

Algorytm indukcji reguł decyzyjnych oparty na modelu EAV

Krzysztof Żabiński

Uniwersytet Śląski
Wydział Nauk Ścisłych i Technicznych
Instytut Informatyki

2021

Spis treści

Model EAV

Podstawowa wersja proponowanego algorytmu

Algorytm ze zautomatyzowanym wyborem atrybutów

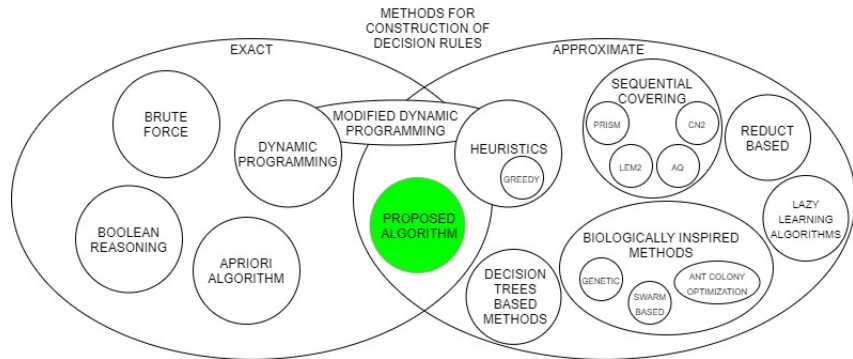
Przykład - podstawowa wersja algorytmu

Przykład - zautomatyzowany wybór atrybutów

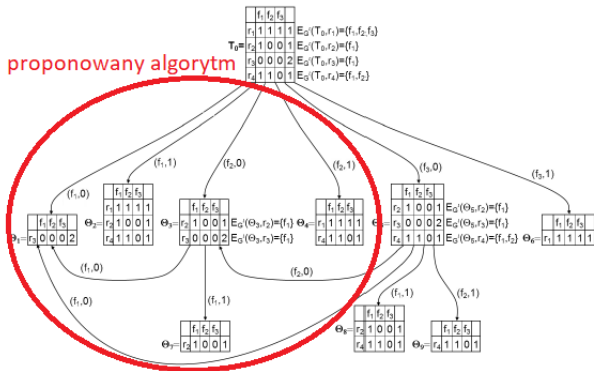
Wyniki eksperymentalne

Wnioski

Podejścia do generowania reguł



Umieszczenie algorytmu względem istniejących metod



Model EAV

Proponowane podejście generowania reguł decyzyjnych jest oparte na reprezentacji tablicy decyzyjnej w postaci EAV (entity-attribute-value), które pojawia się w literaturze¹. W każdym wierszu tabeli w postaci EAV wykorzystanej w podejściu znajduje się:

- | nazwa atrybutu,
- | wartość atrybutu,
- | wartość atrybutu decyzyjnego,
- | numer wiersza w oryginalnej tabeli decyzyjnej.

¹Kowalski, M.; Stawicki, S. SQL-Based Heuristics for Selected KDD Tasks over Large Data Sets. Proceedings of the Federated Conference on Computer Science and Information Systems. IEEE, 2012, pp. 303–310.

Model EAV & SQL

Reprezentacja tablicy decyzyjnej w postaci EAV jest bardzo użyteczna z punktu widzenia wykorzystania bazy danych do generacji reguł decyzyjnych.

RDBMS są dobrym wyborem do analizy dużych zbiorów danych - odpowiednia optymalizacja.

```
CREATE TABLE eav
(
  id serial primary key,
  attribute character varying,
  value character varying,
  decision character varying,
  row bigint
);
```

Rysunek: Komenda SQL tworząca tabelę EAV w PostgreSQL użyta do implementacji proponowanego algorytmu.

Podstawowa wersja proponowanego algorytmu

Proponowany algorytm składa się z 3 podstawowych etapów:

- | przekształcenie wejściowej tablicy decyzyjnej do postaci eav,
- | stworzenie rankingu atrybutów,
- | generowanie reguł decyzyjnych;

Pseudokod algorytmu

Algorithm 1 Pseudokod algorytmu generującego reguły decyzyjne dla tablicy decyzyjnej T .

