Data mining

Nguyen Hung Son

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo

Algorytm
schemat
wybór testu
prunning
null-values

Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

Wykład 6: Drzewa decyzyjne

Nguyen Hung Son

Outline

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo
Algorytm
schemat
wybór testu
prunning
null-values
Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

1. Wprowadzenie
   - Definicje
   - Funkcje testu
   - Optymalne drzewo
2. Konstrukcja drzew decyzyjnych
   - Ogólny schemat
   - Kryterium wyboru testu
   - Przycinanie drzew
   - Problem brakujących wartości
3. Soft cuts and soft Decision tree
   - Soft Decision Tree
   - Searching for soft cuts
   - Discernibility measure:

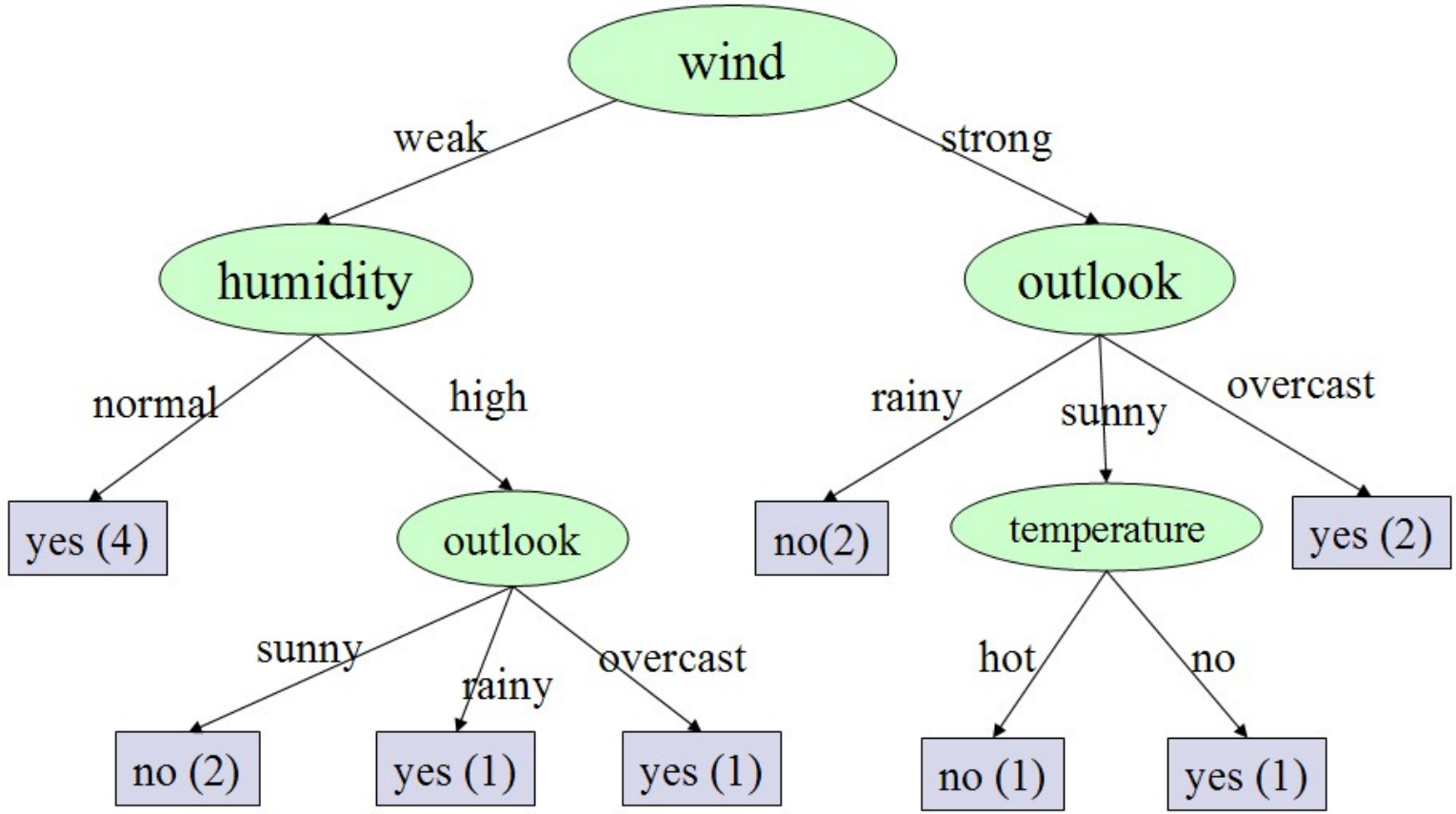
Co to jest drzewo decyzyjne

- **Jest to struktura drzewiasta, w której**
  - **węzły wewnętrzne** zawierają testy na wartościach atrybutów
  - z każdego węzła wewnętrznego wychodzi tyle **gałęzi**, ile jest możliwych wyników testu w tym węzle;
  - **liście** zawierają decyzje o klasyfikacji obiektów
- **Drzewo decyzyjne koduje program zawierający same instrukcje warunkowe**

Przykład tablicy decyzyjnej

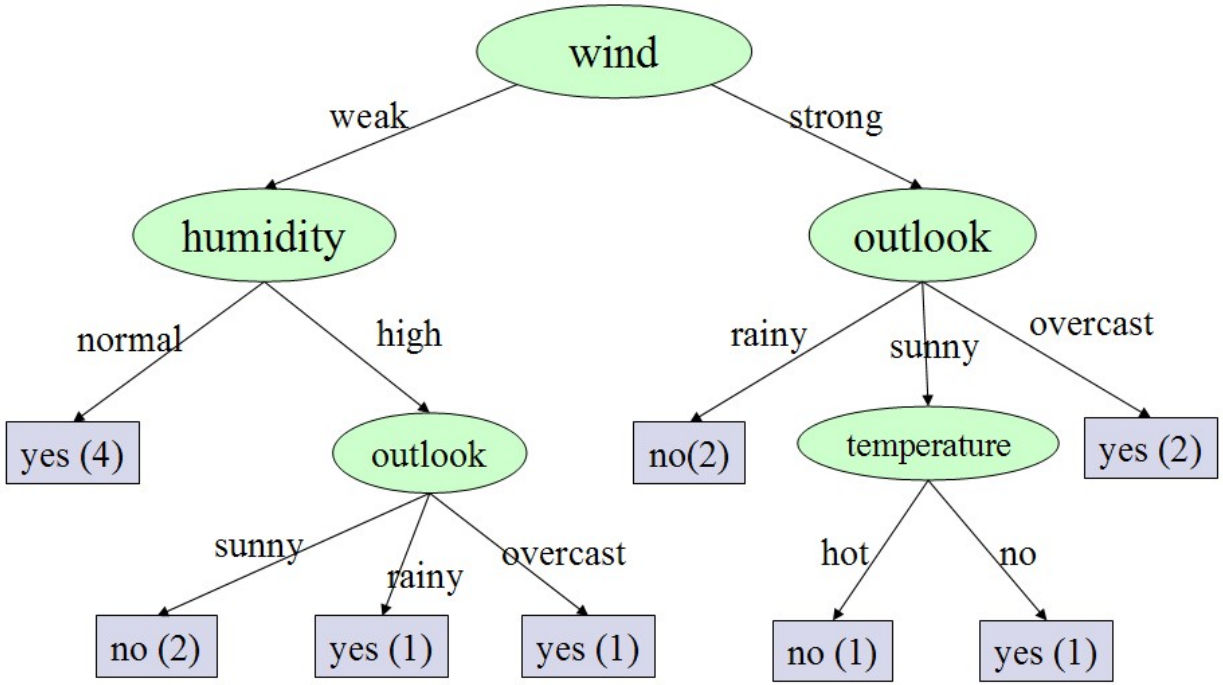
x	outlook	Temperature	humidity	wind	play(x)
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cold	normal	weak	yes
6	rain	cold	normal	strong	no
7	overcast	cold	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cold	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rain	mild	high	strong	no

