Abstracts in alphabetical order

2018 Conference on Optimization and Data Science, TRU
March 9-10, 2018

John Cuzzola, Thompson Rivers University

Introduction to Semantic Annotation and its Uses in Biomedecine

Semantic analysis is concerned with the meaning of some type of unstructured input usually in the form of natural language text. In computing, semantic annotators play the role of recognizing words from the input text that can be directly mapped to an ontology; namely, a large collection of formal concepts from a specific domain. The most well-known repository of concepts is Wikipedia for general knowledge ideas. However, there are many others targeted to very specific domains such as the "Unified Medical Language System (UMLS)" used by medical practitioners to make medical health data more interoperable with other health resources. This presentation will introduce the research area of semantic annotation and named entity linking (mapping words to concepts) using the large ontologies of Wikipedia and UMLS. We will look at the importance of semantic annotation for information search and retrieval within these two domains. We will also describe our work with matching concepts between Wikipedia and UMLS (known as ontology alignment) and finish with a practical use-case currently under development.

Abdel-Monem Ibrahim, Thompson Rivers University

A Novel Binary Water Wave Optimization for Feature Selection

A search method that finds a minimal subset of features (over a feature space) that yields maximum classification accuracy is proposed. This method employs rough set theory along with a newly introduced binary version of the water wave optimization approach (WWO) which is denoted by BWWO. WWO simulates the phenomena of water waves, such as propagation, refraction, and breaking and is one of the newest nature inspired methods for global optimization problems. In our approach, BWWO is used for propagating the updated position and new visions are presented for the other wave phenomena (refraction and breaking). The proposed binary water wave optimization is denoted by BWWO. For experimental verification of the results, two main experiments are carried.
In the first experiment, the robustness of the proposed approach is demonstrated over 16 different datasets using rough set theory. The proposed approach is compared with various typical attribute reduction methods and popular optimizers in the literature, such as ant colony, nonlinear great deluge algorithm, scatter search and others.
For the second experiment, a feature subset that maximizes the classification accuracy using cross-validated *k*-NN classifier while minimizing the number of selected features is obtained over 17 different datasets. Our approach is compared with the binary grey wolf optimization, binary particle swarm optimizer, binary cat swarm optimization, binary dragonfly algorithm and the binary bat algorithm.
The computational results demonstrate the efficiency and effectiveness of the proposed approach in finding a minimal subset of the features that maximize the classification

accuracy.
Furthermore, Friedman test and Wilcoxon's rank-sum nonparametric statistical test are carried out at 5% significance level to judge whether the results of the proposed algorithm differ from those of the other algorithms in a statistically significant way. (Joint Work with M. Tawhid and R. Ward)

Bo Li, Tianjing University of Technology, China and **Hasnat Dewan**, SoBE, TRU
Spatial-temporal Pattern of Manufacturing Firm Agglomeration of Beijing-Tianjin-Hebei Region

Based on the big data of manufacturing firms in Beijing-Tianjin-Hebei region, this study uses Geocoding API of Baidu map to get the geographical latitude and longitude coordinate data. Based on the combined dataset, this study calculates the D-O index to analyze the geographical agglomeration of manufacturing firms in Beijing-Tianjin-Hebei region and investigate its spatial-temporal pattern. Furthermore, this study also visualizes the spatial distribution of each major manufacturing industry firms in the region using GIS. The results of this study will provide a useful reference for the coordinated development of industries and the regional integration in Beijing-Tianjin-Hebei region.

Ning Lu, Thompson Rivers University

Optimal Distributed Scheduling of Real-Time Traffic with Hard Deadlines

In this talk, optimal distributed scheduling of real-time traffic with hard deadlines in an ad hoc wireless network will be presented. We assume the links share a common wireless channel and interference is represented by a conflict graph. Periodic single-hop traffic is considered where packets arrive at the beginning of each frame and need to be delivered by the end of the frame (otherwise, packets will be dropped). Each link is required to guarantee a maximum allowable packet dropping rate. We show that the real-time scheduling problem is combinatorial and tends to be intractable as the network size increases. To solve the real-time scheduling problem, we propose a frame-based carrier-sense multiple access (CSMA) algorithm which is shown to be asymptotically optimal. Moreover, it can be implemented in a distributed manner with low complexity. Simulation results also demonstrate the ability of the proposed algorithm to meet the QoS requirements on deadlines.

Abraham P. Punnen, Simon Fraser University

Representations of quadratic combinatorial optimization problems

The objective function of a quadratic combinatorial optimization problem (QCOP) can be represented by two data points, a quadratic cost matrix Q and a linear cost vector c. Different, but equivalent, representations of the pair (Q, c) for the same QCOP are well known in literature. Research papers often state that without loss of generality we assume Q is symmetric, or upper-triangular or positive semi-definite, etc. These representations however have inherently different properties. Popular general purpose 0-1 QCOP solvers such as GUROBI and CPLEX do not suggest a preferred representation of Q and c. Our experimental analysis discloses that GUROBI prefers the upper triangular representation of the matrix Q while CPLEX prefers the symmetric representation in a statistically

significant manner. Equivalent representation, although preserves optimality, they could alter the corresponding lower bound values obtained by various lower bounding schemes. For the natural lower bound of a QCOP, symmetric representation produced tighter bounds, in general. Effect of equivalent representations when CPLEX and GUROUBI run in a heuristic mode are also explored. Further, we review various equivalent representations of a QCOP from the literature that have theoretical basis to be viewed as 'strong' and provide new theoretical insights for generating such equivalent representations making use of constant value property and diagonalization (linearization) of QCOP instances.

