Deliverable 1: Final Year Dissertation

Sahil Pattni

BSc. Computer Science (Honours)

Supervisor: Neamat El Gayar

November 28, 2020

Declaration

I, Sahil Manojkumar Pattni, confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed: Sahil Manojkumar Pattni

Date: November 28, 2020

Abstract

In this digital age, data is being generated at an exponential rate, with data analytics being used by corporations and small businesses alike to produce insights, reduce costs, optimize operations and increase profits. Generating association rules allow us to find non-intuitive associations that bring new insights to the management, allowing them to leverage this information to maximize their profits. A prime example would be the *'Beers and Diapers'* case [1], where a company looked at their point-of-sale data and found a strong association between beers and diapers being co-purchased, which seems rather unintuitive.

In this study, we will extract a minimum spanning tree (MST) from a data set using machine learning techniques. We will then use this minimum spanning tree to segment products together, and produce association rules and extract the most interesting rules. The resulting rule-set will be compared to rules generated by the established Apriori Algorithm [2], which is the foundation of association rule mining. Additionally, we will study how the grouping of the MST compares to clustering algorithms, and how the structure of our MST would change before and after promotional events, leading to insights that may help the management in such firms make better informed decisions about the type of promotions they would like to run.

Contents

1 Introduction

1.1 What is a Minimum Spanning Tree?

Given an undirected graph G with edges E and vertices V (i.e. $G(E, V)$), a *spanning tree* can be described as a subgraph that is a tree [3] which includes all the vertices V of G with the minimum number of edges required. A *minimum spanning tree* is the spanning tree with the smallest sum of edge weights.

1.2 Context

For business which deal with the sale of a heterogeneous physical assets - such as groceries, hypermarkets and select retail outlets - operations such as inventory management and product placement play an instrumental role in determining the business' financial success. These involve asking questions such as:

- Which products should be placed at the entrance of the store? Which should be placed closer to the exit?
- Which products will benefit the most by being placed at eye-level?
- Which products should be placed next to each other to maximize the purchase volume?

One way to find optimal solutions for such questions is to employ the use of Association Rule Mining techniques. These techniques assess frequent itemsets (e.g. from sales data) and generate association rules between products.
Several algorithms and techniques exist for association rule mining, such as the Apriori Algorithm [2] and FP-Growth [4]. One problem with these algorithms is that they tend to generate an enormous amount of rules, of which many of the rules themselves are large [5]. This makes it grossly inconvenient for the end-user to retrieve any actionable information from the results. What we will do is construct a network of products - where each vertice represents a product or product category, and an edge between two vertices represents the simultaneous occurrence of the the two products [6]. The result is a graph (i.e. a network), whose architecture is visually informative of the products and their associations. The insights that can be gained from the MST can help the management of the mentioned businesses to maximize their profit.

1.3 Aims

The aim of this study is to study the effectiveness of a minimum spanning tree in product classification and association rule mining, in addition to determining how the architecture of the MST will change during and after large events such as promotions. The model will be tested on a relatively large dataset of sales data.

1.4 Objectives

The research objectives for this project have been laid out below, in the order that they will be carried out.

1. Acquire a suitable dataset upon which the MST can be constructed.

2. Explore and evaluate MST extraction algorithms.
   (While Prim's [7][8] and Kruskal's [9] are the most commonly used algorithms to extract the MST from a graph, research has been conducted on more efficient ways to extract the MST using machine learning techniques such as Artificial Neural Networks [10] and K-Nearest-Neighbors [11]).

3. Generate an undirected graph $G(E, V)$ where the edges E are the correlation values between the product vertices.

4. Extract a minimum spanning tree from this graph using the technique determined in Step 2.

5. Analyze this MST and use its architecture to determine product clusters and generate association rules.

6. Generate association rules using the Apriori and FP-Growth algorithm and compare them in both their time complexity and their *'interestingness'* (i.e. the unintuitive rules they generate).

7. Validate the product grouping present in the MST against a clustering algorithm (e.g. such as the K-Means algorithm).

References

[1] D. J. Power, Ed., *Ask Dan! What is the "true story" about data mining, beer and diapers?* Vol. 3 Nov. 2002, ISSN: 23. [Online]. Available: http://www.dssresources.com/newsletters/66.php.

[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *Proceedings of the 20th International Conference on Very Large Databases*, pp. 487–489, 1994.

[3] E. Williamson, *Lists, Decisions and Graphs*. S. Gill Williamson, p. 171.

[4] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 1–12, 2000. DOI: 10.1145/342009.335372.

[5] S. Kotsiantis and D. Kanellopoulos, "Association rules mining: A recent overview," *GESTS International Transactions on Computer Science and Engineering*, vol. 32, pp. 71–82, 2006.

[6] T. Raeder and N. V. Chawla, "Modeling a store's product space as a social network," *International Conference on Advances in Social Network Analysis and Mining*, pp. 164–169, 2009. DOI: 10.1109/ASONAM.2009.53.

[7] R. C. Prim, "Shortest connection networks and some generalizations," *Bell System Technical Journal*, vol. 6, pp. 1389–1401, 6 1957. DOI: 10.1002/j.1538-7305.1957.tb01515.x.

[8] V. Jarník, "O jistém problému minimálním," *Práce Moravské Přírodovědecké Společnosti*, pp. 57–63, 1930.

[9] J. B. K. Jr., "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical Society*, vol. 7, pp. 48–50, 1956. DOI: 10.1090/S0002-9939-1956-0078686-7.

[10] G. Ferilli, P. L. Sacco, E. Teti, and M. Buscema, "Top corporate brands and the global structure of country brand positioning: An autocm ann approach," *Expert Systems with Applications*, vol. 66, pp. 62–75, Dec. 2016. DOI: 10.1016/j.eswa.2016.08.054.

[11] X. Wang, X. L. Wang, Y. Ma, and D. M. Wilkes, "A fast mst-inspired knn-based outlier detection method," *Information Systems*, vol. 48, pp. 89–112, Mar. 2015. DOI: 10.1016/j.is.2014.09.002.