Metody odkrywania wiedzy
Dokumentacja projektowa

Model danych:
Dane dotyczące głosowań sejmowych ze strony Smarter Poland

Michał Kaszlej
Paweł Maj

Listopad 22, 2013

0.1 Temat projektu

- Projekt analityczny
- Zadanie klasyfikacji
- Model danych
  Dane dotyczące głosowań sejmowych ze strony Smarter Poland (dostęp dnia 22.11.2013).

0.2 Interpretacja tematu projektu

0.2.1 Opis zbioru danych

Zbiór danych obejmuje głosowania i skład VI kadencji Sejmu RP (2007-2011). Dane zorganizowane są w następujących plikach:

- `glosowania.txt`
  Plik zawiera wyniki glosowań poszczególnych posłów. Wiersze to nazwiska posłów, kolumny to identyfikatory głosowań.
- `glosowania_metadata.txt`
  Plik zawiera metadane głosowań indeksowane identyfikatorem głosowania. W szczególności sprecyzowana jest data głosownia, temat głosowania, podsumowanie złożonych głosów. Dodatkowo określone zostało czy podczas głosowania posłów danej partii obowiązywała dyscyplina.
- `poslowie_metadata.txt`
  Plik zawiera informacje na temat posłów. W szczególności ostatnią przynależność do partii, pierwsze i ostatnie głosowanie oraz podsumowanie absencji danego posła.
- `party_affiliations.txt`
  Plik zawiera informacje na temat przynależności partynej posła w danym głosowaniu.

Dane pochodzą z systemu orka.sejm.gov.pl, zostały udostępnione jako mechanizm kontroli społeczeństwa nad najwyższym organem władzy ustawodawczej Rzeczypospolitej Polskiej.

0.2.2 Wybór podstawowego atrybutu reprezentującego pojęcie docelowe

W przypadku wybranego przez nas zbioru danych, wybór podstawowego atrybutu nasuwa się samoczynnie - jest nim wynik głosowania. Jest to atrybut dyskretny, przyjmuje w zbiorze danych następujące wartości:

Wartość atrybutu	Interpretacja
1	Głos "za"
-1	Głos "przeciw"
NA	Brak głosu

Dodatkowo wartość "NA" atrybutu zostanie zastąpiona wartością 0. Przekształcenie to ma pozwolić na reprezentację atrybutu jako liczby całkowitej.

0.2.3 Dodatkowe atrybuty

Jako dodatkowe atrybuty można wyróżnić:

- Przynależność do partii politycznej
  W VI kadencji Sejmu RP, atrybut przyjmuje następujące wartości:

Wartość atrybutu	Liczba wystąpień w `party_affiliations.txt`
PO	1736535
PiS	1286816
NA	509520
PSL	259772
Lewica	217725
SLD	125318
niez.	79997
PJN	43630
LiD	22960
SDPL-NL	21657
SDPL	16803
Polska_XXI	14278
Polska_Plus	12816
DKP_SD	12066
DKP	9147

- Dysycyplina partyjna
  Plik `glosowania_metadata.txt` dla każdego głosowania zawiera informację, którę partie polityczne stosowały dyscyplinę partyjną w danym głosowaniu. Atrybut zapisany jest w postaci nazw partii politycznych oddzielonych przeciniem. Prawdopodobnie wymagane będzie przetworzenie atrybutu do prostszej postaci.

- Temat głosowania
  Plik `glosowania_metadata.txt` zawiera pola `nazwa1` i `nazwa2` zawierające informacje o temacie głosowania. Możliwe jest sformułowanie takiego zbioru słów i sformułowań kluczowych, jednozacznie wyróżniających interesujące nas tematy głosowań. Przykładowy zbiór takich sformułowań zaprezentowaliśmy poniżej:

Wartość atrybutu	liczba wystąpień w `glosowania_metadata.txt`
"budżetowej na rok 2008"	214
"budżetowej na rok 2009"	403
"budżetowej na rok 2010"	129
"budżetowej na rok 2011"	141

  Zwrócić należy szczególną uwagę na słowa kluczowe typu "przyjęcie", "odrzucenie" itp. Fakt występowania tych słów w temacie może zmieniać wartość logiczną głosowania.

0.3 Opis algorytmów, które zostaną wykorzystane

Trudno w tej chwili przedstawić dokładną listę algorytmów z których zamierzamy skorzystać.

Chcielibyśmy oprzeć projekt na algorytmie drzew decyzyjnych, jednak ze względu na zastosowanie dodatkowych atrybutów prawdopodobnie niezbędne okaże się skorzystanie dodatkowych metod przetwarzania danych.

