QMDNS-08: Abstracts

Calandra Tate, US Army Research Laboratory/US Military Academy

Statistical Applications in Machine Translation Evaluation

Machine Translation (MT) is the computer translation of text from one language to another. Continual changes in conditions in the world around us produce high demand for foreign language understanding. Particularly in the wake of the September 11th attacks, the Department of Defense has recognized military needs in document processing and evaluation. Today, there are many reputable MT systems and users like the US Government need to make choices as to which engines they should invest in based on which one provides the best output for their specific tasks.

This talk discusses my research in applying nonparametric statistical techniques to Machine Translation Evaluation. I will present results from an experiment using data conducted through a joint Army Research Laboratory (ARL) and Center for the Advanced Study of Language (CASL) project. This includes a description of the experiment design and the use of correlation analysis and generalized linear modeling techniques to identify and characterize the predictive relationship between machine translated document quality, as judged by the four automated translation evaluation measures studied, and the outcome of the information extraction task performed by subjects using the same collection of translated documents.

Kyle Bradbury, Peter Torrione, and Leslie Collins

Ground Tracking in Ground Penetrating Radar

Ground penetrating radar (GPR) is one of the leading sensor modalities for landmine detection. A number of algorithms for landmine detection using GPR rely on the hyperbolic signature of the landmine response. Often the detection of targets in GPR data is hindered by combinations of surface roughness, sensor positional uncertainty, and surface clutter (e.g. rocks, plant life, snow, anthropic objects, etc). These obstacles may lead to the distortion of the hyperbolic response and potentially cause the classification algorithms to miss a potential target. Also, for shallow-buried targets, the ground bounce may obscure the target response. In such situations, if the ground height were known, steps such as ground bounce removal and ground alignment could be taken to remediate these effects. Therefore, a method is proposed here to address the ground tracking problem.

Eric B. Fails, ECE, Duke University

An Autonomous Landmine Detection Application of Random Forests

Current landmine detection technologies typically employ electromagnetic induction (EMI) sensors and/or ground penetrating radar (GPR) antenna array systems. While EMI sensors, which detect metallic content, are highly accurate for detecting shallow metallic anti-personnel landmines, these systems are susceptible to false alarms from buried metallic objects. GPR systems, which detect dielectric discontinuities, excel at detecting deeply buried anti-tank and low metal landmines but suffer false alarms from non-metallic objects in the subsurface such as rocks, roots and soil layering. Our goal in sensor fusion is to exploit the complementary nature of these systems in order to utilize the strengths of the individual sensor platforms while reducing the overall false alarm rate.

Ensemble methods such as Random Forests have been receiving considerable attention in a variety of classification and regression applications. This work will explore the potential for Random Forests to serve as a robust sensor fusion center for the landmine detection problem.

Chris Ratto, Peter Torrione, and Leslie Collins

Context-Dependent Feature Selection for Classification of Simulated Ground-Penetrating Radar

This work investigates the effect of soil moisture and surface roughness on Ground Penetrating Radar (GPR) for landmine detection. Models are proposed for simulating the GPR responses of several objects in a realistic environment using the finite-difference time-domain technique. A layered soil model is implemented using pre-existing measurements of soil water content as a function of depth in various soils. An autoregressive model for surface roughness is also developed using training data from a commercial GPR at several test sites.

After the simulated GPR data are generated, features are extracted using the texture feature coding method. Subsets of features are selected for each environment using sequential search and a mutual information measure. The performance of a k-nearest neighbor classifier and a support vector machine are then evaluated for each subset and compared.

Ron Fricker, Naval Postgraduate School

Optimizing Biosurveillance Systems

Motivated by the threat of bioterrorism, biosurveillance systems are being developed and implemented throughout the United States. Biosurveillance is the regular collection, analysis, and interpretation of real-time and near-real-time indicators of possible disease outbreaks and bioterrorism events by public health organizations. Little is known about how effective these systems will be at quickly detecting a bioterrorism attack, but there is some evidence in the form of excessive false alarm rates that they are being suboptimally employed. This talk will provide an overview of the problem and describe an approach for managing the trade-off between the aggregate *system* false alarm rates and the power to detect a localized bioterrorism attack.

