Applying the Association Rules Mining Technique to Identify Critical Graduation Pathway Courses

2014 CAIR Conference

Afshin Karimi (akarimi@fullerton.edu)
Ed Sullivan, Ph.D. (esullivan@calstate.edu)

November 19, 2014

Data Mining Techniques Used in Higher Education

- Prediction (and/or Classification)
- Clustering
- Relationship Mining

Relationship Mining

- Goal is to discover relationships between variables with data set with large number of variables

- 4 types of Relationship Mining:
  - *Association Rules Mining*
  - *Sequential Pattern Mining*
  - *Correlation Data Mining*
  - *Causal Data Mining*

Association Rules Mining

- Proposed by Agrawal et al in 1993
- If-then rules amongst variables
- Initially used for Market Basket Analysis
- Milk Purchase -> Cereal Purchase (5% support, 80% confidence)
  - 5% support: customers who buy both product (in any order) are 5% of all customers in the database
  - 80% confidence: 80% of those who buy milk also buy cereal
- If student takes courses A and B, she will take course C (not necessarily in that order)

Association Rules Mining Examples

- Walmart study found young males buying beer on Friday afternoons also buy baby diapers
- Amazon recommending items based on your current browsing/buying selections as well as other customers' purchasing patterns
- Google search's auto-complete where after a word is typed in the search box, it suggests a follow-up associated search term

The Apriori Algorithm

- The best known algorithm for Association Rules Mining

- The algorithm is a two step process:
  - Find frequent itemsets
  - Use frequent itemsets to generate rules

Apriori algorithm, continued…

Step 1: Finding frequent itemsets:

Iterative process starting with scanning the database to find frequent 1-itemsets (that meet min. support), then using a Join operation find larger frequent itemsets (through k-itemset)

Step 2: Generating association rules:

Using the found frequent itemsets and minium support and confidence, rules are established

An example

t1:	Beef, Chicken, Milk
t2:	Beef, Cheese
t3:	Cheese, Boots
t4:	Beef, Chicken, Cheese
t5:	Beef, Chicken, Clothes, Cheese, Milk
t6:	Chicken, Clothes, Milk
t7:	Chicken, Milk, Clothes

- Transaction data
- Assume:
  minsup = 30%
  minconf = 80%
- An example frequent *itemset*:

  {Chicken, Clothes, Milk} [sup = 3/7]
- Association rules from the itemset:

  Clothes → Milk, Chicken [sup = 3/7, conf = 3/3]

  … …

  Clothes, Chicken → Milk, [sup = 3/7, conf = 3/3]

Slide from Bing Liu's course material – University of Illinois-Chicago

Input Data (Association Rules Mining)

Customer ID	Beer	Wine	Soda	Cheese	Soap	Apples	Ground Beef	Chips	Pasta Sauce	Gum	Wall Calendar	Ground Coffee	Postcard	Magazine	Mints	Oranges
1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
2	1	0	1	1	0	1	0	1	0	0	1	0	0	0	0	0
3	0	1	0	0	1	0	0	0	0	0	0	0	1	0	1	0
4	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1
5	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0	0	1	0	1	0	0	0	1	1	0	0	0	1	0
8	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1
9	0	1	1	0	0	0	1	0	1	0	0	0	0	1	0	0
10	0	0	1	0	1	1	0	0	0	0	0	1	1	0	0	0
11	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0
12	1	1	1	0	0	0	0	0	0	1	0	1	0	0	0	0
13	0	0	0	1	0	0	1	1	0	0	1	0	0	0	1	0
14	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1
15	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	1
16	0	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0
17	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
18	0	0	0	1	0	0	1	1	0	0	1	0	0	0	1	0
19	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
23	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
26	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

Problem with Association Rules Mining

- **Problem:** Algorithm discovers huge number of association rules (between one or more variables with one or more other variables), many of which are irrelevant

- **A Solution:** Use ‘interestingness’ measures to reduce the rule set

Interestingness

- Objective Interestingness:
  - Support
  - Confidence
  - Cosine
  - Added value
  - Lift
- Subjective Interestingness:
  - Unexpectedness
  - Actionability

Support

Let |X,Y| be the number of transactions that contain both X and Y

Support is the proportion of all transactions that contain both X and Y

Sup (X -> Y) = |X, Y| / n *OR* P (X, Y)

Sup(X -> Y) = Sup (Y -> X)

Confidence

Let |X| the number of transactions that contain X.

