Feature Discovery in Link Mining

Keywords: link mining, feature discovery, machine learning, graph theory, relational data

Background: Traditional data mining approaches attempt to find patterns in a data set characterized by a collection of independent instances of a single relation. This is consistent with the classical statistical inference problem of trying to identify a model given a random sampling of an underlying distribution. A key challenge for machine learning is the problem of mining more richly structured data sets in a way that leverages the linkages between records [1]. In this paradigm, which more accurately resembles real-world data, instances in the data set are relational where different samples are related to each other, either explicitly as typified by friendship relationships in a social network, or on the web by hyperlinks [2]. However, in most large data sets, relationships also exist that are not explicitly annotated. According to Jensen, naively applying traditional machine learning methods to this type of data can lead to inappropriate conclusions [3]. Therefore new approaches are needed to appropriately correlate inherent relationships (i.e. links) in real-world data sets.

In recent years, there has been a growing interest in learning from structured, real-world data. This type of data can be described by a graph where the nodes in the graph represent objects, and edges in the graph represent relationships between objects. Perhaps the most famous example of exploiting link structure is the PageRank algorithm [4] employed by the *Google* search engine.

Link mining is situated at the intersection of graph theory, machine learning, and web mining. This research is potentially useful in a wide range of application areas including bio-informatics, bibliographic analysis, financial analysis, national security, social network analysis, and internet search to name a few. While my research is focused more on the theoretical aspects of this topic than in the applicative possibilities, I was happy to see that my work has already been adapted to the bioinformatics domain to study the interactions of proteins [5].

Research: Despite the recent advances in link mining, this topic is still relatively new and there are many fundamental challenges that remain. Unlike more mature fields of research there does not exist any public package or toolkit that provides a standard baseline from which to explore. Therefore, I propose to create a link mining framework that adapts several of the core principles of link and graph mining into a scalable, shared package. This toolkit would be an essential research and teaching tool similar to the University of Waikato's WEKA toolkit [6] or George Mason's ECJ system [7]. Initially, this project would only incorporate fundamental and highly-extendable principles of link mining, but most importantly it will serve as a launch-pad for more interesting, collaborative theoretical work.

With a core link mining package in place I propose to study the dynamic temporal and graphical nature of relationships within various domains in order to advance the theory of and methodology for determining probabilities of link existence where none are explicitly annotated. This process involves several steps. First a domain must be selected that exhibits the relational attributes applicable to the link mining paradigm. Data from social networks, protein interactions, citations, microarrays, etc. all contain necessary attributes; therefore this step is arguably the most straightforward because many real-world data sets are inherently relational [1].

After the domains are defined, features that describe the relationships need to be extracted. For example, friendship in a social network is annotated by the inclusion of the friend's name on a user's homepage. Pair-dependent features, such as the size of the intersection of interests, etc., offer supplementary evidence for the existence of a friendship. These pair-dependent features will be used to determine the probability for link existence where it is not annotated. Finding the non-obvious pair-dependent features is arguably the most difficult part. Therefore, I propose the use of recent developments in association rule mining and frequent pattern mining by Dr. Jiawei Han et al. [8] to find correlations between data points that best suggest link existence. Furthermore, the general problem of feature selection, extraction and discovery is widely regarded as the most important factor in machine learning [9].

Besides pair-dependent features, I propose to explore the role that graph features have in identifying relationships that lack explicit annotation. In my experience, graph features, such as the shortest path distance between candidate vertices, offer the best support (in terms of entropy) for the existence or absence of links. The major problem with this approach is that extracting graph features is computationally expensive for sufficiently large graphs. Although I have begun work on developing fast, approximate search algorithms I will need to formalize and empirically study these methods. Finally, these features will be used by traditional machine learners to derive information about relationships in data sets.

In each step, theories would be tested using the aforementioned link mining toolkit in order to efficiently derive empirical results. I plan to advertise and freely share my toolkit, and continue to present and publish results at refereed conferences and in refereed journals on a regular basis. While my research generally aims to expand the theoretical and computational potential of machine learning, the implications of link mining research can already been seen in the biological, physical and social sciences, and many researchers believe that the application of link mining techniques will continue to grow as more research is conducted.

With help from the NSF GRF I intend to study at the University of Illinois Urbana-Champaign (UIUC) where the Data Mining Research Group led by Dr. Jiawei Han (reference letter writer) and I already have a working relationship. I believe that Dr. Han and his colleagues at UIUC are among the best researchers in the world, and they would provide the wisdom and expertise necessary for me to continue my work in this fascinating field.

References:

[1] Lu, Q., Getoor, L., "Link-based Classification". ICML'03, Washington DC, 2003.

[2] Sen, P., Getoor, L., "Link-based Classification". University of Maryland CS-TR-4858. 2007.

[3] Jensen, D., "Statistical challenges to inductive inference in linked data". In Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics. 1999.

[4] Page, L., Brin, S., Motwani, R. & Winograd, T. "The pagerank citation ranking: Bring order to the web". Technical Report. Stanford University. 1998.

[5] Paradesi, M.S.R., Caragea, D. and Hsu, W.H., "Structural Prediction of Protein-Protein Interactions in Saccharomyces cerevisiae", IEEE-BIBE'07, vol. 2, Boston, MA, Oct. 2007.

[6] Witten, I. H. and Frank, E., "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[7] ECJ: A Java-based evolutionary computation research system, 2006. http://cs.gmu. edu/eclab/projects/*ecj*/

[8] Han, J., Pei, J., & Yin, Y., "Mining frequent patterns without candidate generation", International Conference on Management of Data ACM-SIGMOD'00, pp. 1-12. 2000.

[9] Caruana, R, Niculescu-Mizil, A., "An empirical comparison of supervised learning algorithms". ICML'06, pp. 161-168, Pittsburgh, PA, 2006.