AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA
W KRAKOWIE

Metody inżynierii wiedzy

Projekt: Asocjacyjna reprezentacja danych i wnioskowanie

Autor:
Jakub Noga

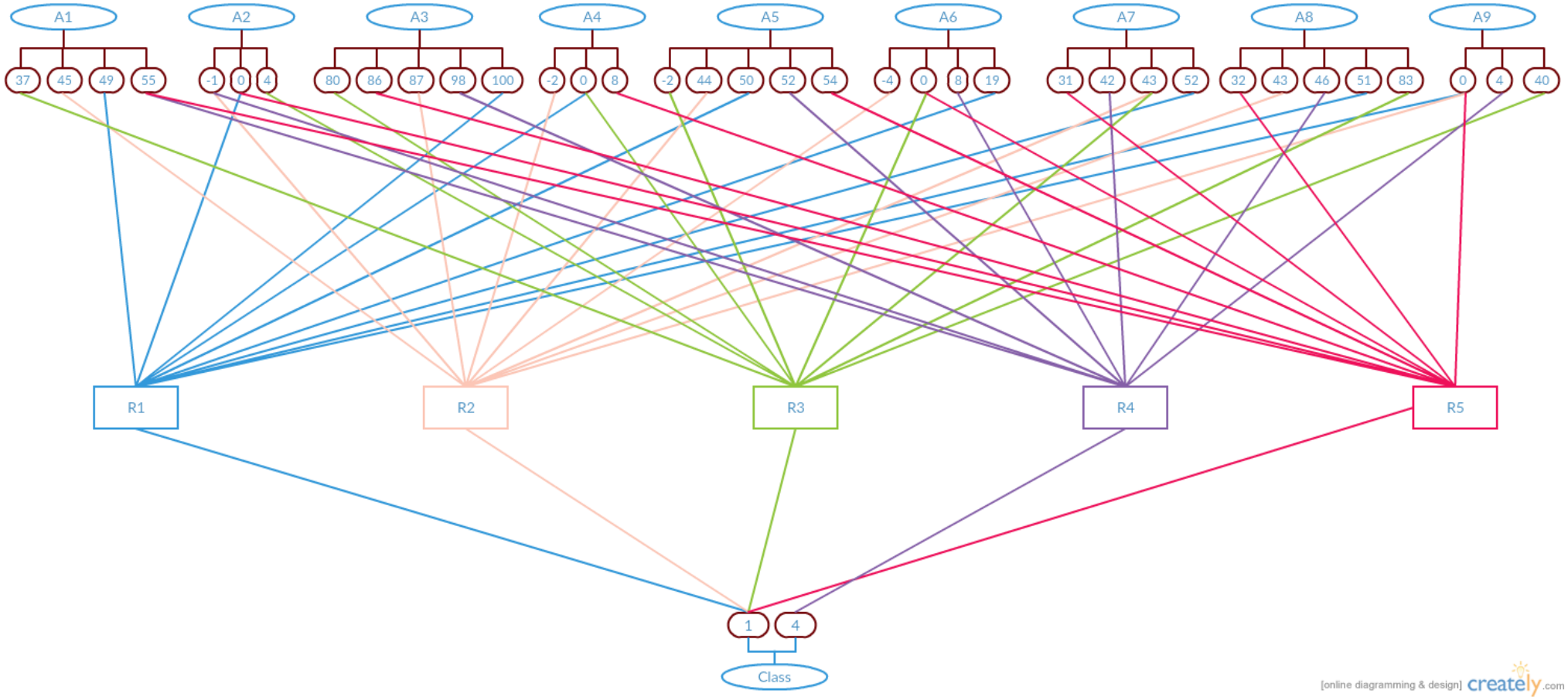
Asocjacyjna grafowa struktura danych

- Zalety asocjacyjnych grafowych struktur danych
  - Sortowanie względem wszystkich atrybutów jednocześnie, w przeciwieństwie do indeksów w bazach relacyjnych
  - Brak duplikatów
  - Powiązanie danych – łatwiejsze odnajdowanie korelacji pomiędzy rekordami (również korelacje obce)
- Wady asocjacyjnych grafowych struktur danych
  - Bardziej skomplikowane dodawanie nowych rekordów
  - Konieczność serializacji i deserializacji