Input: Input decision table T , number p of best attributes to be taken into consideration.

Output: Set of decision rules R

represent T as entity–attribute–value (EAV) form with separate decision;

represent each attribute's value in a discrete numerical form;

obtain attributes' standard deviation per decision class.

take p number of attributes of largest STD—in a descending order;

from T in EAV form select sets v of unique values (including decision) of attributes grouped per decision table's rows;

while there exist sets v_i in v not marked as processed do

generate one-item v_i' set with initial value from v_i which corresponds to creation of separable

subtable $T' = T(f_i, a_i)$;

set v_i is not processed;

while iterations number $< \text{sizeof}(v_i)$ OR separable subtable is not degenerate do

extend v_i' by supplying it with the subsequent element from v_i which corresponds to next partition of T' ;

end while

generate decision rule basing on the values of attributes from v_i' (consequent is the most common decision for T' corresponding to v_i');

supply the set R with the newly created rule;

set v_i being processed.

end while

Ranking atrybutów

Ranking atrybutów jest tworzony na podstawie obliczania odchylenia standardowego względem każdego atrybutu w ramach klasy decyzyjnej, korzystając z polecenia:

```
SELECT attribute, STDDEV(average_value) AS quality FROM
(
  SELECT e.attribute, e.decision, AVG(v.id) AS average_value
  FROM eav e JOIN values v ON e.value = v.value
  GROUP BY attribute, decision
) attribute_average_values
GROUP BY attribute
ORDER BY quality DESC;
```

Rysunek: Komenda SQL obliczająca std dla każdego atrybutu.

Algorytm ze zautomatyzowanym wyborem atrybutów

Algorytm ze zautomatyzowanym wyborem atrybutów różni się od wersji podstawowej funkcją tworzenia rankingu atrybutów.

Tworzenie rankingu atrybutów:

1. obliczenie std per klasa decyzyjna dla każdego atrybutu (w języku SQL grupowanie po atrybucie i klasie decyzyjnej),
2. obliczenie std per klasa decyzyjna dla wszystkich wartości atrybutów z tabeli EAV (w języku SQL grupowanie tylko po klasie decyzyjnej) - jest to możliwe ze względu na zastosowanie jednorodnych numerycznych odpowiedników dla wartości każdego atrybutu,
3. przygotowanie rozkładu skumulowanego dla std ad.1 i dla std ad.2,
4. wybór atrybutów, dla których krzywa rozkładu przyrasta wolniej niż krzywa wspólna dla std ad.2.

Przyk^aad - podstawowa wersja algorytmu - model EAV

Przyk^aad c.d.

Przyk^aad c.d.

Zak^aadajc, »e do analizy bierzemy np. 60% najlepszych atrybutów, wybieramy atrybuty 2 i 3.

mo»liwe warto±ci atrybutów f2 i f3 z wej±ciowej tablicy decyzyjnej:
{1,1}, {1,0}, {0,1}, {0,0}.

Przyk^aad c.d.

Na podstawie przedstawionego wszeźniej grafu, mo»emy opisać regu^ay decyzyjne:

$$| \quad f_2 = 1 \wedge f_3 = 1! \quad d = 1$$

$$| \quad f_2 = 1 \wedge f_3 = 0! \quad d = 2$$


$$| \quad f_2 = 0! \quad d = 3$$

Przykład - zautomatyzowany wybór atrybutów

Różnica względem algorytmu w wersji podstawowej - analiza rozkładu std. Rysunek poniżej pokazuje rozkład skumulowany dla std ad.1 i std ad.2.

Założenie o złożoność obliczeniową algorytmu

średnia złożoność obliczeniowa algorytmu $\Theta(n)$, a pesymistyczna $O(n^2)$.