Przykład drzewa decyzyjnego



Klasyfikacja drzewem decyzyjnym

x	outlook	Temperature	humidity	wind	play(x)
15	rainy	hot	high	weak	???



dec(15) = yes

Outline

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo
Algorytm
schemat
wybór testu
prunning
null-values
Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

1. Wprowadzenie
   - Definicje
   - Funkcje testu
   - Optymalne drzewo
2. Konstrukcja drzew decyzyjnych
   - Ogólny schemat
   - Kryterium wyboru testu
   - Przycinanie drzew
   - Problem brakujących wartości
3. Soft cuts and soft Decision tree
   - Soft Decision Tree
   - Searching for soft cuts
   - Discernibility measure:

Rodzaje testów

Data mining

Nguyen Hung Son

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo

Algorytm
schemat
wybór testu
prunning
null-values

Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

Wyróżniamy 2 klasy funkcji testów

- Testy operują się na wartościach pojedyńczego atrybutu (univariate tree):

$$t : V_a \rightarrow R_t$$

- Testy będące kombinacją wartości kilku atrybutów (multivariate tree).

$$t : V_{a_1} \times V_{a_2} \times ... \times V_{a_k} \rightarrow R_t$$

gdzie

- Va : dziedzina atrybutu a
- Rt : zbiór możliwych wyników testu

Przykłady funkcji testu

Data mining

Nguyen Hung Son

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo

Algorytm
schemat
wybór testu
prunning
null-values

Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

- Dla atrybutów nominalnych a_i oraz obiekt x:
  - test tożsamościowy: $t(x) \rightarrow a_i(x)$
  - test równościowy: $t(x) = \begin{cases} 1 & \text{if } (a_i(x) = v) \\ 0 & \text{otherwise} \end{cases}$
  - test przynależnościowy: $t(x) = \begin{cases} 1 & \text{if } (a_i(x) \in V) \\ 0 & \text{otherwise} \end{cases}$
- Dla atrybutów o wartościach ciągłych:
  - test nierównościowy:
    $t(x) = \begin{cases} 1 & \text{if } (a_i(x) > c) \\ 0 & \text{otherwise, i.e., } (a_i(x) \leq c) \end{cases}$ gdzie c jest wartością progową lub cięciem

Outline

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo
Algorytm
schemat
wybór testu
prunning
null-values
Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

1. Wprowadzenie
   - Definicje
   - Funkcje testu
   - Optymalne drzewo
2. Konstrukcja drzew decyzyjnych
   - Ogólny schemat
   - Kryterium wyboru testu
   - Przycinanie drzew
   - Problem brakujących wartości
3. Soft cuts and soft Decision tree
   - Soft Decision Tree
   - Searching for soft cuts
   - Discernibility measure:

Ocena jakości drzewa

Data mining

Nguyen Hung Son

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo

Algorytm
schemat
wybór testu
prunning
null-values

Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

- Jakość drzewa ocenia się
  - rozmiarem: im drzewo jest mniejsze, tym lepsze
    - mała liczba węzłów,
    - mała wysokość, lub
    - mała liczba liści;
  - dokładnością klasyfikacji na zbiorze treningowym
  - dokładnością klasyfikacji na zbiorze testowym
- Na przykład:

$$Q(T) = \alpha \cdot size(T) + \beta \cdot accuracy(T, P)$$

gdzie α, β są liczbami rzeczywistymi
size(.) jest rozmiarem drzewa
accuracy(.,.) jest jakością klasyfikacji

Problem konstrukcji drzew optymalnych:

Dane są:

- tablica decyzyjna $\mathbb{S}$
- zbiór funkcji testów $\mathbf{TEST}$,
- kryterium jakości Q

Szukane: drzewo decyzyjne $\mathbf{T}$ o najwyższej jakości $Q(\mathbf{T})$.

- Dla większości parametrów, problem szukania optymalnego drzewa jest NP-trudny !
- **Wnioski:**
  Trudno znaleźć optymalne drzewo w czasie wielomianowym;
  Konieczność projektowania heurystyk.
- **Quiz:** Czy drzewo z przykładu jest optymalne?

