TIETS34 Seminar: Data Mining on Biometric identification

Youming Zhang
Computer Science, School of Information Sciences, 33014 University of Tampere, Finland
Youming.Zhang@uta.fi
Course Description

Content of seminar:

• Data mining and the whole procedure.
• The popular biometric data types for identification.
• Some simple methods of pattern classification.
• Evaluation of identification methods and results.
• Other knowledge of statistic.

Grade:

3-4 ECTS and one extra credit will be given if you participate a series measurements of biometric verification.

Measurements have 4 sessions, 4 weeks interval. Each session has 2 tests: one in morning and another in afternoon. Each test will take 10-15mins.
Requirements:

• No exercise, no project, no report and no exam.
• **Minimum 30 minutes individual presentation**
• Acting as a opponent for one presentation
• Participating in 80% sessions

Schedule:

• Two short introductions
• One presentation session per week
Data Mining

What is data mining:
It is the computational process of discovering patterns in large data sets. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.
Involved methods: artificial intelligence, machine learning, statistics, and database systems
Sometimes called Knowledge Discovery.
Data mining be used in many fields: Business, marketing, medical, geography, music etc.

Data mining task:
Descriptive data mining: summary statistics, visualization, association analysis etc.
Predictive data mining: predict unknown or new data (classification).
Association analysis

Discovering interesting relationship hidden in large data set.

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
<th>{Milk} --&gt; {Coke}</th>
<th>{Diaper, Milk} --&gt; {Beer}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Classification

Construct models that describe and distinguish classes and predict classes of new cases.

- Supervised vs. unsupervised.
- 2-classes vs. n-classes.
- Parameters vs. non-parameters.
Cluster analysis

Grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.
Processing of Data Mining

• Preprocessing
  - Be collected for various purposes; Large amount; Noisy or outliers; unreasonable in practice; For less time consuming; To get better result.
  - Data integration; Data cleaning; Data reduction; Data transformation

• Data mining
  - Choose data mining task
  - Choose mining method or algorithm

• Postprocessing
  - The aim is to show the most interesting mining results to the user in an understandable form
  - Knowledge evaluation, interpretation and presentation
More information of data mining

• TIETS11 Data Mining
• TIETS31 Knowledge Discovery
Fingerprint

“Fingerprint Identification is the method of identification using the impressions made by the minute ridge formations or patterns found on the fingertips. No two persons have exactly the same arrangement of ridge patterns, and the patterns of any one individual remain unchanged throughout life. Fingerprints offer an infallible means of personal identification. Other personal characteristics may change, but fingerprints do not.”

http://www.globalsecurity.org/security/systems/biometrics-fingerprint.htm
Face

• “A general statement of the face recognition problem (in computer vision) can be formulated as follows: Given still or video images of a scene, identify or verify one or more persons in the scene using a stored database of faces.”


• Image and Video

• Features selection
  ➤ Distance between the eyes
  ➤ Width of the nose
  ➤ Depth of the eye sockets
  ➤ The shape of the cheekbones
  ➤ The length of the jaw line
Iris recognition is an automated method of biometric identification that uses mathematical pattern-recognition techniques on video images of the irises of an individual's eyes, whose complex random patterns are unique and can be seen from some distance.”------Wikipedia
Electrocardiography (ECG)

- One dimension signals
• Data source (signals) dimension
  - Fingerprint, face image, iris: 2D
  - ECG: 1D
  - Video: 3D

• Data dimension

How many attributes or variables be selected? 5 for face and 21 for ECG in above examples.
Principal component analysis (PCA)

• Find a projection that best represent the data
Linear discriminant analysis (LDA)

- Find a linear combination of features which characterizes or separates two or more classes of objects or events.
Multiple discriminant analysis (MDA)
Independent component analysis (ICA)

- Apply to the problem of blind source separation.