CLASSIFICATION

T.R.Lekhaa
AP/IT

OUTLINE

- Introduction
- Applications
- Decision tree induction
- Algorithm
- Rule based classification

Introduction

2-Steps:

- Learning
- Classification

Step-1 : Learning

Training Data

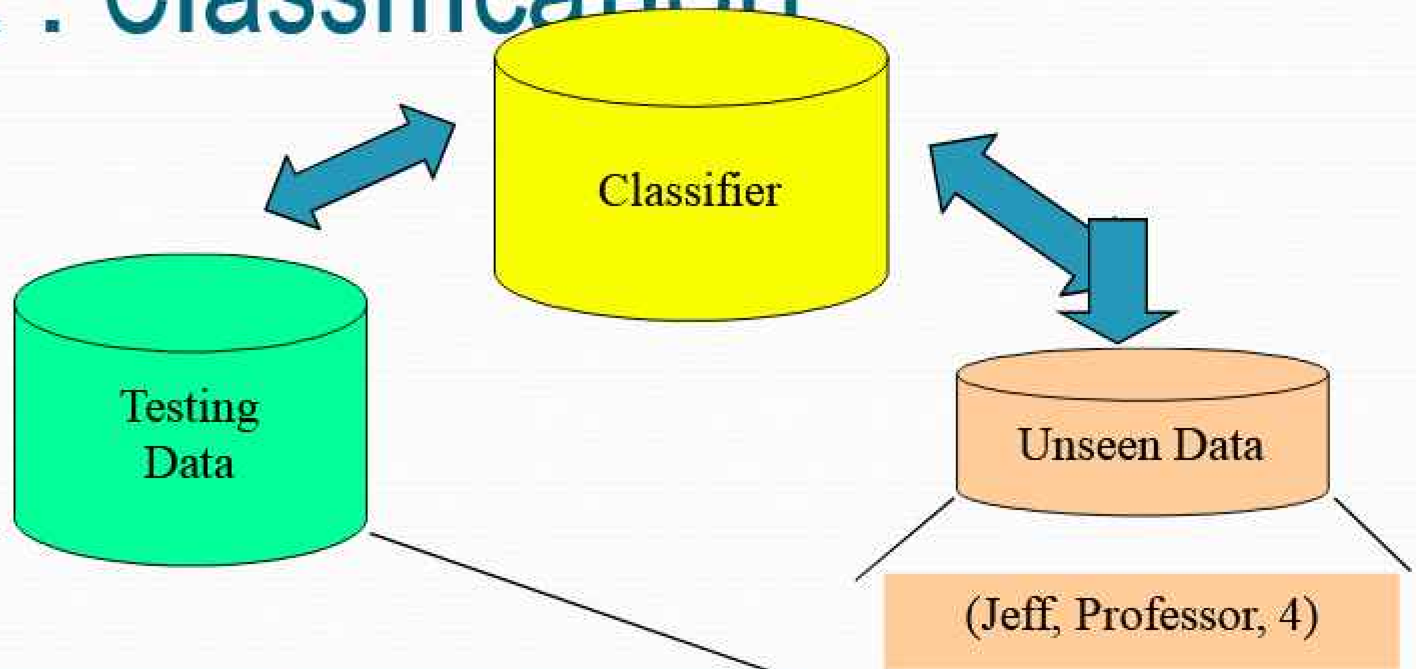
Classification Algorithms

Classifier (Model)

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

Step-2 : Classification



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Tenured?

Yes

Applications

- Credit card approval
- Target marketing
- Medical diagnosis

Classification by decision tree Induction

Decision tree:

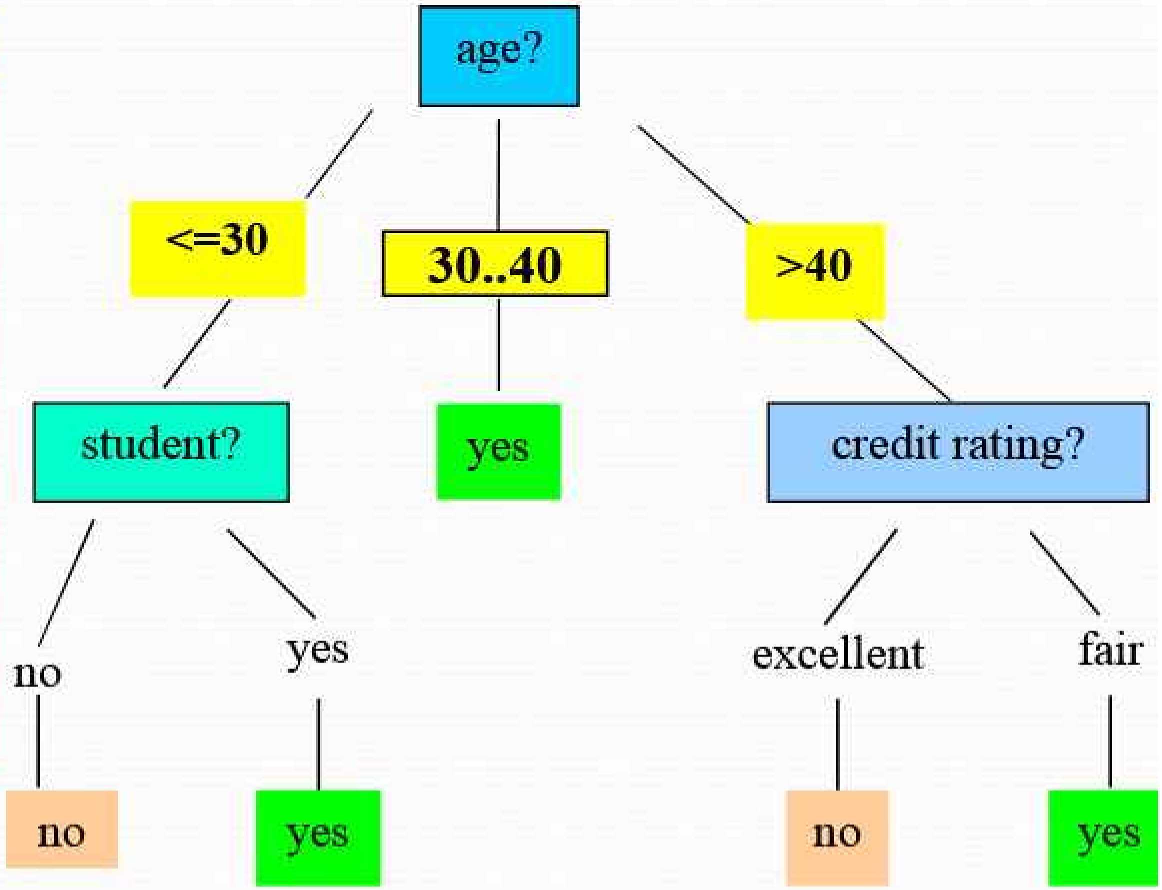
- Flowchart
- Internal node
- Branch
- Leaf node

2-Phases:

- Tree construction
- Tree pruning

Example

age	income	student	credit_rating
<=30	high	no	fair
<=30	high	no	excellent
31…40	high	no	fair
>40	medium	no	fair
>40	low	yes	fair
>40	low	yes	excellent
31…40	low	yes	excellent
<=30	medium	no	fair
<=30	low	yes	fair
>40	medium	yes	fair
<=30	medium	yes	excellent
31…40	medium	no	excellent
31…40	high	yes	fair
>40	medium	no	excellent



Algorithm

Greedy algorithm:

- Top-Down approach
- Training data at root
- Attribute for measurement
    1. Information gain(ID3/C4.5)
    2. Gini index (IBM intelligent miner)

Information gain(ID3/C4.5)

- All attributes are categorical
- Modified for continuous value attribute
- Select the attribute for high information gain
- 2 classes, P&N

Samples sc P element of P class & n element of N class

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Cont..

- Information gain measure used to select S, samples
- $I(s_1, s_2, \ldots s_m) = - \sum_{i=1}^{m} pi \log 2\, pi$

Where pi, probability & m, distinct value

Information encoded in bits pi= $\frac{si}{s}$

Entropy, E(A)= $\sum_{i=1}^{v} I(s1j, s2j, \ldots smj) \frac{s1j, s2j, \ldots smj}{s}$

Subset, $I(s_{1j}, s_{2j}, \ldots s_{mj}) = - \sum_{i=1}^{m} pij \log 2\, pij$

Pij= $\frac{sij}{|sij}$

Cont..

- Gain(A)=I(s1,s2,sm)-E(A)
- I(s1,s2)= -9/14 log2 (9/14) -5/14 log2 (5/14)

= 0.940

Age<=30

Age 31…40

Age>40

Gini index(IBM intelligent miner)

- Attribute having continuous value
- Possible split values for each attribute
- Modified for categorical attributes
- Data set T contains n classas,

Gini (t) – 1- $\sum_{j=1}^{n} p^{\wedge}2j$ Where pj, relative frquency j in T

Dataset splits into 2 subset T1 & T2 with N1 & N2

$Gini_{split}(T) = \frac{N1}{N}$ gini(T1) + $\frac{N2}{N}$ gini(T2)

Gini(d) = 1- (9/14)^2 –(5/14)^2

= 0.459

$gini_{income}$(low,medium)(D) =10/14 gini(D1) + 4/14 gini(D2)

=10/14(1-(7/10)^2 –(3/10)^2) + 4/14 (1-(2/4)^2-(2/4)^2)

=0.443= $gini_{income(high)(D)}$

Tree Pruning

- Reflected noise branches are removed
- 2 types

    1. Post pruning
    2. Pre pruning

Rule based classification

- IF THEN rules are used
- Each attribute- conjunction
- Example:

  IF age<=30 & student=n0 THEN buys-computer=no

  IF age<=30 & student=Yes THEN buys-computer=Yes