Exploring Association Rule Mining

Wei Li

February 25, 2021

1 Introduction

Association rule mining is a technique to identify interesting relations(or pattern) between variables in large database. A typical usage of association rule mining is market basket analysis. We limit the definition and project to market basket analysis, but the problem can be generalized to a wide range of applications from Natural Language Processing, Biologics to Medicals.

1.1 Problem Definition

In the market basket analysis, a retailer has various items in an itemset $I = \{i_1, i_2, \ldots, i_n\}$ and a database of shopper's shopping cart $\Omega = \{R_1, R_2, \ldots, R_m\}$ where $R_i \subseteq I$. A pattern or association rule in Ω can be understood as an if-then relationship(co-occurrence) $A- > B$ that if itemset A is bought by a shopper then itemset B also has high chance to be bought.The goal of Association rule mining is building all rules from the database with certain criterion.

The most common approach to find associations is counting frequency. Support (a.k.a

frequency) measure the absolute frequency of co-occurrence as

$$Support = \frac{freq(AB)}{|\Omega|}$$

Where AB indicates A and B occurs in the same transaction. And confidence is the relative frequency defined as

$$Confidence = \frac{freq(AB)}{freq(A)}$$

Having the two frequency criterion defined, one can generate association rules in Ω with either minimal Support or Confidence. For example,

- generate $A- > B$ with support $c\%$, such that $c\%$ of records in Ω contains AB.
- generate $A- > B$ with Confidence $c\%$, such that $c\%$ of records in Ω containing A also contains B.

When building association rules in such a way, the problem is also known as **Frequent Itemset Mining.**

Association rule mining in I is tough as it requires to generate 2^n subsets of I and count the co-occurence of each of the two subsets.

1.2 Solutions

The Association rule mining has been solved in two perspectives: sequential or parallel; exhaustive or heuristics-based.

For sequential and exhaustive solutions, Apriori (Agrawal, Imielinski, and Swami) was the first proposal solution to association rules mining in 1993, which uses breadth first search and pruning in subsets. After Apriori a variety of Apriori variations had been proposed. Different from, FP-Growth (Han, Pei, and Yin) is another highly influenced method with

an order of magnitude faster than Apriori. The method uses a novel frequent pattern tree structure for storing patterns information. At the same time, ECLAT (Zaki) shows some imporvement over FP-Growth and is also widely used nowadays. There are more modern solution to the problem from 2000s.

For sequential and/or parallel heuristics-based solution, early work (Mata, Alvarez, and Riquelme) firstly used genetic algorithm to solve the problem. A more recent (Yan, Zhang, and Zhang) genetic algorithm implementation allows non-fixed support threshold. More heuristics-based methods were proposed in the 2010s and were mostly genetic algorithm.

Parallel solutions are proposed to meet the large-scale data and distributed computation need. These algorithms are largely based on MapReduce technology. Also, as the GPU computation becomes widely used nowadays, there are works that adapt previous sequential algorithms to GPU architecture.

1.2.1 Scope of project

We are interested in studying three of the benchmark algorithms Apriori (Agrawal, Imielinski, and Swami), FP-Growth (Han, Pei, and Yin), ECLAT (Zaki) and implement these algorithms. Experiments comparing these algorithms will be done using market basket analysis data set. Some small scale are available on Github and large scale data sets can be found on Kaggle: DataA;DataB. Some of the data sets are not well-structured and needs cleaning and reorganization.

We are also interested in looking at one genetic algorithm solution (Yan, Zhang, and Zhang). If time and workload are permitted, we may add a genetic algorithm solution to the experiment.

Also, we will include a more extensive and complete literature review in the report.

References

[AIS93] R. Agrawal, T. Imielinski, and Arun N. Swami. “Mining association rules between sets of items in large databases”. In: *SIGMOD ’93*. 1993.

[HPY00] Jiawei Han, J. Pei, and Y. Yin. “Mining frequent patterns without candidate generation”. In: *SIGMOD ’00*. 2000.

[MAR01] J. Mata, J. L. Alvarez, and J. Riquelme. “Mining Numeric Association Rules with Genetic Algorithms”. In: 2001.

[YZZ09] Xiaowei Yan, C. Zhang, and S. Zhang. “Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support”. In: *Expert Syst. Appl.* 36 (2009), pp. 3066–3076.

[Zak00] Mohammed J. Zaki. “Scalable Algorithms for Association Mining”. In: *IEEE Trans. Knowl. Data Eng.* 12 (2000), pp. 372–390.