Optymalne drzewo decyzyjne

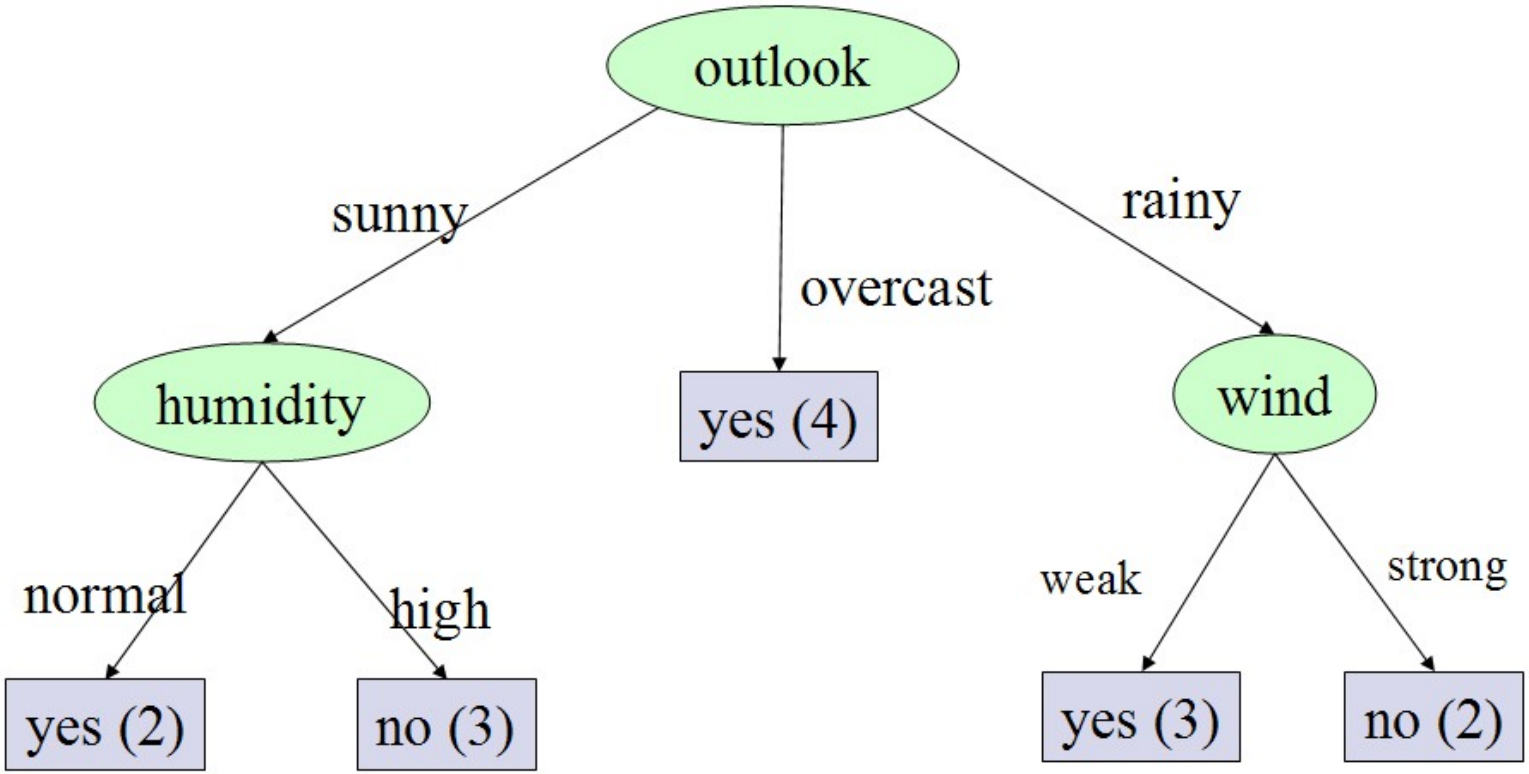
Data mining

Nguyen Hung Son

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo

Algorytm
schemat
wybór testu
prunning
null-values

Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:



Outline

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo
Algorytm
schemat
wybór testu
prunning
null-values
Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

1. Wprowadzenie
   - Definicje
   - Funkcje testu
   - Optymalne drzewo
2. Konstrukcja drzew decyzyjnych
   - Ogólny schemat
   - Kryterium wyboru testu
   - Przycinanie drzew
   - Problem brakujących wartości
3. Soft cuts and soft Decision tree
   - Soft Decision Tree
   - Searching for soft cuts
   - Discernibility measure:

Funkcja rekurencyjna $buduj\_drzewo(U, dec, \mathbf{T})$:

1: **if** $(kryterium\_stopu(U, dec) = \mathbf{true})$ **then**
2: $\quad \mathbf{T}.etykieta = kategoria(U, dec)$;
3: $\quad$ **return**;
4: **end if**
5: $t := wybierz\_test(U, \mathbf{TEST})$;
6: $\mathbf{T}.test := t$;
7: **for** $v \in R_t$ **do**
8: $\quad U_v := \{x \in U : t(x) = v\}$;
9: $\quad$ utwórz nowe poddrzewo $\mathbf{T}'$;
10: $\quad \mathbf{T}.$gałąź$(v) = \mathbf{T}'$;
11: $\quad buduj\_drzewo(U_v, dec, \mathbf{T}')$
12: **end for**

Funkcje pomocnicze

- Kryterium stopu: Zatrzymamy konstrukcji drzewa, gdy aktualny zbiór obiektów:
  - jest pusty lub
  - zawiera obiekty wyłącznie jednej klasy decyzyjnej lub
  - nie ulega podziale przez żaden test
- Wyznaczenie etykiety zasadą większościową:

$$kategoria(P, dec) = \arg \max_{c \in V_{dec}} |P_{[dec=c]}|$$

  tzn., etykietą dla danego zbioru obiektów jest klasa decyzyjna najliczniej reprezentowana w tym zbiorze.
- Kryterium wyboru testu: heurytyczna funkcja oceniająca testy.

Outline

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo
Algorytm
schemat
wybór testu
prunning
null-values
Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

1. Wprowadzenie
   - Definicje
   - Funkcje testu
   - Optymalne drzewo
2. Konstrukcja drzew decyzyjnych
   - Ogólny schemat
   - Kryterium wyboru testu
   - Przycinanie drzew
   - Problem brakujących wartości
3. Soft cuts and soft Decision tree
   - Soft Decision Tree
   - Searching for soft cuts
   - Discernibility measure:

Miary różnorodności zbioru

Każdy zbiór obiektów X ulega podziale na klasy decyzyjne:

$$X = C_1 \cup C_2 \cup ... \cup C_d$$

gdzie $C_i = \{u \in X : dec(u) = i\}$.
Wektor $(p_1, ..., p_r)$, gdzie $p_i = \frac{|C_i|}{|X|}$, nazywamy rozkładem klas decyzyjnych w X.

$$Conflict(X) = \sum_{i<j} |C_i| \times |C_j| = \frac{1}{2}\left(|X|^2 - \sum |C_i|^2\right)$$

$$Entropy(X) = -\sum \frac{|C_i|}{|X|} \cdot \log \frac{|C_i|}{|X|}$$

$$= -\sum p_i \log p_i$$

Własności miar różnorodności

Funkcja $conflict(X)$ oraz $Ent(X)$ przyjmują

- największą wartość, gdy rozkład klas decyzyjnych w zbiorze X jest równomierny.
- najmniejszą wartość, gdy wszystkie obiekty w X są jednej kategorii (X jest **jednorodny**)

W przypadku 2 klas decyzyjnych:

$$Conflict(p, 1-p) = |X|^2 \cdot p(1-p)$$
$$Entropy(p, 1-p) = -p \log p - (1-p) \log(1-p)$$

