POLYTECHNIC UNIVERSITY
Department of Computer Science / Finance and Risk Engineering

# Data Mining: Motivations and Concepts

**K. Ming Leung**

**Abstract:** We discuss here the need, the goals, and the primary tasks of the data mining process.

## Directory




Last Revision Date: October 30, 2007

# Table of Contents

# 1. Introduction

## 1.1. Motivation and Background

A host of technological advances have resulted in generating a huge amount of electronic data, and have enabled the data to be captured, processed, analyzed, and stored rather inexpensively. This capability has enabled industries and innovations such as

- Banking, insurance, financial transactions - electronic banking, ATMs, credit cards, stock market data
- Supermarket check-out scanner data, point-of-sale devices, bar-code readers
- Healthcare - pharmaceutical records
- Communications - telephone-call detail records
- Location data - GPS, cell phones
- Internet and e-commerce - Web logs, click-streams

that generate huge volumes of electronic data. For example, Wal-mart has 20 million transactions/day and a 10 terabyte database, and

Blockbuster has over 36 million household customers.

The need to understand huge, complex, information-rich data sets is important to virtually all fields in business, science and engineering. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming vital in today's increasingly competitive world. