Introduction to Prediction, Classification, Clustering and Association

Francisco Herrera

Research Group on Soft Computing and Information Intelligent Systems (SCI²S)
Dept. of Computer Science and A.I.
University of Granada, Spain

Email: herrera@decsai.ugr.es
http://sci2s.ugr.es
http://decsai.ugr.es/~herrera
Summary

1. Introduction to Data Mining and Knowledge Discovery
2. Data Preparation
3. Introduction to Prediction, Classification, Clustering and Association
4. Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation
5. Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning
7. Some Advanced Topics : Classification with Imbalanced Data Sets, Subgroup Discovery. Data Complexity
Slides used for preparing this talk:

CS490D: Introduction to Data Mining
Prof. Chris Clifton

Association Analysis: Basic Concepts and Algorithms
Lecture Notes for Chapter 6
Introduction to Data Mining
by Tan, Steinbach, Kumar

DATA MINING
Introductory and Advanced Topics
Margaret H. Dunham
Introduction to Prediction, Clustering, Classification and Association

Outline

- Introduction
- Classification
- Prediction
- Clustering
- Association
- Data Mining Systems / Data Set Repositories
- Concluding Remarks
Introduction to Prediction, Clustering, Classification and Association

Outline

✓ Introduction
✓ Classification
✓ Prediction
✓ Clustering
✓ Association
✓ Data Mining Systems / Data Set Repositories
✓ Concluding Remarks
Introduction

- Knowledge
- Interpretation
- Evaluation

Data Mining

- Patterns
- Data

Preprocessing & cleaning

Selection

Target data

Processed data

Selection & cleaning

Data
What Is The Input?

Concepts
Instances/Examples
Attributes
nominal v.s. numeric attributes
Preparing inputs
What to do in data mining

Classification
Find the class a new instance belong to
e.g. whether a cell is a normal cell or a cancerous cell

Numeric prediction
Variation of classification where the output is numeric classes
e.g. frequency of cancerous cell found
What to do (contd.)

**Clustering**
Process to cluster/group the instances into classes ➔ before existence of any classes
- e.g. deriving/classify a new disease into different possible types/groups

- **Association**
Finding rules/conclusions among attributes
- e.g. a high-blood-pressure patient is most likely to have heart-attack disease
Introduction to Prediction, Clustering, Classification and Association

Outline

✓ Introduction
✓ Classification
✓ Prediction
✓ Clustering
✓ Association
✓ Data Mining Systems / Data Set Repositories
✓ Concluding Remarks
Classification Problem

• Given a database $D=\{t_1, t_2, \ldots, t_n\}$ and a set of classes $C=\{C_1, \ldots, C_m\}$, the **Classification Problem** is to define a mapping $f:D \rightarrow C$ where each $t_i$ is assigned to one class.

• **Prediction** is similar, but may be viewed as having infinite number of classes.
Defining Classes

Partitioning Based

Distance Based
Classification—A Two-Step Process

- **Model construction**: describing a set of predetermined classes
  
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  
  - The set of tuples used for model construction is the training set
  
  - The model is represented as classification rules, decision trees, or mathematical formulae
Classification—A Two-Step Process

- **Model usage**: for classifying future or unknown objects
  - **Estimate accuracy of the model**
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set, otherwise over-fitting will occur
  - If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known
Classification Process (1): Model Construction

Training Data

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>YEARS</th>
<th>TENURED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mike</td>
<td>Assistant Prof</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>Mary</td>
<td>Assistant Prof</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Bill</td>
<td>Professor</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>Jim</td>
<td>Associate Prof</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Dave</td>
<td>Assistant Prof</td>
<td>6</td>
<td>no</td>
</tr>
<tr>
<td>Anne</td>
<td>Associate Prof</td>
<td>3</td>
<td>no</td>
</tr>
</tbody>
</table>

Classification Algorithms

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’
Classification Process (2): Use the Model in Prediction

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>YEARS</th>
<th>TENURED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom</td>
<td>Assistant Prof</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>Merlisa</td>
<td>Associate Prof</td>
<td>7</td>
<td>no</td>
</tr>
<tr>
<td>George</td>
<td>Professor</td>
<td>5</td>
<td>yes</td>
</tr>
<tr>
<td>Joseph</td>
<td>Assistant Prof</td>
<td>7</td>
<td>yes</td>
</tr>
</tbody>
</table>

(Tenured?)

Yes
# Dataset

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>student</th>
<th>credit_rating</th>
<th>buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31…40</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31…40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>31…40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td></td>
</tr>
<tr>
<td>31…40</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td></td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td></td>
</tr>
</tbody>
</table>
A Decision Tree for “buys_computer”

- **age?**
  - <=30
  - student?
    - no
    - yes
  - 30..40
    - yes
  - >40
    - credit rating?
      - excellent
      - no
      - yes
      - fair
      - no
      - yes
Classification example

Given a collection of annotated data. (in this case 5 instances of **Katydid**s and five of **Grasshoppers**), decide what type of insect the unlabeled example is.

(c) Eamonn Keogh, eamonn@cs.ucr.edu

Spanish: Grillo - saltamontes
Classification example

Color \{Green, Brown, Gray, Other\}

Abdomen Length

Thorax Length

Has Wings?

Antennae Length

Mandible Size

Spiracle Diameter

Leg Length

(c) Eamonn Keogh, eamonn@cs.ucr.edu
Classification example

The classification problem can now be expressed as:

Given a training database predict the class label of a previously unseen instance

<table>
<thead>
<tr>
<th>Insect ID</th>
<th>Abdomen Length</th>
<th>Antennae Length</th>
<th>Insect Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.7</td>
<td>5.5</td>
<td>Grasshopper</td>
</tr>
<tr>
<td>2</td>
<td>8.0</td>
<td>9.1</td>
<td>Katydid</td>
</tr>
<tr>
<td>3</td>
<td>0.9</td>
<td>4.7</td>
<td>Grasshopper</td>
</tr>
<tr>
<td>4</td>
<td>1.1</td>
<td>3.1</td>
<td>Grasshopper</td>
</tr>
<tr>
<td>5</td>
<td>5.4</td>
<td>8.5</td>
<td>Katydid</td>
</tr>
<tr>
<td>6</td>
<td>2.9</td>
<td>1.9</td>
<td>Grasshopper</td>
</tr>
<tr>
<td>7</td>
<td>6.1</td>
<td>6.6</td>
<td>Katydid</td>
</tr>
<tr>
<td>8</td>
<td>0.5</td>
<td>1.0</td>
<td>Grasshopper</td>
</tr>
<tr>
<td>9</td>
<td>8.3</td>
<td>6.6</td>
<td>Katydid</td>
</tr>
<tr>
<td>10</td>
<td>8.1</td>
<td>4.7</td>
<td>Katydid</td>
</tr>
</tbody>
</table>
Classification example

Grasshoppers  (c) Eamonn Keogh, eamonn@cs.ucr.edu

Katydid
Classification example

Linear classifier

(c) Eamonn Keogh, eamonn@cs.ucr.edu
Classification models

• Interval rules based classifier

• Instance based classifier

• Linear classifier
Classification Accuracy: Estimating Error Rates

- **Partition: Training-and-testing**
  - use two independent data sets, e.g., training set (2/3), test set (1/3)
  - used for data set with large number of samples
- **Cross-validation**
  - divide the data set into $k$ subsamples
  - use $k-1$ subsamples as training data and one subsample as test data—$k$-fold cross-validation
  - for data set with moderate size
- **Bootstrapping (leave-one-out)**
  - for small size data
Introduction to Prediction, Clustering, Classification and Association

Outline

- ✔ Introduction
- ✔ Classification
- ✔ Prediction
- ✔ Clustering
- ✔ Association
- ✔ Data Mining Systems / Data Set Repositories
- ✔ Concluding Remarks
Prediction Problem

Prediction is different from classification

Classification refers to predict categorical class label
Prediction models continuous-valued functions
How to work?