Kryteria wyboru testu

Data mining

Nguyen Hung Son

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo

Algorytm
schemat
wybór testu
prunning
null-values

Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

Niech t definiuje podział X na podzbiory: $X_1 \cup ... \cup X_r$.
Możemy stosować następujące miary do oceniania testów:

- liczba par obiektów rozróżnionych przez test t.

$$disc(t, X) = conflict(X) - \sum conflict(X_i)$$

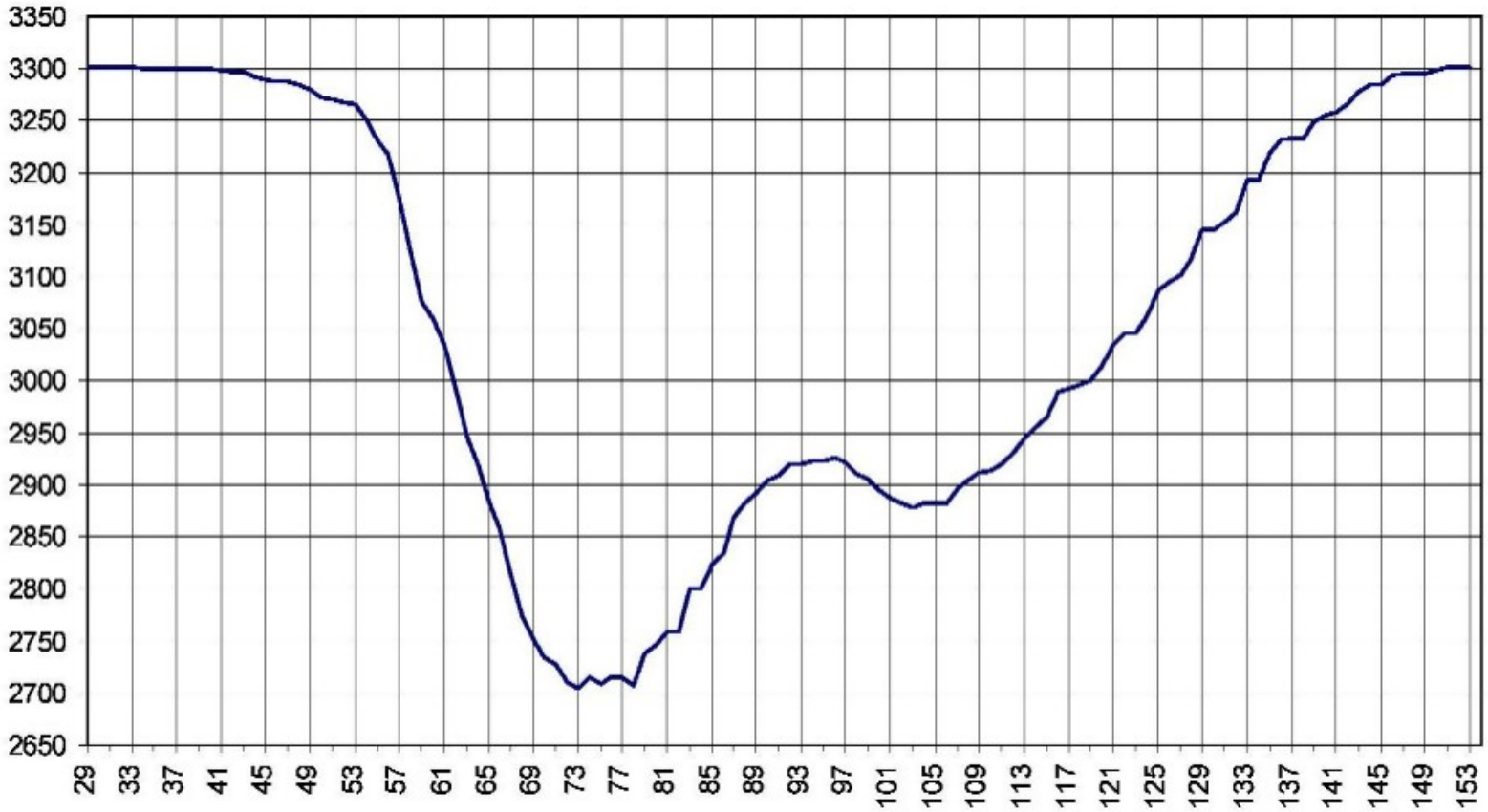
- kryterium przyrostu informacji (ang. Inf. gain).

$$Gain(t, X) = Entropy(X) - \sum_i p_i \cdot Entropy(X_i)$$

Im większe są wartości tych ocen, tym lepszy jest test.

