MINING INTERESTING SUBGRAPHS BY OUTPUT SPACE SAMPLING

By

Mohammad Al Hasan

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Dr. Mohammed Zaki, Thesis Adviser
Dr. Boleslaw Szymanski, Member
Dr. Sanmay Das, Member
Dr. John Mitchell, Member
Dr. Jeffrey T. Kreulen, Member

Rensselaer Polytechnic Institute
Troy, New York

July 2009
(For Graduation August 2009)

ABSTRACT

Lack of scalability of the mining process and the enormous size of the output set are two significant bottlenecks of Frequent Subgraph Mining (FSM). The first restricts the applicability of FSM to large datasets. The second makes it difficult for the user to analyze the frequent patterns for subsequent usage in typical knowledge discovery tasks, such as classification, clustering, outlier detection, etc. However, given the definition and the algorithmic mechanism, both the above problems are, in a way, inherent to FSM, so no immediate solution for them is perceivable.

The first problem, namely the lack of scalability is due to the combinatorial subgraph space which grows exponentially with the size of the database graphs. Another contributing factor to this problem is the complexity of the subgraph isomorphism test. Since this test is an essential sub-task of any subgraph mining algorithm, the well known result that it is NP-Hard dashes any hope of finding an effective solution to the lack of scalability problem.

The other problem, sometimes known as *information overload* can be solved to some extent. For that one needs to design effective summarization or filtering techniques that take the large output set of a graph mining algorithm and return a small set of subgraphs. But, typically for graph patterns these techniques are costly and when processing over a large data set, the aggregated cost is overwhelming. Another important point to note is that the two-step solution that finds all patterns and then summarizes or filters, fails implicitly, when the first step is infeasible due to the lack of scalability problem.

In this thesis, I propose output space sampling (OSS) to alleviate the above two problems. In this paradigm, the objective is to sample frequent patterns instead of complete enumeration. The sampling process automatically performs the interestingness based selection by embedding the interestingness score of the patterns in the desired target distribution. This obviates a two-step mechanism since the sampling automatically prefers the patters that are interesting. Another important point to note is that OSS is a generic method that applies for any kind of patterns

such as a set, a sequence, a tree and of-course a graph.

OSS is based on Markov Chain Monte Carlo (MCMC) sampling. It performs a random walk on the candidate subgraph partial order and returns subgraph samples when the walk converges to a desired stationary distribution. The transition probability matrix of the random walk is computed locally to avoid a complete enumeration of the candidate frequent patterns, which makes the sampling paradigm scalable to large real life graph datasets.

Output space sampling is an entire paradigm shift in frequent pattern mining (FPM) that holds enormous promise. While traditional FPM strives for completeness, OSS targets to obtain a few interesting samples. The definition of interestingness can be very generic, so user can sample patterns from different target distributions by choosing different interestingness functions. This is very beneficial as mined patterns are subject to subsequent use in various knowledge discovery tasks, like classification, clustering, outlier detection, etc. and the interestingness score of a pattern varies for various tasks. OSS can adapt to this requirement just by changing the interestingness function. OSS also solves *pattern redundancy* problem by finding samples that are very different from each other. Note that, pattern redundancy hurts any knowledge based system that builds metrics based on the structural similarity of the patterns.

Output space sampling is a general idea that has various applications. In this thesis, we utilize this general idea to solve specific problems. We consider two different problems: (1) frequent pattern summarization (2) sampling discriminatory patterns. For both the above problems, we assume that the direct mining task is infeasible, so that the user wants to adopt sampling to find few interesting patterns in a reasonable amount of time. We also use the concept of OSS to find representative patterns that are very different from each other. OSS naturally supports this requirement as it obtains random samples which are very different from each other. In this thesis, two different algorithms are proposed for representative pattern mining, which are introduced in the next two paragraphs.

The first algorithm for the representative pattern mining that is proposed in this thesis is called Musk. It is based on a uniform sampling of the output space. It

obtains representative patterns by sampling uniformly from the pool of all frequent maximal patterns; uniformity is achieved by a variant of Markov Chain Monte Carlo (MCMC) algorithm. MUSK follows the concept of OSS by sampling from a target distribution where the maximal patterns have uniform value for the interestingness score.

The second algorithm that we propose for this task is ORIGAMI. It defines the representative pattern-set ($\mathcal{R}$) in a novel manner that attempts to reduce structural similarities among patterns in $\mathcal{R}$ while extending the coverage of frequent pattern space as much as possible. Intuitively, two patterns are α-orthogonal if their similarity is bounded above by α. Each α-orthogonal pattern is also a representative for those patterns that are at least β similar to it. Given user defined $\alpha, \beta \in [0, 1]$, the goal of ORIGAMI is to mine an α-orthogonal, β-representative set that minimizes the set of unrepresented patterns. Similar to OSS paradigm, ORIGAMI uses a randomized algorithm to randomly traverse the pattern space, seeking previously unexplored regions, to return a set of maximal patterns. But, uncharacteristic to OSS, ORIGAMI employs a second-step to extract an α-orthogonal, β-representative set from the mined maximal patterns using a local optimal algorithm. The second step is essential to provide the α-orthogonal, β-representative guarantee.

For all the proposed algorithm, we show the effectiveness on a number of real and synthetic datasets. In particular, We show that the proposed algorithms are able to extract high quality patterns even in cases where existing enumerative pattern mining methods fail to do so.