Uppsala University
Department of Information Technology
Kjell Orsborn

DATA MINING I - 1DL360

Assignment 3 - Frequent itemset and association rule mining

1 Frequent itemset and association rule mining

The purpose of this assignment is to study and experiment with frequent itemset and association rule mining [AIS93, AS94] using AmosMIner (under Windows and AmoxMiner for Mac OS and Linux). Here we will use an association analysis method especially suitable for implementation in a database management system environment. This method uses a vertical projection of the transactions database [SSG04] as opposed to the classical Apriori method [AIS93, AS94] that uses the original horizontal representation of transactions.

You will use an implemented version of the projection-based method [SSG04] within the AmosMiner application using the Amos II database management system. The instructions to install Amos II and AmosMiner are provided on the assignments home page. It can be used either in the computer labs at Uppsala University (Windows) or on your own laptop or desktop computer (Windows, Mac OS and Linux supported).

2 Preparation

We suggest that you read about `association analysis` in Chapter 6 in Tan et al. [Tan06]. It can also be valuable to read the background material in [SSG04, AIS93, AS94].

You can find and download the AmosMiner and Amos II system from the assignments home page and you will also get a chance to familiarize yourself with those systems in the introductory tutorials that you are recommended to attend. There is also one introductory tutorial to this assignment, which is not mandatory, but advisable to attend.

We also strongly recommend you to run the tutorials for Amos II and AmosMiner before you start with the actual assignment.

3 Assignment

In this assignment you will use transactions data that again is synthetically generated using the QUEST data generation tool [IBM96] to provide data with controlled properties. The transactions data to be used is found in the data file `transactions1000.nt`. The data consists of text that typically looks as follows:

```
1 3 4
1 2 3 5
2 3 5
2 5
1 2 3 6
```

In the file, blanks separate items (identified by integers) and new lines separate transactions. For example, the above file contains information about a total of 5 transactions and its second transaction consists of 4 items.

You should perform association analysis on the data set in the transactions data file using the PROPAD projection-based method [SSG04] within the AmosMiner application. Furthermore, you should experiment with the different parameters used in the algorithm to see how they influence the results of the analysis. The different steps in the assignment are described in more detail below.

There is also a result form (available on the assignments home page) to be filled in by you and handed in to your assistant. Also see Section 4 on more details on examination and how you should report the assignment.

Thus, in your assignment you should do the following:

START-UP

Under the section for Assignment 3 on the assignment course page, you will find the script file for the this assignment named 'a3.osql'. This file includes queries and calls to predefined functions for importing data and for performing and tuning association analysis on the data set using the PROPAD projection-based method.

You should explore the various phases of association analysis by interactively exploring and modifying the query language statements found in the script file. This is done most easily by opening the script file

in a text editor (e.g. Notepad in Windows and TextEdit in Mac OS), copy a query expression from the editor and paste it into the AmosMiner terminal window were it will be executed.

The result from executing an association analysis would typically look as follows:

```
<{131,207,443,489},{104},0.975,0.156>
<{207,443,489},{104},0.9765,0.166>
<{131,207,443},{104},0.9765,0.166>
<{131,443,489},{104},0.976,0.163>
<{131,207,489},{104},0.9702,0.163>
<{207,443},{104},0.9778,0.176>
<{131,443},{104},0.9777,0.175>
<{131,207},{104},0.9722,0.175>
<{207,489},{104},0.9721,0.174>
<{443,489},{104},0.9721,0.174>
<{131,489},{104},0.9714,0.17>
<{131},{104},0.9735,0.184>
```

It is now your turn to explore the area of association analysis by applying and modify the query language statements in the scripts file by following the rest of these instructions.

1. PERFORMING ASSOCIATION ANALYSIS BY MEANS OF THE PROPAD ALGORITHM

   (a) Define the `transactions` stored function to store transaction data, and load the data from file `transactions1000.nt` in the `data` directory.

   (b) Set the parameters for the minimum support (i.e. the `:minsupp` parameter) and and the minimum confidence (i.e. the `:minconf` parameter).

   (c) Define the `frequent_itemsets` function that stores frequent itemsets together with support counts.

   (d) Execute the PROPAD algorithm implemented in AmosMiner as the `propad` function:

   ```
   create function
       propad(Bag of Vector of Number tr, Number minsupp)
   ```

```
-> Bag of (Vector of Number, Number)
```

(e) Define the `association_rules` function to store association rules, together with their support and confidence values.