Miara Entropii dla cięć

$$N \times \sum_i p_i \cdot Entropy(X_i)$$



Rozróżnialność dla cięć

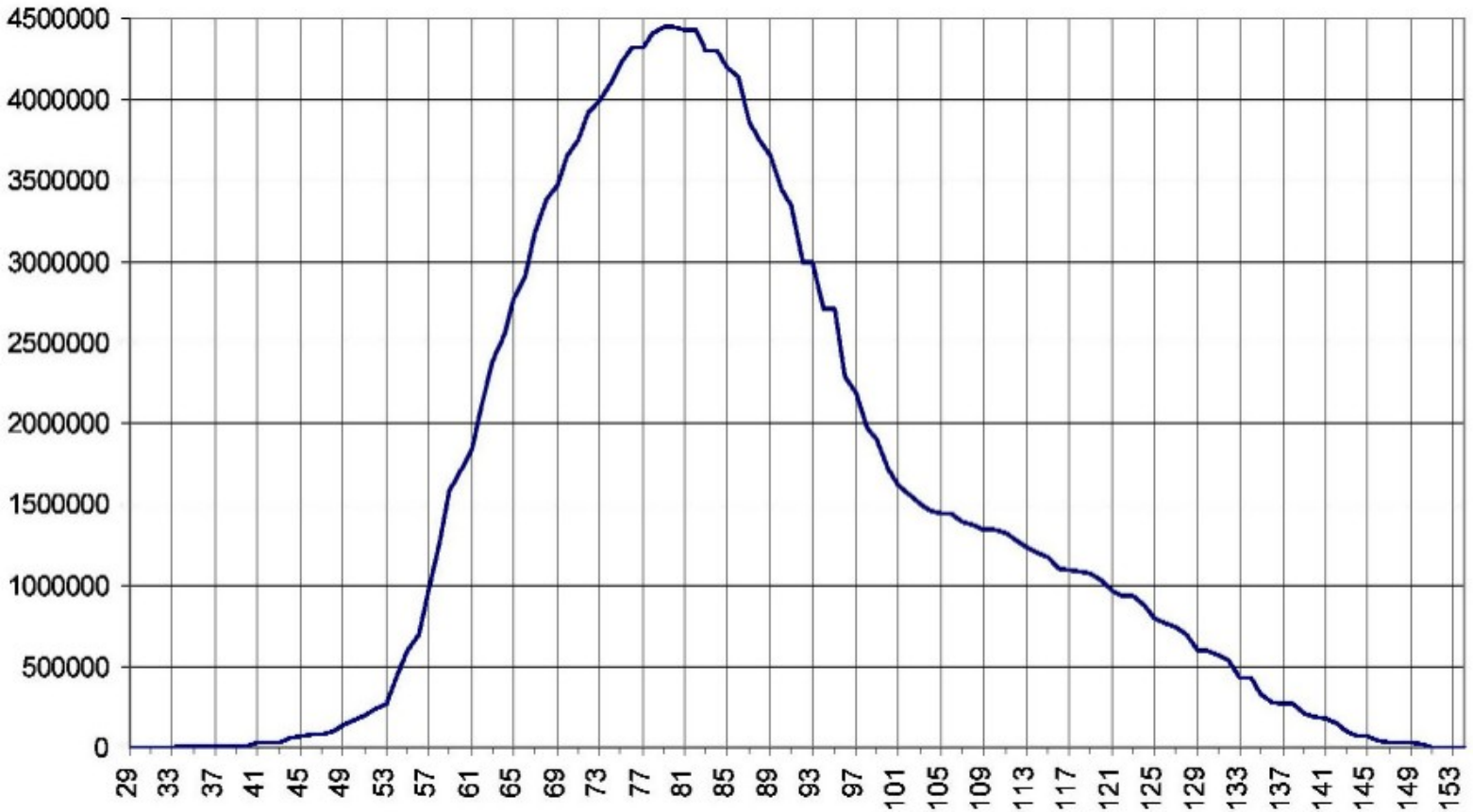
Data mining

Nguyen Hung Son

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo

Algorytm
schemat
wybór testu
prunning
null-values

Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:



Własności funkcji ocen:

- Monotoniczność: Jeśli t' definiuje drobniejszy podział niż t to

$$Gain(t', X) \geq Gain(t, X)$$

(analogiczną sytuację mamy dla miary $conflict()$.

- Funkcje ocen testu t przyjmują małe wartości jeśli rozkłady decyzyjne w podzbiorach wyznaczanych przez t są zbliżone.

Uniwersalne oceny

Zamiast bezwzględnego przyrostu informacji, stosujemy współczynnik przyrostu informacji

$$Gain\_ratio = \frac{Gain(t, X)}{iv(t, X)}$$

gdzie $iv(t, X)$, zwana wartością informacyjną testu t (information value), jest definiowana jak nast.:

$$iv(t, X) = -\sum_{i=1}^{r} \frac{|X_i|}{|X|} \cdot \log \frac{|X_i|}{|X|}$$

Ocena funkcji testu

- Rozróżnialność:

$$disc(t, X) = conflict(X) - \sum conflict(X_i)$$

- Przyrostu informacji (Information gain).

$$Gain(t, X) = Entropy(X) - \sum_i p_i \cdot Entropy(X_i)$$

- Współczynnik przyrostu informacji (gain ratio)

$$Gain\_ratio = \frac{Gain(t, X)}{-\sum_{i=1}^{r} \frac{|X_i|}{|X|} \cdot \log \frac{|X_i|}{|X|}}$$

- Inne (np. Gini's index, test χ^2, ...)

Outline

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo
Algorytm
schemat
wybór testu
prunning
null-values
Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

1. Wprowadzenie
   - Definicje
   - Funkcje testu
   - Optymalne drzewo
2. Konstrukcja drzew decyzyjnych
   - Ogólny schemat
   - Kryterium wyboru testu
   - Przycinanie drzew
   - Problem brakujących wartości
3. Soft cuts and soft Decision tree
   - Soft Decision Tree
   - Searching for soft cuts
   - Discernibility measure:

Przycinanie drzew

Data mining

Nguyen Hung Son

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo

Algorytm
schemat
wybór testu
prunning
null-values

Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

- Problem nadmiernego dopasowania do danych trenujących (prob. przeuczenia się).
- Rozwiązanie:
  - zasada krótkiego opisu: skracamy opis kosztem dokładności klasyfikacji w zbiorze treningowym
  - zastąpienie podrzewa nowym liściem (przycinanie) lub mniejszym podrzewem.
- Podstawowe pytania:

  Q: Kiedy poddrzewo może być zastąpione liściem?

  A: jeśli nowy liść jest niegorszy niż istniejące poddrzewo dla nowych obiektów (nienależących do zbioru treningowego).

  Q: Jak to sprawdzić?

  A: testujemy na próbce zwanej zbiorem przycinania!

Ogólny schemat algorytmu przycinania

Funkcja $przytnij(\mathbf{T}, P)$

1: **for all** $n \in \mathbf{T}$ **do**
2: utwórz nowy liść l etykietowany kategorią dominującą w zbiorze P_n
3: **if** (liść l jest niegorszy od poddrzewa o korzeniu w n pod względem zbioru P) **then**
4: zastąp poddrzewo o korzeniu w n liściem l;
5: **end if**
6: **end for**
7: return $\mathbf{T}$

