

POLLUTION → CANCER

(OR, ASSOCIATION RULE MINING ON CANCER
MORTALITY RATE AND POLLUTION EMISSIONS
DATA)

DANIEL COUVERTIER

EDEN ELOS

ANU PAKANATI

[HTTP://CANCER-POLLUTION.APPSPOT.COM/](http://cancer-pollution.appspot.com/)

Motivation

- Numerous studies exist linking lung cancer and air pollution-- particularly particulate matter.
- Beeson, Abbey, Knutson (1998) found a positive correlation between lung cancer and air pollution
- Landmark study by Cohen and Pope (2002) that studied 500,000 people and found that the risk of death from lung cancer went up 8% for every 10 micrograms of fine particles in a cubic meter, heart disease 6%, and all causes 4%
- It's not necessarily **just** lung cancer-- there may be other non-obvious relationships between cancer and pollution.
- **Question**
 - Can we find any other evidence of links between pollution and cancer via data mining?

Motivation

- Answer

- Association Rule Mining

- Data on the internet

- Cancer mortality rates on 39 types of cancers

- derived from the National Cancer Institute Database

- contains up to 2 categories for race (white/black), 2 categories for sex, and up to 4 for age groups

- Pollution emissions data on 8 types of pollutants

- derived from EPA, 1990 statistics

- contains data by county for CO, NOX, VOC, SO₂, PM₂₅, PM₁₀, NH₃, TOTEMIS

Software Workflow

- Platform independent
 - Pre-processing
 - Python
 - Weka
 - Data Mining
 - Borgelt's Apriori Algorithm
 - Post-processing
 - Python
 - Visualization
 - Python
 - Google App Engine
 - Google Charts
 - Google Maps
 - AM Charts

Background

- Association Rule Mining

- Does not necessarily imply causation

- Antecedent(s) \rightarrow Consequent(s)

- Ex: $A \rightarrow C$

- Support = # of records with A and C / total # of records

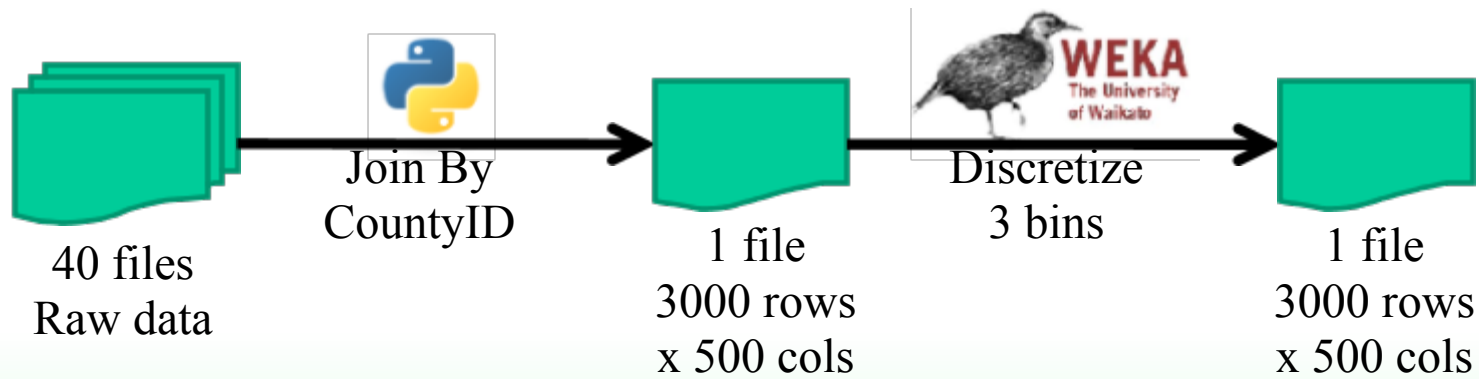
- Confidence = # of records with A and C / # of records with A

- Lift = $P(C|A) / P(C)$

Pre-Processing

Cancer, County, Pollution Data

Counties



Data Mining

- **Weka**

- Apriori
- Slow
- Many unwanted rules
- Didn't really come through for us in the end-- but Dr. Tan suggested an alternative:

- **Borgelt's Apriori**

- By Christian Borgelt
- Generated rules with only 1 consequent
- Allows restrictions on antecedents and consequents

Data Mining

- Parameters for Apriori Algorithm
 - Support: 5%
 - Confidence: 50%
 - Lift: 10%
- Borgelt's software worked well, without issues
- Run on 3.2 GhZ machine with 4 GB RAM (but probably could have been run most anywhere).

Results

- After mining with the apriori algorithm, we get 163,476 rules with precisely one cancer as the consequent and anywhere from one to eight pollutants as antecedent(s).
- Q: How do we evaluate the quality of these rules?
- A: Examine rules for 'interestingness'. Specifically we're interested in clear rules with clear relationships that support a positive correlation between cancer and pollution.
- An interesting rule may have low pollutant levels (for all antecedents) and a low cancer rate, medium pollutants -> medium cancer, or high pollutants -> high cancer

Results (cont.)

- Calculate the quantity of 'interesting' rules
- Count rules that are interesting, sort by # antecedents.
- Key assumption: Since we are using equal frequency binning, if we assumed each attribute to be independent, we should be getting scattered rules
- Interesting rules are $3 / 3^{(\#antecedents + 1)} = 1 / 3^{(\#antecedents)}$ of the total.
- Use χ^2 Tests to determine significance
- Recall χ^2 Tests with Binomial Approximation:
 - $\chi^2 = (\text{Observed} - \text{Expected})^2 / (np(1-p))$
- 1 degree of freedom : rule interesting or not interesting

Results (cont.)

- Chart below describes results across all rules.

Antecedents	Total Rules	Interesting Rules	Expected Rules
1	3794	1679	1265
2	23817	6468	2643
3	42262	13568	1565
4	44269	17594	547
5	31003	14540	128
6	13994	7475	19
7	3728	2181	2
8	447	277	0.05

Results are significant w/ p value .01 for all #s of antecedents!

Results (cont.)

- This holds true substantially across all the cancers-- a vast majority of cancers had rules which were interesting with p values of .01 for all #s of antecedents, with a very very few exceptions.
- Also working on other methods to evaluate the validity of our rules
 - Hold out training set, test set and see if rules from each are similar.

Work Distribution

- **Anu**
 - Pre-processing, Post-processing, Visualization (rules)
- **Daniel**
 - Pre-processing evaluation, Visualization (charts)
- **Eden**
 - Data Mining planning, Visualization (maps)

Visualization Demo

<http://cancer-pollution.appspot.com/>

Conclusions

- Overall, pretty pleased with results
- Association rule mining presents some challenges-- the main issue is deciding how "good" the rules that we found really are.
- We are getting "interesting" rules in disproportionate numbers -- this may imply heavy linkage between pollution and cancer rates across nearly all cancers.
- Still trying to figure out how to evaluate rules further more quantitatively.
- **Lessons learnt:**
 - Google is not the promised land.
 - Web services and technologies can take a lot of time to learn!

Future Work

- Raw data has a temporal component which we flattened out for the purposes of this work. But pollution-cancer interactions clearly also share this temporal component.
- Analysis on more/different discretizations of bins

The background features a vertical gradient from light blue at the top to light green at the bottom. In the upper corners, there are clusters of semi-transparent, light blue bubbles of varying sizes, some overlapping.

Questions?