Saeed Rahmati, Thompson Rivers University

Hypermatrix Complementarity Problems on a Set of Hypermatrices.

A hypermatrix is a multi-array of real entries. For a hypermatrix A, a hypermatrix complementarity problem, HMCP(q, A), is to find a vector x in the n-dimensional real space such that x and $Ax+q$ are non-negative but their dot product is zero, for every q in the n-dimensional real space. In this talk, we characterize uniqueness, feasibility, and strict feasibility of the solution of a complementarity problem induced by a (compact) set of hypermatrices in terms of the hypermatrices involved.

Mateen Shaikh, Thompson Rivers University

Association Rule Mining Adverse Drug Reactions from Spontaneous Reporting Databases

The government of Canada has records for adverse health events that occur while taking drugs, including in combination. This database is filled with observations from pharmaceutical companies (mandatory reporting) and from health care practitioners (voluntary reporting). Association rules are a data mining concept which find associations in such databases and can be used to discover if there are drugs which have more adverse reactions than expected. This can be particularly useful when two drugs with common side effects are jointly prescribed because the estimated benefit is thought to outweigh the estimated risks. This talk discusses (re)evaluation of the estimated risks of adverse events that may be attributed to drug interactions using the spontaneous database. Some focus is spent on concerns of biased reporting and rare events.

Mark Schmidt, CRC, University of British Columbia

Is Greedy Coordinate Descent a Terrible Algorithm?

There has been significant recent work on the theory and application of randomized coordinate descent algorithms, beginning with the work of Nesterov, who showed that a random-coordinate selection rule achieves the same convergence rate as the Gauss-Southwell selection rule.
This result suggests that we should never use the Gauss-Southwell rule, as it is typically much more expensive than random selection. However, the empirical behaviours of these algorithms contradict this theoretical result: in applications where the computational costs of the selection rules are comparable, the Gauss-Southwell selection rule tends to perform substantially better than random coordinate selection. We give a simple analysis of the Gauss-Southwell rule showing that---except in extreme cases---it's convergence rate is

faster than choosing random coordinates. Further, we (i) show that exact coordinate optimization improves the convergence rate for certain sparse problems, (ii) propose a Gauss-Southwell-Lipschitz rule that gives an even faster convergence rate given knowledge of the Lipschitz constants of the partial derivatives, and (iii) analyze proximal-gradient variants of the Gauss-Southwell rule.

Xiaoping Shi, Thompson Rivers University

Modeling High-dimensional Time Series

Modeling high-dimensional time series is necessary in many fields such as neuroscience, signal processing, network evolution, text analysis, and image analysis. Such a time series may contain unknown multiple change-points. For example, the time of cell divisions can be accessed using an automatic embryo monitoring system by a time-lapse observation. When a cell divides at some time point, the distribution of pixel values in the corresponding frame will change, and hence the detection of cell divisions can be formulated as a multiple change-point problem. In this talk, a powerful graph-based change-point detection is introduced.

Richard Santiago Torres, University of McGill

Multi-agent Submodular Optimization

Recent years have seen many algorithmic advances in the area of submodular optimization: $Min/Max\, f(S)$: $S \in \mathcal{F}$, where f is submodular and $\mathcal{F}$ a family of feasible sets. This progress has been coupled with a wealth of new applications for these models. Our focus is on a more general class of *multi-agent submodular optimization* problems introduced by Goel et al. in the minimization setting: $Min \sum_i f_i(S_i) : S_1 \uplus S_2 \uplus \ldots \uplus S_k \in \mathcal{F}$.

Here $\uplus$ denotes disjoint union and hence this model is attractive where resources are being allocated across k agents. In this talk we explore the extent to which the approximability of the multi-agent problems are linked to their single-agent versions.

Yan Xu, University of Victoria

A Model-Based Clustering to Identify Disease-Associated SNPs

Genome-wide association studies (GWASs) aim to detect genetic risk factors for complex human diseases by identifying disease-associated single-nucleotide polymorphisms (SNPs). The most commonly-used GWAS method is the SNP-wise-test approach, in which an association test is performed for each SNP, and then the p-values are adjusted for multiple testing. However, this approach is often lack of power after multiple testing adjustments due to a huge number (> 1 million) of tests in GWAS. To address this problem, we propose a model-based clustering via a mixture of Bayesian hierarchical models, which could borrow information across SNPs to group SNPs to different clusters having different mean genotype levels between cases and controls. Simulation studies and real data studies showed that the proposed model-based clustering outperformed SNP-wised-test approach. Joint work with Xuekui Zhang and Weiliang Qiu.

Xuekui Zhang, CRC, University of Victoria

Temporal Curve Alignment for Normalizing Different Batches of Time Course Genomic Experiments

Motivation: Time course experiments are powerful tools to study dynamic systems. In order to determine the temporal order of different genomic events, researchers often need to compare different high- throughput genomic data types along a time course. However, when different data types are generated using cells grown in different batches, their time axes may not be aligned. As a result, one cannot meaningfully compare these data to determine the temporal order of different genomic events.
Results: Temporal curve alignment (TCA) is a computational method for normalizing time course experiments from different batches. TCA fits multiple smooth curves to the data to describe the major temporal patterns. It then maps different batches of experiments to a common time axis by finding a non-linear transformation of time to optimize the alignment of temporal curves of the same data type between batches. We demonstrate this approach using an analysis of time course ChIP-seq and RNA-seq data in yeast metabolic cycle (YMC). This analysis highlights the importance of time alignment for downstream inference.