

Classification Lecture Notes (2) **(cse634)**

**Review, Training, Testing, Predictive
Accuracy**

Professor Anita Wasilewska

Classification Data

- **Data format:** a data table with key attribute removed. Special attribute- class attribute must be distinguished

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

Classification (Training) Data with objects

rec	Age	Income	Student	Credit_rating	Buys_computer (CLASS)
r1	<=30	High	No	Fair	No
r2	<=30	High	No	Excellent	No
r3	31...40	High	No	Fair	Yes
r4	>40	Medium	No	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	No
r7	31...40	Low	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	<=30	Low	Yes	Fair	Yes
r10	>40	Medium	Yes	Fair	Yes
r11	<=30	Medium	Yes	Excellent	Yes
r12	31...40	Medium	No	Excellent	Yes
r13	31...40	High	Yes	Fair	Yes
r14	>40	Medium	No	Excellent	No

CHARACTERISTIC DESCRIPTIONS (Review)

Example:

- Some of the ***characteristic descriptions*** of the concept **C** with description: **buys_computer= no** are

- Age= \leq 30 & income=high & student=no & credit_rating=fair
- Age= $>$ 40& income=medium & student=no & credit_rating=excellent
- Age= $>$ 40& income=medium
- Age= \leq 30
- student=no & credit_rating=excellent

Characteristic Formula (Review)

Any formula (of a proper language) of a form

IF concept description **THEN** characteristics

is called a characteristic formula

Example:

- **IF** buys_computer= no **THEN** income = low & student=yes & credit=excellent
- **IF** buys_computer= no **THEN** income = low & credit=fair

Characteristic Rule (Review)

- A characteristic formula

IF concept description **THEN** characteristics

is called **a characteristic rule** (for a given database)

if and only if it is **TRUE** in the given database, i.e.

{r: concept description} \wedge {r: characteristics} = not empty set

Characteristic Rule (Review)

EXAMPLE:

The formula

- IF buys_computer= no THEN income = low & student=yes & credit=excellent

is a characteristic rule for our database because

$\{r: \text{buys_computer} = \text{no}\} = \{r1, r2, r6, r8, r16\},$

$\{r: \text{income} = \text{low} \ \& \ \text{student} = \text{yes} \ \& \ \text{credit} = \text{excellent}\} = \{r6, r7\}$

and

$\{r1, r2, r6, r8, r16\} \wedge \{r6, r7\} = \text{not emptyset}$

Characteristic Rule (Review)

EXAMPLE:

The formula

- IF buys_computer= no THEN income = low & credit=fair

Is NOT a characteristic rule for our database because

$\{r: \text{buys_computer} = \text{no}\} = \{r1, r2, r6, r8, r16\},$

$\{r: \text{income} = \text{low} \ \& \ \text{credit} = \text{fair}\} = \{r5, r9\}$

and

$\{r1, r2, r6, r8, r16\} \wedge \{r5, r9\} = \text{emptyset}$

Discrimination (Review)

- *Discrimination is the process which aim is to find rules that allow us to **discriminate** the objects (records) belonging to a given concept (one class) from the rest of records (classes)*
- **DISCRIMINANT RULES** have a form:

If characteristics then concept

Example

- ***If*** Age= \leq 30 & income=high & student=no & credit_rating=fair
then buys_computer= no

Discriminant Formula

A discriminant formula is any formula

If characteristics then concept

- Example:
- IF Age=>40 & inc=low THEN buys_comp= no

Discriminant Rule (Definition)

- A discriminant formula

If characteristics then concept

is a ***DISCRIMINANT RULE*** (in a given database)

iff

{r: Characteristic} \sqsubseteq {r: concept}

Discriminant Rule

- **Example:**

- *A discriminant formula*

**IF Age=>40 & inc=low THEN buys_comp=
no**

IS NOT a discriminant rule in our data base
because

*{o: Age=>40 & inc=low} = {o5, o6} is not a
subset of the set {o:buys_comp= no} = {o1,
o2, o6, o8, o14}*

Characteristic and discriminant rules

- The inverse implication to the characteristic rule is usually NOT a discriminant rule
- Example : the inverse implication to our characteristic rule: ***If*** buys_computer= no **then** income = low & student=yes & credit=excellent

is:

- ***If*** income = low & student=yes & credit=excellent **then** buys_computer= no
- The above rule is NOT a discriminant rule as it can't discriminate between concept with description buys_computer= no and buys_computer= yes
- (see records r6 and r8 in our training dataset)

Descriptive DM Classification Goal (1)

- Given a data set and a concept (class) **c** defined in this dataset **FIND a minimal set (or as small as possible set) characteristic, and/or discriminant rules, or other descriptions** for the concept **c**, or class, or classes.