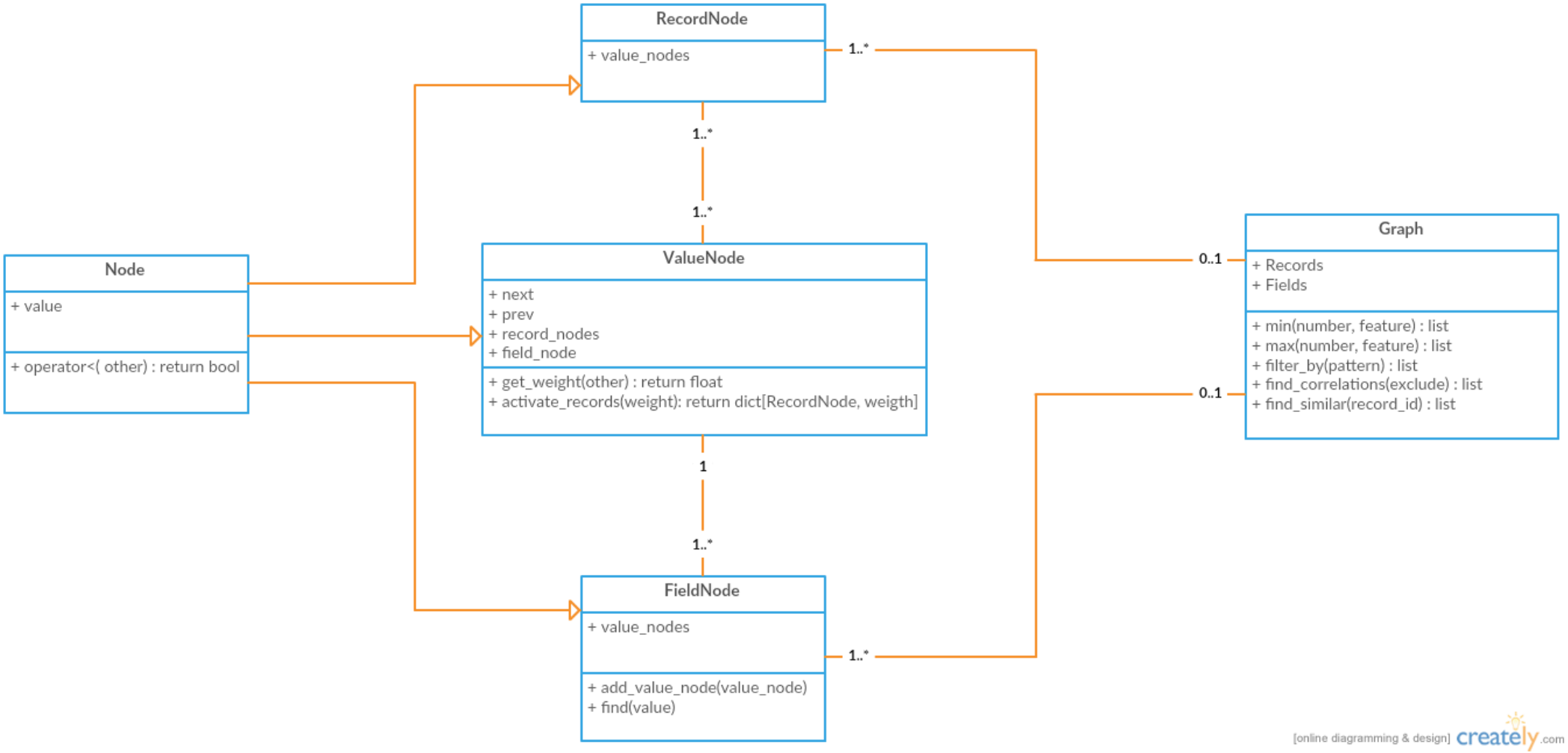
Dataset Shuttle

- Wspomaganie wyboru pomiędzy lądowaniem manualny i automatycznym wahadłowca.
- 9 (anonimowych) parametrów – liczba całkowita ze znakiem
- 7 klas – {1,2,3,4,5,6,7}
  - 1: Rad Flow
  - 2: Fpv Close
  - 3: Fpv Open
  - 4: High
  - 5: Bypass
  - 6: Bpv Close
  - 7: Bpv Open
- 80% danych w klasie 1.

Dataset Shuttle - AGDS



Implementacja



AGDS - Wnioskowanie

- Korelacje
- Minima
- Maksima
- N obiektów o wartościach maksymalnych/minimalnych
- Filtracja względem cechy i zestawu cech (wartość i przedział wartości)
- Wyszukiwanie obiektów podobnych

Pomiary czasu

Akcja	AGDS			Baza danych		
	Min	Max	Avg	Min	Max	Avg
Budowanie grafu AGDS	23.97	25.93	24.50	-	-	-
Minima	0.000005	0.00002	0.000006	0.0001	0.004	0.0002
Maksima	0.000005	0.00003	0.000008	0.0001	0.0004	0.0002
100 elementów o maks. wart.	0.00004	0.0003	0.0001	0.0009	0.001	0.001
100 elementów o min. wart.	0.00003	0.0002	0.0001	0.0009	0.002	0.001
Filtracja	0.04	0.08	0.05	0.3	0.33	0.3
Filtracja wielokryterialna	0.08	0.09	0.08	0.07	0.09	0.08

Cwiczenia laboratoryjne

- KNN + cross-validation
- Eksploracja danych

