

# Reguły asocjacyjne w analizie danych

K. Król  
M. K. Karpiński

Uniwersytet Wrocławski  
Instytut Informatyki  
SPRAWOZDANIE Z PROJEKTU

Wrocław, dnia 22 czerwca 2012 r.

# Spis treści

<b>1</b>	<b>Wstęp</b>	<b>3</b>
<b>2</b>	<b>Dane</b>	<b>3</b>
<b>3</b>	<b>Oprogramowanie</b>	<b>5</b>
<b>4</b>	<b>Algorytmy</b>	<b>5</b>
4.1	Apriori . . . . .	6
4.2	FPGrowth . . . . .	6
4.3	Tetrius . . . . .	7
<b>5</b>	<b>Analiza Danych</b>	<b>7</b>
5.1	Apriori 1 . . . . .	7
5.2	Apriori 2 . . . . .	8
5.3	Apriori 3 . . . . .	8
5.4	FPGrowth 1 . . . . .	9
5.5	FPGrowth 2 . . . . .	9
5.6	FPGrowth 3 . . . . .	9
5.7	Tertius 1 . . . . .	10
<b>6</b>	<b>Podsumowanie</b>	<b>10</b>

# 1 Wstęp

Dokument ten został sporządzony jako sprawozdanie do projektu z przedmiotu: Eksploracja Danych. Projekt ma na celu zbadanie danych ze świata rzeczywistego w poszukiwaniu ciekawych reguł asocjacyjnych z nimi związanych.

Kolejne rozdziały niniejszego sprawozdania zawierają opis sporządzonych eksperymentów, poczynając od znalezionych danych i wykorzystanego oprogramowania, poprzez testy właściwe, kończąc krótkim podsumowaniem wykonanego zadania.

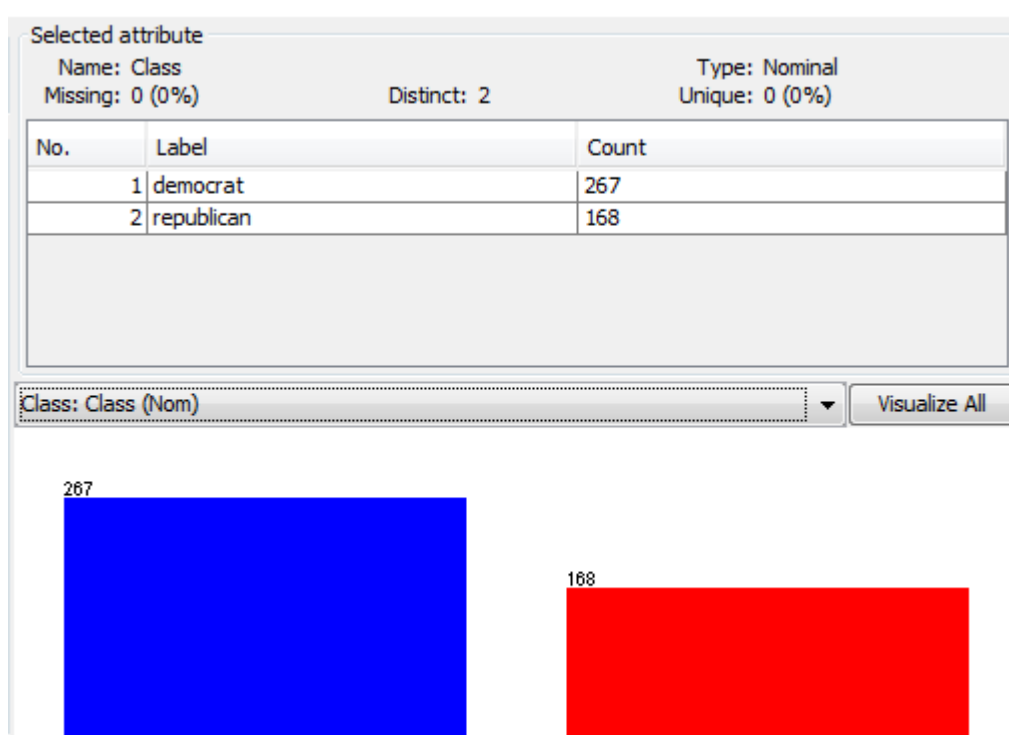
# 2 Dane

Dane wykorzystane w projekcie zostały pobrane ze strony <http://repository.seasr.org/Datasets/UCI/arff/vote.arff>. Dane te zawierają głosy każdego z kongresmenów U.S. House of Representatives dotyczących 16 istotnych problemów publicznych w roku 1984. Docelowo ten zbiór danych wykorzystywano do klasyfikacji ludzi na demokratów/republikanów w zależności od stosunku do podanych problemów:

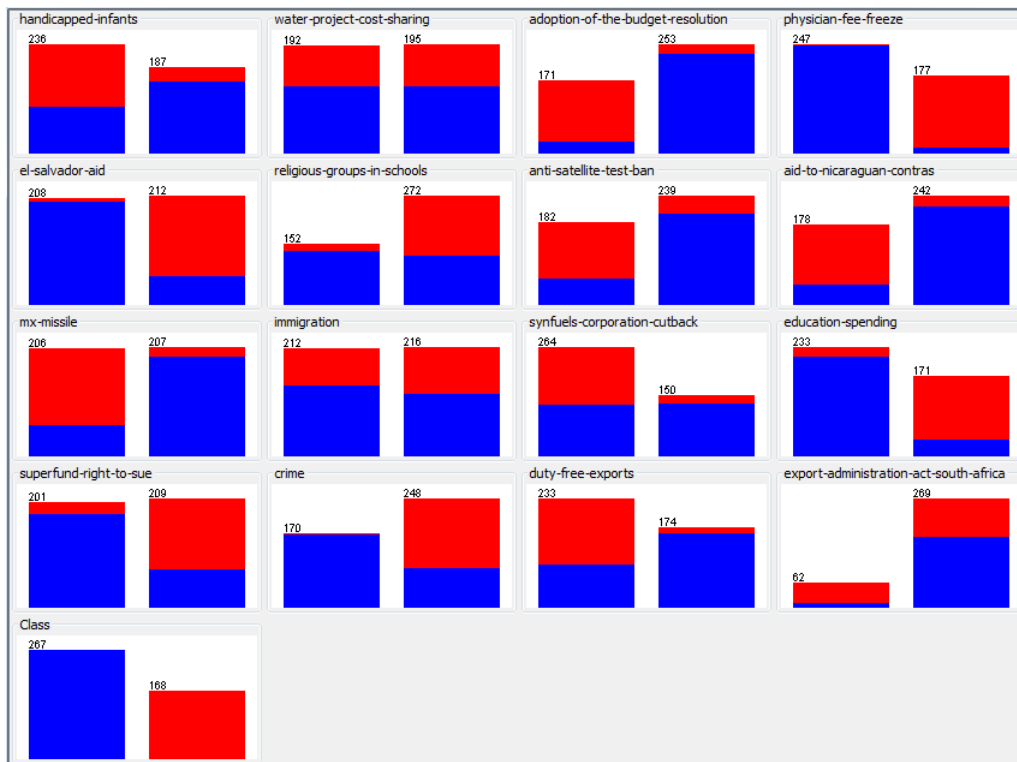
1. handicapped-infants
2. water-project-cost-sharing
3. adoption-of-the-budget-resolution
4. physician-fee-freeze
5. el-salvador-aid
6. religious-groups-in-schools
7. anti-satellite-test-ban
8. aid-to-nicaraguan-contras
9. mx-missile
10. immigration
11. synfuels-corporation-cutback
12. education-spending

13. superfund-right-to-sue
14. crime
15. duty-free-exports
16. export-administration-act-south-africa

Liczba demokratów i republikanów kształtowała się następująco (demokraci - kolor niebieski, republikanie - kolor czerwony):



Głosy typu 'voted for', 'paired for', 'announced for' zostały sklasyfikowane jako głosy 'yes', 'voted against', 'paired against', 'announced against' jako głosy 'no', a głosy 'voted present', 'voted present to avoid conflict of interest' i brak głosu zostały sklasyfikowane jako 'unknown' (oznaczone przez '?'). Poniżej widać tabele, na których oznaczone są głosy za i przeciw z podziałem na demokratów i republikanów (lewa kolumna - głosy 'no', prawa kolumna - głosy 'yes'):



Jak już wspomniano, danych tych użyto do klasyfikacji głosujących. Można jednak spojrzeć na te dane inaczej i spróbować wyszukiwać pewnych tendencji w głosowaniu na niektóre problemy. Chcemy dokonać tego poprzez wydobywanie ciekawych reguł asocjacyjnych związanych z powyższymi danymi.

### 3 Oprogramowanie

Do wykonania eksperymentów posłużyliśmy się programem wspomagającym eksplorację danych WEKA dostępnym na stronie: <http://www.cs.waikato.ac.nz/ml/weka/>.

### 4 Algorytmy

Wykorzystaliśmy niektóre algorytmy dostępne w programie WEKA:

1. Apriori
2. FPGrowth
3. Tertius

## 4.1 Apriori

Algorytm apriori był szeroko omawiany na wykładzie, dlatego też pominiemy jego opis w niniejszej pracy.

