

CSE5230/DMS/2004/3

# Data Mining - CSE5230

## Pre-processing for Data Mining

CSE5230 - Data Mining, 2004 Lecture 3.1

# Lecture Outline

- ◆ **Data Preparation**
  - ❖ Accessing data
  - ❖ Data characterization
  - ❖ Data selection
  - ❖ Useful operations for data clean-up and conversion
  - ❖ Integration Issues
- ◆ **Data Modeling**
  - ❖ Motivation
  - ❖ Ten Golden Rules
  - ❖ Object modeling
  - ❖ Data Abstraction
  - ❖ Working with Meta Data

CSE5230 - Data Mining, 2004 Lecture 3.2

## Lecture Objectives

- ◆ **By the end of this lecture, you should be able to:**
  - ❖ Explain why data preparation is necessary before data mining can commence
  - ❖ Give examples of useful operations during the process of data clean-up and conversion, and show how these operations are applied in specific cases
  - ❖ Explain why modeling is important in data preparation for data mining, and give examples of such models
  - ❖ Explain the notion of data abstraction and why it is useful

## Data Preparation for Data Mining - 1

- ◆ Before starting to use a data mining tool, the data has to be transformed into a suitable form for data mining
- ◆ Many new and powerful data mining tools have become available in recent years, but the law of GIGO still applies:

**Garbage In → Garbage Out**

- ◆ Good data is a prerequisite for producing effective models of any type

## Data Preparation for Data Mining - 2

- ◆ Data preparation and data modeling can therefore be considered as setting up the proper environment for data mining
- ◆ Data preparation will involve
  - ❖ Accessing the data (transfer of data from various sources)
  - ❖ Integrating different data sets
  - ❖ Cleaning the data
  - ❖ Converting the data to a suitable format

## Accessing the data - 1

- ◆ Before data can be identified and assessed, two major questions must be answered:
  - ❖ Is the data accessible?
  - ❖ How does one get it?
- ◆ There are many reasons why data might not be readily accessible, particularly in organizations without a data warehouse:
  - ❖ legal issues
  - ❖ departmental access
  - ❖ political reasons
  - ❖ data format
  - ❖ connectivity
  - ❖ architectural reasons
  - ❖ timing

## Accessing the data - 2

- ◆ **Transferring from original sources**
  - ❖ may have to access from: high density tapes, email attachments, FTP as bulk downloads
- ◆ **Repository types**
  - ❖ **Databases**
    - » Obtain data as separate tables converted to flat files (most databases have the facility).
  - ❖ **Word processors**
    - » Text output without any formatting would be the best
  - ❖ **Spreadsheets**
    - » Small applications/organizations will store data in spreadsheets. Already in row/column format, so easy to access. Most problems due to inconsistent replications
  - ❖ **Machine to Machine**
    - » Possible problems due to different computing architectures
  - ❖ **The Web**
    - » **Structured, semi-structured or structureless data**
      - HTML, XML, free text, images, video, audio, etc.

## Data characterization - 1

- ◆ **After obtaining all the data streams, the nature of each data stream must be characterized**
  - ❖ This is not the same as the data format (i.e. field names and lengths)
- ◆ **Detail/Aggregation Level (Granularity)**
  - ❖ all variables fall somewhere between detailed (e.g. transaction records) and aggregated (e.g. summaries)
  - ❖ in general, detailed data is preferred for data mining
  - ❖ the level of available in a data set determines the level of detail that is possible in the output
  - ❖ usually the level of detail of the input stream must be at least one level below that required of the output stream

## Data characterization - 2

### ◆ Consistency

- ❖ Inconsistency can defeat any modeling technique until it is discovered and corrected
  - » different things may have the same name in different systems
  - » the same thing may be represented by different names in different systems
  - » inconsistent data may be entered in a field in a single system, e.g. auto\_type:

Merc, Mercedes, M-Benz, Mrcds

## Data characterization - 3

### ◆ Pollution

- ❖ Data pollution can come from many sources. One of the most common is when users attempt to stretch a system beyond its intended functionality, e.g.
  - » “B” in a gender field, intended to represent “Business”. Field was originally intended to only even be “M” or “F”.
- ❖ Other sources include:
  - » copying errors (especially when format incorrectly specified)
  - » human resistance - operators may enter garbage if they can't see why they should have to type in all this “extra” data

## Data characterization - 4

### ◆ Objects

- ❖ precise nature of object being measured by the data must be understood
  - » e.g. what is the difference between “consumer spending” and “consumer buying patterns”?