(f) Execute the association rule mining algorithm, implemented in AmosMiner as the `arm` function:

```
create function
   arm(Bag of (Vector of Number, Number)
       frequent_itemsets, Number minconf)
      -> Bag of (Vector of Number head,
                 Vector of integer tail,
                 Number rsupp,
                 Number conf)
```

(g) Examine the results and try different parameter values.

2. VISUALIZING NUMBER OF FREQUENT ITEMSETS FOR DIFFERENT SUPPORT THRESHOLDS

(a) Define the `fis_count` function to store the count of frequent itemsets for each value of the minimum support.

(b) Execute the PROPAD association analysis for a set of different parameter values, store and plot the results.

(c) What values for the minimum support might be interesting?

3. VISUALIZING NUMBER OF RULES FOR DIFFERENT CONFIDENCE THRESHOLDS

(a) Perform association analysis (i.e. frequent item set generation and rule generation) for a set of frequent itemsets using a reasonable minimum support value.

(b) Discover all association rules when choosing a very low confidence threshold.

(c) Count and plot the number of association rules above the confidence threshold for a set of different thresholds.

(d) What values for the minimum confidence level might be interesting?

4. FURTHER EXPLORATION OF THE DATASET

You should now explore the dataset further by investigating som of the items found in the list below:

(a) How many different items are there?

(b) Lower the support to 50, and try association rule mining with very high confidence.

(c) What does support 50 mean?

(d) How many maximum confidence rules do you discover?

(e) Count the number of frequent item sets found with a support of 50.

(f) Count the number of maximum confidence association rules with a support of 50.

(g) How can the number of rules discovered be greater than the number of frequent item sets?

(h) Try association rule mining on with a very high minimum confidence.

(i) Only one maximum confidence rule was found on a frequent item set with support 100. Look at the frequent item sets containing these items.

(j) What conclusions can be drawn from these queries?

(k) Why was not the rule $280 \rightarrow 487$ induced?

(l) Look at all transactions containing these items.

(m) Suppose that a shop manger wants to sell more of item 487. Help him figure out if there are any association rules with item 487 on the right hand side, apart from 280 → 487!

(n) If you found no such rules, explain why.

(o) On the other hand, are there any association rules with item 280 on the right hand side?

Note the distribution of confidence values in these rules.

You can also find these instructions and hints in the TODO list of the `a3.osql` script file.

Furthermore, you are referred to the Amos II manual for further reading, which is available on the lab course home page and that you look at the tutorial slides. You might also find the following sections in the manual useful: Collections in 2.6, count in 2.6.1, and groupby in 2.6.2.

4 Examination

At the examination, your might be saked to run your script file in AmosMiner to show that your association analysis works as expected.

For the association analysis, we expect you to present the outcome of your analysis including the choice of your parameters `minsupp`, `minconf`, etc.

You should also have correctly answered the questions on the assignment report form (available at the assignment home page). Furthermore, you should be prepared to answer questions as exemplified in the Assignment section above.

Acknowledgements

Contributions to the material for this assignment has been supplied by several of the former and current colleagues at the department, in alphabetical order,

including Andrej Andrejev, Gyözö Gidofalvi, Per Gustavsson, Tobias Lindahl, Kjell Orsborn, Tore Risch, Thanh Troung and Erik Zeitler.

References

[Tan06] Tan, P-N, Steinbach, M. and Kumar, V.: Introduction to Data Mining, Addison-Wesley, 2006.

[SSG04] X. Shang, K.-U. Sattler, and I. Geist, "Efficient Frequent Pattern Mining in Relational Databases". 5. Workshop des GI-Arbeitskreis Knowledge Discovery (AK KD) im Rahmen der LWA 2004 (the paper is available on the lab course homepage).

[AIS93] R. Agrawal, T. Imielinski, and A. Swami, "Mining Associations between Sets of Items in Large Databases". In Proceedings of the ACM SIGMOD International Conference on the Management of Data, pp. 207-216, May 1993 (the paper is available on the lab course homepage).

[AS94] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules". In Proceedings of the 20th International Conference on Very Large Databases, pp. 487-499, September 1994 (the paper is available on the lab course homepage).

[IBM96] `http://www.almaden.ibm.com/cs/projects/iis/hdb/Projects/data_mining/mining.shtml`