

# Алгоритми дрвета одлучивања

Ненад Митић

Математички факултет  
[nenad@matf.bg.ac.rs](mailto:nenad@matf.bg.ac.rs)





# C4.5

- Quinlan 1993. године, проширење ID3
- Скуп алгоритама (C4.5, C4.5 без поткресивања, C4.5-правила)
- Користи ентропију као меру и однос информационе добити као критеријум поделе
- n-арно дрво (категорички атрибути)
- Критеријум заустављања
  - број различитих инстанци је испод задате границе
  - сви подаци у чвору припадају истој класи

## C4.5 (nastavak)

- Користи нумеричке атрибуте (дискретизација, сортирање ради најбоље поделе)
- За класу у листу бира се најбројнија класа (метода гласања)
- Врши се накнадно поткресивање дрвета
  - смањење нивоа грешке (користи се посебна тест датотека)
  - песимистичка процена грешке (укључени су сви подаци)
  - интервали поверења (Бернулијева случајна променљива)

## C4.5 недостајуће вредности

### Недостајуће вредности

- Избор атрибута за деобу у чвору који садржи НВ?
  - искључити инстанцу
  - заменити најчешћу вредност уместо НВ
  - **информациона добит  $\times$  % познатих вредности**
  - попунити вредност ("различита" вредност, предвиђање на основу осталих атрибута)

## C4.5 недостајуће вредности

Атрибут за поделу садржи НВ - где се смешта инстанца у подели?

- искључити инстанцу
- инстанца има најчешћу вредност атрибута
- придружити инстанцу свакој грани процентуално према броју познатих вредности атрибута у гранама
- придружити је свакој грани
- направити посебну грану са инстанцама са НВ
- одредити најсличнију вредност и доделити је одговарајућој грани
- доделити инстанцу само једној грани у % познатих вредности атр. које су у тој грани

## C4.5 недостајуће вредности

Како се класификује нова инстанца са НВ атрибута?

Смешта се у

- посебну грана за НВ, ако постоји
- грану са најчешћом вредношћу атрибута
- грану са предвиђеном најсличнијом вредношћу
- дистрибуира у гране према релативној вероватноћи добијеној комбиновањем резултата симултаног претраживања за све могуће исходе тестирања. Коначна класа - класа са највећом вероватноћом у укупној дистрибуцији у дрвету



## C4.5 наследници

### Наследници

- J4.8 (Јава) – изворни код нпр. у Weki
- C5.0 (комерцијална верзија, на Windows-у се користи и име See 5)
  - Мањи скуп правила са истом прецизношћу
  - Побољшање засновано на додатном подстицању (боостинг) – комбиновање различитих класификатора ради повећања прецизности
  - Смањено коришћење меморије за 90% (?)
  - Ради између 5.7 и 240 (?) пута брже од C4.5
  - Нижи ниво грешке од C4.5 и мање дрво одлучивања
  - Ради са тежинама атрибута

# C4.5 Алгоритам

```
/* Skup vrednosti D; e1 - min. br. podataka u cvoru */
```

```
C4.5(D, e1)
```

```
begin
```

```
  Trazeno drvo T={}
```

```
  if svi elementi D pripadaju istoj klasi  
    ili ih ima manje od e1
```

```
    then zavrshi algoritam
```

```
  Za svaki atribut x iz D
```

```
    izracunati informacionu dobit u slucaju podele po x
```

```
  Formirati cvor Y u kome se za podelu koristi atribut x  
    koji ima najveću informacionu dobit
```

```
  Bira se podskup  $D_y = f(x)$ 
```

```
  Za svaki skup  $D_y$ 
```

```
    {
```

```
      Odredi  $T_y = C4.5(D_y)$ 
```

```
      Dodaj  $T_y$  kao granu T u cvoru Y
```

```
    }
```

```
  Vрати T
```

```
end
```

# CART

- CART - Classification And Regression Trees
- Breiman, Friedman, Olshen, Stone 1984
- Теоријски заснован, свеобухватан
- Класификациони проблем – предвиђа се вредност атрибута (класе) који је категорички на основу непрекидних и/или категоричких атрибута
- Регресиони проблем – предвиђа се вредност атрибута (класе) који је непрекидан на основу непрекидних и/или категоричких атрибута

# CART (наставак)

Различите методе за проверу резултата

- класификација: Гини, "twoing", уређени "twoing", симтерични Гини
- "twoing"
  - није везан за меру нечистоће чвора
  - користи се за категоричке атрибуте
  - пореди дистрибуцију у дете-чворовима - дели скуп на два  $\approx$  једнака дела
  - уређени "twoing" - групише заједно инстанце са суседним класама редних циљних атрибута
- касније додате ентропија,  $\chi^2$ , ...

