlabeled (LPU). Secondly, only a small portion of the documents are relevant documents. The target machine learning task is learning for imbalanced text classification.

Biography

Dr. Hongfang Liu is an assistant professor in Department of Biostatistics, Bioinformatics, and Biomathematics at Georgetown University. She devotes most of her time conducting research on biomedical text mining and ontology as well as "-omics" data analysis. She received her bachelor and master education in Mathematics and Statistics from University of Science and Technology of China (1994, 1996), and her master and PhD education in Computer Science from Fordham University (1998) and City University of New York (2002), respectively.