- Prediction work is similar to classification
  - First, construct a model
  - Second, use model to predict unknown value
    - Major method for prediction is regression
      - Linear and multiple regression
      - Non-linear regression
Regression Analysis in Prediction

• **Linear regression**: $Y = \alpha + \beta X$
  - Two parameters, $\alpha$ and $\beta$ specify the line and are to be estimated by using the data at hand.
  - using the least squares criterion to the known values of $Y_1, Y_2, \ldots, X_1, X_2, \ldots$.

• **Multiple regression**: $Y = b_0 + b_1 X_1 + b_2 X_2$.
  - Many nonlinear functions can be transformed into the above.

• **Neural networks, fuzzy rule based systems, \ldots**
Introduction to Prediction, Clustering, Classification and Association

Outline

✓ Introduction
✓ Classification
✓ Prediction
✓ Clustering
✓ Association
✓ Data Mining Systems / Data Set Repositories
✓ Concluding Remarks
Clustering Problem

• Given a database $D = \{t_1, t_2, \ldots, t_n\}$ of tuples and an integer value $k$, the **Clustering Problem** is to define a mapping $f: D \rightarrow \{1, \ldots, k\}$ where each $t_i$ is assigned to one cluster $K_j$, $1 \leq j \leq k$.

• A **Cluster**, $K_j$, contains precisely those tuples mapped to it.

• Unlike classification problem, clusters are not known a priori.
Clustering Examples

• *Segment* customer database based on similar buying patterns.
• Group houses in a town into neighborhoods based on similar features.
• Identify new plant species
• Identify similar Web usage patterns
Clustering Problem
What is Similarity?
Clustering vs. Classification

• No prior knowledge
  – Number of clusters
  – Meaning of clusters

• Unsupervised learning
Levels of Clustering

- a) Six Clusters
- b) Four Clusters
- c) Three Clusters
- d) Two Clusters
- e) One Cluster
Levels of Clustering

Size Based
Clustering Example

<table>
<thead>
<tr>
<th>Income</th>
<th>Age</th>
<th>Children</th>
<th>Marital Status</th>
<th>Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>$25,000</td>
<td>35</td>
<td>3</td>
<td>Single</td>
<td>High School</td>
</tr>
<tr>
<td>$15,000</td>
<td>25</td>
<td>1</td>
<td>Married</td>
<td>High School</td>
</tr>
<tr>
<td>$20,000</td>
<td>40</td>
<td>0</td>
<td>Single</td>
<td>High School</td>
</tr>
<tr>
<td>$30,000</td>
<td>20</td>
<td>0</td>
<td>Divorced</td>
<td>High School</td>
</tr>
<tr>
<td>$20,000</td>
<td>25</td>
<td>3</td>
<td>Divorced</td>
<td>College</td>
</tr>
<tr>
<td>$70,000</td>
<td>60</td>
<td>0</td>
<td>Married</td>
<td>College</td>
</tr>
<tr>
<td>$90,000</td>
<td>30</td>
<td>0</td>
<td>Married</td>
<td>Graduate School</td>
</tr>
<tr>
<td>$200,000</td>
<td>45</td>
<td>5</td>
<td>Married</td>
<td>Graduate School</td>
</tr>
<tr>
<td>$100,000</td>
<td>50</td>
<td>2</td>
<td>Divorced</td>
<td>College</td>
</tr>
</tbody>
</table>
Types of Clustering

- **Hierarchical** – Nested set of clusters created.
- **Partitional** – One set of clusters created.
- **Incremental** – Each element handled one at a time.
- **Simultaneous** – All elements handled together.
- **Overlapping/Non-overlapping**
Types of Clustering

Hierarchical

Partitional

(c) Eamonn Keogh, eamonn@cs.ucr.edu
Outline

✓ Introduction
✓ Classification
✓ Prediction
✓ Clustering
✓ Association
✓ Data Mining Systems / Data Set Repositories
✓ Concluding Remarks
Association Rule Problem

• Given a set of items $I=\{I_1, I_2, \ldots, I_m\}$ and a database of transactions $D=\{t_1, t_2, \ldots, t_n\}$ where $t_i=\{I_{i1}, I_{i2}, \ldots, I_{ik}\}$ and $I_{ij} \in I$, the **Association Rule Problem** is to identify all association rules $X \Rightarrow Y$ with a minimum support and confidence.

