Negative rule mining

EDAMI project

Paweł Rybal and Karolina Borkowska

1 Aim of the project

The aim of this project was to implement and test a negative rule mining algorithm that was designed and described in the article titled ,,Mining Positive and Negative Association Rules: An Approach for confirmed Rules" by Maria-Luiza Antonie and Osmar R. Zaïane.

2 Project assumptions

It was established that the project will be done using Python programming language and tested with various data sets and if possible, those that were used by algorithms original creators.

3 Implemented algorithm

At the beginning data is read form a file and then prepared so that each line in file is considered as one transaction. Also the first line is a header that describes what each element in line represents.

When dataset is ready proper algorithm starts. Its consecutive steps are following:

1. set k as 1, this is the iteration number;
2. find all frequent item sets of size 1 and save them in the set f_1;
3. initialize set f_{k-1} with items form f_1;
4. set f_k as empty set;
5. generate frequent candidates list ck form the f_{k-1} and f_1 items:
   - for each item set in f_{k-1}, for each item in f_1, check whether item is part of item set; if it is not, union of item set and this item become a new candidate;
6. for each ck's item set:
   - check whether its relative support is above a given threshold (i.e. is a frequent item set), if it is so add it to f_k;
   - for each split combination of this item set (combinations of subsets of each possible length against rest of the items, subset becomes an antecedent, rest is consequent), check the correlation of the antecedent and consequent:

- it if is above a given threshold check arisen rules confidence, again if it's above threshold assign it to positive rules set;
- if it is above minimum correlation, but the support is too low, negate the antecedent and the consequent, again the new rule is checked for support and confidence, to establish whether it should be added to the set of negative rules;
- if it is not, but the correlation is below negative minimal correlation two new rules are made: one with negated antecedent and one with negated consequent; both rules have their confidences check and if they are high enough, they are added to negative rules set.

7. change f_{k-1} to have f_k's values;
8. if f_k is not empty, or algorithm reached maximum length of rules, go back to point 4.

4 Using implemented scripts

To run the implemented algorithm `run.py` script has to be called from *sh* terminal with specified parameters. Those parameters are as follows:

Table 1: Parameters of `run.py` script

name	description	parameter	type	obligatory
file_path	Path to data file	–file, -f	string	yes
min_supp	Minimal support	–minsupp, -s	float	yes
min_conf	Minimal confidence	–minconf, -c	float	yes
min_corr	Minimal correlation coefficient	–mincorr, -d	float	no (default = 0.5)
max_len	Maximal rule length	–length, -l	integer	no (default = none)
verbose	Print to output some info during algorithm work	–verbose, -v	none	no

5 Tests

5.1 Datasets

Unfortunately due to uncertainty of what type of preprocessing was done on the dataset used in the original research by Maria-Luiza Antonie and Osmar R. Zaïane, it was impossible to compare results of this implementation with those given in the article. Six other data sets were used to test the algorithm:

dataset	description	size	classes
adults	data extracted from the census bureau database	32561 entries	age workclass education marital-status occupation relationship race sex capital-gain capital-loss hours-per-week native-country
contraception	subset of the 1987 National Indonesia Contraceptive Prevalence Survey	1473 entries	Wifes-age Wifes-education Husbands-education Number-of-children-ever-born Wifes-religion Wifes-now-working? Husbands-occupation Standard-of-living-index Media-exposure Contraceptive-method-used

mushrooms	mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981)	8124 entries	decision cap-shape cap-surface cap-color bruises odor gill-attachment gill-spacing gill-size gill-color stalk-shape stalk-root stalk-surface-above-ring stalk-surface-below-ring stalk-color-above-ring stalk-color-below-ring veil-type veil-color ring-number ring-type spore-print-color population habitat
nursery	derived from a hierarchical decision model originally developed to rank applications for nursery schools	12960 entries	parents has_nurs form children housing finance social health
supermarket	sets of transactions form a supermarket	4627 entries	124 classes
titanic	Titanic's passengers' data	887 entries	survived pclass name sex age siblings/spouses aboard parents/children aboard fare

