**大数据与机器学习国际研讨会**

**内容简介**

1. **Wang Fei：Integrative Network Analytics for Insights Generation from Massive Healthcare Data.**

The arrival of the Precision Medicine age brings tremendous opportunities to scientific discovery and quality improvement in medicine and healthcare. However, it also raises big challenges in dealing with large and massive healthcare data from heterogeneous sources. In this talk, I will present a computational framework called integrative network analytics to generate insights from complex healthcare data including Electronic Health Records (EHR), drug development data, genomic data, etc. I will also demonstrate how such framework can be used in computational drug discovery and personalized treatment recommendation.

1. **Xia "Ben'' Hu: Social Spammer Detection: A Data Mining Perspective.**

With the growing popularity of social media, social spamming has become rampant on all platforms. Many (fake) accounts, known as social spammers, are employed to overwhelm legitimate users with unwanted information. The social spammers are a special kind of spammers who coordinate among themselves to launch attacks such as distributing ads to generate sales, disseminating pornography and viruses, executing phishing attacks, or simply sabotaging a system's reputation. In this talk, I will introduce a novel and systematic analysis of social spammers from data mining perspective to tackle the challenges raised in social media data for spammer detection. Specifically, I will formally define the problem of social spammer detection and discuss the unique properties of social media data that make this problem challenging. By analyzing the two most important types of information, network and content information, I will introduce a unified framework by collectively using heterogeneous information in social media. To tackle the labeling bottleneck in social media, I will show how we can take advantage of the existing information about spam in email, SMS, and on the web for spammer detection in microblogging. I will also present a solution for efficient online processing to handle fast-evolving social spammers.

1. **Jianke Zhu: Scalable Image Retrieval by Sparse Product Quantization.**

Fast Approximate Nearest Neighbor~(ANN) search technique for high-dimensional feature indexing and retrieval is the crux of large-scale image retrieval. A recent promising technique is Product Quantization, which attempts to index high-dimensional image features by decomposing the feature space into a Cartesian product of low dimensional subspaces and quantizing each of them separately. Despite the promising results reported, their quantization approach follows the typical hard assignment of traditional quantization methods, which may result in large quantization errors and thus inferior search performance. Unlike the existing approaches, in this paper, we propose a novel approach called Sparse Product Quantization~(SPQ) to encoding the high-dimensional feature vectors into sparse representation. We optimize the sparse representations of the feature vectors by minimizing their quantization errors, making the resulting representation is essentially close to the original data in practice. Experiments show that the proposed SPQ technique is not only able to compress data, but also an effective encoding technique. We obtain state-of-the-art results for ANN search on four public image datasets and the promising results of content-based image retrieval further validate the efficacy of our proposed method.

1. **Irwin King: Practical Learning Algorithms for Big Data Processing.**

In the big data era, high volume of data and fast velocity of data generation lead to great challenges in traditional machine learning algorithms for practical applications. In this talk, we propose two practical algorithms that are suitable for big data processing. One is online algorithm for dictionary learning, and the other is parallel distance metric learning algorithm.

Online learning algorithms receive samples sequentially and update models incrementally. Based on this paradigm, we present a novel algorithm of online non-negative dictionary learning for sparse Poisson coding. Specifically, online dictionary learning for sparse coding is an effective tool for data analysis. It incrementally learns a set of basis vectors with sparse linear combinations of these vectors when new samples appear. Previous work assumes that the samples embed Gaussian noises, which weaken the power of these methods in handling real applications with non-negative data (e.g., frequency data in word counts). Differently, in this talk, we propose an algorithm of online learning for non-negative dictionary by using moment information in sparse Poisson coding. With Poisson distributions, we present sufficient convergence analyses to guarantee the performance of the proposed algorithm. We finally conduct a series of experiments on word-counts data and image data to show merits of the online algorithm.

We also propose a parallel algorithm for distance metric learning (DML) to tackle big data processing. DML is an effective similarity learning tool to learn a distance function from examples to enhance the model performance in applications of classification, regression, and ranking, etc. Most DML algorithms need to learn a Mahalanobis matrix, a positive semidefinite matrix that scales quadratically with the number of dimensions of input data. This brings huge computational cost in the learning procedure, and makes all proposed algorithms infeasible for extremely high-dimensional data even with the low-rank approximation. In this talk, we propose a novel distributed distance metric learning algorithm based on a state-of-the-art DML algorithm, Information-Theoretic Metric Learning (ITML). We present a rigorous theoretical analysis to upper bound the Bregman divergence between the sequential algorithm and the parallel algorithm. Our experiments on datasets demonstrate the competitive scalability and the performance compared with the original ITML algorithm.

1. **Kaizhu Huang: Memory Network: Learning with Few Samples.**

Neural Networks (NN) have achieved great success in pattern recognition and machine learning. However, the success of NNs usually relies on a sufficiently large number of samples. When fed with limited data, NN's performance may be degraded significantly. In this paper, we introduce a novel neural network called Memory Network, which can learn better from limited data. Taking advantages of the memory from previous samples, the new model could achieve remarkable performance improvement on limited data. We demonstrate the memory network in Multi-Layer Perceptron (MLP). However, it keeps straightforward to extend our idea to other neural networks, e.g., Convolutional Neural Networks (CNN). We detail the network structure, present the training algorithm, and conduct a series of experiments to validate the proposed framework. Experimental results show that our model outperforms the traditional MLP and other competitive algorithms in real data.

1. **Kun Zhang: Causality and Learning**

Can we determine the causal direction between two variables? How can we make optimal predictions in the presence of distribution shifts? We are often faced with such causal modeling or prediction problems in science, management, and engineering. Recently causal discovery has benefited a great deal from machine learning, statistics, and information theory, and on the other hand, causal information has been demonstrated to be able to facilitate understanding and solving certain machine learning problems.

In this talk I will first focus on causal discovery, i.e., learning causal information from purely observational data, and talk about conditional independence-based and functional causal model-based approaches. In particular, the latter type of approaches is able to distinguish cause from effect given two variables. Moreover, some practical issues, including causal discovery from biased data and from nonstationary data, will also be discussed. Secondly, I will consider two machine learning problems--semi-supervised learning and domain adaptation (or transfer learning)--from a causal point of view, and briefly discuss why and how the underlying causal knowledge helps to solve learning problems when the i.i.d. assumption is dropped.