



Sztuczna Inteligencja

(laboratorium 6)

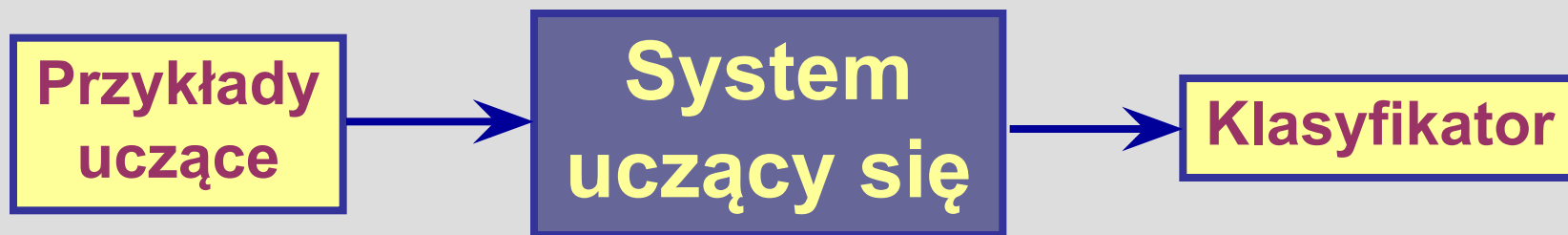
Katedra Systemów Ekspertowych i Sztucznej Inteligencji

Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie

Wprowadzenie

- **Konstruowanie systemu klasyfikującego**
- **Budowa regułowych baz informacyjnych**
- **Parametry oceny bazy reguł**
- **Proces klasyfikacji przypadków nieznanych**
- **Optymalizacja reguł**

- **Konstruowanie systemu klasyfikującego**



Uczenie się klasyfikatora ze zbioru uczącego



Klasyfikowanie nowych obiektów

Tablica decyzji – Fruits.tab

Lp	color	size	shape	taste	weight	fruit
1.	green	medium	round	sour	0.1	Lemon
2.	green	medium	round	sweet	0.5	Apple
3.	green	big	round	sweet	0.4	Apple
4.	green	medium	round	sour	0.3	Apple
5.	green	small	round	sweet	0.1	Grape
6.	yellow	big	round	sweet	0.5	Grapefruit
7.	yellow	big	round	sour	0.6	Grapefruit
8.	green	big	round	sour	0.5	Grapefruit
9.	yellow	medium	elongated	sweet	0.3	Banana
10.	green	medium	elongated	sweet	0.2	Banana
11.	green	big	elongated	sweet	0.3	Banana
12.	yellow	medium	round	sour	0.2	Lemon
13.	yellow	small	round	sour	0.1	Lemon
14.	red	small	round	sweet	0.1	Cherry
15.	red	medium	round	sweet	0.4	Apple
16.	red	big	round	sweet	0.5	Apple
17.	maroon	small	round	sweet	0.2	Grape
18.	green	small	round	sour	0.1	Grape
19.	yellow	big	elongated	sweet	0.3	Banana
20.	green	medium	round	sour	0.1	Lemon
21.	green	medium	round	sour	0.1	Lemon
22.	red	medium	round	sweet	0.3	Apple
23.	red	medium	round	sour	0.4	Apple

Budowa baz regułowych – algorytm GTS

CEL_KONSULTACJI fruit

REGULA 1

JEZELI size JEST medium

ORAZ weight $\geq$ 0.1000

ORAZ weight $<$ 0.1500

TO fruit JEST Lemon

REGULA 2

JEZELI size JEST medium

ORAZ weight $\geq$ 0.4500

ORAZ weight $<$ 0.5500

TO fruit JEST Apple

-
-
-

REGULA 12

JEZELI color JEST red

ORAZ weight $\geq$ 0.4500

ORAZ weight $<$ 0.5500

TO fruit JEST Apple

Parametry oceny reguł

Ocena zbioru reguł:

- liczba reguł
- liczba warunków w regule (średnia liczba warunków w regułach)
- dokładność zbioru reguł (ang. *accuracy*) – liczba poprawnie klasyfikowanych przypadków do liczby wszystkich klasyfikowanych przypadków
- błąd klasyfikacji (ang. *error rate*) - liczba błędnie sklasyfikowanych przypadków w stosunku do liczby wszystkich klasyfikowanych przypadków