• Link Analysis

• **NOTE:** Support of $X \Rightarrow Y$ is same as support of $X \cup Y$. 
Example: Market Basket Data

- Items frequently purchased together:
  
  \[ \text{Bread} \rightarrow \text{PeanutButter} \]

- Uses:
  
  - Placement
  - Advertising
  - Sales
  - Coupons

- Objective: increase sales and reduce costs
Association Rule Definitions

- **Set of items**: \( I = \{ I_1, I_2, ..., I_m \} \)
- **Transactions**: \( D = \{ t_1, t_2, ..., t_n \}, t_j \subseteq I \)
- **Itemset**: \( \{ I_{i1}, I_{i2}, ..., I_{ik} \} \subseteq I \)
- **Support of an itemset**: Percentage of transactions which contain that itemset.
- **Large (Frequent) itemset**: Itemset whose number of occurrences is above a threshold.
Association Rule Definitions

- **Association Rule (AR):** implication $X \Rightarrow Y$ where $X,Y \subseteq I$ and $X \cap Y = \emptyset$.
- **Support of AR (s) $X \Rightarrow Y$:** Percentage of transactions that contain $X \cup Y$.
- **Confidence of AR ($\alpha$) $X \Rightarrow Y$:** Ratio of number of transactions that contain $X \cup Y$ to the number that contain $X$. 
Association Rules Example

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>Bread, Jelly, PeanutButter</td>
</tr>
<tr>
<td>$t_2$</td>
<td>Bread, PeanutButter</td>
</tr>
<tr>
<td>$t_3$</td>
<td>Bread, Milk, PeanutButter</td>
</tr>
<tr>
<td>$t_4$</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>$t_5$</td>
<td>Beer, Milk</td>
</tr>
</tbody>
</table>

$I = \{ \text{Beer, Bread, Jelly, Milk, PeanutButter} \}$

Support of $\{\text{Bread, PeanutButter}\}$ is 60%
## Association Rules Ex (cont’d)

<table>
<thead>
<tr>
<th>( X \Rightarrow Y )</th>
<th>( s )</th>
<th>( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bread ( \Rightarrow ) PeanutButter</td>
<td>60%</td>
<td>75%</td>
</tr>
<tr>
<td>PeanutButter ( \Rightarrow ) Bread</td>
<td>60%</td>
<td>100%</td>
</tr>
<tr>
<td>Beer ( \Rightarrow ) Bread</td>
<td>20%</td>
<td>50%</td>
</tr>
<tr>
<td>PeanutButter ( \Rightarrow ) Jelly</td>
<td>20%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Jelly ( \Rightarrow ) PeanutButter</td>
<td>20%</td>
<td>100%</td>
</tr>
<tr>
<td>Jelly ( \Rightarrow ) Milk</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>
Association Rule Techniques

1. Find Large Itemsets.
2. Generate rules from frequent itemsets.

**Apriori (1993):** Apriori is a classic algorithm for learning association rules

- **Large Itemset Property:**
  Any subset of a large itemset is large.
- **Contrapositive:**
  If an itemset is not large, none of its supersets are large.
Measuring Quality of Rules

- Support
- Confidence
- Interest
- Conviction
- Chi Squared Test
DILBERT reprinted by permission of United Feature Syndicate, Inc.
Introduction to Prediction, Clustering, Classification and Association

Outline

✓ Introduction
✓ Classification
✓ Prediction
✓ Clustering
✓ Association
✓ Data Mining Systems / Data Set Repositories
✓ Concluding Remarks
Data Mining Systems