Myron Katzoff, CDC

A Further Consideration of Two Problems Related to Biosurveillance

For the detection of a catastrophic public health event and the subsequent collection of information to monitor progress in addressing its consequences, we may expect that the statistical methods employed for those purposes will draw upon experience acquired in other applications. The first part of this talk will consider the application of ideas from extreme value theory to the detection of outbreaks and the estimation of the probabilities of occurrence of values for disease incidence rates that might be viewed as extreme. Since my study of these ideas is at a very early stage, there will be more *vigor* than *rigor* in this part of my talk. In the remaining time, I will visit the application of some adaptive sampling techniques that I believe will have promise in obtaining information about the health status of individuals affected by the types of public health events of interest. With the occurrence of such events, the affected individuals may be *hidden* or hard-to-locate because it is unlikely that there will be a sampling frame of them available for our immediate use and they may be well-disperse throughout other populations. The adaptive sampling techniques were originally developed as probability sampling design alternatives for collecting information on populations at risk for AIDS/HIV.

Michael Last, U.S. Government

Social Network Analysis for National Security

Social Network Analysis can help commanders understand how an insurgent organization operates. Insurgent networks often do not behave like normal social networks. However, SNA can help commanders determine what kind of social network an insurgent organization is. That knowledge helps commanders understand what the network looks like, how it is connected, and how to best defeat it.

Dave Marchette, Naval Surface Warfare Center

Identifying Aliases in Graphs

Social network analysis operates on graphs defined by entities and relationships: the vertices of the graphs correspond to the entities, and the edges encode the relationships. In most applications, the set of entities is known, and the graph is either completely known, or is the result of some sampling scheme aimed at identifying a useful subset of the relationships. Consider an email communications graph, where the entities are email addresses and the edges correspond to the existence of an email between the two addresses in a given time window (the edge may be either directed or undirected, but we will ignore multiple emails within the time window). Each graph that we observe contains a subset of all possible email addresses, and so there are addresses (vertices) in the graph at time t that were not there at time $t-1$, and vice versa. Suppose that an individual changes email addresses, so at time $t-1$ she uses addr1 and at time t she uses addr2: addr2 is an alias she has taken on, and we wish to identify the original entity associated with this alias. Since there are typically several email addresses observed for the first time in each graph, and several addresses that are not observed in all graphs, it is not trivial to determine that the person using addr2 is the same as the one who previously used addr1. I will describe a method for matching the alias to the previous graph, which uses a random graph model, the random dot product graph. This model allows us to match the activity associated with addr2 to those in the previous graph (or set of graphs) to determine the best match. The method will be evaluated through simulation, and on a collection of email graphs from the Enron dataset.

Aparna Huzurbazar, Los Alamos National Laboratory

Incorporating System Reliability Estimates into Prognostics and Health Management (PHM)

Prognostics and Health Management (PHM) is increasingly important for understanding and managing today's complex systems. PHM develops the capability to make decisions about maintenance, based on prognostics information, resources, and operational conditions. As part of munitions stockpile surveillance, we have many potential sources of data, including full system pass/fail testing, quality assurance testing at the component or sub-system level, repair data, sensor data, and accelerated testing data. Our focus to date has been on Bayesian modeling for integrating such data into a unified model for system reliability estimates. This talk presents how to incorporate this information into a decision making framework for PHM.

David Siroky, Political Science, Duke University

Random Forest Analysis of Secessionist Movements

National security requires closer attention to the problems and opportunities that secessionist movements around the world present. Random Forests, in combination with other methodologies, can help target that attention. Random Forests are ensembles of trees grown from bootstrapped training data. Using two layers of randomness-a random sample of predictors and a random sample of observations- Random Forests offer significant improvements in accuracy over any single classifier or regression tree, reduce the dependence between covariates and possess a high degree of interpretability. These features make Random Forests not only accurate, but also useful for domain scientists concerned with national security affairs. This talk analyzes the empirical record of secession over the twentieth century in order to ascertain why some secessions produce peace, prosperity and order, while others lead to violence, irredentism and instability.