

Klasyfikacja – wykład dla PiESI (2007)

1. **Wprowadzenie: co to jest klasyfikacja nadzorowana.**
2. **Różne metody odkrywania wiedzy klasyfikacyjnej**
3. **Metody oceny wiedzy o klasyfikacji obiektów –
Miary.**
4. **Eksperymentalna ocena klasyfikatorów.**
5. **Porównanie wielu metod i realizacja procesu
odkrywania wiedzy.**



1

Uczenie się klasyfikacji nadzorowanej



Przykład	x_1	x_2	x_3	x_4	y
0	0	1	1	0	0
1	0	0	0	0	0
2	0	0	1	1	1
3	1	0	0	1	1
4	0	1	1	0	0
5	1	1	0	0	0
6	0	1	0	1	0

- C - target function \rightarrow funkcja klasyfikacyjna $c(x)$ nieznana algorytmowi odkrywającemu wiedzy. Algorytm poszukuje hipotezy h , najlepiej przybliżającej $c(x)$.

2

Wiedza klasyfikacyjna

- Problem określania zasad przydziału obiektów do znanych wstępnie klas na podstawie analizy danych o przykładach klasyfikacji.

Wiek	Zawód	dochód	...	Decyzja
21	Prac. fiz.	1220	...	Nie kupi
26	Menedżer	2900	...	Kupuje
44	Inżynier	2600	...	Kupuje
23	Student	1100	...	Kupuje
56	Nauczyciel	1700	...	Nie kupi
...
45	Lekarz	2200	...	Nie kupi
25	Student	800	...	Kupuje

Przykłady uczące

Algorytm eksploracji



Reprezentacja wiedzy:
np. reguły
R1. Jeżeli student to kupuje komputer
R2. Jeżeli dochód > 2400 ...

3

Reprezentacja przykładów wejściowych

zatrudn.	przeznacz. kredytu	stan cywilny	stan konta	staż pracy	Klasa
P	Komp.	S	...	20000	1	tak
P	sprzet	S	...	4000	2	tak
B	sprzet	R	...	600	0	nie
P	samochod	R	30000	7	tak
P	samochód	R	...	3000	2	nie
B	meble	R	...	2500	0	nie
P	wakacje	R	...	5600	15	tak
P	sprzet	S	...	4000	2	tak

4

Wiedza klasyfikacyjna

- **Typowe zastosowania**

- przydział kredytów,
- analiza danych finansowych,
- diagnoza i dobór terapii w medycynie,
- „target marketing”,
- diagnostyka techniczna i systemy sterowania,
- klasyfikacja dokumentów, wiadomości,
- przewidywanie struktur drugorzędowych w biologii,
- oraz wiele innych ...