Table 2: Datasets description

5.2 Time tests

To check whether it is possible to conduct research using presented algorithm and to check its time consumption, for each dataset script was run ten times with support, confidence and correlation coefficient equal to 0.5. This numbers were chosen due to fact that some datasets take too long with smaller numbers, but this configuration at least provides some rules and iterates over all of datasets. For this experiment it is enough, in the next section all datasets will be tested with specialized settings. Results are shown below (datasets are shown form the smallest to the largest):

Table 3: Average times of conducting an experiment on a given dataset

Dataset	average time	rules found	longest rule
titanic	0.2311s	1081	4
contraception	0.0549s	0	0
supermarket	0.3056s	0	0
mushrooms	19.1735s	504	4
nursery	4.3349s	104	3
adult	3.0311s	0	0

As it was suspected, larger data sets take longer time to compute. This was concluded considering that some datasets do not produce rules, hence time taken to iterate over *adult* dataset takes longer then *contraception* or *supermarket* proves this, as much as time taken for *mushrooms* and *titanic*, that produce some rules. Further sections will consider individual datasets' tests.

5.2.1 Adult dataset

Adult dataset was tested with settings form table 4 and the results are shown in table 5.

Table 4: Settings used for testing on *adult* dataset

support	confidence	correlation coefficient	average time	rules found	longest rule
0.3	0.3	0.5	87.0020s	532	6
0.2	0.2	0.5	203.3814s	1304	7

Table 5: Results for *adult* dataset for setting in a previous table

rule type	is present	rule example
x→ y	yes	'race= White', 'native-country= United-States', 'relationship= Husband', 'capital-loss=none' → 'marital-status= Married-civ-spouse', 'sex= Male' (support: 10315, confidence: 0.9990305380513815)
$\neg x \rightarrow y$	yes	'¬capital-gain=none', '¬sex= Female' → 'marital-status= Married-civ-spouse', 'native-country= United-States', 'relationship= Husband' (support: 2090, confidence: 0.6947368421052632)
$x \rightarrow \neg y$	yes	'sex= Female', 'native-country= United-States', 'capital-gain=none' → '¬relationship= Husband' (support: 9113, confidence: 0.9998902666520355)
$\neg x \rightarrow \neg y$	no	———

This result prove that both ,,normal" rules (x→ y) are produced by algorithm as wall as those with one element negated. The negated rules are a bit tautological in their nature but are true nevertheless.

Lowering minimal support and confidence to 0.2 increased number of rues to 1304 an the longest one was of size 7.

5.2.2 Contraception dataset

Contraception dataset was tested with settings form table 6 and the results are shown in table 5.

Table 6: Settings used for testing on *contraception* dataset

support	confidence	correlation coefficient	average time	rules found	longest rule
0.3	0.3	0.5	0.2183s	6	3
0.1	0.1	0.5	13.9701s	18	5

Table 7: Results for *contraception* dataset for setting in a previous table

rule type	is present	rule example
x→ y	yes	'Wifes-education=very-high' → 'Media-exposure=Good', 'Husbands-education=very-high' (support: 577, confidence: 0.9306759098786829)
$\neg x \rightarrow y$	no	———
$x \rightarrow \neg y$	no	———
$\neg x \rightarrow \neg y$	yes	'¬Husbands-education=very-high' → '¬Standard-of-living-index=4', '¬Wifes-education=very-high', '¬Number-of-children-ever-born=three-to-five', '¬Wifes-age=early20s' (support: 575, confidence: 0.34608695652173915)

In the first test (0.3 support and confidence) produces only 6 rules and all of them of first type (everything positive). Lowering minimal values produces additional negative rules. This proves that it is possible to achieve fourth type of rules ($\neg x \rightarrow \neg y$).

5.2.3 Mushrooms dataset

Mushrooms dataset was tested with settings form table 8 and the results are shown in table 9. Due to time restriction and the fact that all tests were run on simple PC, this set was tested only with high minimum value settings. Testing this dataset was stopped when lower minimal values (0.3 support and confidence) after 17 minutes into first iteration.

