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] in Amos II. 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. This means that you have to install Amos II and AmosMiner according to the instructions on the assignments home page.

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 two introductory tutorials.

There will also be an introductory tutorial to this assignment, which

is not mandatory, but advisable to attend.

The computer labs include a limited number of computers and, to avoid overcrowded labs, each group should sign up for one lab time.

Lab lists will be posted on the board outside 1346.

3 Assignment

In this assignment you will use transactions data that again is synthetically generated using the QUEST data generation tool [IBM] 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:

1. START-UP

   Once you have installed the AmosMiner application, you should have the script file and data files for assignment 3 available (if some file is missing in your AmosMiner directory you should download and install AmosMiner again).

   Load the script file for this assignment, `a3.osql`, in AmosMiner by the following command:

   ```
   < 'a3.osql';
   ```

   When the script file has been loaded correctly, you have already been executing a association analysis that can be tracked by studying the content of the script file itself and the results that were generated during its execution. The result from executing an association analysis would typically look as follows:

   ```
   <{104,131,207},{489},163,0.931428571428571>
   <{104,131,207},{443},166,0.948571428571429>
   <{104,131,207},{443,489},156,0.891428571428571>
   <{104,207},{489},174,0.935483870967742>
   <{104,207},{131,489},163,0.876344086021505>
   <{104,207},{443},176,0.946236559139785>
   <{104,207},{131,443},166,0.89247311827957>
   <{104,207},{443,489},166,0.89247311827957>
   ```

   It is now your turn to modify these scripts by following the rest of these instructions.

2. SUPPORT AND CONFIDENCE VALUES

   You should experiment with minimum support and minimum confidence values. Determine how these parameters influence:

   - the number of frequent itemsets generated
   - the number of association rules generated

   (a) How many different items are there?

```
count(select distinct o
from vector v, number o
where v in read_ntuples("data/transactions1000.nt")
and o in v);
```

(b) Plot the frequent item set size for min support.

```
create function fis_size(integer minsupp) -> integer;

set fis_size(i) = count(propad(transactions(), i))
from integer i where i in 10*iota(3, 30);

plot(vectorof(select {i, s} from integer i, integer s
where s = fis_size(i)));
```

(c) Why does the frequent item set size increase so sharply when minimum support approaches 0?

```
dropfunction(#'fis',1);

store_fis(propad(transactions(), 100), #'fis');
```

(d) What does it mean when the number of rules discovered is greater than the number of frequent item sets?

```
create function rulecount(number conf)->number;

add rulecount(i) = count(arm(#'fis', i))
from number i
where i in iota(0, 11)/10.0;
```

(e) Investigate how the number of discovered rules depend on the confidence, by plotting the number of rules vs confidence.

```
select plot(w)
from vector of vector w
where w = sort(select {i, s}
from integer i, integer s
where s = rulecount(i));
```

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

(g) What does support 50 mean?

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

```
dropfunction(#'fis',1);

store_fis(propad(transactions(), 50), #'fis');
```

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

```
count(select v, fis(v) from vector of integer v);
```

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

```
count(arm(#'fis', 1.0));
```

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

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

```
dropfunction(#'fis',1);
```

```
store_fis(propad(transactions(), 100), #'fis');

arm(#'fis', 1.0);
```

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

```
select v, fis(v) from vector of integer v
where 280 in v;

select v, fis(v) from vector of integer v
where 487 in v;
```

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

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

(p) Look at all transactions containing these items.

```
select v from vector of integer v
where v in transactions() and 487 in v;

select v from vector of integer v
where v in transactions() and 280 in v;
```

(q) 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!

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

```
dropfunction(#'fis',1);

store_fis(propad(transactions(), 50),#'fis');
```

```
select lhs, rhs, conf, abs_supp
from vector of integer lhs, vector of integer rhs,
number conf, integer abs_supp
where <lhs, rhs, abs_supp, conf> in arm(#'fis', 0.3)
and 487 in rhs;
```

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

```
select lhs, rhs, conf, abs_supp
from vector of integer lhs, vector of integer rhs,
number conf, integer abs_supp
where <lhs, rhs, abs_supp, conf> in arm(#'fis', 0.3)
and 280 in rhs;
```

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 script file will be executed in AmosMiner using:

```
< 'a3.osql';
```

- When the analysis has been executed correctly, 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.

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).

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