Data Mining & Feature Selection

M.M. Pedram
pedram@tmu.ac.ir
Faculty of Engineering, Tarbiat Moallem University
The 11th Iranian Conference on Fuzzy Systems, 5-7 July, 2011
Contents

- History
- KDD
  - Data Mining
    - Classification
    - Estimation (Regression)
    - Clustering
    - Market Basket Analysis
      - Association Rule Mining
      - Sequence Mining
  - Feature Selection
    - Filter
    - Wrapper
  - Rough set Theory
A Taxonomy

- wisdom
- knowledge
- information
- data
- ignorance

“Data is not information; information is not knowledge; knowledge is not wisdom.” Gary Flake, Principal Scientist & Head of Yahoo! Research Labs, July 2004.
History

- The birth of Data Mining/KDD:
  - 1989 IJCAI Workshop on *Knowledge Discovery in Databases*.
  - 1991–1994 Workshops on Knowledge Discovery in Databases

- 1995 – date: International Conferences on *Knowledge Discovery in Databases and Data Mining* (KDD)

- 2001 – date: *IEEE ICDM*
Data mining: the core of knowledge discovery process

Data Warehouse

Data Integration

Data Cleaning

Task-relevant Data

Selection

Data Mining

Pattern Evaluation

Knowledge
KDD Process

Increasing potential to support business decisions

- **Data Sources**
  - Paper, Files, Web documents, Scientific experiments, Database Systems

- **Data Preprocessing/Integration, Data Warehouses**

- **Data Exploration**
  - Statistical Summary, Querying, and Reporting

- **Data Mining**
  - Information Discovery

- **Data Presentation**
  - Visualization Techniques

- **Decision Making**

- This is a view from business intelligence communities

- **End User**
- **Business Analyst**
- **Data Analyst**
- **DBA**
Data Mining Definition

The search for interesting patterns and models, in large data collections, using statistical and machine learning methods, and high-performance computational infrastructure.

Data → DM → Knowledge
Why mine data – commercial viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at department/grocery stores
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful

- Competitive Pressure is Strong
  - Provide better, customized services (e.g. in Customer Relationship Management)
Why mine data – scientific viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data

- Traditional techniques infeasible for raw data

- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation
Data Mining, A Multidisciplinary Field

- Machine Learning
- Pattern Recognition
- Statistics
- Applications
- Visualization
- Algorithm
- Database Technology
- High-Performance Computing

Data Mining
Data Mining Techniques

- Data Mining Strategies
  - Unsupervised Clustering
  - Supervised Learning
    - Classification
    - Estimation
  - Market Basket Analysis
    - Prediction
Data Mining Techniques

- **Descriptive**
  - Clustering
  - Market Basket Analysis

- **Predictive**
  - Classification
  - Regression

- **Prediction Methods**
  - Use some variables to predict unknown or future values of other variables.

- **Description Methods**
  - Find human-interpretable patterns that describe the data.
Data Mining Techniques

- Classification
- Estimation (Regression)
- Clustering
- Market Basket Analysis:
  - Association Rule Discovery (Mining)
  - Sequential Pattern Discovery (Mining)
Model

Variable

- numeric
- categorical
Classification

- Decision Tree
- Multi Layer Perceptron & other Neural Networks
- Production System Rules/Fuzzy Extensions
- ...

\[ X \xrightarrow{\text{Model}} Y \]

\[ X_1, X_2, \ldots, X_n \]

\[ \text{Class } #1, \text{Class } #m \]
Estimation

- Output variables are numeric.

Rules
- Neural Networks
- Time Series
- Regression/Fuzzy Extensions
- ...

![Diagram of model with inputs X and outputs Y]
Prediction

- Classification or Estimation for future

- Decision Trees
- Neural Networks
- Rules
- Regression/Fuzzy Extension
- Time Series
- ...

Model

\[ X \rightarrow \text{Model} \rightarrow Y \]

\[ x_1, x_2, \ldots, x_n \rightarrow \text{Model} \rightarrow y_1, y_2, \ldots, y_m \]
Clustering

- A number of similar individuals that occur together, for example
Clustering

- Customer database segmentation based on similar buying patterns.
- Identify similar Web usage patterns
- Group houses in a town into neighborhoods based on similar features.
Market Basket Data

- Extracting the pattern of Items that frequently happened together:

- Uses:
  - Placement
  - Advertising
  - ...