Eksploracja danych

- $T = \{t_1, t_2, \dots, t_n\}$
- $s(X, Y) = \frac{|\{t: X \in t \; lub \; Y \in t\}|}{|T|}$
- $c(X, Y) = \frac{|\{t: X \in t \; oraz \; Y \in t\}|}{|\{t: X \in t\}|}$
- Eksploracja reguł asocjacyjnych polega na wyszukiwaniu takich par $\mathrm{p} = (X, Y)$, że $s(p) \geq s_{min}$ oraz $c(p) \geq c_{min}$

Ekwiwalentna transformacja klas - ECLAT

- Przyporządkowanie ID transakcji do wzorca
- Prosta definicja funkcji $hasAll(a, \dots, y, z) = t(a) \cap \cdots \cap t(y) \cap t(z)$ – zbiór transakcji zawierących wszystkie poszukiwane wzorce
- Wtedy $s(X, Y) = \frac{|hasAll(x_1,x_2\dots,x_n) \cup hasAll(y_1,y_2,\dots,y_n)|}{|T|}$
- Oraz $c(X, Y) = \frac{|hasAll(x_1,x_2\dots,x_n,y_1,y_2,\dots,y_n)|}{|hasAll(x_1,x_2\dots,x_n)|}$

- Transakcje:

1	kawa,mleko,cukier,orzeszki
2	kawa,cukier,jajka
3	kawa,cukier,chleb,masło,ser
4	cukier,orzeszki,jajka,miód,płatki
5	mleko,jajka,masło
6	kawa,orzeszki,chleb
7	mleko,miód,płatki
8	jajka,chleb,masło,ser
9	mleko,chleb,ser

- ECLAT

Płatki	3,6
Mleko	0,4,6,8
Cukier	0,1,2,3
Jajka	1,3,4,7
Masło	2,4,7
Kawa	0,1,2,5
Chleb	2,5,7,8
Miód	3,6
Orzeszki	0,3,5
Ser	2,7,8