Some data mining systems .....
Weka

Data Mining System

- The University of Waikato, New Zealand
- Machine learning software in Java implementation

http://www.cs.waikato.ac.nz/ml/weka/
KEEL Data Mining System

- Machine learning software in Java implementation

http://www.keel.es/
Rapid Miner
Data Mining System

- Rapid Miner YALE: Yet Another Learning Environment

http://rapid-i.com/
Data Mining Repositories

Most of the commercial datasets used by companies for data mining area not available for others to use. However there area a number of “libraries” of datasets that are readily available for downloading from the World Wide Web free of charge by any one. The best known of these is the “Repository” of datasets maintained by the University of California at Irvine, generally known as the “UCI Repository”. The URL for the Repository is: http://archive.ics.uci.edu/ml
Data Mining Repositories

It contains approximately 120 datasets on topics as diverse as credit risks, patients classification, sensor data of a mobile robot, ... Datasets with missing values and noise are included.

A recent development is the creation of the UCI “Knowledge Discovery in Data Bases Archive” at [http://kdd.ics.uci.edu/](http://kdd.ics.uci.edu/). This contains a range of large and complex datasets as a challenge to the data mining research community to scale up its algorithms as the size of stored datasets.
Data Mining Repositories

It contains approximately 120 datasets on topics as diverse as credit risks, patients classification, sensor data of a mobile robot, ... Datasets with missing values and noise are included.

A recent development is the creation of the UCI “Knowledge Discovery in Data Bases Archive” at http://kdd.ics.uci.edu/. This contains a range of large and complex datasets as a challenge to the data mining research community to scale up its algorithms as the size of stored datasets.
## Data Mining Repositories

### UCI Most popular data sets

<table>
<thead>
<tr>
<th>Number</th>
<th>Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>41057:</td>
<td>Iris</td>
</tr>
<tr>
<td>33055:</td>
<td>Adult</td>
</tr>
<tr>
<td>27764:</td>
<td>Wine</td>
</tr>
<tr>
<td>24353:</td>
<td>Breast Cancer Wisconsin (Diagnostic)</td>
</tr>
<tr>
<td>19211:</td>
<td>Poker Hand</td>
</tr>
<tr>
<td>19161:</td>
<td>Abalone</td>
</tr>
</tbody>
</table>
### Data Mining Repositories

#### Iris Data Set

<table>
<thead>
<tr>
<th>Data Set Characteristics:</th>
<th>Multivariate</th>
<th>Number of Instances:</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute Characteristics:</td>
<td>Real</td>
<td>Number of Attributes:</td>
<td>4</td>
</tr>
<tr>
<td>Associated Tasks:</td>
<td>Classification</td>
<td>Missing Values?:</td>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Area:</th>
<th>Life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date Donated:</td>
<td>1988-07-01</td>
</tr>
<tr>
<td>Number of Web Hits:</td>
<td>41063</td>
</tr>
</tbody>
</table>

**Attribute Information:**

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
   -- Iris Setosa
   -- Iris Versicolour
   -- Iris Virginica
Data Mining Systems/Repositories

Other links to Data Mining Systems and Repositories at: http://sci2s.ugr.es/keel/links.php
Introduction to Prediction, Clustering, Classification and Association

Outline

✓ Introduction
✓ Classification
✓ Prediction
✓ Clustering
✓ Association
✓ Data Mining Systems / Data Set Repositories
✓ Concluding Remarks
Some data mining tasks:

• **Prediction Methods**
  – Use some variables to predict unknown or future values of other variables.
    (classification, regression)

• **Description Methods**
  – Find human-interpretable patterns that describe the data.
    (clustering, association, ..)
J. Han, M. Kamber.  
Data Mining: Concepts and Techniques  
Morgan Kaufmann, 2006 (Second Edition)  
http://www.cs.sfu.ca/~han/dmbook

I.H. Witten, E. Frank.  
Data Mining: Practical Machine Learning Tools and Techniques,  

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar  
Introduction to Data Mining (First Edition)  
Addison Wesley, (May 2, 2005)  

Margaret H. Dunham  
Data Mining: Introductory and Advanced Topics  
Prentice Hall, 2003  
http://lyle.smu.edu/~mhd/book