Statistically sound pattern discovery
Part II: Itemsets

Wilhelmiina Hämäläinen

Geoff Webb

http://www.cs.joensuu.fi/~whamalai/ecmlpkdd13/sspdtutorial.html

Overview

- Most association discovery techniques find rules
- Association is often conceived as a relationship between two parts
  - so rules provide an intuitive representation
- However, when many items are all mutually interdependent, a plethora of rules results
- Itemsets can provide a more intuitive representation in many contexts
- However, it is less obvious how to identify potentially interesting itemsets than potentially interesting rules

Rules

bruises?=true → ring-type=pendant
[Coverage=3376; Support=3184; Lift=1.93; p<4.94E-322]

ring-type=pendant → bruises?=true
[Coverage=3968; Support=3184; Lift=1.93; p<4.94E-322]

stalk-surface-above-ring=smooth & ring-type=pendant → bruises?=true
[Coverage=3664; Support=3040; Lift=2.00; p=6.32E-041]

stalk-surface-below-ring=smooth & ring-type=pendant → bruises?=true
[Coverage=3472; Support=2848; Lift=1.97; p=9.66E-013]

stalk-surface-above-ring=smooth & stalk-surface-below-ring=smooth & ring-type=pendant → bruises?=true
[Coverage=3328; Support=2776; Lift=2.01; p=0.0166]

stalk-surface-above-ring=smooth & stalk-surface-below-ring=smooth → ring-type=pendant
[Coverage=4156; Support=3328; Lift=1.64; p=5.89E-178]

stalk-surface-above-ring=smooth & stalk-surface-below-ring=smooth → bruises?=true
[Coverage=4156; Support=2968; Lift=1.72; p=1.47E-156]

stalk-surface-above-ring=smooth → ring-type=pendant
[Coverage=5176; Support=3664; Lift=1.45; p<4.94E-322]

ring-type=pendant → stalk-surface-above-ring=smooth
[Coverage=3968; Support=3664; Lift=1.45; p<4.94E-322]

stalk-surface-below-ring=smooth & ring-type=pendant → stalk-surface-above-ring=smooth
[Coverage=3472; Support=3328; Lift=1.50; p=3.05E-072]

Rules continued

stalk-surface-above-ring=smooth & ring-type=pendant → stalk-surface-below-ring=smooth
[Coverage=3664; Support=3328; Lift=1.49; p=3.05E-072]
bruises?=true → stalk-surface-above-ring=smooth
[Coverage=3376; Support=3232; Lift=1.50; p<4.94E-322]
stalk-surface-above-ring=smooth → bruises?=true
[Coverage=5176; Support=3232; Lift=1.50; p<4.94E-322]
stalk-surface-below-ring=smooth → ring-type=pendant
[Coverage=4936; Support=3472; Lift=1.44; p<4.94E-322]
ring-type=pendant → stalk-surface-below-ring=smooth
[Coverage=3968; Support=3472; Lift=1.44; p<4.94E-322]
bruises?=true & stalk-surface-below-ring=smooth → stalk-surface-above-ring=smooth
[Coverage=3040; Support=2968; Lift=1.53; p=1.56E-036]
stalk-surface-below-ring=smooth → stalk-surface-above-ring=smooth
[Coverage=4936; Support=4156; Lift=1.32; p<4.94E-322]
stalk-surface-above-ring=smooth → stalk-surface-below-ring=smooth
[Coverage=5176; Support=4156; Lift=1.32; p<4.94E-322]
bruises?=true & stalk-surface-above-ring=smooth → stalk-surface-below-ring=smooth
[Coverage=3232; Support=2968; Lift=1.51; p=1.56E-036]
bruises?=true → stalk-surface-below-ring=smooth
[Coverage=3376; Support=3040; Lift=1.48; p<4.94E-322]
stalk-surface-below-ring=smooth → bruises?=true
[Coverage=4936; Support=3040; Lift=1.48; p<4.94E-322]

Itemsets

bruises?=true,
stalk-surface-above-ring=smooth,
stalk-surface-below-ring=smooth,
ring-type=pendant
[Coverage=2776; Leverage=0.1143; p<4.94E-322]

http://linnet.geog.ubc.ca/Atlas/Atlas.aspx?sciname=Agaricus%20albolutescens

Are rules a good representation?

- An association between two items will be represented by two rules
  - three items – nine rules
  - four items – twenty-eight rules
  - …
- It may not be apparent that all the resulting rules represent a single multi-item association

Reference: Webb 2011

But how to find itemsets?

- Association is conceived as deviation from independence between two parts
- Itemsets may have many parts

Main approaches

- Consider all partitions
- Randomisation testing
- Incremental mining
- Information theoretic

Main approaches

- Consider all partitions
- Randomisation testing
- Incremental mining
  - models
- Information theoretic
  - mainly models
  - not statistical

All partitions

- Most rule-based measures of interest relate to difference between the joint frequency and expected frequency under independence between antecedent and consequent
- However itemsets do not have an antecedent and consequent
- Does not work to consider deviation from expectation if all items are independent of each other
  - If $P(x, y) \neq P(x)P(y)$ **and** $P(x, y, z) = P(x, y)P(z)$ **then** $P(x, y, z) \neq P(x)P(y)P(z)$
  - AttendingKDD14, InNewYork, AgeIsEven

References: Webb, 2010; Webb, & Vreeken, 2014

Productive itemsets

- An itemset is unlikely to be interesting if its frequency can be predicted by assuming independence between any partition thereof
- *Pregnant, Oedema, AgeIsEven*
  - *Pregnant, Oedema*
  - *AgeIsEven*
- *Male, PoorEyesight, ProstateCancer, Glasses*
  - *Male, ProstateCancer*
  - *PoorEyesight, Glasses*

References: Webb, 2010; Webb, & Vreeken, 2014

Measuring degree of positive association

Measure degree of positive association as deviation from the maximum of the expected frequency under an assumption of independence between any partition of the itemset

eg $\text{leverage}(I) = \text{P}(I) - \max_{X \subset I} [\text{P}(X)\text{P}(I \setminus X)]$

$$\text{lift}(I) = \text{P}(I) / \max_{X \subset I} [\text{P}(X)\text{P}(I \setminus X)]$$

References: Webb, 2010; Webb, & Vreeken, 2014

Statistical test for productivity

- Null hypothesis: $\exists X \subset I,\ \mathrm{P}(I) \leq \mathrm{P}(X)\mathrm{P}(I \backslash X)$
- Use a Fisher exact test on every partition
- Equivalent to testing that every rule is significant
  - $p(I) = \max_{X \subset I}\{p_F(X \rightarrow I \backslash X)\}$
- No correction for multiple testing
  - null hypothesis only rejected if corresponding null hypothesis is rejected for **every** partition
  - this increases the risk of Type 2 rather than Type 1 error

References: Webb, 2010; Webb, & Vreeken, 2014

Redundancy

- If item X is a necessary consequence of another set of items Y then $\{X\} \cup Y$ should be associated with everything with which Y is associated.
- Eg *pregnant* $\rightarrow$ *female* and *pregnant* $\rightarrow$ *oedema*,
  - *female*, *pregnant*, *oedema* is not likely to be interesting if *pregnant*, *oedema* is known
- Discard itemsets I where $\exists X \subset I,\ Y \subset X,\ \mathrm{P}(Y) = \mathrm{P}(X)$
  - I = {*female*, *pregnant*, *oedema* }
  - X = {*female*, *pregnant* }
  - Y = {*pregnant* }
- Note, no statistical test...

References: Webb, 2010; Webb, & Vreeken, 2014

Independent productivity

- Suppose that *heat*, *oxygen* and *fuel* are all required for *fire*.

- *heat*, *oxygen* and *fuel* are associated with *fire*
- So too are:
  - *heat*, *oxygen*
  - *heat*, *fuel*
  - *oxygen*, *fuel*
  - *heat*
  - *oxygen*
  - *fuel*
- But these six are potentially misleading given the full association

References: Webb, 2010; Webb, & Vreeken, 2014

Independent productivity

- An itemset is unlikely to be interesting if its frequency can be predicted from the frequency of its specialisations
- If both X and $X \cup Y$ are non-redundant and productive then X is only likely to be interesting if it holds with respect to $\neg Y$.
- $p(I) = \max_{X \subset I}\{p_F\left(X \rightarrow I \backslash \mathrm{X} \mid \Lambda_{Y \in P, I \subset Y} \neg(Y \backslash \mathrm{I})\right)\}$
  - P is the set of all non-redundant and productive patterns

References: Webb, 2010; Webb, & Vreeken, 2014

Assessing independent productivity

Given *fuel, oxygen, heat, fire*

- to assess *fuel, oxygen, fire*
- check whether association holds in data without *heat*

fuel	Oxygen	Heat	Fire
fuel	Oxygen	Heat	Fire
fuel	Oxygen	Heat	Fire
fuel	Oxygen	~Heat	~Fire
fuel	Oxygen	~Heat	~Fire
fuel	~Oxygen	Heat	~Fire
fuel	~Oxygen	~Heat	~Fire
~fuel	Oxygen	Heat	~Fire
~fuel	Oxygen	~Heat	~Fire
~fuel	~Oxygen	Heat	~Fire
~fuel	~Oxygen	~Heat	~Fire