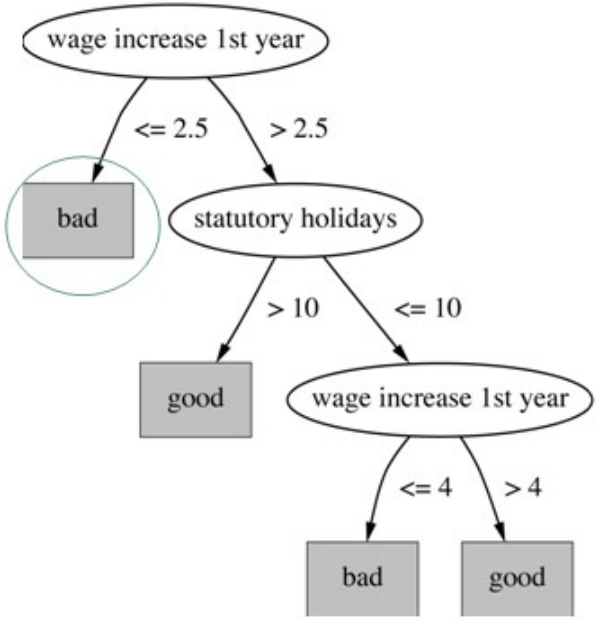
Kryterium przycinania

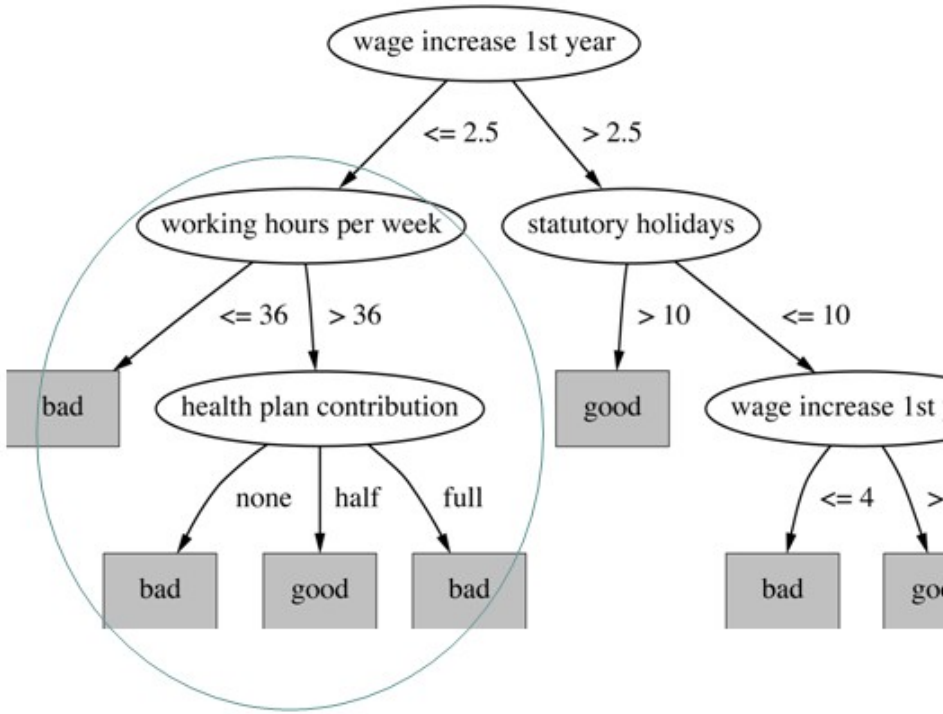
- Niech
  $e_T(l)$ - błąd klasyfikacji kandydującego liścia l,
  $e_T(n)$ - błąd klasyfikacji poddrzewa o korzeniu w n.
- Przycinanie ma miejsce, gdy

$$e_T(l) \leq e_T(n) + \mu \sqrt{\frac{e_T(n)(1 - e_T(n))}{|P_{T,n}|}}$$

na ogół przyjmujemy $\mu = 1$.

Przykład


wage increase 1st year
<= 2.5
> 2.5
bad
statutory holidays
> 10
<= 10
good
wage increase 1st year
<= 4
> 4
bad
good


wage increase 1st year
<= 2.5
> 2.5
working hours per week
statutory holidays
<= 36
> 36
> 10
<= 10
bad
health plan contribution
good
wage increase 1st
none
half
full
<= 4
bad
good
bad
bad
go

Outline

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo
Algorytm
schemat
wybór testu
prunning
null-values
Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

1. Wprowadzenie
   - Definicje
   - Funkcje testu
   - Optymalne drzewo
2. Konstrukcja drzew decyzyjnych
   - Ogólny schemat
   - Kryterium wyboru testu
   - Przycinanie drzew
   - Problem brakujących wartości
3. Soft cuts and soft Decision tree
   - Soft Decision Tree
   - Searching for soft cuts
   - Discernibility measure:

Brakuje danych podczas uczenia się

Możliwe są następujące rozwiązania:

- Zredukowanie wartości kryterium wyboru testu (np. przyrostu informacji) dla danego testu o współczynnik równy:

$$\frac{\text{liczba obiektów z nieznanymi wartościami}}{\text{liczba wszystkich obiektów}}$$

- Wypełnienie nieznanych wartości atrybutu najczęściej występującą wartością w zbiorze obiektów związanych z aktualnym węzłem
- Wypełnienie nieznanych wartości atrybutu średnią ważoną wyznaczoną na jego zbiorze wartości.

Brakuje danych podczas klasyfikowania

Data mining

Nguyen Hung Son

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo

Algorytm
schemat
wybór testu
prunning
null-values

Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

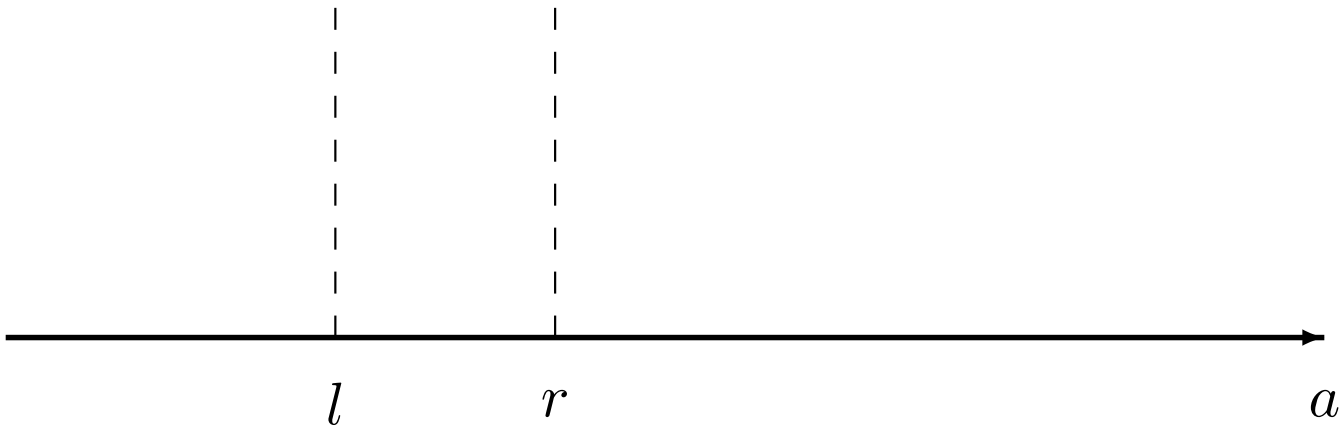
Możliwe rozwiązania:

- Zatrzymanie procesu klasyfikacji w aktualnym węźle i zwrócenie większościowej etykiety dla tego węzła (etykiety, jaką ma największą liczbę obiektów trenujących w tym węźle)
- Wypełnienie nieznanej wartości według jednej z heurystyk podanych wyżej dla przypadku konstruowania drzewa
- Uwzględnienie wszystkich gałęzi (wszystkich możliwych wyników testu) i połączenie odpowiednio zważonych probabilistycznie rezultatatów w rozkład prawdopodobieństwa na zbiorze możliwych klas decyzyjnych dla obiektu testowego.