- Objective: increase sales and reduce costs
Association Rule Definitions

- **Association Rule:**
  
  implication : \( A \Rightarrow B \) where \( A \cap B = {} \)

- **Support of \( A \Rightarrow B \):**
  
  Percentage of transactions that contain \( A \cap B \)
  
  \[
  Supp(A \Rightarrow B) = \frac{|A \cap B|}{|D|} = P(A \wedge B)
  \]
  
  \[
  = \frac{5000}{14000} = 0.357 \equiv 35.7\%
  \]

- **Confidence of \( A \Rightarrow B \):**
  
  \[
  Conf(A \Rightarrow B) = P(B \mid A) = \frac{|A \wedge B|}{|A|} = \frac{Supp(A \Rightarrow B)}{supp(A)} ; \quad supp(A) = \frac{|A|}{|X|}
  \]
  
  \[
  P(B \mid A) = \frac{5000}{10000} = 0.5 \equiv 50\%
  \]
### 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>

<table>
<thead>
<tr>
<th>$A \Rightarrow B$</th>
<th>Supp</th>
<th>Conf</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>
Sequences

A sequence is an ordered list of symbols.

Example of time interval:

\[
\langle \text{Milk}, \text{Cheese}, \text{Ice Cream}, \text{Water}, \text{Coke}, \text{Chocolate} \rangle
\]
Sequences

Example

The best breakfast
And maximum energy

Sequence 1: ((milk, honey, coffee, bread, butter))

It provides about 80% of our energy

Sequence 2: (coffee, chocolate, ice cream, orange juice, water)

It provides about 30% of our energy

Sequence 3: (eggs, orange juice, cheese, chocolate, ice cream)

It provides about 50% of our energy
Sequential Pattern Mining

- Discovery of frequent sequential patterns (subsequences) in a database of sequences.

\[
\begin{align*}
\langle 1 \ 2 \ 3 \ 4 \rangle \\
\langle 1 \ 5 \ 6 \ 7 \ 8 \ 9 \rangle \\
\langle 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \rangle
\end{align*}
\]

\[\text{min}_\text{sup} \geq 50\%\]
What Is Sequential Pattern Mining?

A sequence database

<table>
<thead>
<tr>
<th>SID</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>&lt;a(abc)(ac)d(cf)&gt;</td>
</tr>
<tr>
<td>20</td>
<td>&lt;(ad)c(bc)(ae)&gt;</td>
</tr>
<tr>
<td>30</td>
<td>&lt;(ef)(ab)(df)cb&gt;</td>
</tr>
<tr>
<td>40</td>
<td>&lt;eg(af)cbc&gt;</td>
</tr>
</tbody>
</table>

Given \( \text{min\_sup} = 2 \), \(<\text{ab}\>\)c is a **sequential pattern**.
Example (newspaper articles)

A series of daily newspaper articles

<typhoon (flood, landslide)>
Feature Selection

M.M. Pedram, pedram@tmu.ac.ir

The 11th Iranian Conference on Fuzzy Systems, 5-7 July 2011
Feature Selection

Finding the most compact and **informative set of features**, to improve the efficiency or data storage and processing.
Classify Fruits

(Øx, Øy, Øx/Øy, curvature, color, hardness, weight, smell, …)
Risk Functional

- A function of the parameters of the learning machine, assessing how much it is expected to fail on a given task.
  - Classification: the error rate
  - Regression: the mean square error
Overfitting

- Fit / Robustness Tradeoff
Ockham’s Razor

- Principle proposed by William of Ockham in the fourteenth century.
- Of two theories providing similarly good predictions, prefer the simplest one.
- Shave off unnecessary parameters of your models.
Why Feature Selection?

