




WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME



Introduction to Data Mining

Marzena Kryszkiewicz




HUMAN CAPITAL
HUMAN - BEST INVESTMENT




EUROPEAN UNION
EUROPEAN
SOCIAL FUND

Project is co-financed by European Union within European Social Fund




WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME




Need for Data Mining...

- Many information systems generate and store huge amounts of data up to thousands Petabytes (e.g. NASA's Earth Observation System (EOSDIS) cumulates data of the volume 10^{15} bytes per year (photograms from the satellite observations)).

2




WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME




Need for Data Mining

- The data may hide knowledge of great value.
- But... the usage of such massive amounts of data is beyond human ability to analyze and elicit meaningful patterns.
- Data mining* (DM) is to discover novel, interesting and useful knowledge from large data resources in an efficient way.

3




WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME




Knowledge Discovery versus Data Mining

- Knowledge Discovery* (KD) is a process consisting of several phases, one of which is *data mining*.

4




WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME




Phases in Knowledge Discovery

- Data collecting (e.g. in databases, data warehouses or in flat files)
- Data transforming (cleaning, filtering, summarizing, attribute selecting, ...)
- Data mining (which results in revealing the knowledge hidden in the mined data)**
- Results' evaluation
- Results' presentation and/or usage

5



WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME



Data Mining Tasks

- Classification
- Prediction
- Clustering (and discovering outliers)
- Association rules, episode rules, and sequential patterns
- Rough Sets
- ...

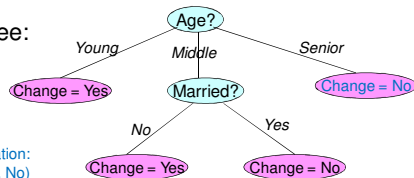
6



DM Tasks: (Eager) Classification

Customer id	Calls' No.	Age	Married	Change
1	Low	Middle	No	Yes
2	High	Middle	Yes	No
3	High	Young	No	Yes
4	Low	Senior	No	No
5	High	Senior	No	No
6	High	Senior	Yes	No
7	Low	Middle	Yes	No

Decision tree:



Subject to classification:
object (Low, Senior, No)

7



DM Tasks: (Eager) Classification

Customer id	Calls' No.	Age	Married	Change
1	Low	Middle	No	Yes
2	High	Middle	Yes	No
3	High	Young	No	Yes
4	Low	Senior	No	No
5	High	Senior	No	No
6	High	Senior	Yes	No
7	Low	Middle	Yes	No

Decision rules

- If (Age = Young), then (Change = Yes).
- If (Age = Senior), then (Change = No).
- If (Age = Middle) and (Married = No), then (Change = Yes).
- If (Age = Middle) and (Married = Yes), then (Change = No).

Subject to classification: object (Low, Senior, No)

8



DM Tasks: (Lazy) Classification

Customer id	Calls' No.	Age	Married	Change
1	Low	Middle	No	Yes
2	High	Middle	Yes	No
3	High	Young	No	Yes
4	Low	Senior	No	No
5	High	Senior	No	No
6	High	Senior	Yes	No
7	Low	Middle	Yes	No

k nearest neighbours

Subject to classification: object (Low, Senior, No).

Its 3 nearest neighbours:

- in decision class Change = Yes: 1 object (#1),
- in decision class Change = No: 2 objects (#4 and #5).

Predicted decision: Change = No

9



DM Tasks: Prediction

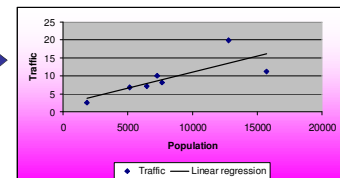
Cell id	Population	Income	TypeOfLand	Traffic
2	1847	1800	9870	2.53
1	5146	2727	1250	6.79
3	6465	1500	1500	7.22
6	7653	1850	1780	8.22
7	7257	1900	2500	9.95
4	15702	1700	3900	11.18
5	12799	1750	5000	19.93

Linear regression:

$$\text{Traffic} = (\alpha \times \text{Population}) + \beta$$

Non-linear regression:

$$\text{Traffic} = (\alpha_0 \times \text{Population}^2) + (\alpha \times \text{Population}) + \beta$$



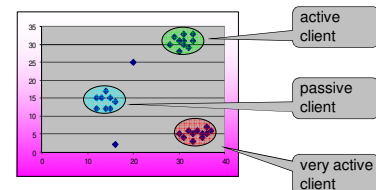
DM Tasks: Clustering...

- *Objects closely located* in the multidimensional space should be assigned to *the same cluster* (group).
- *Objects located far from each other* in the multidimensional space should be assigned to *different clusters*.
- *Outliers* - objects that are distant to (almost) all other objects.
- Outliers - noise or ... potentially very interesting anomalies (e.g. useful for fraud detection).

11



DM Tasks: Clustering



Semantics of the discovered clusters can be provided by a domain expert.

12



DM Tasks: Association Rules

- If a customer buys a PC, then s/he buys Microsoft Windows and Microsoft Office with the support 25% and confidence 75%.
- If patient's symptoms are headache and fever, then s/he suffers from flu with the support 10% and the confidence 82%.

13



DM Tasks: Episode Rules

- Sample episode rule:
 - IF (*link alarm*) AND NEXT(*connection failure*), THEN (*high damage coefficient alarm*) [5] [60] confidence [90%] frequency[151/168].
- Meaning of the rule:
 - In 90% of cases, if a link alarm occurred first and then a connection failure was registered within 5s., then a high damage coefficient alarm occurred within 60s. All three events occurred together 151 times, and two events from the IF part of the rule occurred together 168 times.

14



DM Tasks: Sequential Patterns

- If a client of telecom operator X uses service "to mobiles for 19 PLN", then s/he is likely to change it **within two months** to service "all for 29 PLN", and **within the next two months** – to service "add the second number".

Customer	Time stamp	Event
1	01.05.21	to mobiles for 19 PLN
1	20.06.21	all for 29 PLN
1	12.07.21	all for 39 PLN
1	15.08.21	add the second number
2	01.06.21	to mobiles for 19 PLN
2	01.08.21	all for 29 PLN
2	01.10.21	add the second number
3	01.04.21	all for 29 PLN
...		

15



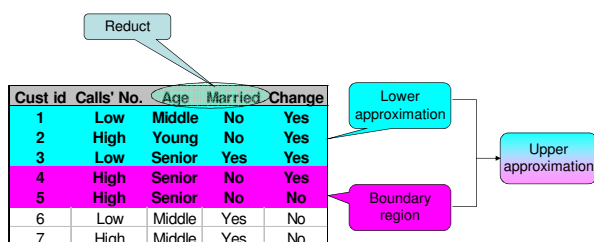
DM Tasks: Functional and Approximate Dependencies

- Sample functional dependency:
 - Social security number determines first name, last name, date of birth and gender of a person.
- Sample approximate dependency:
 - In Polish language, a first name and middle name determines gender with high probability.

16



DM Tasks: Rough Sets



17



Conclusions

- DM is an application-driven field, where research questions tend to be motivated by real world data sets.
- It often involves collaborations among domain experts, computer scientists, statisticians.
- Applying data mining may lead to the discovery of valuable knowledge which, if applied, may considerably increase business profits and decrease the business costs.

18