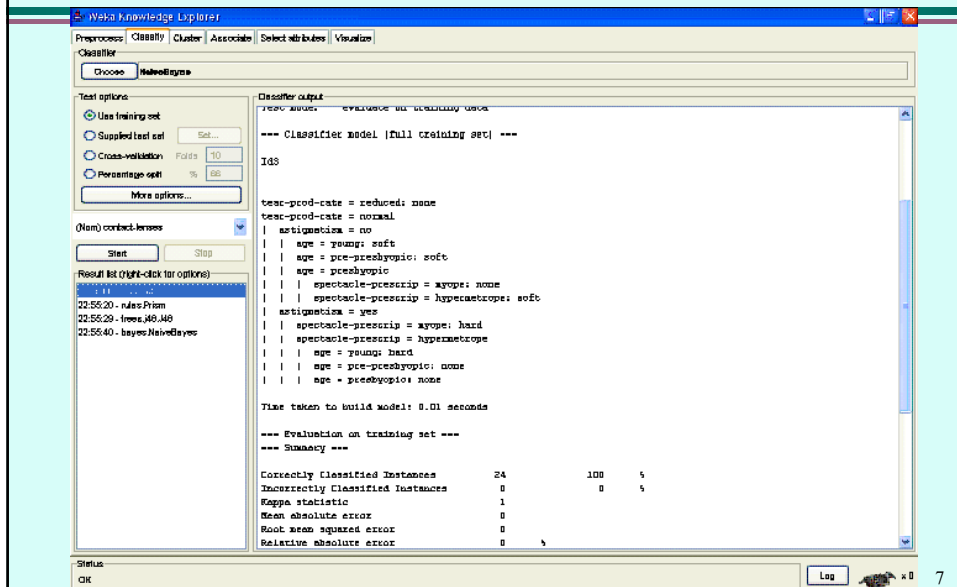
5

Podstawowe metody klasyfikacyjne

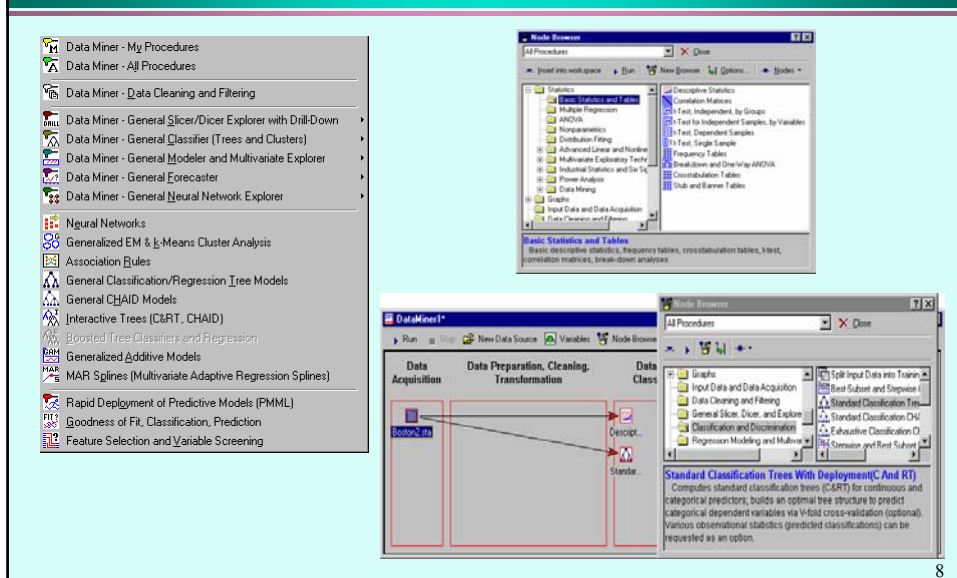
- metody symboliczne (drzewa i reguły decyzyjne),
- metody oparte na logice matematycznej (ILP),
- sztuczne sieci neuronowe,
- metody k-najbliższych sąsiadów,
- klasyfikacja bayesowska (Naive Bayes),
- analiza dyskryminacyjna (statystyczna),
- metody wektorów wspierających,
- regresja logistyczna,
- klasyfikatory genetyczne.
- ...

6

WEKA – „Classifier panel”



Data Miner (Statistica Statsoft) – przykład metod dostępnych w systemie



SAS – przykładowe algorytmy

- Przykładowe algorytmy eksploracji danych (dostępne w tzw. węzłach SAS Enterprise Miner)
 - wiele metod statystyki opisowej,
 - metody przekształceń danych (przeskalowania, uwzględnianie nieznanych wartości, wykrywanie nietypowych obserwacji),
 - modele predykcyjne (liniowa, nieliniowa, logistyczna regresja, drzewa regresji)
 - drzewa klasyfikacyjne (CART, CHAID, C4.5like)
 - sztuczne sieci neuronowe (liniowe/nieliniowe sieci wielowarstwowe, różne wersje RBF).
 - modele złożonych klasyfikatorów (bagging, boosting, combiners,...)
 - modele k-NN
- Oferuje przetwarzanie danych za pomocą specjalnego języka oraz interfejsy graficznego

9

Poszukiwanie i ocena wiedzy klasyfikacyjnej

- **Perspektywy odkrywania wiedzy**
 - **Predykcja** – przewidywanie przydziału nowych obiektów do klas / reprezentacja wiedzy wykorzystywana jako tzw. **klasyfikator** (ocena zdolności klasyfikacyjnej – na ogół jedno wybrane kryterium).
 - **Opis klasyfikacji obiektów** – wyszukiwanie wzorców charakteryzujących właściwości danych i prezentacja ich użytkownikowi w zrozumiałej formie (ocena wielokryterialna i bardziej subiektywna).

10

Tworzenie i ocena klasyfikatorów

Jest procesem **trzyetapowym**:

1. Konstrukcja modelu w oparciu o zbiór danych wejściowych (przykłady uczące).

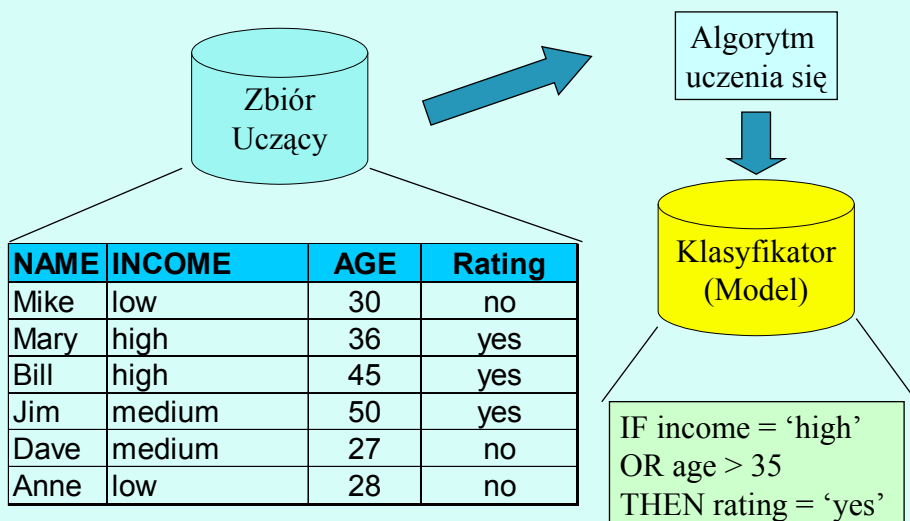
Przykładowe modele - klasyfikatory:

- drzewa decyzyjne,
- reguły (IF .. THEN ..),
- sieci neuronowe.

2. Ocena modelu (przykłady testujące)
3. Użycie modelu (np. klasyfikowanie nowych faktów lub interpretacja regularności)

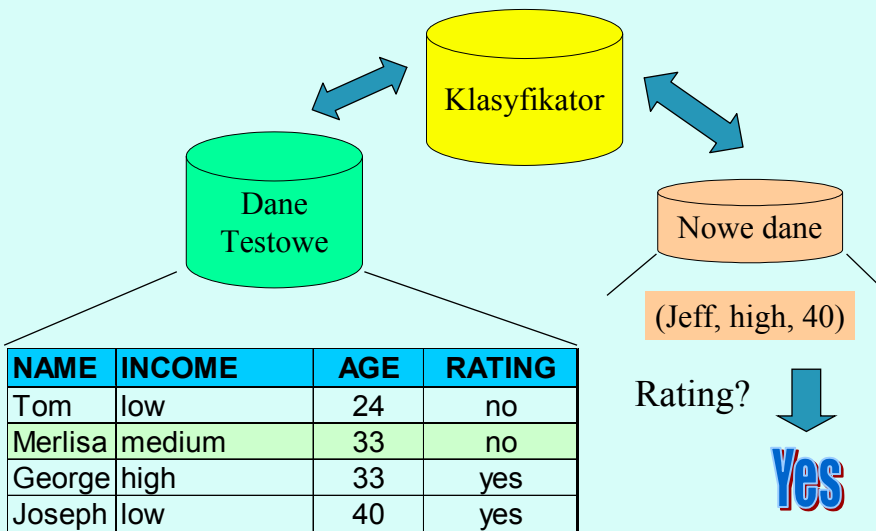
11

Proces Klasyfikowania (I) – Uczenie się



12

Proces Klasyfikowania (II)



13

Kryteria oceny metod klasyfikacyjnych

- Trafność klasyfikacji (Classification/Predictive accuracy)
- Szybkość i skalowalność:
 - czas uczenia się,
 - szybkość samego klasyfikowania
- Odporność (Robustness)
 - szum (noise),
 - missing values,
- Zdolności wyjaśniania: np., drzewa decyzyjne vs. sieci neuronowe
- Złożoność struktury, np.,
 - rozmiar drzew decyzyjnego,
 - miary oceny reguły

14

Trafność klasyfikowania

- Użyj przykładów testowych nie wykorzystanych w fazie indukcji klasyfikatora:
 - N_t – liczba przykładów testowych
 - N_c – liczba poprawnie sklasyfikowanych przykładów testowych
- Trafność klasyfikowania:

$$\eta = \frac{N_c}{N_t}$$

- Alternatywnie błąd klasyfikowania.