## 4.2 FPGrowth

Algorytm FPGrowth korzysta z oszczędnej (pod względem pamięciowym) struktury FP-drzewa. Definicja FP-drzewa:

- ukorzeniony, etykietowany graf acykliczny
- korzeń posiada etykietę 'null', pozostałe wierzchołki reprezentują 1-elementowe zbiory częste
- każdy wierzchołek zawiera ponadto liczbę transakcji wspierających dany zbiór częsty

Algorytm opiera się na dwóch krokach:

1. Kompresja bazy danych  $D$  i przekształcenie do FP-drzewa
2. Eksploracja FP-drzewa

W 1 kroku znajduje się najpierw wszystkie 1-elementowe zbiory częste w bazie  $D$ . Następnie, każdą transakcję  $T_i \in D$  zamienia się na skompresowaną transakcję  $T_{r_i} \in D$  usuwając z  $T_i$  wszystkie elementy, które nie są częste. W ostatniej fazie, skompresowane transakcje są sortowane malejąco po wartości wsparcia. Następnie konstruowane jest FP-drzewo:

1. utwórz korzeń z etykietą 'null'
2. Dla każdej transakcji  $T_{r_i}$  utwórz ścieżkę w FP-drzewie. Transakcje o wspólnym prefiksie, współdziela istniejące ścieżki. W przypadku wystąpienia różnicy, ścieżka jest rozdzielana. Ostatni węzeł ścieżki zawiera liczbę transakcji wspierających zbiór elementów reprezentowany przez całą ścieżkę.

Ponadto przechowuje się 'tablicę nagłówków elementów', która to jest tablicą wskaźników na poszczególne elementy w drzewie, co przyspiesza i ułatwia przeszukiwanie drzewa.

W 2 kroku, następuje eksploracja uprzednio przygotowanego FP-drzewa. Proces ten opiera się na obserwacji, że dla każdego 1-elementowego zbioru częstego  $\alpha$ , wszystkie częste nadzbiory zbioru  $\alpha$  są reprezentowane przez ścieżkę zawierającą ścieżkę dla  $\alpha$ . Zatem eksploracja drzewa przebiega następująco:

1. Dla każdego 1-elementowego zbioru częstego  $\alpha$  znajdujemy wszystkie ścieżki w drzewie, które kończą się wierzchołkiem  $\alpha$
2. Ścieżką prefiksową wzorca  $\alpha$  nazwiemy ścieżkę, która kończy się wierzchołkiem  $\alpha$ . Zbiór wszystkich ścieżek prefiksowych wzorca tworzy warunkową bazę wzorca. Na podstawie warunkowej bazy wzorca tworzymy warunkowe FP-drzewo wzorca  $\alpha$ :  $Tree - \alpha$ .
3. Następnie wywołujemy się rekurencyjnie dla drzewa  $Tree - \alpha$  w celu znalezienia wszystkich zbiorów częstych zawierających  $\alpha$ .

Konstrukcja FP-drzewa wymaga  $O(|D|)$  czasu, natomiast 2. krok wymaga  $O(|\text{tablica nagłówek}|^2 \times \text{głębokość FP-drzewa})$ . W odróżnieniu od *Apriori*, *FPGrowth* nie wymaga załadowania do pamięci całej bazy danych.

### 4.3 Tetrius

Algorytm Tetrius

W testach wykorzystaliśmy różne formy parametryzacji w celu dogłębnej analizy danych.

## 5 Analiza Danych

Poniżej znajduje się podsumowanie 7-miu testów wykonanych różnymi algorytmami z różnym stopniem parametryzacji. Po wykonaniu testów liczba znalezionych reguł była zbyt duża, aby w czytelnej formie zaprezentować je w niniejszej pracy. Dlatego pod każdym z testów znajduje się krótkie podsumowanie wraz kilkoma ręcznie wybranymi, ciekawymi (według autorów) regułami asocjacyjnymi.

### 5.1 Apriori 1

- Parametry: -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
- Metryka: confidence
- Ciekawsze reguły:
  - adoption-of-the-budget-resolution=y physician-fee-freeze=n – > Class=democrat
  - physician-fee-freeze=n aid-to-nicaraguan-contras=y – > Class=democrat
  - el-salvador-aid=n aid-to-nicaraguan-contras=y – > Class=democrat

- Wniosek: standardowy zestaw parametrów polecany przez program nie daje spodziewanych rezultatów. Otrzymaliśmy same reguły klasyfikujące demokratów. Nie mniej jednak z otrzymanego zestawu reguł można wydobyć jakąś wiedzę: demokraci mają podobne zdanie na temat niektórych problemów społecznych. Republikanie bardziej różnią się w swoich poglądach.