Confidence is the proportion of transactions that contain Y amongst the ones that contain X.

$conf(X \rightarrow Y) = |X, Y| \,/\, |X|$ *OR* $P(X, Y) / P(X)$

$conf(X \rightarrow Y) \neq conf(Y \rightarrow X)$

Cosine

(borrowing from cosine of angle between two vectors…)

$$\text{Cosine}(X \rightarrow Y) = |X, Y| / \sqrt{|X| . |Y|}$$

- The closer cosine (X -> Y) is to 1, the more transactions containing item X also contain Y
- The closer cosine (X -> Y) is to 0, the more transactions contain item X without containing Y
- Cosine is a symmetric measure: cosine(X -> Y) = cosine (Y -> X)

Lift

lift(X -> Y) = conf(X -> Y)/P(Y)

If P(X, Y) = P(X) . P(Y), lift is 1. This is the worst case (occurrence of X and occurrence of Y in the same transactions are independent events)

Subjective Interestingness

Subjective Interestingness is application domain- specific. Two such measures are:

- **Unexpectedness**: Grocery chain already knows about (Beer -> Chips) association rule, but not about the (Beer -> Diapers) association rule.
- **Actionability:** Rules that offer strategic information on which user can act on.

Association Rules Example

- Transfer Student Success Project in the Mihaylo College of Business & Economics

- Identify the gateway courses that prevent MCBE transfer students from timely graduation

Association Rules Example Continued…

MCBE Transfer Students Success:

- Examine CBE courses that new transfer students take AND fail during 1st term at Fullerton
- Find all Association Rules between all the variables (course failures) and a new variable that represents graduation in 4 years or less
- Use interestingness measures to focus on the relevant associations

Association Rules Example Continued…

Input File Format

- Rows: fall 08 & 09 new transfer MCBE students who took at least one MCBE course during their 1st term (1,807 students)
- Columns: MCBE courses above students took during their 1st term PLUS Graduation variable that indicates if student graduated in 4 years or less (43 columns)
- Values:
  - 1: failed the course in 1st term (grade of C- thru F, including WU)
  - 0: passed the course in 1st term (grade of C or above) *OR* didn't take course in 1st term

Example Input File

cwid	NotGraduated	ACCT201A	ACCT201B	ACCT301A	ACCT302	ACCT307	ACCT364	BUAD201	BUAD210	BUAD301	ECON201	ECON202	ECON310	ECON315	ECON320	ECO
[redacted]8906	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]0797	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
[redacted]8860	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]9249	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]3102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]1375	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]1853	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]7215	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]0532	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
[redacted]5824	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]6247	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]6475	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]8455	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]6961	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]9981	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]4777	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]7802	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]5741	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]6024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]6881	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]046	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]174	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]874	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]963	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
[redacted]592	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]5482	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	
[redacted]5836	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]896	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]987	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
[redacted]473	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]347	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]982	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]5975	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]2369	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]5081	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]7762	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
[redacted]2135	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

