Mining Interesting Subgraphs by Output Space Sampling*

[Ph.D. Thesis Abstract]

Mohammad Al Hasan
Indiana University - Purdue University, Indianapolis
alhasan@cs.iupui.edu

1. INTRODUCTION AND MOTIVATION

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, we 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 patterns that are interesting as defined by the user. 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.

In this thesis, we also present concrete sampling examples to illustrate the way the concept of OSS can be utilized to solve practical data mining problems. For instance, for exploratory data analysis, we suggest uniform sampling of frequent subgraphs and for frequent subgraph summarization, we suggest uniform sampling of *maximal* frequent subgraphs (Musk algorithm). We also show that by using *delta score* as the interestingness function, output space sampling concept can sample subgraph patterns that are effective for the task of subgraph classification.

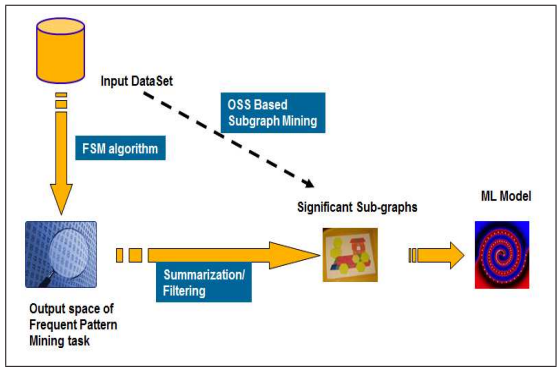


Figure 1: The scope of Output Space Sampling in Frequent Subgraph Mining

2. TECHNICAL SUMMARY

Figure 1 shows the scope of OSS in frequent pattern mining. As it is shown in this figure, a frequent subgraph (pattern) mining step is traditionally followed by a summarization or filtering step that identifies the significant patterns that are useful for the subsequent machine learning job. But, for large datasets both the mining task and the summarization task can be infeasible. OSS skips the complete enumeration of all the frequent patterns, as shown by the dotted line in this figure and obtains the significant (interesting) patterns in a direct step.

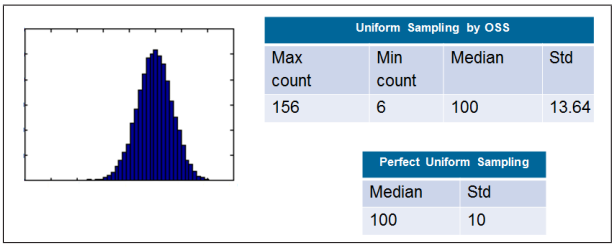
Output space sampling is an entire paradigm shift in frequent pattern mining (FPM) that holds enormous promise. While traditional FPM strives for completeness, OSS targets to obtain a few interesting samples. The definition of interestingness can be very generic, so user can sample patterns from different target distributions by choosing different interestingness functions. This is very beneficial as mined

*This thesis work is done at Rensselaer Polytechnic Institute, Troy, NY

This dissertation was selected as the winner of the 2010 ACM SIGKDD Ph.D. Dissertation Award. Complete version can be downloaded from the following web location: `http://www.kdd.org/awards_dissertation.php`

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 the *pattern redundancy* problem by finding samples that are very different from each other.

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 stationary distribution is defined implicitly by defining an *interestingness* function $f : \mathcal{F} \rightarrow \mathbb{R}_+ \cup 0$, where $\mathcal{F}$ is the set of all the patterns in the output space and $\mathbb{R}_+$ is the set of positive real numbers. For instance, if output space of a mining task contains n subgraphs, with the interestingness scores, such as, $s_1, s_2, \ldots, s_n$, then the desired sampling distribution is defined as $\pi(i) = \frac{s_i}{\sum_i s_i}$. 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.



Uniform Sampling by OSS

Max count	Min count	Median	Std
156	6	100	13.64

Perfect Uniform Sampling

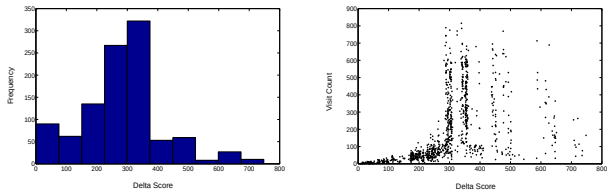
Median	Std
100	10

Figure 2: Uniform Sampling performance on an itemset dataset, Distribution of visit counts (left), statistics(right)

Output space sampling is a general idea that can be adapted for various applications by changing the desired sampling distribution. In this thesis, we show the examples of various sampling distributions of which the following two are particularly interesting: (1) uniform sampling of frequent patterns (2) sampling discriminatory patterns for classification. For (1), the interestingness function is a constant function; i.e. all the patterns are equally interesting, which yields a uniform sampling. For (2), the constant function performs poorly, since all frequent subgraphs are not good feature for graph classification. We used *delta score* for (2), which is the difference between the *positive support* and *negative support*, i.e. for a frequent subgraph g, we considers the labels of its support-list and partition them in two classes based on the label value. Delta score is just the difference between the sizes of these two sets.

In Figure 2, we show the performance of uniform sampling on an itemset dataset (Chess dataset from UCI Library). For this experiment we ran the uniform sampler for sufficiently large number of times and found the sampling count for each itemset in the output space. For a uniform sampling, the frequency distribution of the sampling count values follow a normal distribution, which is successfully mirrored by the OSS uniform sampler as shown on the distribution plot on the left of Figure 2. The statistics between the ideal and achieved distribution (by OSS) is also shown (right).

In Figure 3, we show the performance of OSS while finding subgraph features for graph classification for a chemical graph dataset. For this task, subgraph patterns with high delta score are desirable. On the left, we plot the frequency distribution of delta score of all the frequent patterns. On the right, we show a scatter plot between sampling (visit) count and delta scores for the patterns returned by OSS. It is easy to see that patterns with low delta score are undersampled but the patterns with high delta score are oversampled by OSS.



(a) (b)

Figure 3: (a) Frequency distribution of Delta Score in frequent patterns (b) Scatter plot of visit count vs delta score

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 obtains representative patterns by sampling uniformly from the pool of all frequent maximal patterns. 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 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 algorithms presented in this thesis, 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.

Ph.D. Dissertation Committee

Dr. Mohammed Zaki (Chair), Dr. Boleslaw Szymanski, Dr. Sanmay Das, Dr. John Mitchell, and Dr. Jeffrey Kreulen