Graph-based Data Mining in Epidemia and Terrorism Data

Courtney D Corley[1], Diane Cook[2], Lawrence B. Holder[2] and Karan P. Singh[1].

Background:

Graph-based data mining (GDM) is the task of finding novel, useful, and understandable graph-theoretic patterns in a graph representation of data. Our approach to graph-based data mining, Subdue, focuses on identifying novel, but not necessarily most frequent, patterns in a graph representation of data [Holder et al., 2003]. Subdue searches the space of subgraphs by extending candidate subgraphs by one edge. Each candidate is evaluated using a minimum description length metric [Rissanen, 1989], which measures how well the subgraph compresses the input graph if each instance of the subgraph were replaced by a single vertex. Data mining has been applied to a number of databases, but the range of applications is frequently limited to those that can be represented as transactions or as a collection of attribute/value pairs. Since tools such as Subdue have been created to process and mine structural data, new critical databases and applications can be pursued.

Methods:

We demonstrate application of this data-mining technique in two domains, epidemia and terrorist communication networks. Providing analysis of epidemiological data has the potential to provide greater insight to the nature of epidemics and other health crises and to provide timely recognition and classification of epidemic onsets. Epidemics are conventionally recognized through the spatial/temporal clustering of patients with similar illnesses or unusually high incidents of a specific disease and analysis with Subdue is well-suited for this type of structural data. Similarly, analysis of communication networks for finding and classifying erroneous behavior is beneficial to our nation's preparedness and averting future attacks.

As part of the U.S. Air Force program on Evidence Assessment, Grouping, Linking and Evaluation (EAGLE), a domain has been built to simulate the evidence available about terrorist groups and their plans prior to their execution. This domain is motivated from an understanding of the real problem of intelligence data analysis. The data consists of threat and non-threat actors and groups, targets, exploitation modes, capabilities, resources, communications, visits to targets, and transfer of resources.

The first step will be to decide on the best representation for the data that highlights critical pieces of information as well as critical relationships between data points that form the backbone of the graph representation. The data provided to us by the U.S. Air Force is well-described in graph-form and we need only to translate this into a Subdue readable format. The epidemiological data we analyze, specifically Syphilis report cases in Tarrant County Texas in the previous six years, does not contain structural information in its native format. Hence, we transform the report cases by adding structural information, such as connecting cases with neighboring geographic locations and temporally similar report dates.

Results:

In preliminary experiments, Subdue has been used to learn patterns distinguishing vulnerability exploitation cases (terrorist attacks) from productivity exploitation cases (legitimate uses) and to distinguish threat groups from non-threat groups. The learned pattern indicates that the same individual both visited the target and was involved in a generalized transfer of a resource. The initial experiments show that Subdue achieved 58-68% accuracy and on the threat group data, Subdue achieved up to 93% classification accuracy. Subdue's pattern discovery mode was applied to the Syphilis report data and in preliminary evaluations resulted in learning several substructures. The substructure with the highest informative value (not necessarily the most frequent) was male Hispanic reports cases in Fort Worth Texas with early latent Syphilis that were connected to other Hispanic report cases and closely related in age. Disseminating greater information and promoting awareness to this demographic stratum will help curtail the current Syphilis epidemic.

Conclusions:

We demonstrate an effective representation for epidemiological and terrorist data, and show that Subdue can discover patterns in the separate and combined databases that are useful for analysis of the data, classification of data into predefined classes, and prediction of new onset epidemic cases.

Acknowledgements:

We would like thank the National Science Foundation for support under grant NSF IIS-0505819.

References:

[Holder and Cook, 2003] L. B. Holder and D. J. Cook, Graph-based relational learning: Current and future directions. SIGKDD Explorations special issue on Multi-relational Data Mining, 5(1), pages 90-93, 2003.

[Rissanen, 1989] J. Rissanen, Stochastic Complexity in Statistical Inquiry. World Scientific, 1989.

1 University of North Texas – Health Science Center, School of Public Health, Department of Biostatistics, Fort Worth, TX

2 University of Texas at Arlington, Department of Computer Science and Engineering, Arlington, TX