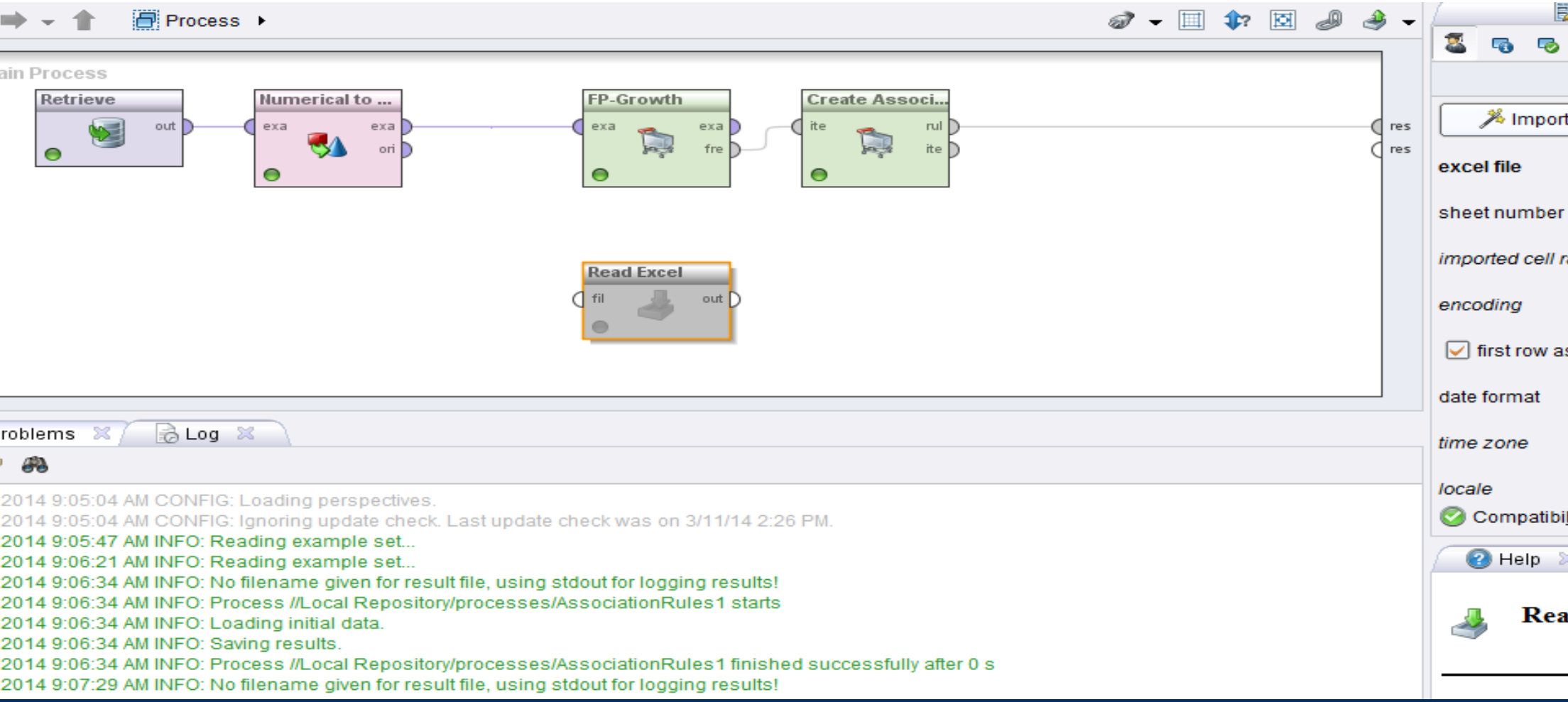
Association Rules Example Continued…

- Algorithm finds large number of rules between one or more variables with one or more (other) variables

- Here we focus on association rules between different course variables and graduation variable:
  (X -> Grad in 4 Yrs) where X is any of the 42 CBE courses.

- Furthermore, narrow the list by using Support & Confidence measures

RapidMiner 5 Software Demo



Association Rules Example Continued…

Results:

- Top 3 identified gateway courses are all 200 level courses (lower division core courses) that new transfer students take AND fail
- Graduation variable not really the ‘target’ variable

Future Work/Summary

- Further study of the identified gateway courses
- If order of events is important, use Sequential Mining method instead (not covered in this presentation)
- No need to have intimate knowledge of the algorithm used. Just need to compile model's input data file

Questions/Comments?

Contact: akarimi@fullerton.edu