

Lecture 2: Data Issues in Data Mining

- Data size and complexity
- Data sources
- Measurements and types of data
- Formats of data
- Data warehousing
- Meta-data / Describing data
- *Real world data is dirty*

Data size and complexity

- *We are drowning in data, but starving of knowledge* (J. Han)
- Automated data collection and mature database technology (allows data to be stored efficiently, cheap, persistent; in databases, data warehouses and other repositories)
- Large and massive data sets
 - Millions to billions of records
 - Tens to thousands of attributes (or variables)
 - Data is rarely collected for data mining purposes! (rather for online transaction processing)
- A lot of data is *write only*

Data sources

- Relational databases (transactional data, mostly normalised into many tables, with keys between them, continuous and frequent updates)
- Data warehouses (decision support data, processed and cleaned, historical data, aggregated, updated at certain intervals)
- Internet (click-stream data, log files, HTML, XML, e-mails)
- Files (portable text (e.g. CSV) or non-portable, proprietary binary files)
- Scientific instruments and experiments (astronomy, genomics, seismology, physics, chemistry, etc.)

Measurements and types of data (1)

- Numerical data
 - Integer, floating-point, binary
 - Interval, ratio
 - Non-scalar (e.g. velocity – speed and direction)
- Non-numerical data
 - Nominal data (just naming things, e.g. names)
 - Categorical data (grouping things, e.g. postcodes, university course codes)
 - Ordinal data (ordering things, e.g. wine tasting)
- Series data
 - Ordering is an important feature (otherwise not series data)
 - One attribute must always be monotonic (increasing or decreasing)
 - Most common are *time series* (others are e.g. geographic location)

Measurements and types of data (2)

- Multimedia data (different standards, often compressed)
 - Images
 - Video
 - Audio
- Different mappings and conversions between data types are possible and often needed
- Different data mining techniques can handle different types of data
(or are restricted to certain types of data)

Formats of data

- Structured data
(relational database tables, integrated data warehouses, images, video, audio, etc.)
- Semi-structured data
(XML, HTML, e-mails, log files)
- Free format data
(free format text – ASCII, Unicode)

Data warehousing (1)

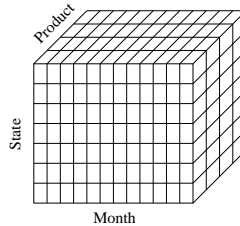
- A data warehouse is a decision support database that is maintained separately from an organisation's operational database(s)
- Provides a solid platform of consolidated, historical data for analysis
- Organised around major subjects, like customers, products, or sales (and provides a simple and concise view around these entities)
- Often constructed by cleaning, standardising and integrating multiple heterogeneous data sources
(to ensure consistency in coding, naming, measuring, etc.)
(→ Lectures 3 and 4)

Data warehousing (2)

- Longer time horizon than operational systems (used for transaction processing) (historical data is important for analysis)
- Contains a time element
(e.g. new data is loaded into a data warehouse on a weekly or monthly basis)
- Only two operations on data: Initial loading and querying of data (read) (while transaction processing systems have reads, writes, updates)
- Differences between a data warehouse and an operational database
 - Database: OLTP (On-Line Transaction Processing)
 - Data warehouse: OLAP (On-Line Analytic Processing)
 - Separate data warehouse due to performance, data representation, consistency, integration, and data quality

Data warehousing (3)

- Data warehouse architecture
 - Data cubes (multi-dimensional aggregated data views)
 - Dimension tables (details of the dimensions) and fact tables (values and names of the facts, e.g. `items_sold`, as well as keys into dimension tables)
 - Data is stored at different levels of details (e.g. `country/state/city`, or `item/item_group/item_category`)



Slide 9 of 14

MATH3346 – Data Mining, S2/2006

Data warehousing (4)

- For data warehouses, a *multi-dimensional data model* is most popular (compared to entity-relationship model for relational databases)
- Implemented as
 - Star schema
(a large central *fact* table containing bulk of the data, and a set of smaller *dimension* tables)
 - Snowflake schema
(variant of star schema with normalised dimension tables)
 - Fact constellation schema
(multiple fact tables who share dimension tables), can be viewed as a collection of star schemas

Slide 10 of 14

MATH3346 – Data Mining, S2/2006

Data warehousing (5)

- Data warehouse operations
 - Roll-up (summarise data)
 - Drill-down or roll-down (get detailed view)
 - Slice and dice (project and select)
 - Pivot (rotate), reorient the cube, 2D to 2D visualisation
- Applications of data warehousing
 - Information processing (basic statistics, reporting, tables, charts, graphs, Web-based reporting, etc.)
 - Analytic processing (further drill down, multi-dimensional analysis, on both summarised and detailed data)
 - Data mining: Use a clean, stable, high-quality source for data mining algorithms

Slide 11 of 14

MATH3346 – Data Mining, S2/2006

Meta-data / Describing data

- Meta-data: Data about data
 - Structure of the data (types, names, format, etc.)
 - Summary statistics (e.g. missing values, histograms)
 - Quality of the data
 - Information about pre-processing done
 - Data source and owner, business information, charging policies
 - Information about data access, retrieval, updates
- Sometimes called *data dictionary*
(within a large organisation, or between organisations, e.g. health departments)
- Stored in a database, as XML schema, etc.
- Meta-meta-data: Data about meta-data...?

Slide 12 of 14

MATH3346 – Data Mining, S2/2006

Real world data is dirty (1)

- Various sources of errors
 - Misinterpretation of the data
 - Errors during data entry
 - Missing data
 - Out-of-date data
- Personal information (names and addresses) are especially prone to data entry errors
- A great effort is often needed to *clean* and *standardise* raw data (data preprocessing)

Real world data is dirty (2)

- What does *dirty data* mean?
 - Incomplete data
(missing attributes, missing attribute values, only aggregated data, etc.)
 - Inconsistent data
(different coding, impossible values or out-of-range values)
 - Noisy data
(data containing errors, outliers, not accurate values)
- For quality mining results, quality data is needed
- Transactions databases systems should be designed with data mining in mind
- Pre-processing is an important step for successful data mining