New Sampling-based Approximation Algorithms for Frequent Itemsets Mining

Shiyu Ji, Kun Wan
{shiyu, kun}@cs.ucsb.edu

1 Problem Statement

1.1 Introduction

Frequent Itemsets (FI) mining has been popular in research recently [1, 7, 13]. The goal of FI mining is to find out the items that most frequently appear in the observed transactions, e.g., the researchers who are the most prolific in writing papers with others, the patterns that appear frequently in a long piece of genetic code, etc.

In the era of big data, to compute the exact frequencies can be very time consuming. Thus in many cases approximated values are also acceptable [13]. For FI mining in large scaled transactional dataset, we often take samplings on the transactions, and compute the frequencies of the itemsets among the sampled transactions as approximated results of their true frequencies among all the transactions. Usually the sampling size is much less than the scale of all the transactions, and the approximations can achieve acceptable precision. Thus FI approximation can be useful in practice.

The state-of-art progressive sampling based FI approximation algorithms [13] need an upper bound of the approximation error for the worst case, i.e., the maximum error the algorithm can generate among all the items. The algorithms keep taking new samples until the upper bound is less than the acceptable threshold. Hence how to bound the maximum error as tightly as possible is an interesting problem. The current bounds use some results of Rademacher average in statistical learning theory [16, 17, 3, 2]. However, we find that based on the idea given by [3], we can develop a new upper bound *without* Rademacher average. We also find that this new bound is asymptotically tighter than the existing bounds, i.e., given the chosen samples, as the allowed error probability approaches zero, the new bound is roughly only half of the existing ones. This implies that by using the new bound, a progressive sampling based FI approximation algorithm can reach the guaranteed accuracy with much fewer samples. We also notice that there is no parameter in the new bound that needs to be progressively computed. Hence the sample size that will guarantee the worst error can be precomputed.

Based on the similar idea, we also consider the top-k FI mining problem, which seeks for the k most frequent itemsets in the observed ones. We need to decide when the sampling should stop. The number of the samples transactions is enough if the worst-case error upper bound is less than the frequency gap between the k-th and the $(k+1)$-th most frequent itemsets. Hence we propose a progressive approximation algorithm to address the top-k FI mining problem.

Our Plan. We plan to give a new worst-case error upper bound that is asymptotically tighter than the state-of-art bounds, and propose an approximation algorithm which can guarantee the worst-case error upper bound with precomputed sample size. We will also give a progressive sampling algorithm to find the top-k most frequent itemsets. We will need experiments on real data sets to evaluate the performance of our algorithms, and compare our results with the state-of-art approximation algorithms.

1.2 Related Works

Sampling-based frequent itemset approximation has been studied extensively by the researchers. The first works on this problem used heuristic methods to progressively approximate the frequencies [4, 5, 8]. There

Name	No. of Transactions	No. of items
accidents	340183	468
chess	3196	75
connect	67557	129
kosarak	990002	41270
mushroom	8124	119
pumsb	49046	7116
pumsb star	49046	7116
retail	88162	16470

Table 1: Dataset characteristics

were no guarantee on the worst-case error upper bound. To fix this, Riondato and Upfal were the first to propose FI approximation algorithms that could guarantee the worst error bounds by using results of Vapnik-Chervonenkis (VC) dimension [11, 12] and Rademacher average [13]. Note that in statistical learning theory VC dimension and Rademacher average are usually used to address the worst-case error upper bound for *infinite case*, i.e., the number of possible functions in the learning model is infinite [3]. However in the case of FI mining, since there are only *finite* itemsets, it is possible to develop bounds without VC dimension or Rademacher average [3]. In this paper we apply this idea on FI mining problem. Riondato et al. also considered using parallelism in FI mining [10], which is an orthogonal topic to sampling-based FI approximation.

In practice we are often only interested in the most frequent itemsets. Thus top-k FI mining is a popular research topic with many research works [9, 15, 13, 14].

2 Data Sets

For consistency, we choose FIMI'03 data repository [6], the real-world data set from [13] (the data repository is available at `http://fimi.ua.ac.be/data/`). The item and transaction data sizes of the FIMI datasets are given in Table 1. We will use these data sets to evaluate the samples sizes and worst case errors of our algorithms, and compare our results with the state-of-art algorithms.

3 Deliverables

We plan to deliver the documents as follows:

- A report or paper about our new bounds and algorithms, and experiment results.
- Source code of our experiments on real data sets.

4 Milestones

As our plan, we list the possible milestones in the future as follows:

- Establish a new worst-case error upper bound and do the proof work.
- Propose sampling-based approximation algorithms, i.e., estimating frequency for each itemset, approximating the top-k frequent itemsets, etc.
- Evaluate our algorithms with real-world data sets, and compare our results with the existing works like [13].

References

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[2] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[3] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[4] Bin Chen, Peter Haas, and Peter Scheuermann. A new two-phase sampling based algorithm for discovering association rules. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–468. ACM, 2002.

[5] Kun-Ta Chuang, Ming-Syan Chen, and Wen-Chieh Yang. Progressive sampling for association rules based on sampling error estimation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 505–515. Springer, 2005.

[6] Bart Goethals and Mohammed J Zaki. Advances in frequent itemset mining implementations: report on fimi'03. *ACM SIGKDD Explorations Newsletter*, 6(1):109–117, 2004.

[7] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

[8] Srinivasan Parthasarathy. Efficient progressive sampling for association rules. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 354–361. IEEE, 2002.

[9] Andrea Pietracaprina, Matteo Riondato, Eli Upfal, and Fabio Vandin. Mining top-k frequent itemsets through progressive sampling. *Data Mining and Knowledge Discovery*, 21(2):310–326, 2010.

[10] Matteo Riondato, Justin A DeBrabant, Rodrigo Fonseca, and Eli Upfal. Parma: a parallel randomized algorithm for approximate association rules mining in mapreduce. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 85–94. ACM, 2012.

[11] Matteo Riondato and Eli Upfal. Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 25–41. Springer, 2012.

[12] Matteo Riondato and Eli Upfal. Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(4):20, 2014.

[13] Matteo Riondato and Eli Upfal. Mining frequent itemsets through progressive sampling with rademacher averages. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1005–1014. ACM, 2015.

[14] Matteo Riondato and Fabio Vandin. Finding the true frequent itemsets. In *SDM*, pages 497–505. SIAM, 2014.

[15] Tobias Scheffer and Stefan Wrobel. Finding the most interesting patterns in a database quickly by using sequential sampling. *Journal of Machine Learning Research*, 3(Dec):833–862, 2002.

[16] Vladimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[17] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.