Szczególną rolę będzie dla nas miało przetwarzanie tematów głosowań. Pozwoli to powiązać dane statystyczne z rzeczywistymi, nurtującymi nas w życiu codziennym problemami. Takie przetwarzanie wymagało będzie zdefinowania zbioru interesujących nas słów kluczowych i sformułowań. Dodatkowym problemem będzie też wyznaczanie wartości logiczej głosowań (można głosować za przyjęciem bądź za odrzuceniem). Sposób wyznaczania tego zbioru zdefiniowany zostanie w finalnej wersji dokumentacji.

0.4 Plan eksperymentów

0.4.1 Plan projektu

Z opisu projektu nie wynika czy przeprowadzić mamy jedną czy wiele klasyfikacji. Postanowiliśmy więc zdefiniować szereg eksperymentów, z których każdy zakończy się przyporządkowaniem każdego posła do jakiejś klasy.

Na wypadek gdyby wielokrotna klasyfikacja nie była punktowana postanowiliśmy także wyróżnić jeden eksperyment jako podstawowy.

Eksperymenty dodaktowe zostaną zrealizowane w drugiej kolejności, po zrealizowaniu eksperymentu podstawowego.

0.4.2 Eksperyment podstawowy

Naszą propozycją na eksperyment podstawowy, jest klasyfikacja posłów do partii politycznych.

Zastosowane przez nas podejście różni się tym od przypisania obecnego w danych, że klasyfikacja ta nie będzie zależna od wyboru posłów, lecz od wyników ich głosowań. Porównywać będziemy to, jak głosował w danych głosowaniach poseł - z tym jak głosowały partie polityczne. Pomijane będą dyscypliny partyjne dla tych posłów, których obowiązują.

Dodatkowe komplikacje wynikają z faktu, że część partii ulega rozwiązaniu w trakcie kadencji oraz to, że powstają nowe partie i koła poselskie.

W wyniku tej klasyfikacji posłowie zostaną przydzieleni do partii politycznych najbardziej zbieżnych z ich wynikami ich głosowań.

0.4.3 Eksperymenty dodatkowe

O ile uda się przeprowadzić eksperyment podstawowy, w następnej kolejności przeprowadzone zostaną poniższe eksperymenty:

1. Klasyfikacja posłów na podstawie tematów głosowań

   - Przygotowanie zbioru sformułowań kluczowych
     Przygotowanie zbioru polegać będzie na określeniu sformułować opisujących wybrane ważne głosowania VI kadencji Sejmu RP. Sformułowania dobrane muszą zostać w taki sposób, aby jednoznacznie określały zbiór głosowań związanych z danym zaganieniem.
   - Zdefiniowanie kategorii
     Przygotowane zostaną zależności między wartościami głosów w tematach kluczowych a kategoriami do których przydzieleni zostaną posłowie.
   - Stworzenie drzewa decyzyjnego
     Stworzone zostanie drzewo decyzyjne dla ustalonych reguł.

2. Głosy posła kontra głosy jego partii. Czy dany poseł głosuje zgodnie z linią swojej partii politycznej?

   - Porównianie głosowań posła do średniej głosowań jego partii. Jak zbieżne są głosy danego posła z głosami innych posłów jego partii?
   - Przedstawienie 'lojalności' posłów, średniej lojalności poszczególnych partii.
   - Wykrywanie przypadków dyscypliny partyjnej. Takie przypadki, nie uwzględniane, uwzględniane z mniejszą wagą.
   - Wykrywanie przypadków złamania dyscypliny partyjnej.

3. Analiza głosów typu "wstrzymuję się"

   - Sformułowanie zbioru sformułowań kluczowych, odpowiadającego trudnym głosowaniom VI kadencji Sejmu RP.
   - Eliminacja głosowań z dyscypliną partyjną.
   - Zaklasyfikowanie tematów jako tych z wysoką liczbą głosów wstrzymuję się oraz pozostałych.

0.5 Otwarte kwestie wymagające późniejszego rozwiązania (wraz z wyjaśnieniem powodów, dla których ich rozwiązanie jest odłożone na później)

- Sprecyzowanie szczegółów dotyczących zastosowanych algorytmów
  Tak jak zostało napisane wcześniej, projekt pragniemy zrealizować korzystając z algorytmu drzew decyzyjnych. W przypadku eksperymentu podstawowego mamy jednak do czynienia z zależnością od wielu atrybutów. Nalezy wziąć pod uwagę dyscypliny partyjne, oraz pojawianie się i znikanie partii politycznych.

Występowanie dodatkowych atrybutów, znacznie komplikuję zastosowanie drzewa decyzyjnego, dlatego też dokładniejszy opis algorytmu przedstawiony zostnie razem z realizacją.

- Kryteria oceny modeli
  Kryteria oceny modeli zależne będą od szczegółów algorytmu i zostaną sprecyzowane wraz z nimi.