Inne możliwości analizy:

- macierz pomyłek (ang. confusion matrix),
- koszty pomyłek i klasyfikacja binarna,
- miary Sensitivity i Specificity / krzywa ROC

15

Trafność klasyfikowania: techniki estymacji

- **Techniki podziału: „Training-and-testing”**
 - Użyj dwóch niezależnych zbiorów., uczącego (2/3), testowego (1/3)
 - Jednokrotny podział losowy stosuje się dla dużych zbiorów
- **„Cross-validation” - Ocena krzyżowa**
 - Podziel losowo dane w k podzbiorów (równomierne lub warstwowe)
 - Użyj $k-1$ podzbiorów jako części uczącej i pozostałej jako testującej (k -fold cross-validation).
 - Oblicz wynik średni.
 - Stosowane dla danych o średnich rozmiarach (najczęściej $k = 10$)
Uwaga opcja losowania warstwowego (ang. stratified sampling).
- **Bootstrapping i leaving-one-out**
 - Dla małych rozmiarów danych.
 - „Leaving-one-out” jest szczególnym przypadkiem, dla którego liczba iteracji jest równa liczbie przykładów

16

Macierz pomyłek

- Analiza pomyłek w przydziale do różnych klas przy pomocy tzw. macierz pomyłek (ang. *confusion matrix*)
- Macierz $r \times r$, gdzie wiersze odpowiadają poprawnym klasom decyzyjnym, a kolumny decyzjom przewidywanym przez klasyfikator; na przecięciu wiersza i oraz kolumny j - liczba przykładów n_{ij} należących oryginalnie do klasy i -tej, a zaliczonej do klasy j -tej

Przykład:

Oryginalne klasy	Przewidywane klasy decyzyjne		
	K_1	K_2	K_3
K_1	50	0	0
K_2	0	48	2
K_3	0	4	46

17

Klasyfikacja binarna

- Niektóre zastosowania → jedna z klas posiada szczególne znaczenie, np. diagnozowanie poważnej choroby. Problem → klasyfikacja binarna.

Oryginalne klasy	Przewidywane klasy decyzyjne	
	Pozytywna	Negatywna
Pozytywna	TP	FN
Negatywna	FP	TN

- Nazewnictwo (inspirowane medycznie):
 - TP (ang. *true positive*) – liczba poprawnie sklasyfikowanych przykładów z wybranej klasy (ang. *hit*),
 - FN (ang. *false negative*) – liczba błędnie sklasyfikowanych przykładów z tej klasy, tj. decyzja negatywna podczas gdy przykład w rzeczywistości jest pozytywny (błąd pominięcia - z ang. *miss*),
 - TN (ang. *true negative*) – liczba przykładów poprawnie nie przydzielonych do wybranej klasy (poprawnie odrzuconych z ang. *correct rejection*),
 - FP (ang. *false positive*) – liczba przykładów błędnie przydzielonych do wybranej klasy, podczas gdy w rzeczywistości do niej nie należą (ang. *false alarm*).

18

Miary stosowane w analizie klasyfikacji binarnej

- Dodatkowe miary oceny rozpoznawania wybranej klasy:
 - **wrażliwość** (ang. *sensitivity*) = $TP / (TP+FN)$,
 - **specyficzność** (ang. *specificity*) = $TN / (FP+TN)$.
- Inne miary:
 - **False-positive rate** = $FP / (FP+TN)$, czyli **1 – specyficzność**.
- Wnikliwszą analizę działania klasyfikatorów binarnych dokonuje się w oparciu o analizę krzywej ROC, ang. *Receiver Operating Characteristic*.

Oryginalne klasy	Przewidywane klasy decyzyjne	
	Pozytywna	Negatywna
Pozytywna	<i>TP</i>	<i>FN</i>
Negatywna	<i>FP</i>	<i>TN</i>

19

Analiza macierzy... spróbuj rozwiązać...

$$Sensitivity = \frac{TP}{TP+FN} = ?$$

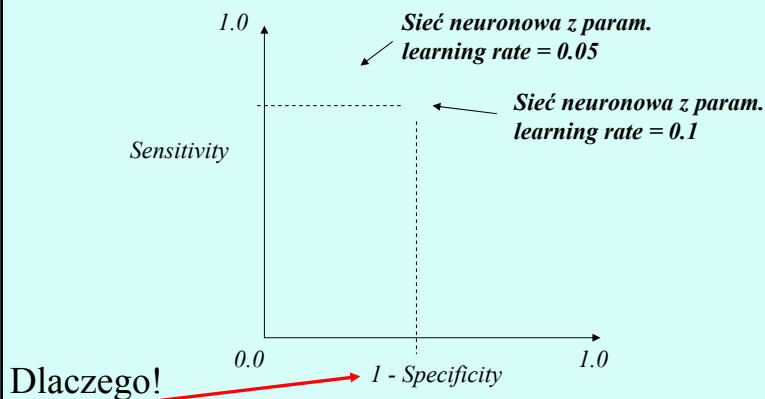
$$Specificity = \frac{TN}{TN+FP} = ?$$

		Co przewidywano		
		1	0	
Rzeczywista Klasa	1	60	30	60+30 = 90 przykładów w danych należało do Klasy 1
	0	80	20	80+20 = 100 przykładów było w Klasy 0
90+100 = 190 łączna liczba przykładów				