### ◆ Domain

- ❖ Every variable has a domain: a range of permitted values
- ❖ Summary statistics and frequency counts can be used to detect erroneous values outside the domain
- ❖ Some variables have conditional domains, violations of which are harder to detect
  - » e.g. in a medical database a diagnosis of ovarian cancer is conditional on the gender of the patient being female

## Data characterization - 5

### ◆ Default values

- ❖ if the system has default values for fields, this must be known. Conditional defaults can create apparently significant patterns which in fact represent a lack of data

### ◆ Integrity

- ❖ Checking integrity evaluates the relationships permitted between variables
  - » e.g. an employee may have multiple cars, but is unlikely to be allowed to have multiple employee numbers
- ❖ related to the domain issue

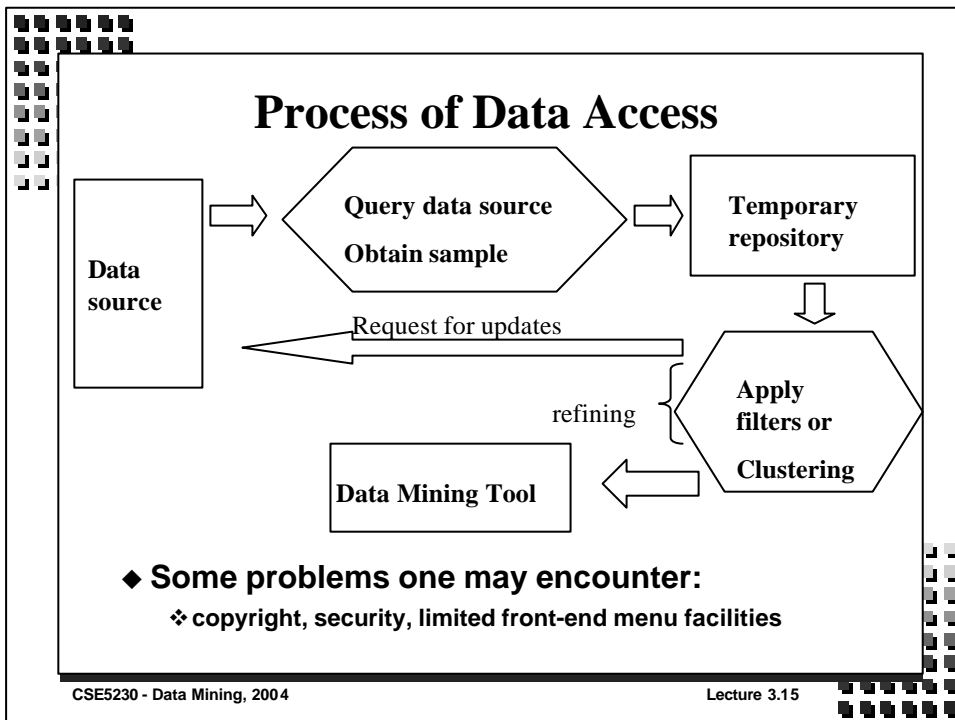
## Data characterization - 6

### ◆ Duplicate or redundant variables

- ❖ redundant data can easily result from the merging of data streams
- ❖ occurs when essentially identical data appears in multiple variables, e.g. “date\_of\_birth”, “age”
- ❖ if not actually identical, will still slow building of model
- ❖ if actually identical can cause significant numerical computation problems for some models — even causing crashes

## Extracting part of the available data

- ◆ In most cases original data sets would be too large to handle as a single entity. There are two ways of handling this problem:
  - ❖ Limit the scope of the the problem
    - » concentrate on particular products, regions, time frames, dollar values etc. OLAP can be used to explore data prior to such limiting
    - » if no pre-defined ideas exist, use tools such as Self-Organizing Neural Networks to obtain an initial understanding of the structure of the data
  - ❖ Obtain a representative sample of the data
    - » Similar to statistical sampling
- ◆ Once an entity of interest is identified via initial analysis, one can follow the lead and request more information (“walking the data”)



- ## Useful operations during data access/preparation - 1
- ◆ **Text Standardization**
    - ❖ convert all text to upper- or lowercase
      - » This helps to avoid problems due to case differences in different occurrences of the same data (e.g. the names of people or organizations)
    - ❖ remove extraneous characters e.g. !&%\$#@ etc.
    - ❖ Remove punctuation (in some applications)
    - ❖ Word stemming (for text mining/text retrieval applications)
  - ◆ **Concatenation**
    - ❖ combine data spread across multiple fields e.g. names, addresses. The aim is to produce a unique representation of the data object
  - ◆ **Representation formats**
    - ❖ some sorts of data come in many formats
      - » e.g. dates: 12/05/03, 05-Dec-03, 20031205
    - ❖ transform all to a single, simple format
      - » and one that is future-proof (e.g. Y2K)
- CSE5230 - Data Mining, 2004 Lecture 3.16

## Useful operations during data access/preparation - 2

- ◆ **Abstraction**
  - ❖ it can sometimes be useful to reduce the information in a field to simple yes/no values: e.g.
    - » flag people as having a criminal record, rather than having a separate category for each possible crime
- ◆ **Unit conversion**
  - ❖ choose a standard unit for each field and enforce it: e.g.
    - » \$A, € ® \$US
    - » yards, feet ® metres
  - ❖ This can have dramatic consequences if not done!
    - » The loss of the Mars Climate Orbiter (admittedly not a DM example<sup>©</sup>)
- ◆ **Exclusion**
  - ❖ data processing takes up valuable computation time, so one should exclude unnecessary or unwanted fields where possible
  - ❖ fields containing bad, dirty or missing data may also be removed

## Useful operations during data access/preparation - 3

- ◆ **Numeric Encoding**
  - ❖ Many data mining tools require numerical input data (e.g. neural networks) — but not all data is numeric! Data variables can be:
    - » **Numeric** (integer or floating point), e.g.
      - 1, 2.6, 56.7, 10e-7,...
    - » **Nominal**: the names of categories or classes, e.g.
      - cat, sheep, goldfish, duck, elephant, goat,...
    - » **Ordinal**: samples from an *ordered* list of categories, e.g.:
      - Terrible, Bad, OK, Good, Excellent
  - ❖ Numerical variables must sometimes be normalized (e.g. to the range [0,1]) before being presented to a data mining tool
    - » e.g. to prevent saturation of a neural network node

## Useful operations during data access/preparation - 4

- ◆ **Numeric Encoding cont.**
  - ❖ Care must be taken when encoding nominal variables in numeric form
    - » Do not introduce relationships that are not present in the original data. For example, the mapping
      - cat ® 1, sheep ® 2, goldfish ® 3, duck ® 4, elephant ® 5, goat ® 6, ...implies that the “distance” between duck and elephant is 1, whereas between sheep and goat it is 4
      - This is not a real relationship – it is an artefact introduced by the encoding
    - » One solution is to convert the single original variable into multiple binary variables, one for each category
  - ❖ Care must also be taken with ordinal variables, particularly when one value implies the presence of others.
    - » Consider a variable describing how often someone watches television, e.g.
      - At least once a day, at least once a week, at least once a month, etc.

CSE5230 - Data Mining, 2004

Lecture 3.19

## Data integration issues - 1

- ◆ **Multi-source**
  - ❖ Oracle, Excel, Informix, DB2, MySQL, etc.
    - » Standardized database drivers help (e.g. ODBC)
    - » Data Warehousing helps
- ◆ **Multi-format**
  - ❖ relational databases, hierarchical structures, XML, HTML, free text, etc.
- ◆ **Multi-platform**
  - ❖ DOS, MS Windows, UNIX, etc.
    - » Issues such as end-of-line characters, bigendian/littleendian binary file formats, etc.
- ◆ **Multi-security**
  - ❖ copyright, privacy, personal records, government data, etc.

CSE5230 - Data Mining, 2004

Lecture 3.20

## Data integration issues - 2

### ◆ Multimedia

- ❖ text, images, audio, video, etc.
  - » Features of interest must be defined and extracted from the raw data
  - » Cleaning might be required when formats are inconsistent

### ◆ Multi-location

- ❖ LAN, WAN, dial-up connections, etc.

### ◆ Multi-query

- ❖ whether query format is consistent across data sets
  - » again, database drivers useful here
- ❖ whether multiple extractions are possible
  - » i.e. whether large number of extractions are possible — some systems do not allow batch extractions, one has to obtain records individually, etc.

## Modeling Data for Data Mining - 1

### ◆ A major reason for preparing data is so that mining can discover *models*

### ◆ What is modeling?

- ❖ it is assumed that the data set (available or obtainable) contains information that would be of interest if only we could understand what was in it
- ❖ Since we don't understand the information that is in the data just by looking at it, some tool is needed which will turn the information lurking in the data set into an understandable form

## Modeling Data for Data Mining - 2

- ◆ Object is to transfer the raw data structure to a format that can be used for mining
- ◆ The models created will determine the type of results that can be discovered during the analysis
- ◆ With most current data mining tools, the analyst has to have some idea what type of patterns can be identified during the analysis, and model the data to suit these requirements
- ◆ If the data is not properly modeled, important patterns may go undetected, thus undermining the likelihood of success

## Modeling Data for Data Mining - 3

- ◆ To make a model is to express the relationships governing how a change in a variable or set of variables (inputs) affects another variable or set of variables (outputs)
- ◆ we also want information about the reliability of these relationships
- ◆ the expression of the relationships may have many forms:
  - ❖ charts, graphs, equations, computer programs

## **Ten Golden Rules for Building Models -1**

- 1. Select clearly defined problems that will yield tangible benefits**
- 2. Specify the required solution**
- 3. Define how the solution is going to be used**
- 4. Understand as much as possible about the problem and the data set (the domain)**
- 5. Let the problem drive the modeling (i.e. tool selection, data preparation, etc.)**

## **Ten Golden Rules for Building Models -2**

- 6. State any assumptions**
- 7. Refine the model iteratively**
- 8. Make the model as simple as possible - but no simpler (paraphrasing Einstein)**
- 9. Define instability in the model (critical areas where change in output is very large for small changes in inputs)**
- 10. Define uncertainty in the model (critical areas and ranges in the data set where the model produces low confidence predictions/insights)**

## Object modeling

- ◆ The main approach to data modeling assumes an object-oriented framework, where information is represented as *objects*, their *descriptive attributes*, and *relationships* that exist between object *classes*.
- ◆ Examples object classes
  - ❖ Credit ratings of **customers** can be checked
  - ❖ **Contracts** can be renewed
  - ❖ **Telephone calls** can be billed
- ◆ Identifying attributes
  - ❖ In a medical database system, the class **patient** may have the attributes *height*, *weight*, *age*, *gender*, etc.

## Data Abstraction

- ◆ Information can be abstracted such that the analyst can initially get an overall picture of the data and gradually expand in a top-down manner
- ◆ Will also permit processing of more data
- ◆ Can be used to identify patterns that can only be seen in grouped data, e.g. group patients into broad age groups (0-10, 10-20, 20-30, etc.)
- ◆ Clustering can be used to fully or partially automate this process

## Working with Metadata - 1

- ◆ Traditional definition of metadata is “data *about data*”
- ◆ Some data miners include “data *within data*” in the definition
- ◆ Example: Deriving metadata from dates:
  - ❖ identifying seasonal sales trends
  - ❖ identifying pivot points for some activity
    - » e.g. happens on the 2nd Sunday of July
  - ❖ Note: “July 4th, 1976” is potentially:  
7th Month of the Year, 4th Day of the Month, 1976,  
Sunday, 1st Day of the Week, 186th Day of Year, 1st  
Quarter of the Financial Year, Winter (in the southern  
hemisphere), etc.

## Working with Metadata - 2

- ◆ Metadata can also be derived from
  - ❖ ID numbers
  - ❖ passport numbers
  - ❖ drivers' licence numbers
  - ❖ post codes
  - ❖ etc.
- ◆ data can be modeled to make use of these
- ◆ Example: Metadata derived from addresses and names
  - ❖ identify the general make up of a shop's clients
    - » e.g. correlate addresses with map data to determine the distances customers travel to come to the shop



## References

- ◆ Dorian Pyle, “Data Preparation for Data Mining”, Morgan Kaufmann Publishers, 1999.