References: Webb, 2010

Assessing independent productivity

Given *fuel, oxygen, heat, fire*

- to assess *fuel, oxygen, fire*
- check whether association holds in data without *heat*

fuel	Oxygen	Heat	Fire
fuel	Oxygen	Heat	Fire
fuel	Oxygen	Heat	Fire
fuel	Oxygen	~Heat	~Fire
fuel	Oxygen	~Heat	~Fire
fuel	~Oxygen	Heat	~Fire
fuel	~Oxygen	~Heat	~Fire
~fuel	Oxygen	Heat	~Fire
~fuel	Oxygen	~Heat	~Fire
~fuel	~Oxygen	Heat	~Fire
~fuel	~Oxygen	~Heat	~Fire

References: Webb, 2010

Assessing independent productivity

Given *fuel, oxygen, heat, fire*

- to assess *fuel, oxygen*
- check whether association holds in data without *heat* or *fire*

fuel	Oxygen	Heat	Fire
fuel	Oxygen	Heat	Fire
fuel	Oxygen	Heat	Fire
fuel	Oxygen	~Heat	~Fire
fuel	Oxygen	~Heat	~Fire
fuel	~Oxygen	Heat	~Fire
fuel	~Oxygen	~Heat	~Fire
~fuel	Oxygen	Heat	~Fire
~fuel	Oxygen	~Heat	~Fire
~fuel	~Oxygen	Heat	~Fire
~fuel	~Oxygen	~Heat	~Fire

References: Webb, 2010

Assessing independent productivity

Given *fuel, oxygen, heat, fire*

- to assess *fuel, oxygen*
- check whether association holds in data without *heat* or *fire*

fuel	Oxygen	Heat	Fire
fuel	Oxygen	Heat	Fire
fuel	Oxygen	Heat	Fire
fuel	Oxygen	~Heat	~Fire
fuel	Oxygen	~Heat	~Fire
fuel	~Oxygen	Heat	~Fire
fuel	~Oxygen	~Heat	~Fire
~fuel	Oxygen	Heat	~Fire
~fuel	Oxygen	~Heat	~Fire
~fuel	~Oxygen	Heat	~Fire
~fuel	~Oxygen	~Heat	~Fire

References: Webb, 2010

Mushroom

- 118 items, 8124 examples
- 9676 non-redundant productive itemsets (α=0.05)
- 3164 are not independently productive

edible=e, odor=n
[support=3408, leverage=0.194559, p<1E-320]

gill-size=b, edible=e
[support=3920, leverage=0.124710, p<1E-320]

gill-size=b, edible=e, odor=n
[support=3216, leverage=0.106078, p<1E-320]

gill-size=b, odor=n
[support=3288, leverage=0.104737, p<1E-320]

False Discoveries

- We typically want to discover associations that generalise beyond the given data
- The massive search involved in association discovery results in a massive risk of false discoveries
  - associations that appear to hold in the sample but do not hold in the generating process

Reference: Webb 2007

Bonferroni correction for multiple testing

- Divide critical value by size of search space
  - Eg retail
    - 16470 items
    - Critical value = $0.05/2^{16470} < 5E^{-4000}$
- Use layered critical values
  - Sum of all critical values cannot exceed familywise critical value
  - Allocate different familywise critical values to different itemset sizes
  - $\frac{\alpha}{2^{|I|-1}}$ divided by all itemsets of size $|I|$
  - The critical value for itemset I is thus $\frac{\alpha}{2^{|I|-1}\binom{n}{|I|}}$
  - Eg retail
    - 16470 items
    - Critical value for itemsets of size 2 = $\frac{0.05}{2^1\binom{16470}{2}}$ = 1.84E-10

Reference: Webb 2007, 2008

Randomization testing

- Randomization testing can be used to find significant itemsets.
- All the advantages and disadvantages enumerated for dependency rules.
- Not possible to efficiently test for productivity or independent productivity using randomisation testing.

References: Megiddo & Srikant, 1998; Gionis *et al* 2007

Incremental and interactive mining

- Iteratively find the most informative itemset relative to those found so far
- May have human-in-the-loop
- Aim to model the full joint distribution
  - will tend to develop more succinct collections of itemsets than self-sufficient itemsets
  - will necessarily choose between one of many potential such collections.

References: Hanhijarvi *et al* 2009; Lijffijt *et al* 2014

Belgian lottery

- {43, 44, 45}
- 902 frequent itemsets (min sup = 1%)
  - All are closed and all are non-derivable
- KRIMP selects 232 itemsets.
- MTV selects no itemsets.

