

Data Mining

cse392

Professor Anita Wasilewska
Computer Science Department
State University of New York
at Stony Brook

Introduction Part Two

- Data Mining Methods (Approaches, Models) and their Functionalities:
- Classification
- Association
- Clustering
- Statistical Models
- Data Mining vs Machine Learning

Summary of Introduction Part One

- **Data Mining** intuitive definition:
 - DM is a process of discovering interesting patterns from large amounts of data
- **DM** is a natural evolution of database technology due to a accumulation of a vast amounts of data
- **DM** is in a great demand, with wide applications

Introduction 1: Summary

- **DM** is a **KDD (Knowledge Discovery in Databases) process** and includes: data cleaning, data integration, data selection, data transformation, **data mining proper**, pattern evaluation, and knowledge presentation
- **DM** can be performed in a variety of information repositories

Approaches to Data Mining (I)

- **Descriptive Methods:** Consist in the creation of mathematical models, algorithms, methods, to extract knowledge descriptions, rules, regularities and other patterns
- **Rough Sets** is the most precise descriptive model
- **Statistical Methods:** They are focused in the creation of statistical models to analyze data. (Regression, Bayesian networks, NN, Clustering)

Descriptive Methods

What is a Description

- Given a relational table (the format of our data) with a non empty set
- A of attributes, and a set V of values of attributes.
- We denote, for any attribute $a \in A$, its value of attribute by $v_a \in V$.
- A description is any conjunction of expressions $a = v_a$, for $a \in A$, $v_a \in V$
- Formally we write it
- $D: \bigwedge a = v_a$

Which kind of Description

- Descriptions of a form **D: $\bigwedge a=Va$**

Are also called **Characteristics** (of a given set of data, called a CONCEPT)

Other forms of descriptions

Are **Characteristic or Discriminant**

Rules (formulas) and we define them as follows.

Discriminant and Characteristic Rules

The RULES have a form of an implication

IF D1 THEN D2

where D1, D2 are descriptions.

A DISCRIMINANT RULE (for a CONCEPT C) is:

IF CHARACTERISTICS (of the concept) THEN CONCEPT

A CHARACTERISTIC RULE (for a CONCEPT C) is:

- **IF CONCEPT THEN CHARACTERISTICS (of the concept)**

Association Rules

- $I = \{i_1, i_2, \dots, i_n\}$ a set of **items**
- **Transaction T**: set of items, **T** is subset of **I**
- **Data Base**: set of transactions
- **An association rule** is an implication of the form : **$X \rightarrow Y$** , where **X, Y** are **disjoint** subsets of **T**
- **Association Problem**: Find rules that have support and confidence greater than user-specified minimum support and minimum confidence

Association Rules Presentations

- **Association rules presentation (predicate presentation)**

Multi-dimensional vs. single-dimensional association

Multi-dimensional:

age(X, "20..29") ^ income(X, "20..29K") → buys(X, "PC")

[support = 2%, confidence = 60%]

Single-dimensional:

buys(x, "computer") → buys(x, "software") [1%, 75%]

Association rules presentation (non-predicate presentation)

Age = 20..29 ∧ income=20..29K → buys=PC (2%, 60%)

Buys=computer → buys=software (1%,75%)

Association A priori Algorithm

- Agrawal (IBM S. José. California), Imielinski (Rutgers).
- It is an intuitive and efficient algorithm to extract associations from a set of transactions
- Algorithm Iterates until the associations obtained don't have the required support

Descriptive Data Mining Problem

A CLASSIFICATION PROBLEM:

- Given a Data Table and a set of records called a CONCEPT,

FIND the smallest and most concise set of its DESCRIPTIONS or RULES describing the concept.

ASSOCIATION PROBLEM:

Given a SET OF TRANSACTIONS,
Find the smallest and most concise set
Of associations, or association rules.

Statistical methods

- Numerical data are needed
- Descriptive statistics is also often used in preprocessing steps to study the sample
- Hypothesis validation and regression analysis are also used as a part of the data mining steps of the process

Classification Problem Requirements

- Decision attribute
- Condition attributes
- Sometimes numerical data but there are DESCRIPTIVE algorithms to deal with any kind of data.
- Maximum length of descriptions
- Minimum support of the rule

Neural Network and Clustering

- **Neural Network** is a classification method: the neural network is trained to obtain classification patterns
- **MUST HAVE NORMALIZED NUMERICAL DATA**

- **Clustering**: form groups of objects without any previous hypothesis
- **MUST HAVE NUMERICAL DATA**

Genetic Algorithms

- **Genetic Algorithm is an optimization method**
- GA should be used when the goal is to find an optimal solution in solution space
- They often are used together with neural networks, or other methods to produce more understandable (optimal) outputs

Clustering Requirements

- Set of attributes (with numerical attributes values)
- Maximum number of clusters
- Number of iterations
- Minimum number of elements in any cluster

DM Functionalities (1)

Concept, concept description

- **Concept** – is defined semantically as any subset of records.
- We often define the by concept attribute **c** and its value **v**
- **In this case the concept description** is syntactically written as : **c=v** and we define:
- **CONCEPT={objects(records): c=v}**
- For example: *climate=wet* (description of the concept)
- **CONCEPT={records: climate=wet}**
- **We also use words: CLASS, class attribute**
for concept, concept attribute

REMEMBER: all definitions are relative to the database we deal with.

DM Functionalities (2)

Concept characteristics

- **Concept C characteristics** is a set of attributes a_1, a_2, \dots, a_k , and their respective values v_1, v_2, \dots, v_k that are characteristic for a given concept C , i.e.
- $\{\text{objects(records): } a_1=v_1 \ \& \ a_2=v_2 \ \& \ \dots \ \& \ a_k=v_k\} \wedge$
(intersection with) C is a non empty set
- **Concept C characteristics description** is then syntactically written as
 $a_1=v_1 \ \& \ a_2=v_2 \ \& \ \dots \ \& \ a_k=v_k$

Characterization

Characteristic Rule

- ***Characterization describes a process which aim is to find rules that describe characteristic properties of a concept. They are called characteristic rules and take the form***

If concept then characteristics

- ***C=1 → A=1 & B=3*** **25%** (support: there are 25% of the records for which the rule is true)
- ***C=1 → A=1 & B=4*** **17%**
- ***C=1 → A=0 & B=2*** **16%**

Discrimination

Discriminant Rule

- *Discrimination is a 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 (other classes). These rules are called **discriminant rules** and have a form*

If characteristics then concept

- **A=0 & B=1 → C=1** 33% 83% (support, confidence: the conditional probability of the concept given the characteristics)
- **A=2 & B=0 → C=1** 27% 80%
- **A=1 & B=1 → C=1** 12% 76%
- Discriminant rule can be good even if it has a low support (and high confidence)

Data Mining Functionalities (3)

- **Classification and Classification Prediction** is called **Supervised Learning**

It consists of finding models (**rules**) that describe (**characterize**) or/ and distinguish (**discriminate**) classes or concepts for future prediction

Example: classify countries based on climate (characteristics), or classify cars based on gas mileage and use it to predict classification of a new car

Models, algorithms, methods: **decision-tree, neural network, Bayes Network, Rough Sets, genetic algorithms**

Presentation of results: characteristic and /or discriminant rules, converged network (neural, Bayes)

Data Mining Functionalities (4)

- **Statistical Prediction** – is often used to predict some unknown or missing numerical values
- **Cluster analysis (statistical method)**

Class label is unknown: the algorithms group data to form new classes- it is called **unsupervised learning**

For example: cluster houses to find distribution patterns

Clustering is based on the principle:

maximizing the intra-class (clusters) similarity and **minimizing** the interclass similarity

Data Mining Functionalities (5)

- **Outlier analysis (statistical)**

Outlier: a data object that does not comply with the general behavior of the data

It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis

Data Mining Functionalities (6)

- **Trend and evolution analysis (statistical)**
 - Trend and deviation: regression analysis
 - Sequential pattern mining, periodicity analysis
 - Similarity-based analysis
- **Other pattern-directed or statistical analyses**

Classification

- Given a set of objects (**concept, class**) described by a concept attribute or a set of attributes, a classification algorithm builds a set of **discriminant and /or characterization rules** (or other descriptions) in order to be able, as the next step, to classify unknown sets of objects
- This is also called a **supervised learning**

Classification Methods, Models, Algorithms

(Chapter 7)

- Decision Trees (ID3, C4.5)
- Neural Networks
- Rough Sets
- Bayesian Networks
- Genetic Algorithms

Association Model (chapter 6)

Problem Statement

- $I = \{i_1, i_2, \dots, i_n\}$ a set of **items**
- **Transaction T**: set of items, **T** is subset of **I**
- **Data Base**: set of transactions
- An association rule is an implication of the form : **$X \rightarrow Y$** , where **X, Y** are **disjoint** subsets of **T**
- **Problem**: Find rules that have support and confidence greater than user-specified minimum support and minimum confidence

Parameters of the Association Rules

- **Confidence:** a rule $X \rightarrow Y$ holds in the database D with a confidence c if the $c\%$ of transactions in D that contain X also contain Y
- **Support:** a rule $X \rightarrow Y$ has a support s in D if $s\%$ of transactions contain XUY

Association Rules

- **Association rules presentation (predicate presentation)**

Multi-dimensional vs. single-dimensional association

Multi-dimensional:

$\text{age}(X, \text{"20..29"}) \wedge \text{income}(X, \text{"20..29K"}) \rightarrow \text{buys}(X, \text{"PC"})$

[support = 2%, confidence = 60%]

Single-dimensional:

$\text{buys}(x, \text{"computer"}) \rightarrow \text{buys}(x, \text{"software"})$ [1%, 75%]

Association rules presentation (non-predicate presentation)

Age = 20..29 \wedge income=20..29K \rightarrow buys=PC (2%, 60%)

Buys=computer \rightarrow buys=software (1%,75%)

Association Analysis

- The problem of association rule discovery can be split into three sub-problems:
 1. Find the set of products (records, transactions) that have the minimum support required
 2. Find **frequent sets**
 3. Use the frequent sets to generate rules

Clustering

- Database segmentation
- Given a set of objects (records) the algorithm obtains a division of the objects into clusters in which the distance of objects inside a cluster is minimal and the distance among objects of different clusters is maximal
- Unsupervised learning

Other Statistical Methods

- Regression
- Temporal Series

.....

Major Issues in Data Mining (1)

- Mining methodology and user interaction

Mining different kinds of knowledge in databases

Interactive mining of knowledge at multiple levels of abstraction

Incorporation of background knowledge

Data mining query languages and ad-hoc data mining

Expression and visualization of data mining results

Major Issues in Data Mining (2)

Handling noise and incomplete data

Pattern evaluation: the interestingness problem

Performance and scalability

Efficiency and scalability of data mining algorithms

Parallel, distributed and incremental mining methods

Major Issues in Data Mining (3)

- **Issues relating to the diversity of data types**
 - Handling relational and complex types of data
 - Mining information from heterogeneous databases and global information systems (WWW)
- **Issues related to applications and social impacts**
 - Application of discovered knowledge
 - Domain-specific data mining tools
 - Intelligent query answering
 - Process control and decision making
 - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
 - Protection of data security, integrity, and privacy

Preprocessing

- Select, integrate, and clean the data
- Decide which kind of patterns are needed
- Decide which algorithm is the best . It depends on many factors
- Prepare data for algorithms

Implementation Preparation (1)

- Identify the problem to be solved.
- Study it in detail
- Explore the solution space,
- Find one acceptable solution (feasible to implement)
- Specify the solution
- Prepare the data

Data Preparation (2)

- Remember GIGO! (garbage in garbage out)
- Add some data, if necessary
- Structure the data in a proper form
- Be careful with incomplete and noisy data

Some implementation preparation rules to follow

- Select the problem
- Specify the problem
- Study the data
- The problem must guide the search for tools and technologies
- Search for the simplest model (algorithm, method)
- Define for each data the solution is valid, where it is not valid at all and where it is valid with some constraints

Studying the data

- The surrounding world consists of objects , (data) and the DM problem is to find the relationships among objects
- **The objects** are characterized by properties
- (attributes, values of attributes)
that have to be analyzed
- **The results** (rules, descriptions) are valid
(true) under certain circumstances (data) and in
certain moments (available data at the moment)

Measures

- Type of data decides a way in which data are analyzed and preprocessed
 - Names (attributes)
 - Categories, classes, class attributes
 - Ordered values of attributes
 - Intervals of values of attributes
 - Types of values of attributes

General Types of Data

- Generally we distinguish:
 - Quantitative Data
 - Qualitative Data
- Bivaluated: often very useful
- Null Values are not applicable
- Missing data are usually not acceptable

What to take into account

- Eliminate redundant records
- Eliminate out of range values of attributes
- **Decide a generalization level**
- Consistency

Other preprocessing tasks

- Generalization vs specification
- Discretization
- Sampling (of records)
- Reducing number of attributes at the preprocessing stage

Preprocessing Summary

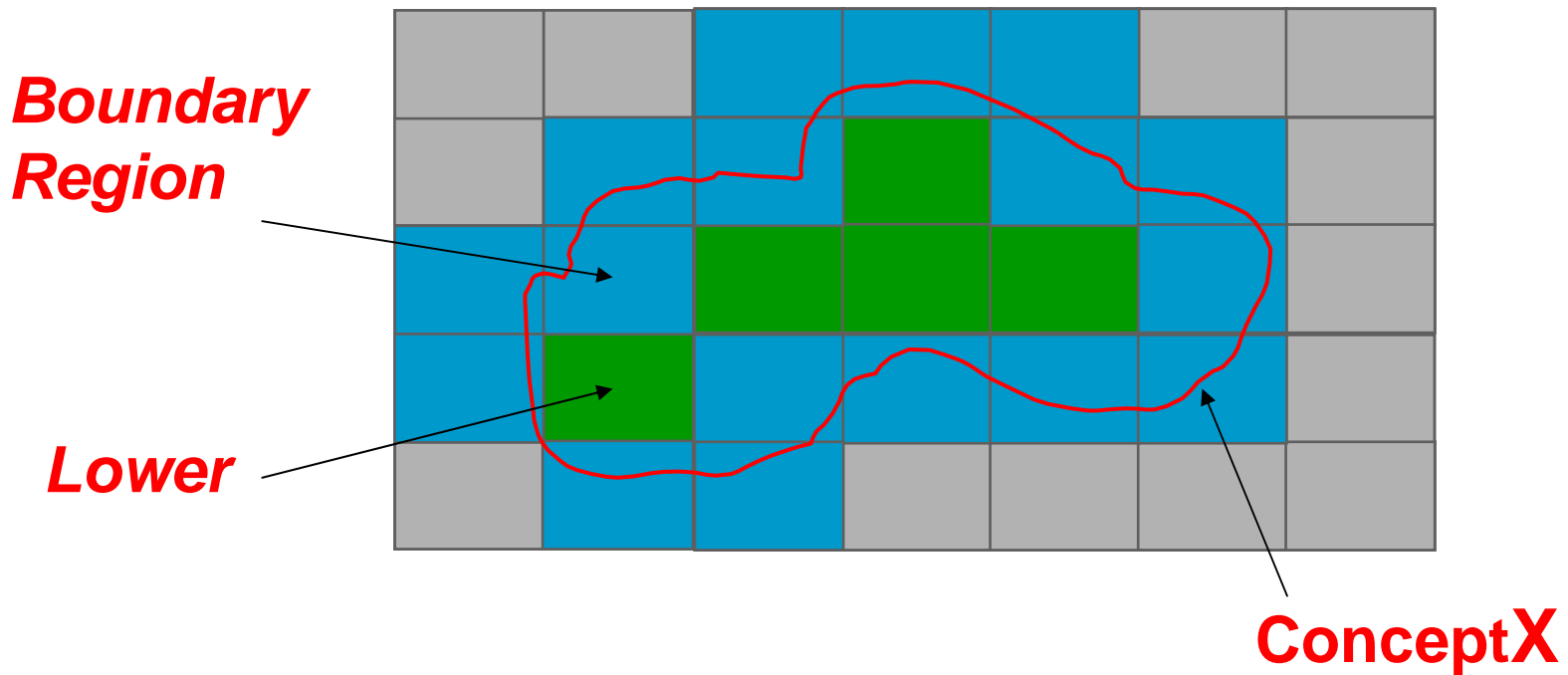
- The preprocessing is usually required and is an essential part of the DM process
- If preprocessing is not performed patterns obtained could be of no use.
- It is a tedious task that could even take more time than DM proper

APPROACHES TO DATA MINING

Rough Sets

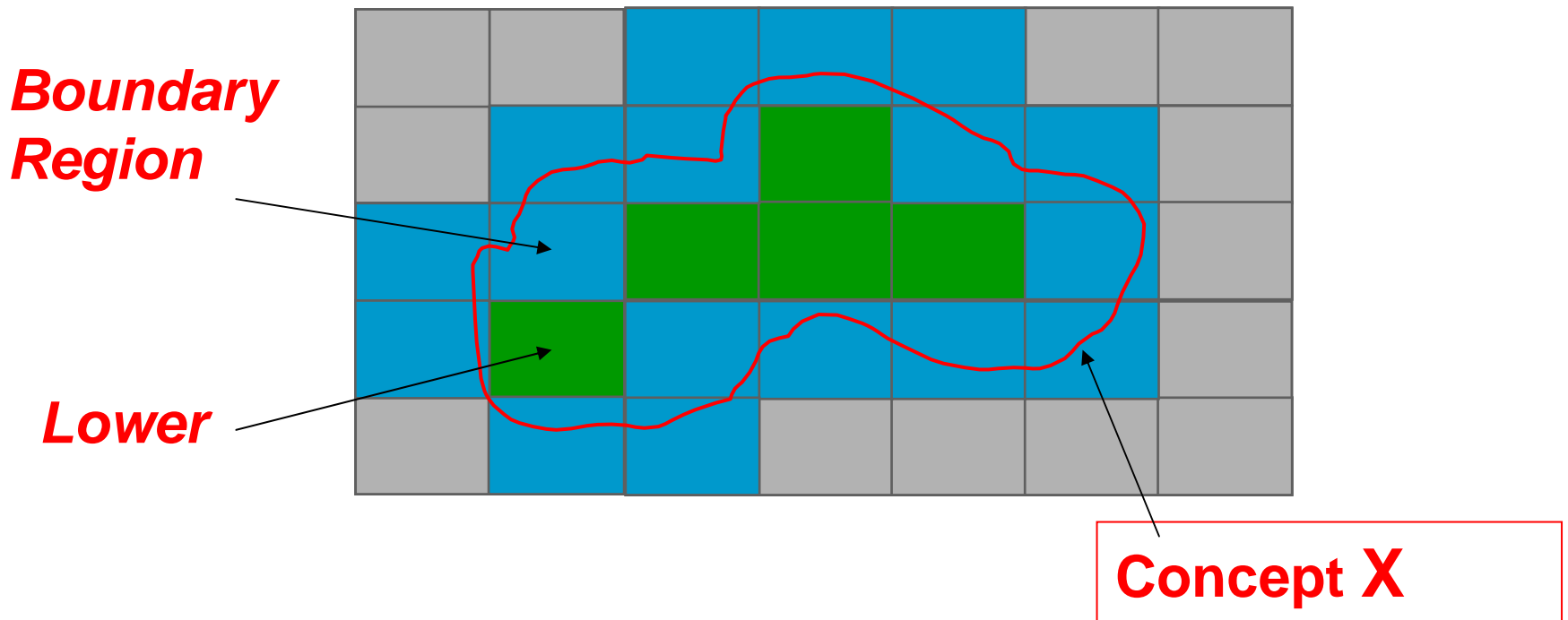
- Approximation space $A=(U,IND(B))$:
 - *Lower Approximation* $\underline{X}_B = \{o \in U / [o] \subseteq X\}$
 - *Upper Approximation* $\overline{X}_B = \{o \in U / [o] \cap X \neq \emptyset\}$
 - *Boundary Region* $Bnd(X)_B = \overline{X}_B - \underline{X}_B$
 - *Positive Region*: $POS_B(D) = \bigcup \{\overline{X} : X \in IND(D)\}$

Rough Sets



$$\text{Boundary} + \text{Lower} = \text{Upper}$$

Rough Sets



$$\text{Boundary} + \text{Lower} = \text{Upper}$$

Variable Precision Rough Set Model



$$c(X, Y) = \begin{cases} 0 \\ 1 - \mathit{card}(X \cap Y) / \mathit{card}(X) \end{cases} \quad \text{if } \begin{cases} \mathit{card}(X) = 0 \\ \mathit{card}(X) > 0 \end{cases}$$

51

Rough Sets in SQL

```
Begin UPPER
  setdb(dbName);
  exec(conn, "BEGIN");

  "DECLARE classes CLASSES FOR
  SELECT C1, . . . . ., CN, D, COUNT (*) AS cnt
  FROM R
  GROUP BY C1, . . . . ., CN, D
  ORDER BY C1, . . . . ., CN, D, CNT desc");

  while not_end_records() do
    equ_class=exec("FETCH 1 IN cursor");
    first_decision_value=get_value(equ_class("D"));
    insert(equ_class, upper[first_decision_value]);
    while (equ_class == exec("FETCH 1 IN cursor")) do
      decision_value=get_value(equ_class("D"));
      insert(equ_class, upper[first_decision_value]);
    end while
  end while
End UPPER
```