- | zbiory eksperymentalne wybrane z UCI Machine Learning Repository,
- | zbiory zostały przygotowane: zduplikowane wartości usunięte, a wartości brakujące zastąpione MCV,
- | cel eksperymentów:
 - | porównanie proponowanego algorytmu z klasycznym DP,
 - | porównanie proponowanego algorytmu z regu^aami wygenerowanymi na podstawie reduktów,
 - | porównanie algorytmu w wersji podstawowej i ze zautomatyzowanym wyborem atrybutów.
- | wykorzystane metody walidacji:
 - | 2-krotna walidacja krzyżowa z podziałem na zbiory treningowy walidacyjny i testowy (30:20:50) przeprowadzona 50 razy,
 - | 10-krotna walidacja krzyżowa.
- | wyniki porównane z wykorzystaniem testu Wilcoxona. 

Wyniki eksperymentalne

Tabela: 'redni b^aid klasy kacji dla proponowanego algorytmu - algorytmu w wersji podstawowej i DP

data set	100% of attributes		80% of attributes		60% of attributes		DP len	DP sup
	error	std	error	std	error	std	error	error
balance-scale	0.45	0.08	0.45	0.08	0.48	0.08	0.29	0.28
breast-cancer	0.00	0.00	0.00	0.00	0.00	0.02	0.31	0.30
cars	0.07	0.11	0.08	0.13	0.18	0.21	0.22	0.21
hayes-roth-data	0.52	0.10	0.52	0.10	0.51	0.12	0.37	0.35
house-votes	0.25	0.12	0.25	0.12	0.26	0.14	0.08	0.05
lymphography	0.16	0.11	0.16	0.10	0.14	0.11	0.35	0.28
nursery	0.00	0.01	0.00	0.01	0.00	0.0	0.05	0.05
shuttle-landing	0.22	0.19	0.18	0.17	0.22	0.18	0.40	0.39
soybean-small	0.21	0.14	0.21	0.14	0.21	0.14	0.17	0.17
zoo-data	0.40	0.26	0.38	0.25	0.36	0.24	0.24	0.18
average	0.21	0.10	0.20	0.10	0.21	0.11	0.25	0.23

Wyniki eksperymentalne

Tabela: Różnica względna długości i wsparcia regu^a decyzyjnych

data set	Number of		100% of attributes		80% of attributes		60% of attributes	
	rows	attributes	len	sup	len	sup	len	sup
balance-scale	625	4	0.14	-0.42	0.14	-0.42	-0.06	-0.07
breast-cancer	266	9	1.03	-0.61	1.00	-0.61	0.81	-0.57
cars	1728	6	0.48	-0.38	0.32	-0.38	0.14	-0.37
hayes-roth-data	69	5	0.23	-0.42	0.23	-0.42	0.09	-0.40
house-votes	279	16	1.60	-0.58	1.60	-0.58	1.45	-0.58
lymphography	148	18	0.97	-0.87	0.97	-0.87	0.94	-0.87
nursery	12960	8	1.29	-1.00	1.25	-1.00	0.60	-0.99
shuttle-landing	15	6	1.19	-0.09	0.85	-0.09	0.69	-0.09
soybean-small	47	35	3.40	-0.78	3.40	-0.78	3.40	-0.78
zoo-data	59	16	2.12	-0.49	1.83	-0.49	1.86	-0.49

$$\text{Roznica_względna} = \frac{\text{wartosc_dla_algorytmu} - \text{wartosc_dla_DP}}{\text{wartosc_dla_DP}}$$

Wyniki eksperymentalne

Tabela: Wyniki klasyfikacji dla regu^a indukowanych na podstawie reduktów.

data set	alpha=0		alpha=0.001		alpha=0.01		alpha=0.1	
	accuracy	std	accuracy	std	accuracy	std	accuracy	std
breast-cancer	0.80	0.35	0.82	0.33	0.83	0.35	0.84	0.36
cars	0.73	0.05	0.75	0.07	0.77	0.08	0.76	0.05
house-votes	0.67	0.04	0.66	0.02	0.69	0.01	0.67	0.07
kr-vs-kp	0.60	0.40	0.61	0.38	0.57	0.41	0.57	0.41
mushroom	0.74	0.05	0.79	0.05	0.79	0.02	0.79	0.02
soybean-small	0.81	0.34	0.84	0.35	0.85	0.32	0.80	0.34
spect-test	0.99	0.02	0.99	0.02	0.99	0.02	0.99	0.02
tic-tac-toe	0.74	0.37	0.74	0.37	0.74	0.37	0.74	0.37

Test Wilcoxon pokaza^a, że wyniki są porównywalne dla różnych wartości α .