20

Analiza krzywej ROC

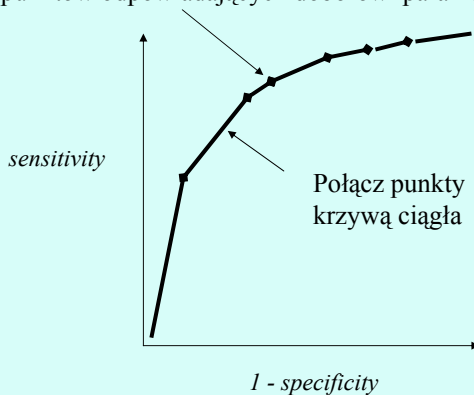
Każda technika budowy klasyfikatora może być scharakteryzowana poprzez pewne wartości miar 'sensitivity' i 'specificity'. Graficznie można je przedstawić na wykresie 'sensitivity' vs. $1 - \text{'specificity'}$.



21

ROC - analiza

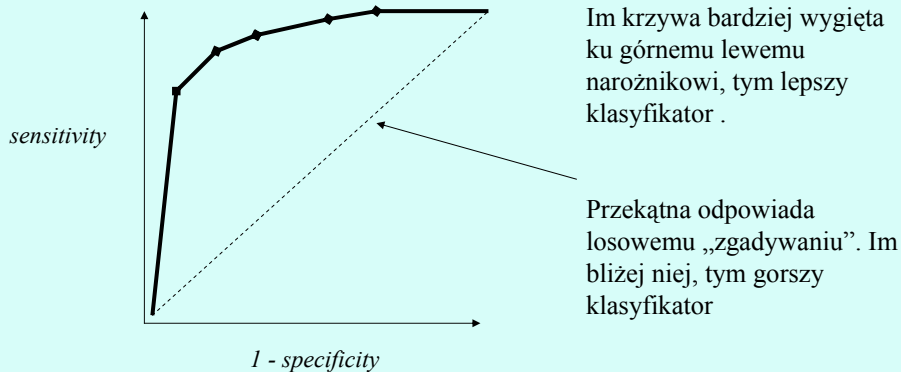
Algorytm może być parametryzowany, i w rezultacie otrzymuje się serie punktów odpowiadających doborowi parametrów



Wykres nazywany
'krzywą' ROC.

22

Krzywa ROC



Można porównywać działanie kilku klasyfikatorów.
Miary oceny np. AUC – pole pod krzywą,...

23

Porównywanie klasyfikatorów

- Jak oceniać skuteczność klasyfikacyjną dwóch różnych klasyfikatorów na tych samych danych?
- Ograniczamy zainteresowanie wyłącznie do trafności klasyfikacyjnej – oszacowanie techniką 10-krotnej oceny krzyżowej (ang. *k-fold cross validation*).
- Zastosowano dwa różne algorytmy uczące *AL1* i *AL2* do tego samego zbioru przykładów, otrzymując dwa różne klasyfikatory *KL1* i *KL2*. Oszacowanie ich trafności klasyfikacyjnej (10-fcv):
 - klasyfikator *KL1* → 86,98%
 - klasyfikator *KL2* → 87,43%.
- Czy uzasadnione jest stwierdzenie, że klasyfikator *KL2* jest skuteczniejszy niż klasyfikator *KL1*?

24

Analiza wyniku oszacowania trafności klasyfikowania

Podział	KI_1	KI_2
1	87,45	88,4
2	86,5	88,1
3	86,4	87,2
4	86,8	86
5	87,8	87,6
6	86,6	86,4
7	87,3	87
8	87,2	87,4
9	88	89
10	85,8	87,2
Srednia	86,98	87,43
Odchylenie	0,65	0,85

- Test statystyczny (t-Studenta dla par zmiennych/zależnych)
- $H_0 : ?$
- $t_{\text{emp}} = 1,733$ ($p = 0,117$) ???

