INFORMS 2011 Data Mining and Health Informatics (DM-HI) Workshop Schedule

Saturday, 8:30am

■ Breakfast

Saturday, 9:10am - 10:00am

■ Parallel Session 1A

Room 209A, Convention Center

Bayesian Modeling and Market Basket Analysis

Cluster: Data Mining and Health Informatics Workshop
Invited Session

1 - A Hierarchical Model for Association Rule Mining of Sequential Events (9:10am-9:35am)

Cynthia Rudin, Massachusetts Institute of Technology, MIT Sloan School of Management, E62-576, 77 Massachusetts Avenue, Boston, MA, 02139, United States of America, rudin@mit.edu, Tyler McCormick, David Madigan

In many healthcare settings, patients visit healthcare professionals periodically and report multiple medical conditions, or symptoms, at each encounter. We propose a statistical modeling technique, called the Hierarchical Association Rule Model (HARM), that predicts a patient's possible future symptoms given the patient's current and past history of reported symptoms. The core of our technique is a Bayesian hierarchical model for selecting predictive association rules (such as "symptom 1 and symptom 2 ? symptom 3 ") from a large set of candidate rules. The hierarchical nature of the method allows "borrowing strength" using the symptoms of many similar patients. Thus, the method provides predictions specialized to any given patient, even when little information about the patient's history of symptoms is available. We evaluate our method using data from a large clinical trial database. Our experiments indicate that HARM outperforms several baseline approaches including a standard "maximum confidence, minimum support threshold" technique used in association rule mining, and also a non-hierarchical version of our Bayesian method (from Rudin et al.,2010) that ranks rules using "adjusted confidence."

2 - A Dynamic Bayesian Network Approach to Market Basket Analysis (9:35am - 10:00am)

Bumsoo Kim, Doctoral Student, George Washington University, 2201 G Street NW, Fungar Hall 415H, Washington, DC, 20054, United States of America, bk4498@gwu.edu, Refik Soyer, Srinivas Prasad

Researchers in multiple fields have been intrigued by the problem of market basket analysis, one of the classic problems in data mining, where different methods and tools are used to find meaningful associations in consumer purchase data. Although there are many different methods used in the literature, the fundamental goal in market basket analysis involves searching and analysis of association rules. These associations could be some type of correlation relationship such as "the sales of product A and B are positively related" or in some cases may be causal relationship such as "promotion efforts on product A increases sales on product A". In this study, we investigate the "Bayesian Network" approach to the market basket analysis problem and focus specifically on the marketing application of the method. First of all, we look at the Bayesian network paradigm and evaluate the method against classical methods in tackling the market basket problem. We also demonstrate how Bayesian networks can be used to identify and incorporate direct and indirect effects of promotion in marketing datasets.

■ Parallel Session 1B

Room 209B, Convention Center

Analysis of Populations

Cluster: Data Mining and Health Informatics Workshop
Invited Session

1 - Exploring the Frequency of Quality Health Information In Online Search Results (9:10am - 9:35am)

Brent Kitchens, University of Florida, 355c Stuzin Hall, University of Florida, Gainesville, FL, United States of America, brentkitchens@ufl.edu, Shengli Li, Christopher Harle

Health consumers increasingly use online health information to inform their decisions and behavior. Because there may be serious consequences related to decisions based on inaccurate health-related information, it is important to understand the quality of health information that is retrieved by the public in their daily use of the Internet. This study empirically investigates the proportion of quality information that may be expected to be returned by a popular search engine (Google) for typical health-related searches. Results show that approximately half or more of the information returned is of a high quality, as judged by third parties that review and certify online health information. That said, first-page search engine results still include practically significant proportions of web sites whose quality may be questioned. Future research should seek to refine understanding of the amount and variation of online health information quality. Most importantly, this work should attempt to understand and analyze information quality in the context of realistic consumer information search and use behavior.