Przykładowe reguły asocjacyjne

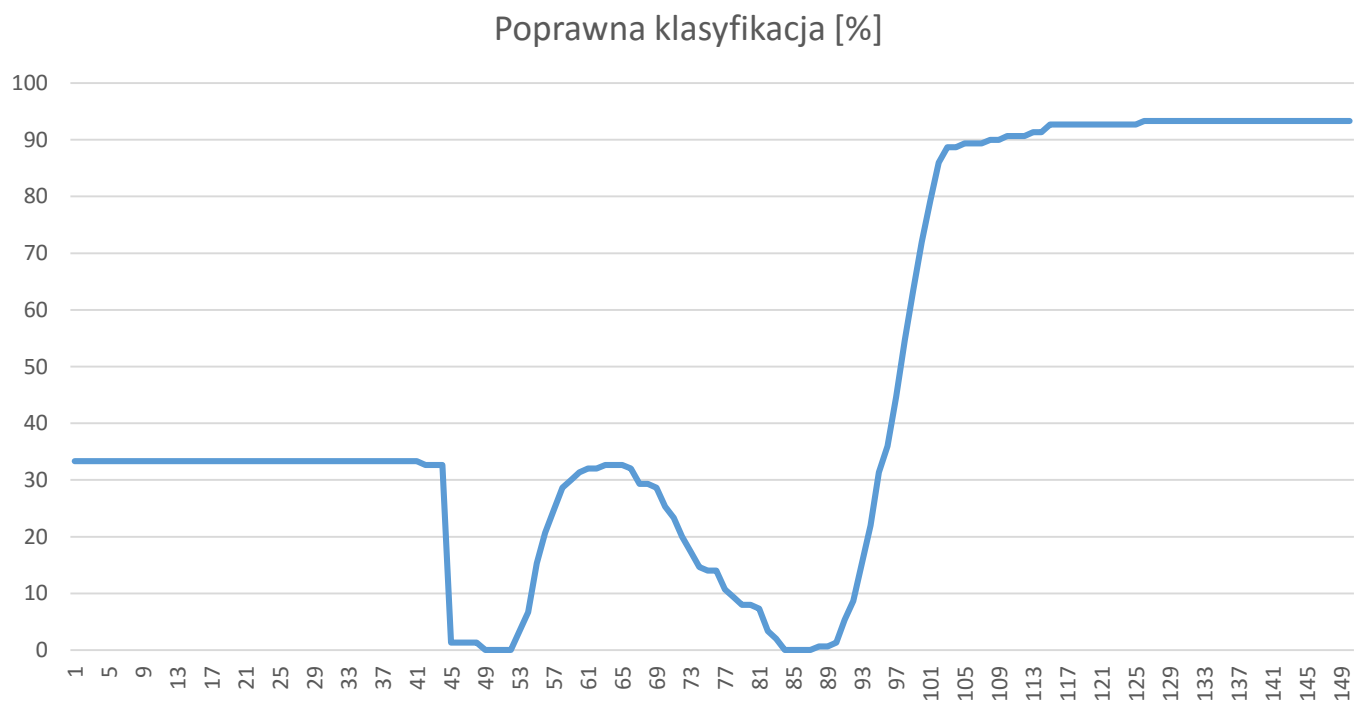
- [„Chleb”, „Masło”] -> [„Ser”] (33%,100%)
- [„Ser”] -> [„Chleb”, „Masło”] (33%, 66%)
- [„Chleb”, „Ser”] -> [„Masło”] (44%, 66%)
- [„Masło”] -> [„Chleb”, „Ser”] (44%, 66%)

KNN + cross-validation

- Walidacja krzyżowa pozwala dobrać optymalną wartość parametru k w metodzie kNN
- Walidacja zbiorem złożonym z reprezentatów klas w liczbie proporcjonalnej do wielkości poszczególnych klas
- Powtórzenie walidacji dla każdego k

Wyniki walidacji

- Najlepsze wyniki dla k >= 103
- Zwiększanie k do 150 nie przynosi poprawy klasyfikacji



Dziękuję z uwagę

Proszę o pytania