25

Porównanie działania dwóch klasyfikatorów DT oraz n^2 na wielu zbiorach danych (wyniki średnie z 10-oceny krzyżowej wraz z przedziałem ufności)

Data set	Classification accuracy DT (%)	Classification accuracy n^2 (%)	Improvement n^2 vs. DT (%)
Automobile	85.5 ± 1.9	87.0 ± 1.9	1.5*
Cooc	54.0 ± 2.0	59.0 ± 1.7	5.0
Ecoli	79.7 ± 0.8	81.0 ± 1.7	1.3
Glass	70.7 ± 2.1	74.0 ± 1.1	3.3
Hist	71.3 ± 2.3	73.0 ± 1.8	1.7
Meta-data	47.2 ± 1.4	49.8 ± 1.4	2.6
Primary Tumor	40.2 ± 1.5	45.1 ± 1.2	4.9
Soybean-large	91.9 ± 0.7	92.4 ± 0.5	0.5*
Vowel	81.1 ± 1.1	83.7 ± 0.5	2.6
Yeast	49.1 ± 2.1	52.8 ± 1.8	3.7

26

Perspektywa opisowa

- Trudniejsza niż ocena zdolności klasyfikacyjnych.
- Rozważmy przykład reguł:
 - Klasyfikacyjne (decyzyjne).
 - Asocjacyjne.
- Pojedyncza reguła oceniana jako potencjalny reprezentant „interesującego” wzorca z danych
 - W literaturze propozycje tzw. ilościowych miar oceny reguł oraz sposoby definiowania „interesujących” reguł, także na podstawie wymagań podawanych przez użytkownika.

27

Opisowe miary oceny reguł

- Miary dla reguły r (jeżeli P to Q) definiowane na podstawie zbioru przykładów U , z którego została wygenerowana.
- Tablica kontyngencji dla reguły *jeżeli* P *to* Q :

	Q	$\neg Q$	
P	n_{PQ}	$n_{P\neg Q}$	n_P
$\neg P$	$n_{\neg P Q}$	$n_{\neg P \neg Q}$	$n_{\neg P}$
	n_Q	$n_{\neg Q}$	n

- Przegląd różnych miar, np.: Yao Y.Y, Zhong N.: An analysis of quantitative measures associated with rules, w: Proc. of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining, LNAI 1574, Springer, 1999, s. 479-488.
- Także rozprawa habilitacyjna J.Stefanowski: Algorytmy indukcji reguł w odkrywaniu wiedzy.

28

Popularne miary oceny reguł

- **Wsparcie reguły** (ang. *support*) zdefiniowane jako:

$$G(P \wedge Q) = \frac{n_{PQ}}{n}$$

- **Dokładność** (ang. *accuracy*) / wiarygodność (ang. *confidence*) reguły (bezwzględne wsparcie konkluzji Q przez przesłankę P):

$$AS(Q | P) = \frac{n_{PQ}}{n_P}$$

- **Względne pokrycie** reguły zdefiniowane jako:

$$AS(P | Q) = \frac{n_{PQ}}{n_Q}$$

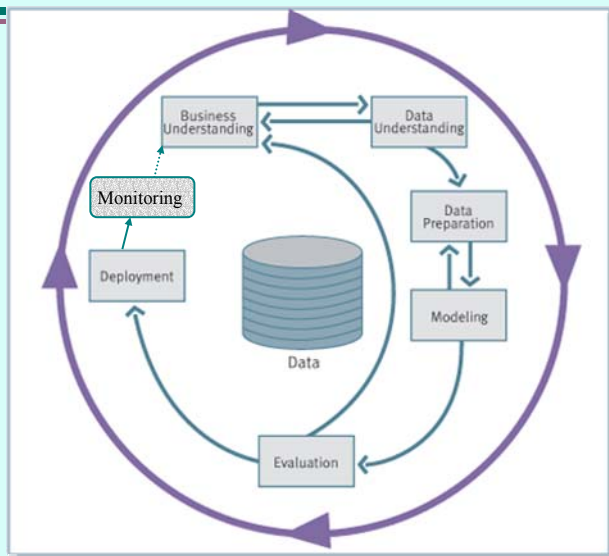
29

Proces odkrywania wiedzy

- Uruchomienie algorytmu budowy i oceny klasyfikatora to nie wszystko!
- Dane rzeczywiste na ogół nie są dostępne w formacie dogodnym dla aplikacji indukującej klasyfikator.
- W praktycznych zastosowaniach wiele wysiłku jest potrzebne dla pozyskania, przygotowania i przetwarzania wstępnego danych.
- Spojrzenie procesowe na KDD.