Soft cuts

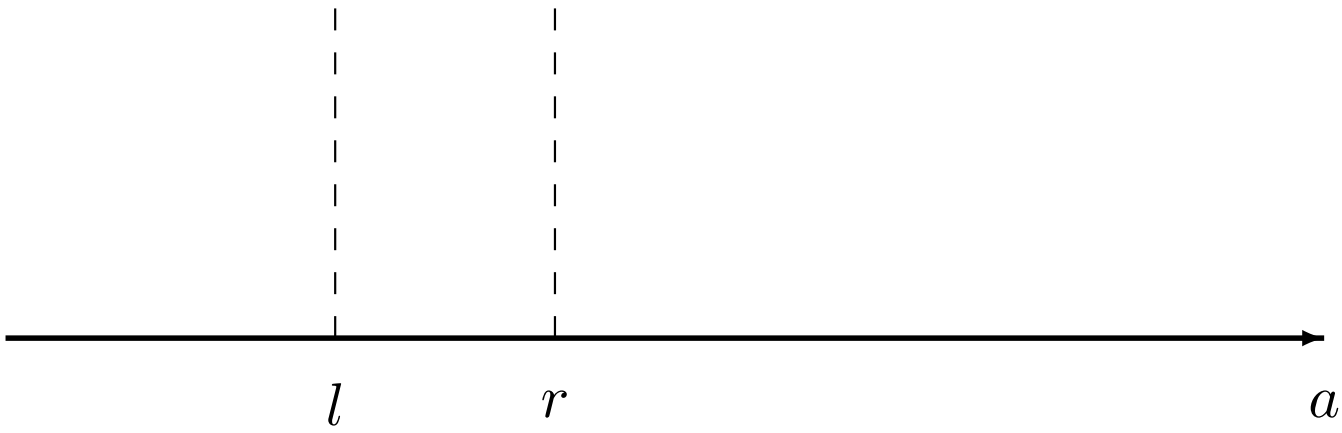
A soft cut is any triple $p = \langle a, l, r \rangle$, where

- $a \in A$ is an attribute,
- $l, r \in \Re$ are called the left and right bounds of p ;
- the value $\varepsilon = \frac{r-l}{2}$ is called the uncertain radius of p.
- We say that a soft cut p discerns a pair of objects x_1, x_2 if $a(x_1) < l$ and $a(x_2) > r$.



Some interpretations of $p = \langle a, l, r \rangle$:

- *there is a real cut somewhere between* l *and* r.
- *for any value* $v \in [l, r]$ *we are not able to check if* v *is either on the left side or on the right side of the real cut.*
- $[l, r]$ *is an uncertain interval of the soft cut* p.
- *normal cut can be treated as soft cut of radius* 0.



Outline

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo
Algorytm
schemat
wybór testu
prunning
null-values
Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

1. Wprowadzenie
   - Definicje
   - Funkcje testu
   - Optymalne drzewo
2. Konstrukcja drzew decyzyjnych
   - Ogólny schemat
   - Kryterium wyboru testu
   - Przycinanie drzew
   - Problem brakujących wartości
3. Soft cuts and soft Decision tree
   - Soft Decision Tree
   - Searching for soft cuts
   - Discernibility measure:

Soft Decision tree

Data mining

Nguyen Hung Son

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo

Algorytm
schemat
wybór testu
prunning
null-values

Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

- The test functions can be defined by soft cuts
- Here we propose two strategies using described above soft cuts:
  - *fuzzy decision tree*: any new object u can be classified as follows:
    - For every internal node, compute the probability that u turns left and u turns right;
    - For every leave L compute the probability that u is reaching L;
    - The decision for u is equal to decision labeling the leaf with largest probability.
  - *rough decision tree*: in case of uncertainty
    - Use both left and right subtrees to classify the new object;
    - Put together their answer and return the answer vector;
    - Vote for the best decision class.

Outline

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo
Algorytm
schemat
wybór testu
prunning
null-values
Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

1. Wprowadzenie
   - Definicje
   - Funkcje testu
   - Optymalne drzewo
2. Konstrukcja drzew decyzyjnych
   - Ogólny schemat
   - Kryterium wyboru testu
   - Przycinanie drzew
   - Problem brakujących wartości
3. Soft cuts and soft Decision tree
   - Soft Decision Tree
   - Searching for soft cuts
   - Discernibility measure:

STANDARD ALGORITHM FOR BEST CUT

- For a given attribute a and a set of candidate cuts $\{c_1, ..., c_N\}$, the best cut (a, c_i) with respect to given heuristic measure

$$F : \{c_1, ..., c_N\} \to \mathbb{R}^+$$

  can be founded in time $\Omega(N)$.
- The minimal number of simple SQL queries of form

  `SELECT COUNT`
  `FROM data_table`
  `WHERE (a BETWEEN` c_L `AND` c_R`) GROUPED BY` dec.

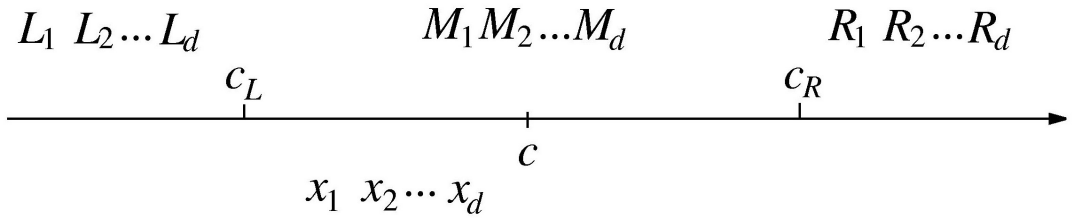
  necessary to find out the best cut is $\Omega(dN)$

OUR PROPOSITIONS FOR SOFT CUTS

- Tail cuts can be eliminated
- Divide and Conquer Technique

Divide and Conquer Technique

- The algorithm outline:
  1. *Divide the set of possible cuts into k intervals*
  2. *Chose the interval to which the best cut may belong with the highest probability.*
  3. *If the considered interval is not STABLE enough then Go to Step 1*
  4. *Return the current interval as a result.*
- The number of SQL queries is $O(d \cdot k \log_k n)$ and is minimum for $k = 3$;
- How to define the measure evaluating the quality of the interval $[c_L; c_R]$?



This measure should estimate the quality of the best cut from $[c_L; c_R]$.

We construct estimation measures for intervals in four cases:

	Discernibility measure	Entropy Measure
Independency assumption	?	?
Dependency assumption	?	?