2 - Methods for the Analysis of Human Microbiome Data (9:35am - 10:00am)

Paul Brooks, Virginia Commonwealth University, PO Box 843083, Richmond, VA, 23284, United States of America, jpbrooks@vcu.edu, Ed Boone, Qin Wang, Patricio S LaRosa, Bill Shannon, David Edwards

High-throughput sequencing technology is currently being applied to study the human microbiome. The human microbiome is the community of microbes that populate habitats on and around various body sites. A sample from a subject is processed to create profiles that characterize the population of microbes; the ultimate goal is to find relationships between these microbiome profiles and subject physiology and health. In this paper, we address some of the challenges that arise in the analysis of data generated in a microbiome experiment. First, we discuss methods for visualizing microbiome data. Second, we describe procedures for assessing bias introduced in the sample processing pipeline so that proper estimates of true abundances may be obtained. Finally, we describe a statistical hypothesis test for confirming that groups of samples arise from different distributions.

Saturday, 10:10am - 11:00am

■ Parallel Session 2A

Room 209A, Convention Center

Microbiology and Networks

Cluster: Data Mining and Health Informatics Workshop
Invited Session

1 - Uncovering Disease Networks using Counterterrorism Approaches (10:10am - 10:35am)

Mary McGuire, Postdoctoral Fellow, UTHealth Medical School, Dept. of Pathology & Lab Medicine, 6431 Fannin St., MSB 2.121, Houston, TX, 77030, United States of America, mary.f.mcguire@uth.tmc.edu, Gino Lim, Robert Brown

Despite an ever-increasing amount of biomedical research data and the increasing capabilities to monitor and analyze patients' data, the underlying mechanisms of disease are still poorly understood. Network models of biological interactions offer insights into disease progression with the goal to improve diagnosis, prognosis and therapeutic interventions. Biological networks can be drawn as directed graphs whose nodes are biological entities and whose arcs are the interactions among those entities. The entities range from molecules or cells at the lower levels, moving up to tissues, organs, organism and populations. The arcs can represent structural or functional entity interactions and can be dynamic over time. Although biological network models based on Internet or social network analogies may be useful, their focus on hubs runs contrary to the distributed nature of progression in diseases such as cancer. We propose that terrorist network models may help uncover hidden disease networks by using qualitative and quantitative methods traditionally used for military, defense, and security applications, including – but not limited to – those used in counterintelligence, counterinsurgency and counterterrorism efforts. Examples will be drawn from disease networks including metastasis of renal cell carcinoma.

2 - Finding the Most Discriminating Sets of Biomarkers by Maximum Weighted Clique (10:35am - 11:00am)

Seyed Javad Sajjadi, PhD candidate, University of South Florida, 4202 E Fowler Ave, ENC 2400, Tampa, FL, United States of America, ssajjadi@mail.usf.edu, Bo Zeng, Xiaoning Qian, Amin Ahmadi Adl

In order to find the most discriminating and reliable subsets of biomarkers for a given disease, we consider a node and edge weighted graph to represent the set of biomarkers and their interactions with respect to their influence on disease development. Then, we formulate a Maximum Weighted Clique Problem (MWCP), which is a large-scale integer programming problem. To compute its solution, we develop a fast heuristic algorithm that solves the MWCP sequentially to identify the highly connected sub-graphs without overlap. A set of computational results will be presented along with discussion of insights on our solution strategy in solving biomarker identification problems.

■ Parallel Session 2B

Room 209B, Convention Center

Data-Driven Modeling

Cluster: Data Mining and Health Informatics Workshop
Invited Session

1 - A New Real-time Parameter Estimation for Virtual Cardiovascular Model using ECG Signal (10:10am - 10:35am)

Trung Le, Oklahoma State University, Industrial Engineering, Stillwater, OK, 74075, United States of America, trung.le@okstate.edu, Satish Bukkapatnam, Ranga Komanduri

This research presents a new approach for real-time simulation of heart dynamics where the real-time electrocardiogram (ECG) measurements are used as external activation. The model uses lumped mass nonlinear differential equations to capture the coupled mechanical and physiological actions of the heart chambers, valves, pulmonary and systemic blood circulation loops. The model parameters are estimated online from the ECG features. The model was implemented in Matlab/Simulink environment and tested using signals in MGH/MF Waveform Database from Physionet. The results show that the model was able to discern the effect of variations in the time and frequency patterns of central venous pressure, pulmonary blood pressure, and respiratory impedance. The model-derived signals correlated well (R2 ~0.78) with the actual hemodynamic signals. These findings support the suitability of developing a virtual instrument platform where the model-derived signals (presented appropriately) can be used for clinical diagnosis in lieu of expensive instrumentation.