30

Knowledge Discovery Process opis procesu zgodnie CRISP-DM

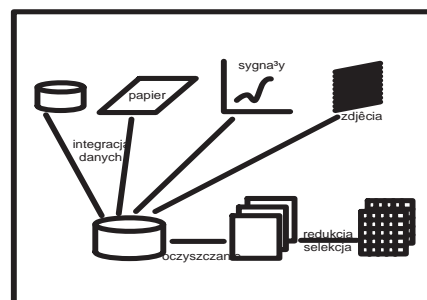


Spójrz na
www.crisp-dm.org
aby dowiedzieć
się więcej.

31

Pozyskiwanie i integracja danych

- Pozyskiwanie danych z różnych źródeł.
 - Współpraca z ekspertami oraz stań się sam „ekspertem”.
 - Dane są dostępne w różnych formatach
- Trudności w integracji danych.
 - Potrzeba wiedzy dziedzinowej oraz pomocy eksperta.
- Oczyszczanie danych z różnego rodzaju niedoskonałości oraz błędów.
 - Kontroluj i sprawdzaj.



Kilka pytań.

- Jakie źródła danych są związane z zadaniem / zastosowaniem?
- Które z dostępnych danych są adekwatne do celów zastosowania (data relevant)?
- Czy mamy dostęp do innych źródeł danych?
- Jakiej wielkości są dane historyczne (obiekty i atrybuty)?
- Kto dobrze zna posiadane dane (who is data expert)?

33

Problem „czyszczenie” i uspoźnienia formatów

- Problem przywrócenia integralności dziedzin atrybutów.
- Przykłady:
 - Numery kont bankowych lub numery telefonów mogą mieć w jednym systemie typ „String”, a w innych typ „Numeric”.
 - Płeć zapisywana na różne sposoby (pełna nazwa, skróty, kody,...)
 - Daty reprezentowane w różnych formatach („ddmmyy”, „yyymmdd”, „yyyy-mm-dd”,...)
 - np. “Sep 24, 2003” , 9/24/03, 24.09.03, itp. → zmień do wspólnego formatu.
 - Pola przechowujące walutę.
 - Różne systemy używają różnych rozmiarów pól wartości tekstowych.
 - Pola tekstowe ukrywają ważne dodatkowe informacje.

34

Źródła trudności

- Bardzo duża liczba obserwacji / przykładów.
- Zbyt duża liczba atrybutów.
- Nieistotność części atrybutów dla klasyfikacji obiektów.
- Wzajemna współzależności atrybutów warunkowych.
- Równoczesna obecności atrybutów różnego typu.
- Występowanie niezdefiniowanych wartości atrybutów.

35

O rozmiarach, ...

- **Liczba przykładów/obserwacji (records)**
 - *Heurystyczna zasada: tysiące obserwacji lub więcej*
 - Mniejsza liczba wymaga specjalistycznych metod statystycznego wnioskowania.
- **Liczba atrybutów (fields)**
 - *Heurystyki: Dla każdego atrybutu, 10 lub więcej przykładów; Liczba przykładów minimum o rząd większa niż atrybutów.*
 - Jeśli za dużo atrybutów, to użyj właściwych metod redukcji rozmiarów (tzw. feature selection).
- **Liczba klas / pojęć decyzyjnych (target concepts)**
 - *Heurystyczna zasada : >100 przykładów na klasę*
 - Niezrównoważone klasy!
 - Specjalne podejścia (ang. imbalanced data) , ocena stosuj metody losowania warstwowego.

36

Proste podejścia do redukcji atrybutów

- Po pierwsze: usuń kolumny z małą zmiennością wartości atrybutów.
- Sprawdź liczbę różnych wartości atrybutu:
 - *Usuń jeśli wszystkie pola zawierają tę samą wartość (e.g. null), poza małą liczbą rekordów (minp % or less of all records).*
 - *minp 0.5% lub ogólniej mniej niż 5% różnych przykładów w najmniejszej klasie*
- Bardziej wyrafinowane podejścia w statystycznej analizie danych lub data mining (ML) jako selekcja atrybutów.
 - WEKA spójrz na zakładkę „attribute selection”.

