Rare Pattern Mining

Ian Smith

April 15, 2017

University of Manitoba

# Introduction

Data mining is defined as the non-trivial extraction of implicit, previously unknown, and potentially useful information from data. Rare Pattern mining is the extraction of patterns that occur very infrequently within a database. Currently, much of the research available on data mining focuses on frequent pattern mining; however, rare pattern mining is an exceedingly interesting, growing field within computer science . My research will demonstrate that a frequent pattern mining algorithm, the Frequent Pattern Growth algorithm, can be modified to retrieve interesting rare itemsets that can then be studied. We hope to also show that our Rare Pattern Growth algorithm can do this efficiently.

There are quite a few different applications for rare pattern mining. For example, rare pattern mining can be employed for fraud detection. In a database of credit card transactions, the rare patterns could turn out to be fraudulent, and the credit card company could use a rare pattern mining algorithm to detect and prevent these transactions from harming their customers. A company that does this might be able to attract more customers by promising them protection.

The Frequent Pattern Growth algorithm uses a divide-and-conquer approach toward mining frequent patterns. The frequent single items of each transaction are added one at a time into a tree data structure, and then each frequent itemset is mined from the tree using a recursive function. My Rare Pattern Growth will operate in much the same way. The more frequent rare items will be added into the tree first, and mined from it last.

The organization of this report is as follows: first I will examine existing literature in data mining that focuses on rare itemset mining, as well as Frequent Pattern Growth. Second, I will describe my algorithm to mine rare itemset using diagrams and pseudo-code. I will also be examining the efficiency of using VIPER. Third, I will analyze the results of my algorithm. And finally I will conclude with an examination of what I learned from running the algorithm, and how it could be applied to the data mining field as a whole.

# Literature Review

Aggrawal and Han wrote one an excellent, comprehensive book about the study of frequent pattern mining. This resource has proved invaluable in examining Frequent Pattern Growth for this project, and other areas of data mining, as it successfully provides an overview of the different frequent pattern mining methods .

Weng proposed an Apriori-based mining approach called Fuzzy Apriori Rare Itemset Mining [FARIM], for mining “specific rare itemsets consisting of quantitative data” .[[1]](#footnote-1) Weng proposed using FARIM for low test or quiz scores in a school setting; if there was a student, or a group of students struggling with class content, then determining exactly what it is they are struggling with would go a long way in finding a solution [3]. Weng believed that his approach would be more successful if it included clustering and classification methods, and if the support parameter was inferred from the data.

Hemalatha, Vaidehi, and Lakshmi wrote about finding rare itemsets in data streams, as opposed to static datasets . To that end, they proposed an algorithm for finding Minimal Infrequent Patterns from Data Streams, defined three measures for outlier detection, and created a Minimal Infrequent Pattern based Outlier Detection algorithm. They found, among other things, that their methods were well suited for extracting useful data from sensor data streams and identifying meaningful outliers from those streams.

Wu, Chen, and Chang wrote about Attribute-Oriented Induction (AOI), and proposed using AOI to mine negative generalized knowledge from datasets . Their reasoning has to do with medical data; for example, if only a few Taiwanese people were infected with the H1N1 flu virus the number of people that are Taiwanese and have contracted H1N1 will be very small, and not considered a frequent itemset. However, if few Taiwanese contracted H1N1, then that might indicate that the Taiwanese were somehow resistant to the disease.

Agrawal and Agrawal presented an overview of how data mining techniques could be used to detect anomalies in datasets . Their Classification approach was to build a model based on the normal behavior of the system, and then feed testing data into that model in order to determine which datasets were anomalous . They tried various different Clustering approaches including k-Means, k-Medoids, and other approaches. They found that hybrid approaches, which combine Clustering and Classification based anomaly detection systems, had the best chance at finding anomalous behaviors in Intrusion Detection Systems.

Lin, Liao, and Chen actually wrote about using the Frequent Pattern Growth algorithm to find frequent itemsets, and in particular to reduce the number of candidate itemsets examined by the algorithm, and reducing the number of times it is necessary to scan the entire database . This is because the Apriori algorithm requires scanning the database repeatedly, and Dynamic Hashing and Pruning algorithm improves the performance of Apriori and lowers the cost of database scanning. But FP-Growth goes even further to improve the performance of Frequent Pattern Mining.

Lin, Lao, and Chen then propose using an Improved FP-Growth algorithm to improve the performance of FP-growth . They do this in part by using an address-table structure to lower the complexity of mapping frequent 1-itemsets in an FP-tree, and by using a hybrid FP-tree mining method that reduced the need to rebuild conditional FP-trees. Their simulation shows that their algorithm improved the performance of FP-growth by an order of magnitude in terms of execution time.

Cagliero and Garza actually use FP-Growth in a similar to fashion to what I am using it for, rare pattern mining .

# Description

# Analysis

# Conclusion

# Bibliography

|  |  |
| --- | --- |
| [1] | L. Cagliero and P. Garza, "Infrequent Weighted Itemset Mining using Frequent Pattern Growth," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, no. 4, pp. 903-915, April 2014. |
| [2] | Y.-Y. Wu, Y.-L. Chem and R.-I. Chang, "Mining negative generalized knowledge from relational databases," *Knowledge-Based Systems,* vol. 24, no. 1, pp. 134-145, February 2011. |
| [3] | C.-H. Weng, "Mining fuzzy specific rare itemsets for education data," *Knowledge-Based Systems,* vol. 24, no. 5, pp. 697-708, July 2011. |
| [4] | C. S. Hemalatha, V. Vaidehi and R. Lakshmi, "Minimal infrequent pattern based approach for mining outliers in data streams," *Expert Systems with Applications,* vol. 42, no. 4, pp. 1998-2012, March 2015. |
| [5] | S. Agrawal and J. Agrawal, "Survey on Anomaly Detection using Data Mining Techniques," *Procedia Computer Science,* vol. 60, pp. 708-713, 2015. |
| [6] | K.-C. Lin, I.-E. Liao and Z.-S. Chen, "An improved frequent pattern growth method for mininig association rules," *Expert Systems with Applications,* vol. 38, no. 5, pp. 5154-5161, May 2011. |
| [7] | C. C. Aggarwal and J. Han, Eds., Frequent Pattern Mining, Springer Cham Heidelberg, 2014. |
| [8] | M. Adda, L. Wu, S. White and Y. Feng, "Pattern Detection with Rare Itemset Mining," *International Journal on Soft Computing, Artificial Intelligence and Applications,* vol. 1, no. 1, pp. 1-17, August 2012. |
| [9] | S. Bhattacharyya, S. Jha, K. Tharakunnel and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Elsevier,* vol. 50, no. 3, pp. 602-613, February 2011. |
| [10] | R. S. Debreceny and G. L. Gray, "Data mining journal entries for fraud detection: An exploratory study," *International Journal of Accounting Information Systems,* vol. 11, no. 3, pp. 157-181, September 2010. |
| [11] | D. Yu, G. Sheikholeslami and A. Zhang, "Finding Outliers in Very Large Datasets," *Knowledge and Information Systems,* vol. 4, pp. 387-412, 2002. |

1. The Apriori algorithm is a frequent pattern mining algorithm where frequent single-item-sets are combined to create larger frequent itemsets, and then the database is scanned to determine the support of the new itemsets. This process continues until there are no more itemsets that can be combined [7] [↑](#footnote-ref-1)