Real-Time Biosurveillance Program

Auton Lab workshop Report

Carnegie Mellon University's Auton Lab is the technology partner in providing analytics and detection software for the Real-Time Biosurveillance pilot project[1] (RTBP). Auton Lab Team leader: Prof. Artur Dubrawski (Director) visited Sri Lanka, April 21 – 25, 2009, to present the Auton Lab solutions, acquire knowledge of the Sri Lankan and Indian disease surveillance systems, understand the available data sets for analysis, and outline the deliverables. His visit included a 1 day workshop with the Indian and Sri Lankan researchers, conducting a colloquium at the University of Colombo School of Computing, and visiting project sites/facilities to experience, at first hand, the working environment. This report summarizes each of the aforesaid activities.

Key Words: Community, m-Health, Disease, Epidemiology, Surveillance, Mobile Phone, Statistical Data Mining, Alerting, Information, Communication, Technology, Capacity building, Sri Lanka

Workshop: Introduction to T-Cube

Researchers from India and Sri Lanka gathered at the Sarvodaya Community Disaster Management Center in Moratuwa, Sri Lanka on April 21, 2009 to participate in the Auton Lab conducted training program. The objectives of this workshop were for -

- Researchers to acquire knowledge on Auton Lab statistical analysis theory
- Receive an overview of the user/administrative manuals or guides
- Practices installation and activation of the software components
- Operate the software for performing various analysis
- Learn to interpret the data for detection of adverse events
- Build a set of Trainers to build capacity in their respective countries

First the audience was introduced to Auton Lab and their business, which predominantly is on data crunching for detecting events. A nutshell version of the underlying theory of the detection algorithms was presented.

The software algorithms are designed to detect events in multivariate and multi-stream datasets. The multivariate data would contain information on disease, syndrome, age, gender, and location of patients' illness. The RTBP at present offers a single stream of data, which is collected through the grassroots level healthcare workers. However, would attempt to include weather (temperature, rainfall, pressure, and wind-factor) as a secondary stream. Other possible data streams can come from the Agriculture, Pharmacy sales, School attendance, Emergency services, etc for determining correlations for identifying abnormal events with respect to disease outbreaks.

[1] RTBP project description - http://lirneasia.net/projects/2008-2010/evaluating-a-real-time-biosurveillance-program/

T-Cube is the name of the AD-Tree data structure based on Hash-tables that stores the multivariate data for fast queries that is three folds faster than normal relational databases. Bayesian networks are used to train the system for generating rapid outputs for real-time inputs. The graphic user interfaces that use the T-Cube data structure are similar to a computer game where the interfaces respond instantaneously to the user controls that define the decision maker's queries; where two sliders are used to define the time window and other check boxes are used for defining the various filters to scrutinize the data from all angels. Few of the controls include viewing the data set in a log scale, linear scale, moving average, Cu-Sum, baseline, and positives.

The statistical data mining algorithms use P-value to set the cutoff for distinguishing between adverse events and none events in the multivariate datasets; where the threshold is typically set to α = 5% (produces 5% of the results). The alerts are ranked according to the P-value. In some cases the 5% may still result in too many alerts and the False Data Recovery (FDR) setting reduces the threshold to about 2%; thus, producing less false positive alerts. The algorithms take in to account the Receiver Operating Characteristics (ROC) and Activity Monitoring Operating Characteristic (AMOR) for statistical process control; where the ideal solution space should be governed by the dimensions: number of alerts (y-axis) and latency (x-axis) should be closer to x = 0. The ideology behind this is that in order to generate less false positive alerts the algorithms must execute more steps (or test more hypothesis) which results in latencies. On the other hand reducing latency results in generating excess alerts and may disagree with the decision makers liking as it requires investigating all the alerts; where most of them may be false positives.

Temporal scan and *spatial scan* are the two prominent algorithms that try to detect anomalies in the data sets. The disease signature is a combination of the disease rates (densities) and the counts. Temporal scan looks at density estimates in the target data in a given time window and tests it against the baseline data. The algorithm performs a Chi-square test or Fisher's exact test on the disease densities rates inside the selected window against the disease densities rates outside the window to determine whether it is an anomaly or not. Spatial scan algorithm performs a similar statistical hypothesis testing but instead of looking at data in a time window it looks at data within and outside of a spatial window, usually a rectangular space. The rectangular area is moved across the geographical map (locations) to find the anomalies.

After presenting the T-Cube tool, the participants were given the chance to present their working environment in relation to public health. Both Indian and Sri Lanka researchers explained the scenarios to outline the strategy for customizing the analysis and detection tools. One of the discrepancies was related to defining locations and the geospatial hierarchies as well as the availability of lat/lon GIS boundary and point data. It may be the case that the project may have to generate these datasets. Another attribute that is missing in the RTBP dataset, but is important, is the on-set datetime of the patient's illness. This attribute may need to be added to the database and the mobile and web applications collecting the data.

Colloquium: University of Colombo School of Computing

Prof. Ruvan Weerasinghe, head of the University of Colombo School of Computing (UCSC), invited Prof. Dubrawski to a colloquium to present the Auton Lab work titled: *Machine Learning in*

Support of Biomedical Security. The audience consisted faculty and students mainly interested in machine learning. Dr. Nalin Ranasinghe introduced the speaker to the audience. The talk was geared towards machine learning algorithms for detecting events. A few key points mentioned the importance of efficiency, fast data structure, scalability, and self-adapting. Auton Labs main focus is on scalability and not adopting new methods. Data structures usually carry corroborating data and as a result the necessary data can be minimized.

The algorithms are geared towards rapid detection of emerging patterns in order to mitigate the impact of the actions by exploiting available early signals. Whether or not to raise an alert is based on Chi square test on the categorical or symbolic datasets. Running supervised learning in detection helps improve the false positives.

USCS faculty member, Mr. Harsha Wijeywardana, mentioned that they were building a data warehousing system to examine migratory birds (wild life) for detecting Bird Flu. Another project involves applying Furrier Transformations for detecting seasonal trends. UCSC also engages in other machine learning projects and is keen in collaborating with Carnegie Mellon University in these endeavors; especially using these additional data streams in the RTBP.

Field Visit: Hettipola District Hospital

Artur Dubrawski joined the Suwadana Center Volunteers, working as Research Assistants in collecting health data from healthcare facilities, in visiting the Hettipola District Hospital. This visit gave a close encounter of the operations in the Sri Lankan context to identify the point of care and services where data can be captured.

Other Remarks

Time lines were set for the Auton Lab deliverables -
May 01 – Software Requirements Specifications
June 15 – Integrate T-Cube with RTBP database
July 15 – Beta release of the Tamil Nadu and Kurunegala T-Cube applications
Sept 01 – Conduct simulated drills
Dec 15 – Artur Dubrawski visit India and Sri Lanka to consult health officials

In addition, Auton Lab will assist RTBP in developing a simulator to be used in the mock drills; where the simulator mimic the spread of disease and produce time lines and densities. The project will use these time lines and densities to inform the participants of the time and symptoms they should report to health workers during the mock drills. Approximately 25 households from each village, total of 32 villages, will participate in this drill. Thereafter, the project will evaluate whether T-Cube can detect the hypothetical disease spread.