DOCWORD.NIPS Top-25 leverage itemsets

kaufmann,morgan
trained,training
report,technical
san,mateo
mit,cambridge
descent,gradient
mateo,morgan
image,images
san,mateo,morgan
mit,press
grant,supported
morgan,advances
springer,verlag

top,bottom
san,morgan
kaufmann,mateo
san,kaufmann,mateo
distribution,probability
conference,international
conference,proceeding
hidden,trained
kaufmann,mateo,morgan
learn,learned
san,kaufmann,mateo,morgan
hidden,training

Reference: Webb, & Vreeken, 2014

WORDOC.NIPS Top-25 leverage rules

kaufmann → morgan

morgan → kaufmann

abstract,morgan → kaufmann

abstract,kaufmann → morgan

references,morgan → kaufmann

references,kaufmann → morgan

abstract,references,morgan → kaufmann

abstract,references,kaufmann → morgan

system,morgan → kaufmann

system,kaufmann → morgan

neural,kaufmann → morgan

neural,morgan → kaufmann

abstract,system,kaufmann → morgan

abstract,system,morgan → kaufmann

abstract,neural,kaufmann → morgan

abstract,neural,morgan → kaufmann

result,kaufmann → morgan

result,morgan → kaufmann

references,system,morgan → kaufmann

neural,references,kaufmann → morgan

neural,references,morgan → kaufmann

abstract,references,system,morgan → kaufmann

abstract,references,system,kaufmann → morgan

abstract,result,kaufmann → morgan

abstract,neural,references,kaufmann → morgan

Reference: Webb, & Vreeken, 2014

WORDOC.NIPS Top-25 frequent (closed) itemsets

abstract,references
abstract,result
references,result
abstract,function
abstract,references,result
abstract,neural
abstract,system
function,references
abstract,set
abstract,function,references
neural,references
function,result
abstract,neural,references
references,system
references,set
neural,result
abstract,function,result
abstract,introduction
abstract,references,system
abstract,references,set
result,system
result,set
abstract,neural,result
abstract,network
abstract,number

Reference: Webb, & Vreeken, 2014

WORDOC.NIPS Top-25 lift self-sufficient itemsets

duane,leapfrog
americana,periplaneta
alessandro,sperduti
crippa,ghiselli
chorale,harmonization
iiiiiiii,iiiiiiiiii
artery,coronary
kerszberg,linster
nuno,vasconcelos
brasher,krug
mizumori,postsubiculum
implantable,pickard
zag,zig
ekman,hager
lpnn,petek
petek,schmidbauer
chorale,harmonet
deerwester,dumais
harmonet,harmonization
fodor,pylyshyn
jeremy,bonet
ornstein,uhlenbeck
nakashima,satoshi
taube,postsubiculum
iceg,implantable

Closure of duane,leapfrog (all words in all 4 documents)

abstract, according, algorithm, approach, approximation, bayesian, carlo, case, cases, computation, computer, defined, department, discarded, distribution, duane, dynamic, dynamical, energy, equation, error, estimate, exp, form, found, framework, function, gaussian, general, gradient, hamiltonian, hidden, hybrid, input, integral, iteration, keeping, kinetic, large, leapfrog, learning, letter, level, linear, log, low, mackay, marginal, mean, method, metropolis, model, momentum, monte, neal, network, neural, noise, non, number, obtained, output, parameter, performance, phase, physic, point, posterior, prediction, prior, probability, problem, references, rejection, required, result, run, sample, sampling, science, set, simulating, simulation, small, space, squared, step, system, task, term, test, training, uniformly, unit, university, values, vol, weight, zero

Itemsets Summary

- More attention has been paid to finding associations efficiently than to which ones to find
- While we cannot be certain what will be interesting, the following probably won't
  - frequency explained by independence between a partition
  - frequency explained by specialisations
- Statistical testing is essential
- Itemsets often provide a much more succinct summary of association than rules
  - rules provide more fine grained detail
  - rules useful if there is a specific item of interest
- Self-Sufficient Itemsets
  - capture all of these principles
  - support comprehensible explanations for why itemsets are rejected
  - can be discovered efficiently
  - often find small sets of patterns (mushroom: 9,676, retail: 13,663)

Reference: Novak *et al* 2009

Software

- OPUS Miner can be downloaded from : http://www.csse.monash.edu.au/~webb/Software/opus_miner.tgz

Statistically sound pattern discovery
Current state and future challenges

- Efficient and reliable algorithms for binary and categorical data
  - branch-and-bound style
  - no minimum frequencies (or 'harmless' like 5/n)
- Numeric variables
  - impact rules allow numerical consequences (Webb)
  - **main challenge**: Numerical variables in the condition part of rule and in itemset
    - How to integrate an optimal discretization into search?
- How to detect all "redundant" patterns?
- Long patterns

End!

- Questions?

- All material: http://cs.joensuu.fi/pages/whamalai/kdd14/sspdtutorial.html