- Lot of inputs $\implies$ Lots of parameters & Large input space
  $\implies$ Curse of dimensionality and risks of overfitting!
## FS Methods Classification

### Unsupervised

<table>
<thead>
<tr>
<th>Selection</th>
<th>Linear</th>
<th>Nonlinear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation between inputs</td>
<td>PCA</td>
<td>Kohonen maps</td>
</tr>
</tbody>
</table>

### Supervised

<table>
<thead>
<tr>
<th>Selection</th>
<th>Linear</th>
<th>Nonlinear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation between inputs &amp; Outputs</td>
<td>Linear Discriminant Analysis, Partial Least Squares</td>
<td>Mutual information between inputs and outputs, Greedy algorithms, Genetic algorithms</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Projection</th>
<th>Linear</th>
<th>Nonlinear</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Projection pursuit</td>
</tr>
</tbody>
</table>
FS Methods Classification

- **Selection:** choosing among the original features
  - easy (+)
  - interpretability of the features (+)

- **Projection:** creating new features from the original ones
  - more general → possibly more efficient (+)
  - more difficult (-)
  - features not interpretable (-)
Supervised selection: filter versus wrapper

- Supervising does not necessarily mean to use the model!

**Filter**

- Note the final quality criterion.

**Wrapper**

- Many (non)linear models are designed.
Important Questions in FS

- **Question 1**: Subset relevance assessment
  - Among all $2^d - 1$ possible subsets, which is the best one?

- **Question 2**: Optimal Subset search
  - How not to consider all $2^d - 1$ possible subsets?

There are two feature qualities that must be considered by FS methods: **relevancy and redundancy**.
Subset relevance assessment

- Relevance is difficult to define!

- **Filter approach** *(model free)*:
  - a variable (or set of ) is relevant if it is statistically dependent on $y$.
  
  $$P(y|x_i) \neq P(y)$$

- **Wrapper approach** *(uses model f)*:
  - a variable (or set of ) is relevant if the model built on it shows good performances.
Subset relevance assessment

- Correlation
- Mutual information
Correlation

❖ Correlation, a linear filter
  ▸ Measures linear dependencies (between -1 and 1)
  ▸ • indicates no linear relation.

❖ Definition:
  ▸ correlation between random variable $X$ and random variable $Y$

$$\rho = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)} \sqrt{\text{Var}(y)}} = \frac{E(xy) - E(x)E(y)}{\sqrt{\text{Var}(x)} \sqrt{\text{Var}(y)}} = \frac{E[(x - E[x]) \cdot (y - E[y])]}{\sqrt{E[(x - E[x])^2]} \cdot \sqrt{E[(y - E[y])^2]}}$$

▸ Suppose a dataset $\{x^j, y^j\}$

$$= \frac{\sum_{j=1}^{N} (x^j - \bar{x}) \cdot (y^j - \bar{y})}{\sqrt{\sum_{j=1}^{N} (x^j - \bar{x})^2} \cdot \sqrt{\sum_{j=1}^{N} (y^j - \bar{y})^2}}$$
Correlation

Strong correlation

Weak correlation
Correlation

Note: Correlation does not measure nonlinear relations!
Correlation

- Correlation:
  - is linear.
  - is parametric (it makes the hypothesis of a ...linear model).
  - does not explain causality.
  - is almost impossible to define between more than 2 variables.
  - is sensitive to outliers.
Mutual information

- Mutual information between random variable $x$ and random variable $y$ measures how the uncertainty on $y$ is reduced when $x$ is known (and vice versa).

- Relevance of a subset $X_S$: mutual information $MI(X_S; y)$ between this subset and the target variable $y$. 
**Mutual information**

- **Some properties:**
  - If $x$ and $y$ are independent, $MI(y;x) = 0$
  - $MI(y;y) = H(y)$
  - $MI(y;x)$ is always non-negative and less than $\min(H(y), H(x))$
Mutual information

- Nonlinear:

\[
MI(X,Y) = \int P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \, dX \, dY
\]

\[= KL\left( P(X,Y) \parallel P(X)P(Y) \right) \]
Taxonomy of FS approaches

Feature Selection

Transformation Based
- Linear
  - PCA
  - PP
  - MDS
- Nonlinear
  - Isomap
  - LLE

Selection Based
- Filter
- Wrapper
- Hybrid
- Rough Set
- Fuzzy Rough Set
Rough Set
Rough Sets

- Modeling imperfect knowledge.
- It can be employed to reduce the dimensionality of datasets.

Zdzislaw Pawlak: Rough Set Theory (1982)
Rough Sets

- A rough set is itself the approximation of a vague concept (set) by a pair of precise concepts, called lower and upper approximations, that are a classification of the domain of interest into disjoint categories.
- It works by exploring and exploiting the granularity structure of the data only.
Information and Decision Systems

- **Information system**: a table of data, consisting of objects (rows in the table) and attributes (columns) = database

- **Decision system**: an information system may be extended by the inclusion of decision attributes.