2 - Data-driven Models for Monitoring Long-term Effects of Drugs (10:35am - 11:00am)

Yihan Guan, Ph.D. Candidate, Stanford University, 475 Via Ortega 212F, Stanford, CA, 94305, United States of America, yihan@stanford.edu, Margret Bjarnadottir

Even if drugs have been proven safe and effective, knowledge about long-term pharmacological and toxicological effects is often limited. Traditional drug studies are short-term clinical trials and therefore long-term benefits and risks cannot be fully determined from these studies. Long-term toxicity effects are key considerations for drug classes taken for chronic conditions, because exposure to these drugs can be years or decades. In addition, recent studies have suggested that it may be cost-saving to administer statins as a preventative regiment to a large part of a population. Therefore, obtaining knowledge of long-term effects of drugs is more urgent than ever, as we are moving away from strictly therapeutic drug use to preventative use. One possible data source to study long-term effects is healthcare claims, which has the advantage of a large volume to study rare outcomes. In this paper, we develop and contrast statistical frameworks to compare longitudinal outcomes associated with long-term use of drugs. We discuss properties of different methods under different toxicity assumptions and evaluate their effectiveness using claims data, focusing on chronic statins use. In particular, we propose a self-control case design as an efficient approach for long-term analysis under increasing toxicity assumptions.

Saturday, 11:00am - 11:30pm

■ Coffee Break

Saturday, 11:30am - 12:30pm

■ Special Invited Talk

Room 208B, Convention Center

Invited Talk- Nick Street

Cluster: Data Mining and Health Informatics Workshop
Invited Session

1 - Data Mining for Personalized Health Care Decision Making

Nick Street, The University of Iowa, S210 John Pappajohn Business Building, Iowa City, IA, 52246, United States of America, nick-street@uiowa.edu

Health care choices are made based on results of large-scale studies. The care you receive is therefore based on what works for most people, not necessarily what will work for you. However, massive datasets of patient records are increasingly being mined to allow personalized health care decision making. This talk will highlight work by our group that addresses the following questions: How can a person choose the best lifestyle choices to minimize the risk of heart disease? How can a nurse minimize a child's distress during a painful medical procedure? And, how can a physician choose the best diagnostic test to reach a fast, cheap, and correct diagnosis?

Saturday, 12:00am - 2:00pm

■ Lunch

Saturday, 2:00pm - 2:50pm

■ Parallel Session 3A

Room 209A, Convention Center

Machine Learning

Cluster: Data Mining and Health Informatics Workshop
Invited Session

1 - A Latent Feature Factor Graph for Cancer Diagnosis using Microarray Gene Expression Data(2:00pm - 2:25pm)

Mohammad Khoshneshin, The University of Iowa, 318 Finkbine Ln. 10, Iowa City, IA, 52246, United States of America, mohammad-khoshneshin@uiowa.edu, Nick Street

In this paper, we propose a latent feature model, Latent feature factor graph (LFFG), for the gene expression microarray classification problem. LFFG is a factor graph in which each entity is represented via a latent feature variable, a node in the factor graph. Then the relationship between entities is captured via a function of latent features, a factor in the factor graph. We use maximum a posteriori estimation to learn the latent features for the gene expression microarray classification problem. The experiments over a small dataset are promising as LFFG outperforms SVM, the state-of-the-art classification method in microarray classification.

2 - Probabilistic Classification using Fuzzy Support Vector Machines (2:25pm - 2:50pm)

Marzieh Parandehgheibi, Graduate Student, Massachusetts Institute of Technology, 70 Pacific Street, Apt. #403C, Cambridge, MA, 02139-4251, United States of America, marp@mit.edu

In medical applications such as recognizing the type of tumor as Malignant or Benign, a wrong diagnosis can be devastating. Methods like Fuzzy Support Vector Machines (FSVM) try to reduce the effect of misplaced training points by assigning a lower weight to the outliers. However, there are still uncertain points which are similar to both classes and assigning a class by the given information will cause errors. Here, we propose a two-phase classification method which probabilistically assigns the uncertain points to each of the classes. First, FSVM is applied to the whole training data such that most of the uncertain points will be placed in the margin. Moreover, the certain points are assigned to appropriate classes. Next, a fuzzy membership function and an appropriate rule with a grade certainty are defined to classify the points that were located in the margin. This will result in assigning uncertain points to each of the classes with a specific probability. The proposed method is applied to the Breast Cancer Wisconsin (Diagnostic) Dataset which consists of 569 instances in 2 classes of Malignant and Benign. This method assigns certain instances to their appropriate classes with probability of one, and the uncertain instances to each of the classes with associated probabilities. Therefore, based on the degree of uncertainty, doctors can suggest further examinations before making the final diagnosis.