37

Redukcja rozmiarów danych – Selekcja atrybutów

- Dany jest n elementowy zbiór przykładów (obiektów). Każdy przykład x jest zdefiniowany na $V_1 V_2 V_m$ gdzie V_i jest dziedziną i -tego atrybutu. W przypadku nadzorowanej klasyfikacji przykłady zdefiniowane są jako $\langle x, y \rangle$ gdzie y określa pożądaną odpowiedź, np. klasyfikację przykładu.
- **Cel selekcji atrybutów:**
 - *Wybierz minimalny podzbiór atrybutów, dla którego rozkład prawdopodobieństwa różnych klas obiektów jest jak najbliższy oryginalnemu rozkładowi uzyskanemu z wykorzystaniem wszystkich atrybutów.*
- **Nadzorowana klasyfikacja**
 - *Dla danego algorytmu uczenia i zbioru uczącego, znajdź najmniejszy podzbiór atrybutów dla którego system klasyfikujący przewiduje przydział obiektów do klas decyzyjnych z jak największą trafnością.*

38

Selekcja w trakcie wstępnego przetwarzania danych

- Ocena pojedynczych atrybutów:
 - testy χ^2 i miary siły związku,
 - miary wykorzystujące względną entropię między atrybutem warunkowym a decyzyjnym (ang. *info gain*, *gain ratio*),
 - ...
- Ocena podzbiorów atrybutów (powinny być niezależne wzajemnie a silnie zależne z klasyfikacją):
 - Miara korelacji wzajemnych,
 - Statystyki λ Wilksa, T2-Hotellinga, odległości D2 Mahalanobisa,
 - Redukty w teorii zbiorów przybliżonych,
 - Techniki dekompozycji na podzbiory (ang. *data table templates*)
 - ...
- Model „filter” vs. „wrapper”

39

Nieznane wartości atrybutów

Sposoby uwzględniania brakujących wartości:

- Stosowane w przetwarzaniu wstępnym (przekształć niekompletne dane w kompletne).
- Zintegrowane z algorytmami odkrywania wiedzy

Przetwarzanie wstępne:

- Podejście naiwne:
 - Zignorowanie przykładów opisanych nieznanymi wartościami.
- **Zastępowanie** brakujących wartości poprzez:
 - Użycie globalnej stałej wartości.
 - Zastąpienie najczęściej występującą wartością atrybutu nominalnego.
 - Zastąpienie wartością średnią atrybutu liczbowego.
 - Użycie najczęstszej lub średniej wartości atrybutu znajdowanej na podstawie rozkładu wartości wśród przykładów należących *tylko* do tej samej *klasy decyzyjnej* co analizowany przykład.
 - Użycie zbioru wszystkich możliwych wartości tego atrybutu.
 - Użycie podzbioru wartości atrybutu wraz z informacją o stopniach możliwości ich realizacji.
 - Wykonanie analizy zależności wartości atrybutu od atrybutów w pełni zdefiniowanych (regresja, drzewa i reguły decyzyjne).

40

Transformacje danych: Dyskretyzacja

- Niektóre metody wymagają danych dyskretnych, np. Naïve Bayes, zbiory przybliżone, reguły asocjacyjne, wzorce sekwencji.
- Ponadto przydatne do podsumowania danych i redukcji rozmiarów.
- Dyskretyzacja jest:
 - procesem zamiany atrybutów liczbowych na atrybuty symboliczne typu porządkowego. Polega ona podziale oryginalnej dziedziny atrybutu liczbowego na pewną liczbę przedziałów i przypisaniu tym przedziałom kodów symbolicznych.
- Wiele różnorodnych podejść:
 - Nadzorowana vs. nienadzorowana,
 - Globalna vs. lokalna (z punktu widzenia atrybutów),
 - Dynamiczna vs. Statyczna (dobór parametrów).

41

Przykładowe popularne metody

- Podział równymi przedziałami (*equal-width interval*)
 - Podziel zakres przedziału atrybutu na N podprzedziałów równej długości.
- Podział przedziałami o równej częstości (*equal-frequency interval*);
 - Podprzedziały zawierają w przybliżeniu taką samą liczbę obserwacji.
- *ChiMerge* –zachowuje podobieństwo względnych częstości klas decyzyjnych w podprzedziałach.
- Minimalizacja entropii warunkowej klas decyzyjnych (*Class Entropy discretization*);
 - Wersja lokalna, wersja wykorzystująca zasadę MDL, wersja globalizowana.
- Modyfikacje algorytmów analizy skupień (aglomeracyjne z warunkiem zatrzymania)

42

No i na razie wystarczy

- ➡ **O innych zagadnieniach procesu odkrywania wiedzy jeszcze porozmawiamy!**