# CART (наставак)

- Бинарна подела
  - Инстанца се прослеђује лево ако је испуњен услов, и десно у супротном
- Претходне вероватноће и небалансиране класе
  - аутоматски уклања дисбаланс класа
  - механизам *претходника* -вероватноћа сваке циљне категорије у материјалу за тренинг
  - користе се као тежине, без утицаја на крајњу дистрибуцију класа
- Недостајуће вредност - сурогати
  - атрибут који најбоље подражава (нпр. иста класа) атрибут по коме се дели

# CART Поткресивање дрвета

- У оригиналној верзији конструише се комплетно дрво а тек иза тога се иде на поткресивање
- Формира се више могућих дрвета са покресаним гранама/подрветом за свака два листа која имају заједничког претка
- *Поткресивање резањем трошкова* (енг. *cost complexity pruning*) знатно смањује број дрвета који се разматрају
- Оптимално дрво је поткресано дрво са најмањом ценом за тестне податке
- Текуће верзије раде пре-поткресивање

# CART - неке карактеристике

- Одређивање важности атрибута
  - Користи се збир унапређења (добити) свих чворова у којима се атрибут користи за деобу помножен са процентом тренинг података у чвору
  - Суругати улазе у разматрање
- Матрица цене
- Тежине

# CHAID

- CHAID (Chi-Squared Automatic Interaction Detector) - Kass 1980. године
- Атрибут класе - само категорички
- Остали атрибути именски, редни, категорички, непрекидни (непрекидни се трансформишу у редне)
- Три корака (упаривање, подела, заустављање)
- Дрво класификације се формира узастопном применом ових правила на сваки чвор, почев од кореног



# CHAID (наставак)

- За сваки атрибут  $a$  одређује се пар вредности  $V_a$  са најмањом значајношћу разлике (тј. најсличнији)
- Рачуна се  $p$  вредност у односу на атрибут класе. За поделу се бира атрибут са најмањом  $p$  вредношћу статистичког теста - чвор садржи хомогене вредности
- Проверава се да ли је  $p$  вредност већа од прага
- Ако јесте, пар се упарује у једну сложену категорију и тражи се следећи пар вредности
- Процес се завршава када нема значајних парова

# CHAID (наставак)

- Статистички тест =  $f(\text{типа атрибута класе})$ 
  - непрекидан - F мера
  - номиналан -  $\chi^2$  тест
  - редни – тест односа вероватноћа
- Критеријум заустављања
  - пре-поткресивање
  - не постоји  $\leq$  задатог прага
  - достигнута највећа (задата) дубина дрвета
  - достигнут најмањи број елемената у чвору
- n-арно дрво

# CHAID (наставак)

Рад са недостајућим вредностима

- Ако се подела вршу на основу атрибута који садржи НВ инстанца се не користи
- Ако сви атрибути имају НВ инстанца се игнорише
- НВ се третира као посебна категорија

Исцрпан CHAID алгоритам (Exhaustive CHAID)

- Модификације CHAID - проверава све могуће поделе укрштањем атрибута
- Захтева више времена од CHAID

# QUEST

- QUEST - Quick, Unbiased, Efficient Statistical Trees
- Подржава униваријантне и линеарне комбинације подела вредности у чвору
- Веза између атрибута класе и атрибута при подели
  - ANOVA F—тест или *Levene*-ов тест за редне и непрекидне атрибуте
  - $\chi^2$  тест за категоријске атрибуте
- Пост-поткресивање дрвета
- Атрибут класе - само категоријски
- Остали атрибути именски, редни, категоријски, непрекидни
- Бинарна подела

# SLIQ

- SLIQ - Supervised Learning In Quest
- Варијанта QUEST-а развијена у ИБМ-у за класификацију великих тренинг података
- Ефикасно ради са подацима који не могу да стану у меморију рачунара
- Није заснован на принципу Хантовог алгоритма
- Користи се техника раста дрвета прво у ширину
- Техника пресортирања у фази раста дрвета
- Користи вертикални формат података - на почетку се сви подаци сортирају и сместе у листу

# SLIQ (наставак)

- Подела на основу гини индекса
- Користи *листу класа* - структуру резидентну у меморији која чува ознаке класа сваке инстанце
- Величина *листе класа* је директно пропорционална броју броју улазних слогова
- Категорички и нумерички атрибути
- Поткресивање - техника заснована на МДЛ
- Имплементација за серијско и паралелно извршавање

# SPRINT

- SPRINT (Scalable PaRallelizable INduction of decision Trees)
- SPRINT је модификација SLIQ алгоритма која уклања меморијска ограничења
- Ознаке класа придружује идентификаторима инстанци
- Категорички и нумерички атрибути
- Имплементација за серијско и паралелно извршавање

# Задатак

Над подацима из базе STUD2020 формирати табелу NAZIV\_PROGRAMA која садржи

- све податке из табеле DOSIJE сем идентификације програма који студент студира. При том искључити студенте који имају признате испите
- називе предмета које је студент положио. Узети податке за 15 предмета који се по програму слушају закључно са шестим семестром. При томе неки од назива могу да буду и непознати, уколико је студент није положио предмет. Предмете изабрати тако да не припадају сви искључиво једном студијском програму, уз услов да у изабраном скупу буде највише 1 изборни предмет. У табелу не укључивати идентификације предмета, већ само називе



# Задатак (наставак)

... табела NAZIV\_PROGRAMA садржи...

- звање које се добија по завршетку студијског програма који студент студира. Укључити звање за све нивое студија и све студијске програме. Овај атрибут представља атрибут класе у класификацији

Применити различите алгоритме класификације на податке из формиране табеле. За циљну класу узети звање које се добија

- 1 Извршити препроцесирање података из табеле, уколико је потребно
- 2 Дискутовати добијене резултате у зависности од опција примењених у алгоритму класификације