ANU MLSS 2010: Data Mining

Part 1: Introduction, data mining challenges, and data issues for data mining

Data Mining module outline

- Part 1:
  - Very short introduction to data mining
  - Data mining process
  - Challenges in data mining
  - Data cleaning, integration and pre-processing
- Part 2:
  - Association rule mining
- Part 3:
  - Data mining application techniques: data streams and link mining
  - Privacy aspects of data mining
  - References and Resources

Many slides are based on 'Data Mining: Concepts and Techniques' by J. Han and M. Kamber, see: http://www.cs.uiuc.edu/~hanj/bk2/

Very short introduction to data mining (1)

- Many government agencies, businesses, and research projects collect massive amounts of data
  - Amazon.com: 42 Terabytes, YouTube (2006): 45 Terabytes, ChoicePoint: 250 Terabyte (information about 250 million people), AT&T: 323 Terabytes (1.9 Trillion phone records), etc. etc.
  - The sizes of databases increases exponentially
  - Source: http://www.focus.com/fyi/operations/10-largest-databases-in-the-world/
  - Also: http://www.businessintelligencelowdown.com/2007/02/top_10_largest_.html
- Questions arise:
  - Is there any new, unexpected and potentially useful information contained in this data?
  - Can we use historical data to predict future outcomes (such as customer purchase behaviour, detect fraud, predict terrorism, etc.)?

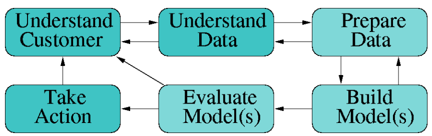
Very short introduction to data mining (2)

- Data mining involves:
  - Database and data warehouse technologies
  - *Machine learning and artificial intelligence*
  - *Statistics* and numerical mathematics
  - Parallel and high-performance computing
  - Visualisation
  - Privacy technologies
- Data mining techniques:
  - *Data cleaning, pre-processing, and integration*
  - Cluster analysis (unsupervised learning)
  - *Rule discovery (association mining)*
  - Classification and prediction (supervised learning)

Very short introduction to data mining (3)

- Application specific data mining techniques:
  - Spatial and temporal data mining
  - Text and Web mining
  - Outlier detection
  - *Data stream*, time series and sequence mining
  - Multimedia data mining (audio, images, video)
  - *Network, link* and graph mining
  - *Privacy preservation*
- Data mining is applied in many areas, including:
  - Bioinformatics and health
  - Governments (statistics, census, taxation, social welfare)
  - Credit card and insurance companies
  - Terror, crime and fraud detection
  - Networking and telecommunications
  - Marketing and retail

The data mining / KDD process



- Data mining is an interactive process
- Data mining = *Build Model(s)*
- Typically up to 90% of time and effort are spent in the first three steps!

(Follows: *CRoss Industry Standard Process for Data Mining*, http://www.crisp-dm.org/)

Major challenges in data mining

- Data size
  - Size of data collections grows more than linear, doubling around every 18 months (similar to Moore's law of CPU speed)
  - Scalable algorithms are needed
- Data complexity
  - Different types of data (databases, free text, HTML, XML, multimedia)
  - Dimensionality of the data increases (more attributes)
  - The *curse of dimensionality* affects many algorithms (for example find nearest neighbours in high dimensions)
- Privacy and confidentiality
  - Data mining can reveal details about people which is not available otherwise
  - Linking and matching data is especially critical / controversial

Ten grand challenges in data mining (U. Fayyad)

- Technical challenges
  - How does the data grow?
  - Scalability (of algorithms)
  - Complexity/understandability trade-off
  - Interestingness
  - A theory for what we do
- Pragmatic challenges
  - Where is the data?
  - Embedding algorithms and solutions within operational systems
  - Integrating domain knowledge
  - Managing and maintaining models
  - Effectiveness measurement

(Source: http://www.acm.org/sigs/sigkdd/explorations/, vol 5, no 2, Dec. 2003)

Data size and complexity

- *We are drowning in data but starving of knowledge*

  (Jiawei Han)
- Automated data collection and mature database technology
  - Allows data to be stored efficiently, cheap, persistent
  - Using databases, data warehouses and other repositories
  - Data is increasingly stored distributed (storage area networks, grids, etc.)
- Large and massive data collections
  - Millions to billions of records
  - Tens to thousands of attributes (sometimes also called *variables*)
  - Data is rarely collected for data mining (rather for online transaction processing - OLTP)
- A lot of data is *write only* (or *read once only*)

Data sources

- Relational databases
  - Transactional data, mostly normalised into many tables, with keys between them, continuous and frequent updates on (single) records
- Data warehouses
  - Decision support data, processed and cleaned, historical data, aggregated, updated at certain intervals *(more later)*
- Internet
  - Click-stream data, log files, HTML, XML, blogs, e-mails, media files, etc.
- Files
  - Portable text (like comma separated, tabulator, fixed column) or non-portable proprietary binary files
- Scientific instruments, experiments and simulations
  - Astronomy, genomics, seismology, physics, chemistry, etc.
- Sensors (often data streams)

Types and measurements of data (1)

- Numerical data
  - Integer, floating-point, binary, interval, ratio
  - Non-scalar (like velocity: speed and direction)
- Non-numerical data
  - Nominal data (just naming things, for example personal names)
  - Categorical data (grouping things, like postcodes, university course codes)
  - Ordinal data (ordering things, for example wine tasting, movie ratings)
- 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*

Types and measurements of data (2)

- Multimedia data
  - Images, video, audio
  - Many standard formats used, binary, often compressed
- Different mappings and conversions between data types are possible and often needed
  - Some conversions are loss-less, others are lossy
- Different data mining techniques can handle different types of data
  - Some are restricted to certain types of data, for example only numerical data

Formats of data

- Structured data
  - Relational database tables, integrated data warehouses
  - Images, video, audio (can be compressed), many different formats
- Semi-structured data
  - XML, HTML, e-mails, SMS, log files
- Free-format data
  - Mainly free-format text - ASCII or Unicode

Real world data is dirty (1)

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

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
  - *Garbage-in garbage-out* principle
- Transactional database systems should be designed with data quality and data mining in mind
- Pre-processing is an important step for successful data mining and data analysis

Root conditions of data quality problems

- Multiple data sources
- Subjective judgment in data production
- Limited computing resources
- Security/accessibility trade-off
- Coded data across disciplines
- Complex data representations
- Volume of data
- Input rules too restrictive or bypassed
- Changing data needs
- Distributed heterogeneous systems

Data quality measures

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Interpretability
- Accessibility

Data pre-processing tasks

- Data cleaning
  - Fill-in missing values, smooth noisy data, identify/remove outliers, resolve inconsistencies
- Data transformation
  - Normalise and/or aggregate data
- Data reduction and discretisation
  - Reduce volume of data, but still produce same or similar analytical result, discretisation in particular for numerical data
- Data integration, matching and linking

Data cleaning

- Data cleaning tasks
  - Fill-in (impute) missing values
  - Detect and correct inconsistent data
  - Identify outliers / smooth noisy data
- Missing data may be due to
  - Attributes not considered important
  - Misunderstanding at data entry
  - Inconsistencies with other data and thus deleted
  - Equipment malfunction (for example EFTPOS down, so only cash transactions)
- Missing data may need to be inferred (data imputation)

Data cleaning – Missing values

- How to handle missing data?
  - Ignore the records that contain missing values
  - Fill in missing value manually (often unfeasible)
  - Fill in with a global constant (e.g. *unknown* or *n/a*). Not recommended as a data mining algorithm might see this as a normal value!
  - Fill in with attribute mean or median
  - Fill in with class mean or median (classes need to be known)
  - Fill in with most likely value (using regression, decision trees, most similar records, etc.)
  - Use other attributes to predict value (e.g. if a *postcode* is missing use *suburb* value and external look-up table – if one-to-one relationship)
  - Data editing/imputation (rules based)

Data cleaning – Inconsistent data / outliers

- Why inconsistent data?
  - Due to data entry errors or data integration (different formats, codes, etc.)
  - Important to have data entry verification (check both format and values of data entered), most of the time only format is checked
  - Correct with help of external reference data (look-up tables, e.g. *Sydney, NSW, 7000 -> Sydney, NSW, 2000*) or rules (e.g. *male / 0 -> M, female / 1 -> F)*
- Identify outliers and noisy data
  - Noise: Random error or variance in a measurement
  - Incorrect attribute values (faulty data collection, data entry problems, data transmission problems, data conversion errors, inconsistent naming, technology limitations, bugs, for example buffer overflow or attribute length limits)
  - Handle noisy data through binning, clustering, regression, manual inspection
  - Don't remove or modify outliers for outlier detection!

Data transformation

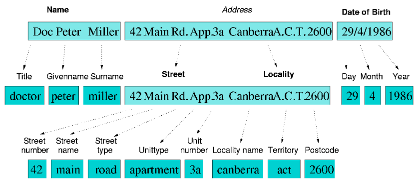
- Consolidate data into forms suitable for data mining
- Smooth data (remove noise)
- Aggregate data (summarisation, *e.g. daily sales → weekly sales → monthly sales)*
- Generalise data (replace data with higher level concepts, e.g. *address details → city → state → country)*
- Normalise data (scale to within a specified range)
  - Min-max (for example into [0...1] interval, or 0%..100%)
  - Z-score or zero-mean (based on mean and standard deviation of an attribute)
  - Decimal scaling (move decimal point for all values)
- Important to save normalisation parameters in meta-data repository

Attribute / feature construction

- Sometimes it is helpful or necessary to construct new attributes or *features*
  - Based on existing attributes in data
  - Helpful for understanding
  - For example: Create attribute *volume* based on attributes *height*, *depth* and *width* (for example in a post or parcel database)
- Construction is based on mathematical or logical operations
- Attribute / feature construction can help to discover missing information about the relationships between the original data attributes

Data parsing and standardisation

- Parse free format data into specific, well defined attributes
- Standardise using rules and look-up tables (correction and replacement tables), or probabilistically (using e.g. *hidden Markov models*)
- Important for data matching and linkage (for addresses, names, etc.)



Why data parsing and standardisation?

- Real world data is often dirty
  - Typographical and other errors
  - Different coding schemes
  - Missing values
  - Data changing over time
- Name and addresses are especially prone to data entry errors
  - Scanned, hand-written, over telephone, hand-typed
  - Same person often provides her/his details differently
  - Different correct spelling variations for proper names (for example *Gail* and *Gayle*, or *Dixon* and *Dickson*)

Data integration and data linkage

- Increasingly, data mining projects require data from more than one data source
- Data is often distributed (different databases or data warehouses)
  - For example an epidemiological study that needs information about hospital admissions and car accidents
- Geographically distributed data or historical data
  - For example, integrate historical data into a new data warehouse
- Enrich data with additional (external) data (to improve data mining accuracy)

Data integration techniques

- Data integration
  - Combines data from multiple sources into a coherent form
  - Schema integration (for example, *A.cust-id* <=> *B.cust-no*)
  - Integrate Metadata from different sources
- Entity resolution (identification) problem
  - Identify real world entities from multiple data sources (for example, *Bill Clinton = William Clinton*, or *Mr Obama = the president*)
  - Also called *record linkage* or *data matching*
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources can be different
  - Possible reasons: different representations, different codings, different scales (for example metric vs. British units)

Schema integration

- Imagine two database tables

PID	Name	DOB
1234	Mayer	01/01/75
4791	Simmons	21-10-1969

PID	Surname	Age
1234	Meyer	32
4791	Simonds	38

- Integration issues
  - The same attribute may have different names
  - An attribute may be derived from another
  - Attributes might be redundant
  - There can be duplicate records (under different keys)
- Conflicts have to be detected and resolved
- Integration is made easier if unique entity keys are available in all the data sets (or tables) to be linked

Data linkage / matching (1)

- Task of linking together records from one or more data sources that represent the same entity
- If there are no unique entity keys in data, the available attributes have to be used
  - Often personal information (like names, addresses, dates of birth, etc.)
  - Privacy and confidentiality becomes an issue *(more later in course)*
- Application areas
  - Health (epidemiology)
  - Census, taxation, immigration, social welfare
  - Business mailing lists, collaborative e-Commerce
  - Crime, fraud and terror detection (US: TIA, MATRIX)

Data linkage / matching (2)

- Different parts of the linked records are of interest
  - Personal information (crime, fraud and terror detection, mailing lists)
  - Non-personal information (epidemiology, census, most data mining) For example:

Age	Disease	*Name*
55	Cancer	*John Miller*
32	Diabetes	*Joe Meyer*
67	Cancer	*Lucy Smith*

Name	DoBirth	DoDeath
J. Miller	04/08/47	12/12/02
J. Meier	11/09/69	26/02/01
L. Smith	01/01/34	08/09/01

Disease	DoBirth	DoDeath	Gender
Cancer	04/08/47	12/12/02	M
Diabetes	11/09/69	26/02/01	M
Cancer	01/01/34	08/09/01	F

Data linkage / matching process