## 5.2 Apriori 2

- Parametry: -N 10 -T 1 -C 1.1 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
- Metryka: lift
- Ciekawsze reguły:
  - physician-fee-freeze=n – > Class=democrat
  - Class=democrat – > physician-fee-freeze=n
  - adoption-of-the-budget-resolution=y – > physician-fee-freeze=n
  - physician-fee-freeze=n – > adoption-of-the-budget-resolution=y
- Wniosek: zmiana metryki i lekka zmiana parametrów zaowocowała dwoma nowymi wnioskami: pierwszy znów dotyczy klasyfikacji. Zauważmy, że dwie pierwsze reguły tworzą równoważność. Dostajemy więc mocną zależność między problemem physician-fee-freeze a byciem demokratą. Drugi wniosek jest dla nas bardziej interesujący. Kolejne dwie reguły również tworzą równoważność, co daje nam mocną zależność między problemami adoption-of-the-budget-resolution a physician-fee-freeze.

## 5.3 Apriori 3

- Parametry: -N 20 -T 2 -C 0.1 -D 0.1 -U 1.0 -M 0.2 -S -1.0 -c -1
- Metryka: leverage
- Ciekawsze reguły:
  - anti-satellite-test-ban=y physician-fee-freeze=n aid-to-nicaraguan-contras=y – > el-salvador-aid=n
  - el-salvador-aid=n – > physician-fee-freeze=n anti-satellite-test-ban=y aid-to-nicaraguan-contras=y



- Wniosek: zwiększenie liczby poszukiwanych reguł oraz kolejna zmiana metryki dały nam ciekawsze wyniki.

## 5.4 FPGrowth 1

- Parametry: -P 2 -I -1 -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1
- Metryka: confidence
- Ciekawsze reguły:
  - el-salvador-aid=y, Class=republican: – > physician-fee-freeze=y
  - crime=y, Class=republican: – > physician-fee-freeze=y
- Wniosek: zmiana algorytmu wyznaczyła nam zmianę opcji politycznej. Algorytm FPGrowth preferuje reguły związane z republikanami.

## 5.5 FPGrowth 2

- Parametry: -P 2 -I -1 -N 10 -T 1 -C 0.9 -D 0.05 -U 1.0 -M 0.1
- Metryka: lift
- Ciekawsze reguły:
  - crime=y – > religious-groups-in-schools=y
  - adoption-of-the-budget-resolution=y – > aid-to-nicaraguan-contras=y
  - aid-to-nicaraguan-contras=y – > adoption-of-the-budget-resolution=y
  - el-salvador-aid=y – > religious-groups-in-schools=y
- Wniosek: z metryką lift nasz algorytm wyzbył się reguł posiadających klasyfikator. Baza wiedzy poszerza się o nowe reguły.

## 5.6 FPGrowth 3

- Parametry: -P 2 -I -1 -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1
- Metryka: conviction
- Ciekawsze reguły:
  - superfund-right-to-sue=y – > crime=y
  - mx-missile=y – > adoption-of-the-budget-resolution=y

- religious-groups-in-schools=y, crime=y – > el-salvador-aid=y
- mx-missile=y – > anti-satellite-test-ban=y
- Wniosek: zmiana parametrów oraz metryka conviction daje nam jeszcze więcej reguł.

## 5.7 Tertius 1

- Parametry: -F 0.5 -C 0.5 -N 1.0 -L 5 -G 0 -c 0 -I 0 -P 0
- Ciekawsze reguły:
  - education-spending = n – > crime = n
  - duty-free-exports = n – > aid-to-nicaraguan-contras = n
  - el-salvador-aid = y and crime = y – > adoption-of-the-budget-resolution = n or physician-fee-freeze = y
  - aid-to-nicaraguan-contras = y and Class = democrat – > physician-fee-freeze = n or mx-missile = y
- Wniosek: algorytm Tertius daje więcej możliwości niż poprzednie algorytmy. Zamiast zbiorów, po obu stronach implikacji znajdują się wyrażenia boolowskie, co czyni wyszukane reguły bardziej interesującymi.

## 6 Podsumowanie

Algorytmy wyszukujące reguły asocjacyjnie niewątpliwie przydają się do wydobywania wiedzy ze zgromadzonych danych. W wykorzystanym zbiorze zauważyliśmy dużą różnorodność znalezionych reguł w zależności od uruchomionego algorytmu. Dlatego oczywistym wnioskiem (a także radą na przyszłość) jest: wykorzystanie kilku algorytmów na jednym zbiorze danych jest dobrym sposobem na uzyskanie różnorodnej wiedzy.

Kolejną obserwacją, jaką wyciągnęliśmy z przeprowadzonych eksperymentów jest brak niektórych atrybutów w wynikowych regułach. Dla przykładu: w żadnej z reguł nie pojawia się problem imigrantów (mimo wielokrotnych powtórzeń testów). Nie oznacza to, że nie możemy wyciągnąć z tego żadnej wiedzy. Wręcz przeciwnie - oznacza to, że nastawienie do problemu imigrantów jest niezależne od innych problemów i opcji politycznej, do której dana osoba należy.