Parametry oceny reguł

$$H = G + \text{sqrt}(A)$$

przy czym:

$$G \text{ (generality)} = (E^c + E^e)/E$$

gdzie:

E^c – jest liczbą przykładów ze zbioru treningowego, poprawnie sklasyfikowanych

E^e - jest liczbą przykładów ze zbioru treningowego, sklasyfikowanych błędnie

E - jest liczbą wszystkich przykładów w zbiorze treningowym

natomiast

$$A \text{ (accuracy)} = E^c / (E^c + E^e)$$

(gdy $A=1$, system GTS tworzy regułę)

Parametry oceny reguł

Ocena wybranej reguły:

- Siła reguły (ang. *Strength*, E^C) jest liczbą poprawnie klasyfikowanych przypadków ze zbioru uczącego. Siła reguły jest używana w obliczaniu pozostałych parametrów reguły.

$$Strength = E^C$$

- Dokładność reguły (ang. *Accuracy*) jest to stosunek liczby poprawnie klasyfikowanych przypadków ze zbioru uczącego (E^C) do sumy liczby poprawnie (E^C) oraz błędnie (E^E) klasyfikowanych przypadków ze zbioru uczącego.

$$Accuracy = \frac{E^C}{E^C + E^E}$$

Parametry oceny reguł

Ocena wybranej reguły:

- Ogólność reguły (ang. *Generality*) jest to stosunek sumy liczby poprawnie (E^C) oraz błędnie (E^E) klasyfikowanych przypadków ze zbioru uczącego do liczby wszystkich przypadków ze zbioru uczącego (E^{ALL}).

$$Generality = \frac{E^C + E^E}{E^{ALL}}$$

- Specyficzność reguły (ang. *Specificity*) jest to stosunek liczby poprawnie klasyfikowanych przypadków ze zbioru uczącego (E^C) do liczby przypadków z danej klasy decyzyjnej (E^{CLASS}).

$$Specificity = \frac{E^C}{E^{CLASS}}$$

Parametry oceny reguł

Ocena wybranej reguły:

- Wsparcie reguły (ang. *Support*) jest to stosunek liczby poprawnie klasyfikowanych przypadków ze zbioru uczącego (E^C) do liczby wszystkich przypadków ze zbioru uczącego (E^{ALL}).

$$Support = \frac{E^C}{E^{ALL}}$$

Klasyfikowanie przypadków za pomocą reguł:

Reguły decyzyjne wygenerowane z przykładów uczących używane są do klasyfikowania nowych obiektów (lub przykładów testowych)

Klasyfikowanie obiektów opiera się na **dopasowaniu** (ang. *matching*) opisu obiektu do części warunkowych reguł decyzyjnych. Wyróżniamy dopasowanie **pełne** i **częściowe**.

Pełne dopasowanie (ang. *complete matching*) – opis klasyfikowanego obiektu spełnia wszystkie warunki elementarne, występujące w części warunkowej reguły

Częściowe dopasowanie (ang. *partial matching*) – istnieje przynajmniej jeden warunek elementarny, który nie jest spełniony przez opis klasyfikowanego obiektu

Klasyfikowanie przypadków za pomocą reguł:

Dla danego rozpatrywanego przypadku nieznanego sprawdzamy czy istnieje dopasowana do niego reguła. W takim przypadku można wyróżnić trzy sytuacje:

1. Jeżeli istnieje tylko jedna reguła to obiekt jest przez nią klasyfikowany.

Klasyfikowanie przypadków za pomocą reguł:

Dla danego rozpatrywanego przypadku nieznanego sprawdzamy czy istnieje dopasowana do niego reguła. W takim przypadku można wyróżnić trzy sytuacje:

2. Jeżeli istnieje więcej niż jedna reguła system sprawdza czy reguły wskazują różne klasy.

- jeżeli tak to obliczane jest poparcie dla danej klasy decyzyjnej:

$$SUP(K_K) = \sum_{r \in R_K} Sila(r) * Liczba_warunków(r)$$

gdzie R_K oznacza reguły $r \in R_K$ (wskazujące klasę K_K) dopasowane do obiektu nieznanego. Przypadek nieznany jest przydzielany do klasy K_K , dla której poparcie $SUP(K_K)$ jest największa

- jeżeli nie to obiekt jest przypisywany do klasy wskazywanej przez reguły.

