Rare Pattern Mining Proposal

Ian Smith

January 25, 2017

University of Manitoba

Data mining is defined as the non-trivial extraction of implicit, previously unknown, and potentially useful information from data. Much of the literature currently available for data mining focuses on frequent itemsets, but a great deal can be learned from the examination of rare itemsets as well. There are already a number of algorithms available for frequent pattern mining, and I believe that they can be adapted to be used in rare pattern mining. To that end, I propose to develop an approach to mine rare itemsets from a database based on frequent pattern mining. My initial review of relevant literature follows.

Weng proposed an Apriori-based mining approach called Fuzzy Apriori Rare Itemset Mining [FARIM], for mining “specific rare itemsets consisting of quantitative data” [1, p. 698].[[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 [1]. 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 [2]. 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 [3]. 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.

I would like to add to what these researchers found and modify an existing frequent pattern mining algorithm, such as Apriori of FP-Growth[[2]](#footnote-2) to mine rare itemsets from databases. My algorithm should return itemsets that are present in the database with very little support. Since there could be a large number of these itemsets, it may also be necessary to employ clustering and classification based approaches to isolate particularly rare or interesting itemsets like Weng suggested, or to focus on a specific kind of rare itemset. I also plan on varying the support variable in different databases.

One result of this project will most likely be a software prototype. To that end it will also be necessary to create or find basic test data. It would be ideal if this software prototype worked with empirical data. I also expect to prove that the itemsets returned by the algorithm are actually rare, first due to a lack of support in the database, and second, that they are significantly dissimilar to other itemsets with little support.

# Bibliography

|  |  |
| --- | --- |
| [1] | C.-H. Weng, "Mining fuzzy specific rare itemsets for education data," *Knowledge-Based Systems,* vol. 24, no. 5, pp. 697-708, July 2011. |
| [2] | 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. |
| [3] | 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. |
| [4] | C. C. Aggarwal and J. Han, Eds., Frequent Pattern Mining, Springer Cham Heidelberg, 2014. |

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 [4] [↑](#footnote-ref-1)
2. Frequent Pattern (FP) Growth algorithms follow a tree-based approach in order to reduce the number of times the database needs to be scanned. [4] The tree is constructed in such a way that the support of any itemset can be inferred easily. [4] [↑](#footnote-ref-2)