Faculty of Electronics and Information Technology

WARSAW UNIVERSITY OF TECHNOLOGY

Course
Data Mining (103B-CSCSN-MSA-EDAMI)

Project Report

Generating non-redundant Association Rules based on Closed Frequent Itemsets.

Bruno Axel Kamere
Student Record Book Number 330749

Supervisor
Robert Bembenik, PhD,

Warsaw, Poland 2023

Contents

1 Introduction **1**

1.1 Problem definition 1

2 Solution Description **3**

3 Implementation Description **5**

4 User Manual **7**

5 Data set Description **9**

6 Results **11**

7 Conclusion **13**

Chapter 1

Introduction

Data mining is the process of discovering patterns in large data sets by using mathematical algorithms and statistical methods. The main goal of data mining is to extract information from a data set and transform it into an understandable structure that can be used to solve business problems and real world problems through data analytics.

This process involves implementing algorithms that define rules <TODO: Read more about rules and reference>. These rules are then applied to the data set to discover patterns that can be used to make predictions. The patterns discovered can be used to answer questions about the data set and to solve problems. The patterns discovered can also be used to make predictions about future events. The most common rules used in data mining are association, sequence, classification, clustering, and forecasting. This project will focus mainly on association rules discovery based on closed frequent itemsets.

<TODO: Do more research and improve this introduction>

1.1 Problem definition

<TODO: Write problem definition>

Chapter 2

Solution Description

This project proposes implementing the Charm algorithm to generate non-redundant association rules based on closed frequent itemsets. Charm (Closed itemset Miner) is a depth-first search algorithm specifically designed to discover closed frequent itemsets efficiently. By leveraging Charm, we can obtain a compact and non-redundant set of itemsets, ensuring the generated association rules are informative and meaningful.

The solution is implemented in Python and involves several steps. First, the Charm algorithm is used to mine closed frequent itemsets from the dataset. The algorithm explores the lattice of itemsets, pruning unnecessary search branches and avoiding redundant itemset generation. This step guarantees that only closed frequent itemsets, which have no supersets with the same support, are considered.

Once the closed frequent itemsets are obtained, association rules can be generated. A pruning technique is applied to eliminate rules that are subsumed by other rules with higher confidence. This pruning step ensures that only the most interesting and distinct rules are retained, further reducing redundancy.

The solution will be experimented on the Traffic Accidents data set (http://fimi.uantwerpen.be/data/a <TODO: Add proper citation with bibtex later> to evaluate its performance. The results will be compared with those obtained by other algorithms like Apriori to determine the effectiveness of the proposed solution.

Below is pseudocode for the Charm algorithm: - Create a function to extract unique items from the data. The function should take in a dataset as a parameter and return a list of unique items in the dataset. - Create a function to calculate the support of an itemset in the data set. The function should take in a dataset and an itemset as parameters and return the support of the itemset in the dataset. - Create a function, Charm, that takes in a dataset and a minimum support threshold as parameters. The function should, somehow (Do be defined in the final document), use the two functions defined above and return a list of closed frequent itemsets. - Once the closed frequent itemsets are obtained, association rules can be generated. A pruning technique is applied to eliminate rules that are subsumed by other rules with higher confidence. This pruning step ensures that only the most interesting and distinct rules are retained, further reducing redundancy.

<TODO: Research and read literature to expand more on the above proposed solution and how to implement>

Chapter 3

Implementation Description

Chapter 4

User Manual

Chapter 5

Data set Description

Chapter 6

Results

Chapter 7

Conclusion