- A decision system is **consistent** if for every set of objects whose attribute values are the same, the corresponding decision attributes are identical.
**Example #1**

*Decision system*: consists of 4 conditional features \((a, b, c, d)\), a decision feature \((e)\), and eight objects.

<table>
<thead>
<tr>
<th>(x \in \mathbb{U})</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(\Rightarrow)</th>
<th>(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(S)</td>
<td>(R)</td>
<td>(T)</td>
<td>(T)</td>
<td></td>
<td>(R)</td>
</tr>
<tr>
<td>1</td>
<td>(R)</td>
<td>(S)</td>
<td>(S)</td>
<td>(S)</td>
<td></td>
<td>(T)</td>
</tr>
<tr>
<td>2</td>
<td>(T)</td>
<td>(R)</td>
<td>(R)</td>
<td>(S)</td>
<td></td>
<td>(S)</td>
</tr>
<tr>
<td>3</td>
<td>(S)</td>
<td>(S)</td>
<td>(R)</td>
<td>(T)</td>
<td></td>
<td>(T)</td>
</tr>
<tr>
<td>4</td>
<td>(S)</td>
<td>(R)</td>
<td>(T)</td>
<td>(R)</td>
<td></td>
<td>(S)</td>
</tr>
<tr>
<td>5</td>
<td>(T)</td>
<td>(T)</td>
<td>(R)</td>
<td>(S)</td>
<td></td>
<td>(S)</td>
</tr>
<tr>
<td>6</td>
<td>(T)</td>
<td>(S)</td>
<td>(S)</td>
<td>(S)</td>
<td></td>
<td>(T)</td>
</tr>
<tr>
<td>7</td>
<td>(R)</td>
<td>(S)</td>
<td>(S)</td>
<td>(R)</td>
<td></td>
<td>(S)</td>
</tr>
</tbody>
</table>
Nomenclature

\( I = (U,A) \): an information system,

\( U \): a nonempty set of finite objects (the universe of discourse),

\( A \): a nonempty finite set of attributes

\( a : U \rightarrow V_a \): an attribute, \( a \in A \).

\( V_a \): the set of values that attribute \( a \) may take.

\( A = \{C \cup D\} \): decision systems,

\( P \): Feature subset, \( P \subseteq A \),

\( C \): the set of input features,

\( D \): the set of class indexes.

\( d \in D \): a class index, itself a variable

\( d : U \rightarrow \{\cdot, 1\} \) such that for \( x \in U \),

\( d(x) = 1 \) if \( x \) has class \( d \) and

\( d(x) = \cdot \) otherwise.
Indiscernibility

- With any $P \subseteq A$ there is an associated equivalence relation $IND(P)$:

$$IND(P) = \{(x, y) \in U \times U \mid \forall a \in P, a(x) = a(y)\}$$

- Note that this relation corresponds to the equivalence relation for which two objects are equivalent if and only if they have the same vectors of attribute values for the attributes in $P$. 
Indiscernibility

- The partition of $U$, determined by $IND(P)$, is denoted $U/IND(P)$ or $U/P$, which is simply the set of equivalence classes generated by $IND(P)$:

$$U/IND(P) = \bigotimes\{ U/IND(\{a\}) \mid a \in P \}$$

- Where

$$A \bigotimes B = \{X \cap Y \mid X \in A, Y \in B, X \cap Y \neq \emptyset \}$$

- The equivalence classes of the indiscernibility relation with respect to $P$ are denoted $[x]_P$, $x \in U$. 

M.M. Pedram, pedram@tmu.ac.ir
**Indiscernibility**

- **Example:** if \( P = \{b, c\} \), then objects 1, 6, and 7 are indiscernible, as are objects 0 and 4.