Descriptive DM Classification Goal (2)

- We also want these rules to involve as few attributes as it is possible, i.e. we want the rules to have **as short as possible length of descriptions.**
- It the supervised LEARNING process the rules must be also TESTED on an independent test data.

Supervised Learning Classification

- The process of creating discriminant and/or characteristic rules and TESTING them
- is called a **learning process**, and when it is finished we say that the concept (class) has been learned (and tested) from examples (records in the dataset that form the the TRAINING set).
- It is called **a supervised learning** because we know the concept description for all examples in the training and test set.

A FULL SET OF DISCRIMINANT RULES for our Training Dataset (Obtained by the Decision Tree Algorithm)

- The rules are:

IF *age* = “<=30” AND *student* = “no” THEN *buys_computer*
= “no”

IF *age* = “<=30” AND *student* = “yes” THEN *buys_computer*
= “yes”

IF *age* = “31...40” THEN
buys_computer = “yes”

IF *age* = “>40” AND *credit_rating* = “excellent” THEN
buys_computer = “no”

IF *age* = “<=30” AND *credit_rating* = “fair” THEN
buys_computer = “yes”

Rules testing

- In order to use rules for testing, and later when testing is done and predictive accuracy is acceptable we write rules in a **predicate form**:

IF *age(x, <=30)* AND *student(x, no)* THEN

buys_computer (x, no)

IF *age(x, <=30)* AND *student (x, yes)* THEN

buys_computer (x, yes)

- Attributes and their values of the new record x are matched with the IF part of the rule and the record is classified accordingly to the THEN part of the rule.

Example of a TEST Data for our TRAINING set

rec	Age	Income	Student	Credit_rating	Buys_computer(CLASS)
r1	<=30	Low	No	Fair	yes
r2	<=30	High	yes	Excellent	No
r3	<=30	High	No	Fair	Yes
r4	31...40	Medium	yes	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	yes
r7	31...40	High	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	31...40	Low	no	Excellent	Yes
r10	>40	Medium	Yes	Fair	Yes

Test dataset and Predictive Accuracy

- The Test Dataset has the same format as the training dataset, i.e. the values of concept attribute are known
- We use it to evaluate the predictive accuracy of our set of rules (discovered by a classification algorithm)
- **PREDICTIVE ACCURACY** of the set of rules, or any classification algorithm is **a percentage of correctly classified data** in the testing dataset.
- If the predictive accuracy is not high enough, or far too high we chose a different learning and testing datasets and start process again
- There are many methods of testing the rules and they will be discussed later

Correctly and Not Correctly Classified Records

- A Record is **correctly classified** if and only if the following conditions hold:
 - (1) we **can classify** the record, i.e. there is a rule such that its LEFT side matches the record,
 - (2) **classification determined by the rule is correct**, i.e. the RIGHT side of the rule matches the value of the record class attribute,

OTHERWISE

- The record is **not correctly classified**
- **WORDS USED: not correctly = incorrectly = misclassified**

Exercise 1

- Assume that we have a following set of rules:
- R1: $a1=1 \wedge a2=0 \Rightarrow \text{class}=\text{yes}$
- R2: $a1=0 \wedge a2=3 \Rightarrow \text{class}=\text{no}$
- R3: $a2=1 \Rightarrow \text{class}=\text{yes}$
- The TEST data has the following 6 records, where the attributes are a1, a2, class
- $r1 = (1, 0, \text{yes})$, $r2 = (0, 3, \text{yes})$, $r3 = (1, 1, \text{no})$,
 $r4 = (2, 1, \text{yes})$, $r5 = (3, 1, \text{yes})$, $r6 = (1, 2, \text{no})$

CALCULATE the Predictive Accuracy of this set of rules with respect to the above TEST data of 6 records.