Klasyfikowanie przypadków za pomocą reguł:

Dla danego rozpatrywanego przypadku nieznanego sprawdzamy czy istnieje dopasowana do niego reguła. W takim przypadku można wyróżnić trzy sytuacje:

3. Jeżeli nie istnieje reguła dopasowana do przypadku, szukamy reguł **częściowo** dopasowanych do obiektu nieznanego **e**. Dla każdej z nich obliczamy dodatkową miarę

$$MF(r, e) = \frac{\text{liczba_warunków_dopasowanych_reguly_r}}{\text{liczba_wszystkich_warunków_w_regule_r}}$$

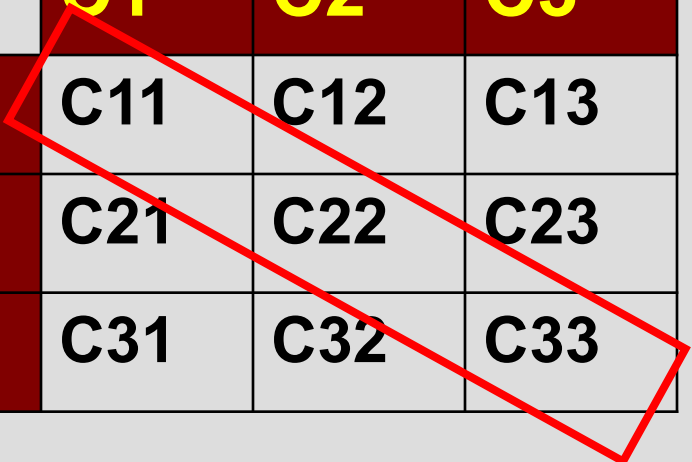
Dla reguł częściowo dopasowanych z danej klasy **K_K** liczymy

$$SUPP(K_K) = \sum_{r \in R_K} MF(r, e) * Sila(r) * Liczba_warunków(r)$$

Gdzie **R_K** oznacza reguły **R_K** częściowo dopasowane do obiektu **e**.
Obiekt **e** jest przypisywany do klasy **K_K** która ma największe **SUPP(K_K)**

Macierz rozproszenia (confusion matrix)

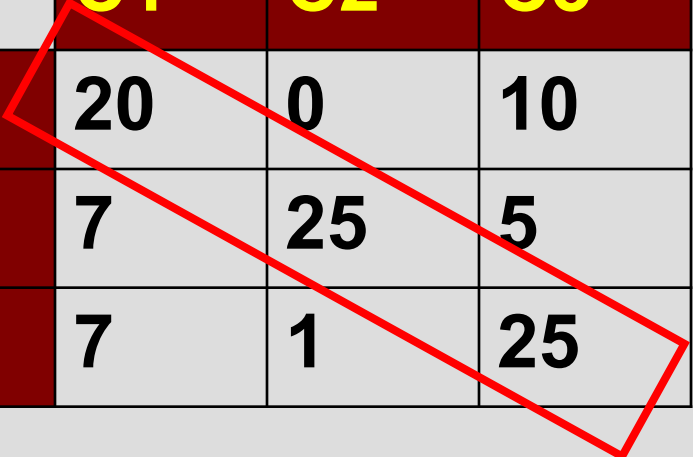
	C1	C2	C3
C1	C11	C12	C13
C2	C21	C22	C23
C3	C31	C32	C33



$$\text{Dokładność zbioru reguł} = \frac{(C11 + C22 + C33)}{\text{liczba wszystkich klasyfikowanych przypadków}}$$

Macierz rozproszenia (confusion matrix)

	C1	C2	C3
C1	20	0	10
C2	7	25	5
C3	7	1	25



$$\text{Dokładność zbioru reguł} = \frac{(20 + 25 + 25)}{20+25+25+10+5+7+7+1} = \frac{70}{100} = 0,7$$

Dokładność zbioru reguł = 70%

Error Rate = 100% - Dokładność zbioru reguł = 30%

Optymalizacja zbioru reguł:

- **Usunięcie reguł redundantnych (RR):**

Operacja polegająca na usunięciu reguł, które posiadają identyczne warunki i ich wartości w części warunkowej w ramach tej samej kategorii decyzji.