<table>
<thead>
<tr>
<th>( x \in U )</th>
<th>( a )</th>
<th>( b )</th>
<th>( c )</th>
<th>( d )</th>
<th>( e )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>S</td>
<td>( \text{R} )</td>
<td>( \text{T} )</td>
<td>T</td>
<td>R</td>
</tr>
<tr>
<td>1</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>T</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>( \text{R} )</td>
<td>( \text{R} )</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>3</td>
<td>S</td>
<td>S</td>
<td>( \text{R} )</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>4</td>
<td>S</td>
<td>( \text{R} )</td>
<td>T</td>
<td>R</td>
<td>S</td>
</tr>
<tr>
<td>5</td>
<td>T</td>
<td>( \text{T} )</td>
<td>( \text{R} )</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>6</td>
<td>T</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>T</td>
</tr>
<tr>
<td>7</td>
<td>R</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>S</td>
</tr>
</tbody>
</table>

- **IND(\( P \))** creates the following partition of \( U \):

\[
\frac{U}{\text{IND}(P)} = \frac{U}{\text{IND}(\{b\})} \otimes \frac{U}{\text{IND}(\{c\})} \\
= \{\{0, 2, 4\}, \{1, 3, 6, 7\}, \{5\}\} \\
\otimes \{\{2, 3, 5\}, \{1, 6, 7\}, \{0, 4\}\} \\
= \{\{2\}, \{0, 4\}, \{3\}, \{1, 6, 7\}, \{5\}\}
\]
Lower and Upper Approximations

Let $X \subseteq U$. $X$ can be approximated using only the information contained within $P$ by constructing the $P$-lower and $P$-upper approximations of the classical crisp set $X$:

$$
P_X = \left\{ x \mid [x]_P \subseteq X \right\}
$$

$$
\overline{P}_X = \left\{ x \mid [x]_P \cap X \neq \emptyset \right\}
$$

The tuple $\langle P_X, \overline{P}_X \rangle$ is called a rough set.
Positive, Negative, and Boundary Regions

- Let $P$ and $Q$ be sets of attributes inducing equivalence relations over $U$, then the positive, negative, and boundary regions are defined as:

$$POS_P(Q) = \bigcup_{X \in U/Q} PX$$

$$NEG_P(Q) = U - \bigcup_{X \in U/Q} \overline{PX}$$

$$BND_P(Q) = \bigcup_{X \in U/Q} \overline{PX} - \bigcup_{X \in U/Q} PX$$
Positive, Negative, and Boundary Regions

Upper Approximation \( \bar{P}X \)

Set \( X \)

Lower Approximation \( P\bar{X} \)

Equivalence Class: \([x]_P\) (Granules)

\([x]_P\) : the set of all points which are *indiscernible* with point \( x \) in terms of feature subset \( P \). (set of all points belonging to the same granule as of the point \( x \) in feature space \( \Omega_p \))
Positive, Negative, and Boundary Regions

- $\text{POS}_P(Q)$: The positive region comprises all objects of $U$ that can be classified to classes of $U/Q$ using the information contained within attributes $P$.

- $\text{BND}_P(Q)$: The boundary region is the set of objects that can possibly, but not certainly, be classified in this way.

- $\text{NEG}_P(Q)$: The negative region is the set of objects that cannot be classified to classes of $U/Q$. 
Let $P = \{b,c\}$ and $Q = \{e\}$, then

$$\mathbb{U}/\text{IND}(P) = \mathbb{U}/\text{IND}({b}) \otimes \mathbb{U}/\text{IND}({c}) = \{2\}, \{0, 4\}, \{3\}, \{1, 6, 7\}, \{5\}$$

$$\text{POS}_P(Q) = \mathbb{U}\{\emptyset, \{2, 5\}, \{3\}\} = \{2, 3, 5\}$$

$$\text{NEG}_P(Q) = \mathbb{U} - \mathbb{U}\{\{0, 4\}, \{2, 0, 4, 1, 6, 7, 5\}, \{3, 1, 6, 7\}\}$$

$$= \emptyset$$

$$\text{BND}_P(Q) = \mathbb{U}\{\{0, 4\}, \{2, 0, 4, 1, 6, 7, 5\}, \{3, 1, 6, 7\}\} - \{2, 3, 5\}$$