Exercise 2

- Evaluate the Predictive Accuracy of the set of rules:

R1: IF *age* = “<=30” AND *student* = “no” THEN *buys_computer* = “no”

R2: IF *age* = “<=30” AND *student* = “yes” THEN *buys_computer* = “yes”

R3: IF *age* = “31...40” THEN
buys_computer = “yes”

R4: IF *age* = “>40” AND *credit_rating* = “excellent” THEN
buys_computer = “no”

R5: IF *age* = “<=30” AND *credit_rating* = “fair” THEN
buys_computer = “yes”

with respect to the TEST data on the next slide .

TEST DATA for Example 2

rec	Age	Income	Student	Credit_rating	Buys_computer(CLASS)
r1	<=30	Low	No	Fair	yes
r2	<=30	High	yes	Excellent	No
r3	<=30	High	No	Fair	Yes
r4	31...40	Medium	yes	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	yes
r7	31...40	High	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	31...40	Low	no	Excellent	Yes
r10	>40	Medium	Yes	Fair	Yes

Predictive Accuracy

- There are 10 records and 5 rules R1, R2 ... R5
- Record r1 is well classified by rule R5
- Record r2 is misclassified
- Record r3 is well classified by rule R5
- Record r4 is well classified by rule R5
- Record r5 is misclassified
- Record r6 is misclassified
- Record r7 is well classified by rule R3
- Record r8 is well classified by rule R1
- Record r9 is well classified by rule R3
- Record r10 is misclassified
- We have 6 correctly classified records out of 10
- **Predictive accuracy is 60%**

Classification and Classifiers

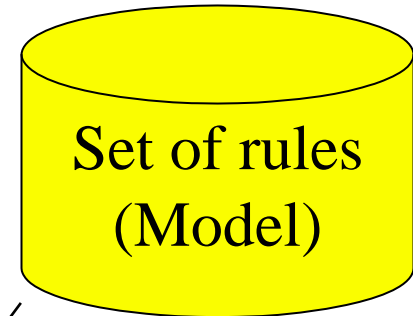
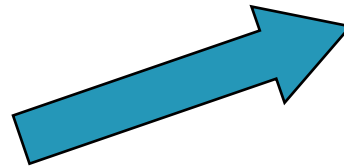
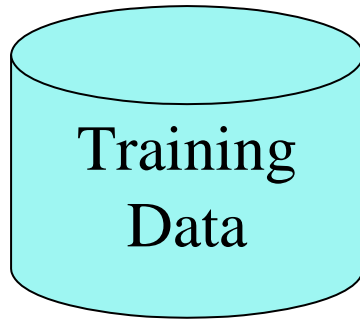
- An algorithm (model, method) is called **a classification algorithm** if it uses the data and its classification to build a set of patterns: discriminant and /or characteristic rules or **other patterns descriptions**.
- Those patterns are structured in such a way that we can use them **to classify unknown sets of objects**- unknown records.
- For that reason, and because of the goal a classification algorithm is often called shortly **a classifier**.
- The name **classifier** implies more than just classification algorithm.
- **A classifier is a final product of the data set and a classification algorithm.**

Classification and Classifiers

- Building a classifier consists of two phases:
training and testing.
- In both phases we use data (**training data set** and disjoint with it **test data set**) for which the class labels are known for ALL of the records.
- **We use** the training data set to create patterns (rules, trees, or a trained Neural or Bayesian network).
- **We evaluate** created patterns with the use of **test data**, which classification is known.
- The measure for a trained classifier accuracy is called **predictive accuracy.**
- **The classifier is build** i.e. we terminate the process if it has been trained and tested and predictive accuracy was on an acceptable level.

Training: a Classifier Construction

(book slide)