Wyniki eksperymentalne

Tabela: Wyniki klasyfikacji dla regu^a indukowanych na podstawie proponowanego algorytmu w wersji podstawowej.

data set	100%		80%		60%	
	accuracy	std	accuracy	std	accuracy	std
breast-cancer	0.78	0.32	0.79	0.34	0.81	0.34
cars	0.79	0.06	0.80	0.02	0.81	0.03
house-votes	0.74	0.07	0.77	0.10	0.78	0.10
kr-vs-kp	0.66	0.32	0.70	0.28	0.74	0.25
mushroom	0.87	0.13	0.87	0.12	0.87	0.12
soybean-small	0.81	0.31	0.82	0.31	0.84	0.32
spect-test	0.88	0.09	0.90	0.11	0.92	0.13
tic-tac-toe	0.71	0.39	0.75	0.39	0.77	0.39

Test Wilcoxon pokaza^a, że wyniki są lepsze dla 60% atrybutów w porównaniu do 100% i 80% atrybutów oraz w porównaniu do klasyfikacji regu^a uzyskanych na podstawie reduktów.

Wyniki eksperymentalne

Porównanie \pm redniej długości regu^a wyindukowanych dla reduktów i proponowanego algorytmu w wersji podstawowej.

Wyniki eksperymentalne

Tabela: Wyniki klasyfikacji dla regu^a indukowanych na podstawie proponowanego algorytmu w obu wersjach.

data set	100%		80%		60%		automatic	
	accuracy	std	accuracy	std	accuracy	std	accuracy	std
breast-cancer	0.78	0.32	0.79	0.34	0.81	0.34	0.83	0.07
cars	0.79	0.06	0.80	0.02	0.81	0.03	0.82	0.05
house-votes	0.74	0.07	0.77	0.10	0.78	0.10	0.81	0.09
kr-vs-kp	0.66	0.32	0.70	0.28	0.74	0.25	0.75	0.12
mushroom	0.87	0.13	0.87	0.12	0.87	0.12	0.92	0.03
soybean-small	0.81	0.31	0.82	0.31	0.84	0.32	0.79	0.10
spect-test	0.88	0.09	0.90	0.11	0.92	0.13	0.99	0.01
tic-tac-toe	0.71	0.39	0.75	0.39	0.77	0.39	0.84	0.09

Test Wilcoxon pokaza^a, że wyniki dla zautomatyzowanej wersji są porównywalne z wynikami dla 60% atrybutów oraz lepsze od wyników dla 100% i 80% atrybutów.

Wyniki eksperymentalne

Porównanie maksymalnej możliwej długości regu^a wyindukowanych dla proponowanego algorytmu w wersji podstawowej (60% atrybutów) i w wersji ze zautomatyzowanym wyborem atrybutów.

Wnioski

- | Zaproponowany algorytm umożliwia generowanie regu^a o jakości klasyfikacji zbliżonej do algorytmu DP oraz o dużości i wsparciu niedalekich od wartości optymalnych.
- | Algorytm w wersji zautomatyzowanej pozwala na lepsze dopasowanie do analizowanego zbioru danych pod kątem wybieranych atrybutów.
- | Algorytm pozwala na pracę z dużymi zbiorami danych - mniejsza złożoność obliczeniowa niż w przypadku DP.

Przyszłe badania

- | Zastosowanie proponowanego algorytmu do selekcji cech dla danych stylometrycznych.
- | Porównanie z heurystykami, które pozwalają na uzyskanie reguł bliskich regułom optymalnym z punktu widzenia reprezentacji wiedzy (długość i wsparcie).