$$= \{0, 1, 4, 6, 7\}$$

<table>
<thead>
<tr>
<th>$x \in \mathbb{U}$</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$d$</th>
<th>$\Rightarrow$</th>
<th>$e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$S$</td>
<td>$R$</td>
<td>$T$</td>
<td>$T$</td>
<td>$T$</td>
<td>$R$</td>
</tr>
<tr>
<td>1</td>
<td>$R$</td>
<td>$S$</td>
<td>$S$</td>
<td>$S$</td>
<td>$T$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$T$</td>
<td>$R$</td>
<td>$R$</td>
<td>$S$</td>
<td>$T$</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$S$</td>
<td>$S$</td>
<td>$R$</td>
<td>$T$</td>
<td>$T$</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$S$</td>
<td>$R$</td>
<td>$T$</td>
<td>$R$</td>
<td>$S$</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$T$</td>
<td>$T$</td>
<td>$R$</td>
<td>$S$</td>
<td>$S$</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$T$</td>
<td>$S$</td>
<td>$S$</td>
<td>$S$</td>
<td>$T$</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>$R$</td>
<td>$S$</td>
<td>$S$</td>
<td>$R$</td>
<td>$S$</td>
<td></td>
</tr>
</tbody>
</table>
Feature Dependency

- A set of attributes $Q$ **depends totally** on a set of attributes $P$, denoted $P \Rightarrow Q$, if all attribute values from $Q$ are uniquely determined by values of attributes from $P$.

- If there exists a functional dependency between values of $Q$ and $P$, then $Q$ depends totally on $P$. 
Feature Dependency and Significance

Dependency In rough set theory:

- For $P, Q \subseteq A$, it is said that $Q$ **depends on $P$ in a degree $k$** ($0 \leq k \leq 1$), denoted $P \Rightarrow_k Q$, if

\[
k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|}
\]

- $|.|$ : the cardinality.

- If $k = 1$, $Q$ depends totally on $P$,
  - if $0 < k < 1$, $Q$ depends partially (in a degree $k$) on $P$,
  - if $k = 0$, $Q$ does not depend on $P$. 

---

M.M. Pedram, pedram@tmu.ac.ir  
The 11th Iranian Conference on Fuzzy systems, 5-7 July, 2011
Example: the degree of dependency of attribute \{e\} on the attributes \{b,c\} is:

\[
\gamma_{\{b,c\}}(\{e\}) = \frac{|POS_{\{b,c\}}(\{e\})|}{|U|} = \frac{|\{2, 3, 5\}|}{|\{0, 1, 2, 3, 4, 5, 6, 7\}|} = \frac{3}{8}
\]
Feature Dependency and Significance

- Given $P, Q$ and a feature $a \in P$, the **significance** of feature $a$ upon $Q$ is defined by:

$$\sigma_P(Q, a) = \gamma_P(Q) - \gamma_{P-\{a\}}(Q)$$

- If the significance of feature $a$ is $\cdot$, then the feature $a$ is **dispensible**.
For example, if \( P = \{a,b,c\} \) and \( Q = e \), then

\[
\gamma_{a,b,c}(\{e\}) = \frac{|\{2, 3, 5, 6\}|}{8} = \frac{4}{8}
\]

\[
\gamma_{a,b}(\{e\}) = \frac{|\{2, 3, 5, 6\}|}{8} = \frac{4}{8}
\]

\[
\gamma_{b,c}(\{e\}) = \frac{|\{2, 3, 5\}|}{8} = \frac{3}{8}
\]

\[
\gamma_{a,c}(\{e\}) = \frac{|\{2, 3, 5, 6\}|}{8} = \frac{4}{8}
\]

And calculating the significance of the three attributes gives

\[
\sigma_P(Q, a) = \gamma_{a,b,c}(\{e\}) - \gamma_{b,c}(\{e\}) = \frac{1}{8}
\]

\[
\sigma_P(Q, b) = \gamma_{a,b,c}(\{e\}) - \gamma_{a,c}(\{e\}) = 0
\]

\[
\sigma_P(Q, c) = \gamma_{a,b,c}(\{e\}) - \gamma_{a,b}(\{e\}) = 0
\]

Thus, the attribute \( a \) is indispensable, but attributes \( b \) and \( c \) can be dispensed.
Reducts

- For many application problems, it is often necessary to maintain a concise form of the information system.

- One way to implement this is to search for a minimal representation of the original dataset.

- For this, the concept of a reduct is introduced and defined as a minimal subset $R$ of the initial attribute set $C$ such that for a given set of attributes $D$, $\gamma_R(D) = \gamma_C(D)$. 
?