- **Usunięcie reguł zbędnych (NU):**

Operacja polegająca na usunięciu reguł, które nie klasyfikują prawidłowo żadnego z przypadków ze zbioru uczącego.

- **Usunięcie reguł pochłaniających się (AR):**

Operacja polegająca na usunięciu reguł, które posiadają wspólną z innymi regułami część warunkową, wzbogaconą dodatkowo warunkami uzupełniającymi.

Optymalizacja zbioru reguł:

- **Usunięcie zbędnych warunków (W):**

Operacja polegająca na usunięciu z danej reguły warunków, które nie powodują zmiany liczby prawidłowo klasyfikowanych przypadków ze zbioru uczącego.

- **Łączenie reguł (Ł):**

Operacja polegająca na połączeniu reguł zawierających ten sam zestaw atrybutów numerycznych w części warunkowej. Wartości tych atrybutów stanowią przedziały liczbowe zawierające się w sobie lub zachodzące na siebie. Zestaw atrybutów symbolicznych i ich wartości w części warunkowej łączonych reguł musi być identyczny. Cała operacja odbywa się w ramach tej samej klasy decyzji.

Optymalizacja zbioru reguł:

- **Utworzenie reguł brakujących (BR):**

Często obserwuje się że opracowany model uczenia (zbiór reguł) nie klasyfikuje wszystkich przypadków ze zbioru uczącego. Na podstawie tych przypadków nieklasyfikowanych tworzone są tzw. **reguły brakujące**. Reguły te tworzone mogą być dwoma metodami:

- **pierwsza metoda (Standardowa)** polega na utworzeniu reguł zawierających warunki utworzone na podstawie wszystkich atrybutów opisujących i ich wartości występujących w przypadkach nieklasyfikowanych,
- **druga metoda (Algorytm GTS)** polega na utworzeniu nowych, dodatkowych reguł przy użyciu algorytmu pokrycia General-To-Specific (GTS) operującego na zbiorze przypadków nieklasyfikowanych.

Optymalizacja zbioru reguł:

- Wybór reguł finalnych (FR):

Operacja polega na wyborze spośród całego zbioru reguł, tzw. reguł finalnych. Reguły te wybierane są na podstawie wartości parametru Istotność reguły (*Importance*) obliczanego dla każdej reguły r

$$Istotność(r) = Siła(r) * Liczba_warunków(r) + Specyficzność(r) - Słabość(r)$$

gdzie Słabość reguły (*Weakness*) jest to stosunek liczby błędnie klasyfikowanych przypadków ze zbioru uczącego (E^E) do liczby poprawnie klasyfikowanych przypadków ze zbioru uczącego (E^C)

$$Weakness = \frac{E^E}{E^C}$$

Optymalizacja zbioru reguł:

- Wybór reguł finalnych (FR):

Następnie reguły są sortowane rosnąco według parametru $H(r)$. Wyłączana jest pierwsza reguła (najmniejszy parametr *Istotność H*) i sprawdzane czy pozostałe reguły klasyfikują wszystkie przypadki z tablicy decyzji.

Jeżeli tak to reguła ta wyłączana jest ze zbioru reguł. Operacja ta jest wykonywana na wszystkich kolejnych regułach. Efektem działania jest zbiór reguł o najwyższym parametrze $H(r)$ – *najbardziej istotnych w zbiorze reguł* - pokrywający wszystkie przypadki ze zbioru uczącego.

Zadania

1. Oceń ten zbiór reguł dla pliku źródłowego **Socz_v0X.tab**, oblicz: siłę, specyficzność, dokładność, ogólność, wsparcie reguł. W pliku SOCZ_REG.txt zamieszczono zbiór reguł decyzji.

2. Wyznacz reguły dla WSZYSTKICH zbiorów uczących Socz_v0X.tab (przy użyciu programu **K:MWNS\SOFT\RuleSEEKER**). Zastosuj krok NR -> Algorytm GTS.

3. Regułami Socz_REG.kb, oblicz błąd klasyfikacji:

- przed optymalizacją
- po optymalizacji

metodą standardową i częściowym dopasowaniem dla swojego zbioru Socz_v0X – testowego.

(X – ostatnia cyfra z numeru indeksu studenta)