■ Parallel Session 3B

Room 209B, Convention Center

Healthcare Logistics

Cluster: Data Mining and Health Informatics Workshop
Invited Session

1 - Planning and Operational Decision Models for Public Health Emergency Response Logistics (2:00pm - 2:25pm)

Kathleen King, Cornell University, 257 Rhodes Hall, Ithaca, NY, 14853, United States of America, kak59@cornell.edu, John Muckstadt

Public health emergencies, such as pandemics and bioterrorist attacks, require rapid responses from federal, state, and local authorities in order to prevent widespread mortality and morbidity. Current response plans seldom account for the variety of risks and uncertainties inherent in making the logistics decisions necessary for this type of coordinated effort. We have developed a suite of models to help public health officials both prepare for potential emergencies and to make operating decisions when actually responding to an event. For the planning phase, we have constructed methods for estimating the inventory, staffing, and transportation, and we show the marginal value of increasing resource capacities in various scenarios.

We have also built a simulation to evaluate the quality of existing response plans, to help policy makers understand how the different parts of the system will affect one another. For the operational phase of an emergency response, we have built methods for estimating patient demands at vaccine dispensing clinics over time, allocating medical supplies throughout the distribution network, and setting staffing plans based on observed demands. Our work provides public health officials with quantitative methods for making policy decisions and analyzing emergency response plans.

2 - Multi-agent Patient Scheduling Through Auctioned Decentralized MDPs (2:25pm - 2:50pm)

Hadi Hosseini, University of Waterloo, Ontario, Waterloo, Canada, h5hossei@uwaterloo.ca, Robin Cohen, Jesse Hoey

In this paper, we consider a novel approach to a multiagent resource allocation problem that arises in medical settings in which a set of consumers (patients) require a finite set of resources and the overall state of all consumers in the system needs to be optimized. With many consumers and resources, it is intractable to compute an optimal solution. Yet, there is a measure of utility or consumer success (e.g. the health state of the patients) that can be leveraged towards determining preferable allocations. We propose an approximate solution in which each consumer is represented by a Markov Decision Process (MDP) based agent that models the expected progression of its utility given the resources, and has an approximate model of its expected access to resources given the global state of all competing agents. A global auction-based mechanism is then proposed that requires each agent to bid for resources based on its expected gains (e.g. its long-term health) from obtaining each resource. The allocation of resources proceeds by maximization over bids in an iterative multi-round auction. We are able to show that the global utility value across all consumers is closer to optimal using our approach compared to competing approaches. In addition, our approach is sufficiently flexible to support the integration of temporal constraints between resources.

Saturday, 3:00pm - 3:50pm

■ Parallel Session 4A

Room 209A, Convention Center

Clinical Studies

Cluster: Data Mining and Health Informatics Workshop
Invited Session

1 - Understanding the Role of Mental Disorders on HIV Patient Outcomes (3:00pm -3:25pm)

Shengfan Zhang, North Carolina State University, 400 Daniels Hall, Raleigh, NC, 27695, United States of America, szhang5@ncsu.edu, Fay Payton, Julie Ivy

This paper studies the impact of comorbidities, with a particular focus on mental disorders, on HIV patient outcomes as measured by inpatient length of stay (LOS) and total charges. Generalized linear models (gamma models) are developed to characterize the effects of selected comorbidities on HIV patient outcomes in the adult 2006 National Inpatient Sample. Comorbid HIV patients experience different LOS and total charges. In particular, having mental disorders results in a decrease in both LOS (by 19%) and total charges (by 15%) for HIV patients. To characterize the role of individual mental disorder, principal component and cluster analyses on ICD-9 codes are used, and eight conditions are found to be most strongly associated with HIV. These results suggest the complexity and combination of these conditions necessitate an integrated approach for diagnoses, treatment and disease management of comorbid HIV patients, particularly those with substance abuse and depressive mood disorders.