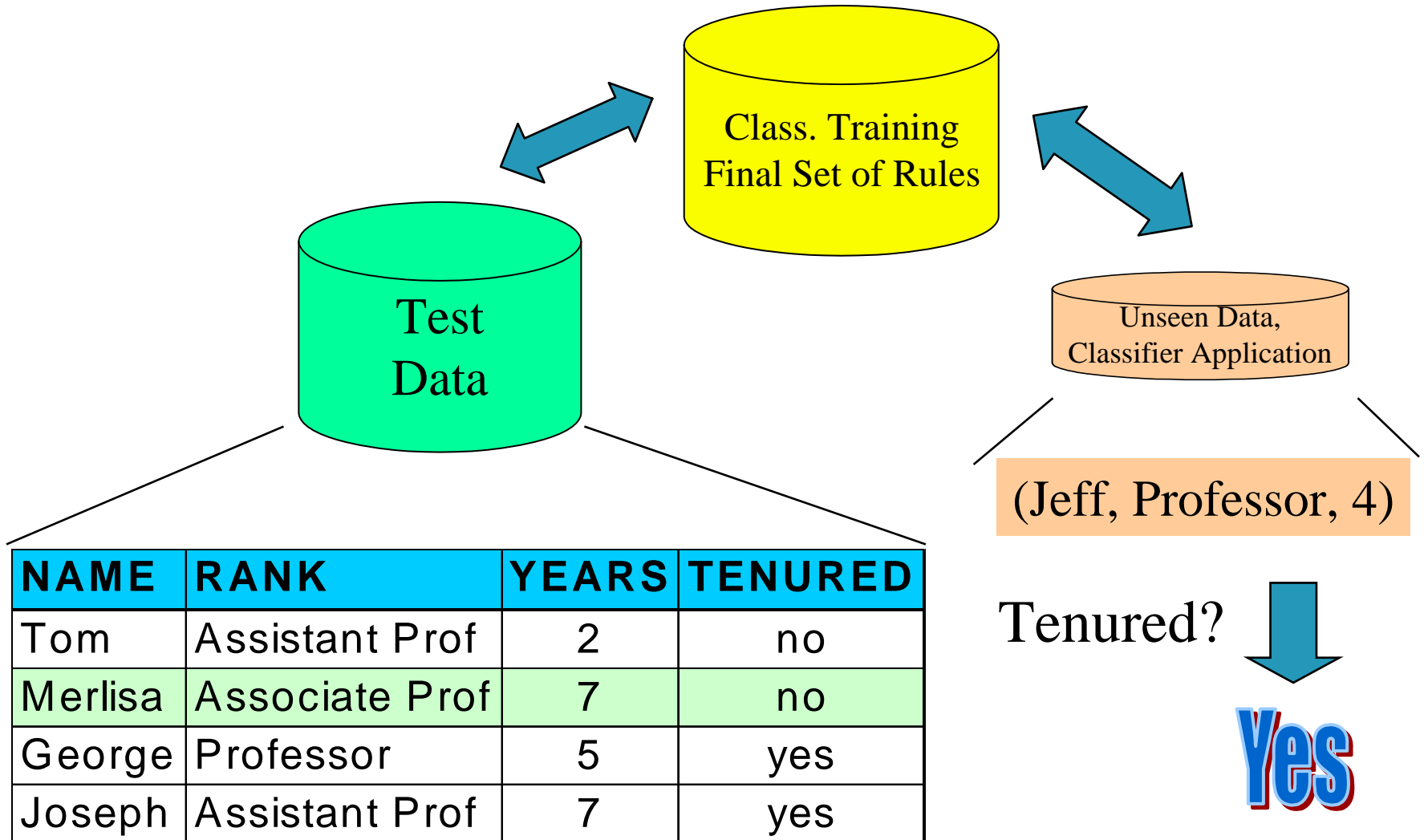


NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

IF rank = 'professor'
OR years > 6
THEN tenured = 'yes'

Testing and Prediction (use of a trained classifier)

(book slide)



Classifiers Predictive Accuracy

- **PREDICTIVE ACCURACY** of a classifier is a percentage of well classified data in the testing data set.
- **Predictive accuracy depends heavily on a choice of the test and training data.**
- There are many methods of choosing test and training sets and hence evaluating the predictive accuracy. This is a separate field of research.

Predictive Accuracy Evaluation

The main methods of predictive accuracy evaluations are (see slides: Classifier Testing):

- **Re-substitution** ($N ; N$)
- **Holdout** ($2N/3 ; N/3$)
- **x-fold cross-validation** ($N - N/x ; N/x$)
- **Leave-one-out** ($N - 1 ; 1$),

where **N** is the number of instances in the dataset (see the separate presentation)

- The process of building and evaluating a classifier is also called a **supervised learning**, or lately when dealing with large data bases a classification method in **Data Mining**.