Table 8: Settings used for testing on *mushrooms* dataset

support	confidence	correlation coefficient	average time	rules found	longest rule
0.6	0.6	0.5	2.3888s	56	4
0.5	0.5	0.5	18.3183s	504	6

Table 9: Results for *mushrooms* dataset for setting in a previous table

rule type	is present	rule example
x→ y	yes	'ring-number=one', 'stalk-surface-below-ring=smooth', 'gill-attachment=free' → 'veil-type=partial' (support: 4304, confidence: 1.0)
$\neg x \rightarrow y$	yes	' decision=edible', ' gill-attachment=free' → 'bruises=no', 'veil-type=partial' (support: 19, confidence: 0.9473684210526315)
$x \rightarrow \neg y$	yes	'stalk-surface-below-ring=smooth', 'veil-color=white', 'veil-type=partial', 'gill-attachment=free' → ' bruises=no' (support: 4744, confidence: 0.6405986509274874)
$\neg x \rightarrow \neg y$	no	———

Mushroom data set does not produce double negative rules ($\neg x \rightarrow \neg y$), this can be due to high support and confidence values, but again it takes too long time to test that theory.

5.3 Nursery dataset

Nursery dataset was tested with settings form table 10 and the results are shown in table 11

Table 10: Settings used for testing on *nursery* dataset

support	confidence	correlation coefficient	average time	rules found	longest rule
0.3	0.3	0.5	0.1454s	0	0
0.1	0.1	0.5	4.1125s	104	3
0.05	0.05	0.5	26.0353s	702	4

Table 11: Results for *nursery* dataset for setting in a previous table

rule type	is present	rule example
x→ y	yes	'finance=convenient', 'health=not_recom' → '=not_recom' (support: 2160, confidence: 1.0)
$\neg x \rightarrow y$	no	———
$x \rightarrow \neg y$	no	———
$\neg x \rightarrow \neg y$	yes	health=not_recom', ' =priority' → ' parents=usual' (support: 4374, confidence: 0.7816643804298126)

Setting 0.1 as minimal support and confidence yields double positive (x→ y) and double negative ($\neg x \rightarrow \neg y$) rules. Lowering requirement does not change that.

5.4 Supermarket dataset

Supermarket dataset was tested with settings form table 12 and the results are shown in table 13

Table 12: Settings used for testing on *supermarket* dataset

support	confidence	correlation coefficient	average time	rules found	longest rule
0.3	0.3	0.5	6.0283s	46	3
0.2	0.2	0.5	112.5376s	229	4

Table 13: Results for *supermarket* dataset for setting in a previous table

rule type	is present	rule example
x→ y	yes	'finance=convenient', 'health=not_recom' → '=not_recom' (support: 2160, confidence: 1.0)
$\neg x \rightarrow y$	no	———
$x \rightarrow \neg y$	no	———
$\neg x \rightarrow \neg y$	yes	health=not_recom', ' =priority' → ' parents=usual' (support: 4374, confidence: 0.7816643804298126)

Only two types of rules are yielded. Lowering requirement does not change that.

5.5 Titanic dataset

Titanic dataset was tested with settings form table 14 and the results are shown in table 15

Table 14: Settings used for testing on *titanic* dataset

support	confidence	correlation coefficient	average time	rules found	longest rule
0.3	0.3	0.5	0.0797s	21	4
0.2	0.2	0.5	0.1512s	22	4

Table 15: Results for *titanic* dataset for setting in a previous table

rule type	is present	rule example
x→ y	yes	'Sex=male' → 'Parents/Children Aboard=0', 'Survived=0' (support: 573, confidence: 0.6980802792321117)
$\neg x \rightarrow y$	no	' Survived=1' → 'Sex=male' (support: 545, confidence: 0.8495412844036697)
$x \rightarrow \neg y$	no	'Parents/Children Aboard=0', 'Survived=0' → ' Sex=female' (support: 441, confidence: 0.9047619047619048)
$\neg x \rightarrow \neg y$	yes	' Sex=female' → ' Survived=1' (support: 573, confidence: 0.8097731239092496)

This experiment proves that in one dataset all types of rules can be present, given the right set.

6 Conclusions

Implementation of this algorithm and tests prove that proposed method works. It has its downsides (producing rules as: male → not wife), but they are hardly only this algorithms problems. Chosen datasets prove all types of rules can be yielded, give right conditions.