2 - Using Gaussian Processes to Monitor Diabetes Development (3:25pm - 3:50pm)

Si-Chi Chin, The University of Iowa, LIB 3087, Iowa City, IA, 52242, United States of America, si-chi-chin@uiowa.edu, Nick Street

This paper uses Gaussian process techniques to model time series data of HbA1c level, a common measure to monitor or screen diabetes. The HbA1c level estimates how well blood sugar is under control. To facilitate the control of diabetes, we develop a patient-level model to individually predict the development of the disease for each patient. Gaussian processes represent a successful machine learning technique known for their flexible modeling abilities and high predictive performances. This approach allows multi-dimensional inputs and assigns a confidence score to the predictions, accounting for temporal uncertainty of time series data. The purpose of this paper is to discuss the use of the Gaussian process technique, previously unseen in diabetes research, to monitor the development of the disease.

■ Parallel Session 4B

Room 209B, Convention Center

Cancer Screening

Cluster: Data Mining and Health Informatics Workshop
Invited Session

1 - Mining the Effects of Changing Demographics on Breast Cancer Screening Practices in Women (3:00pm - 3:25pm)

Kalyan Pasupathy, University of Missouri, CE732 CS&E Building, Five Hospital Drive, Columbia, MO, 65212, United States of America, pasupathyk@missouri.edu, Zeyana Hamid

Breast cancer is a major chronic disease and early identification is necessary for treatment. Mammograms are clinical screening tests that increase the detection of breast cancer and provide diagnosis for treatment. Educating the public and promoting screening practices requires understanding of individual behavior patterns and temporal changes. For instance, various socio-economic subgroups have different behavior patterns. And, the demographics of communities change over time. Understanding these effects is necessary to design programs and implement policy. Partially funded by the Missouri Department of Health and Senior Services, and based on behavioral risk data collected over five years, this research identifies profiles of women's breast cancer screening practices based on socio-economic factors using various data mining techniques. Major factors include age, cost of access, income and number of children in household. Such profiles can be used to model the entire community, study and predict future public health status of communities.

2 - An MDP Approach to Cervical Cancer Screening (3:25pm - 3:50pm)

Diana Sanchez, Master, Universidad de los Andes, Cra 1 #18a - 12, Bogota, Colombia, dm.sanchez57@uniandes.edu.co, Thomas Yeung, Cristina Lopez, Raha Akhavan-Tabatabaei

Cervical cancer, caused by the human papillomavirus (HPV), is the second most common cancer in women. Proper screening through the Pap test can reduce cancer incidences by 60%. Abnormal Pap tests are followed by a colposcopy procedure in order to obtain further information. We propose a finite-horizon Markov Decision Process (MDP) to find the optimal screening policy for cancer prevention. The model chooses whether an additional screening should be made beyond the annual Pap frequency and whether the Pap test should be bypassed in favor of a direct colposcopy to eliminate an additional procedure and trip to the doctor's office. Decisions are made based on the woman's current diagnosis, age, HPV contraction risk, and screening results. The cost function considers the tradeoffs between prevention and treatment procedures and the risk of taking no action while taking into account the life-quality of patients. We apply the model to data collected from a representative sample of 1,141 Colombian women. The results suggest that in order to effectively track disease progress and avoid higher cancer rates and future costs, more frequent screenings should be carried out on patients with higher risk profiles. However, the annual guideline is sufficient for healthy women.

Saturday, 3:50pm - 4:10pm

■ Coffee Break

Saturday, 4:10pm - 5:10pm

■ Panel Discussion

Room 208B, Convention Center

Panel Discussion

Cluster: Data Mining and Health Informatics Workshop
Invited Session

1 - The Future of Data Mining in Health Informatics

Moderator: Cynthia Rudin, Massachusetts Institute of Technology, MIT Sloan School of Management, E62-576, 77 Massachusetts Avenue, Boston, MA 02139, United States of America, rudin@mit.edu.

Please join us for a panel discussion on the future of data-mining in health informatics applications. Panelists will include: J. Paul Brooks (Virgina Commonwealth), Wei Jiang (Shanghai Jiaotong), Fay Payton (North Carolina State), Suchi Saria (Stanford), and Nick Street (University of Iowa).