Outline

Wprowadzenie
Definicje
Funkcje testu
Optymalne drzewo
Algorytm
schemat
wybór testu
prunning
null-values
Soft DT
Soft Decision Tree
Searching for soft cuts
Discernibility measure:

1 Wprowadzenie
- Definicje
- Funkcje testu
- Optymalne drzewo

2 Konstrukcja drzew decyzyjnych
- Ogólny schemat
- Kryterium wyboru testu
- Przycinanie drzew
- Problem brakujących wartości

3 Soft cuts and soft Decision tree
- Soft Decision Tree
- Searching for soft cuts
- Discernibility measure:

Under **dependency assumption**, i.e.

$$\frac{x_1}{M_1} \simeq \frac{x_2}{M_2} \simeq ... \simeq \frac{x_d}{M_d} \simeq \frac{x_1 + ... + x_d}{M_1 + ... + M_d} = \frac{x}{M} =: t \in [0, 1]$$

discernibility measure for $[c_L; c_R]$ can be estimated by:

$$\frac{W(c_L) + W(c_R) + conflict(c_L; c_R)}{2} + \frac{[W(c_R) - W(c_L)]^2}{conflict(c_L; x_R)}$$

Under **dependency assumption**, i.e. $x_1, ..., x_d$ are independent random variables with uniform distribution over sets $\{0, ..., M_1\}$, ..., $\{0, ..., M_d\}$, respectively.

- The mean $E(W(c))$ for any cut $c \in [c_L; c_R]$ satisfies

$$E(W(c)) = \frac{W(c_L) + W(c_R) + conflict(c_L; c_R)}{2}$$

- and for the standard deviation of $W(c)$ we have

$$D^2(W(c)) = \sum_{i=1}^{n} \left[\frac{M_i(M_i+2)}{12} \left(\sum_{j \neq i} (R_j - L_j) \right)^2 \right]$$

- One can construct the measure estimating quality of the best cut in $[c_L; c_R]$ by

$$\boxed{Eval\left([c_L; c_R], \alpha\right) = E(W(c)) + \alpha\sqrt{D^2(W(c))}}$$

Example of tail cut elimination

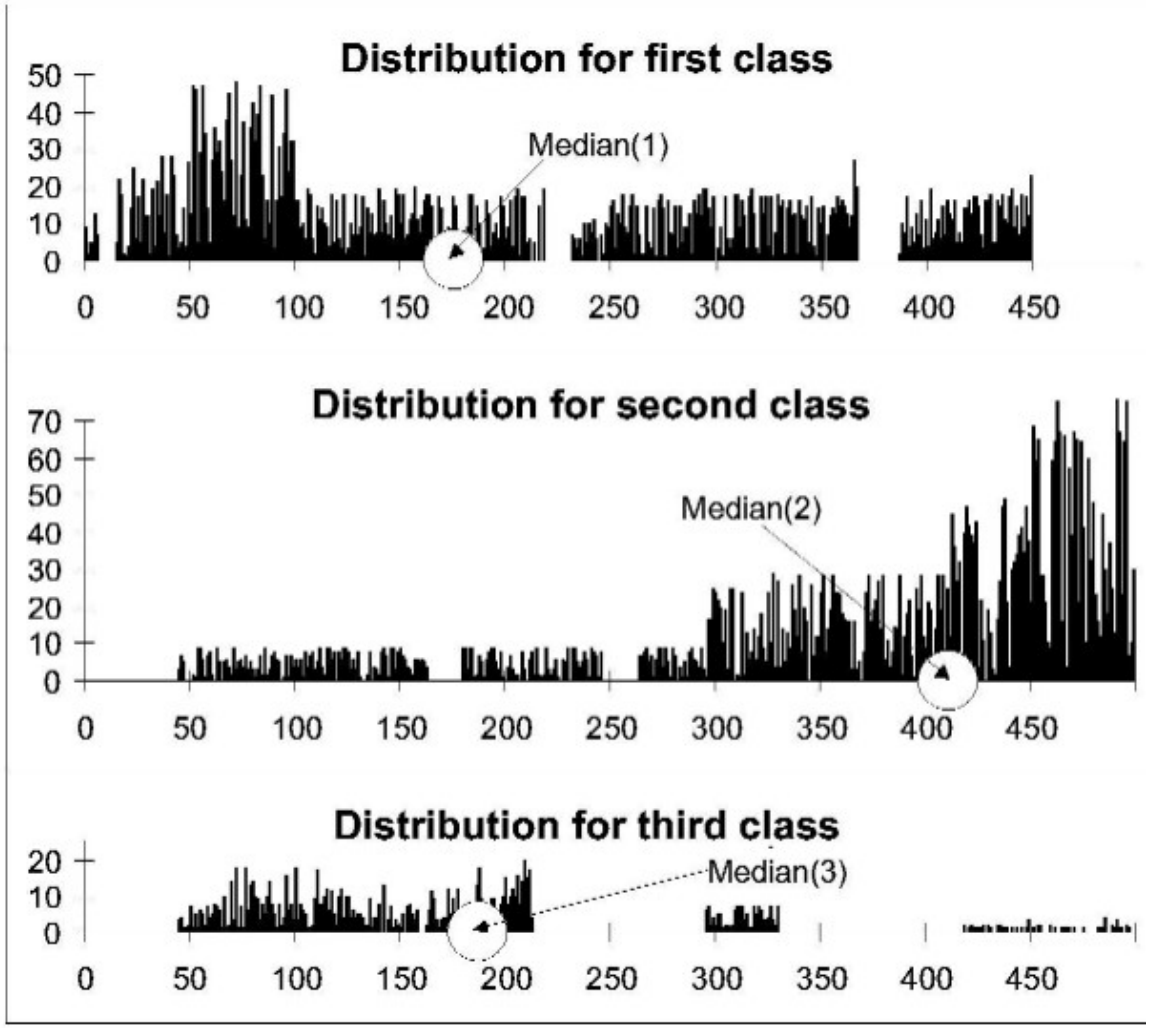
Data mining

Nguyen Hung Son

Wprowadzenie
- Definicje
- Funkcje testu
- Optymalne drzewo

Algorytm
- schemat
- wybór testu
- prunning
- null-values

Soft DT
- Soft Decision Tree
- Searching for soft cuts
- Discernibility measure:



Searching for best cuts

Data mining

Nguyen Hung Son

Wprowadzenie
- Definicje
- Funkcje testu
- Optymalne drzewo

Algorytm
- schemat
- wybór testu
- prunning
- null-values

Soft DT
- Soft Decision Tree
- Searching for soft cuts
- Discernibility measure: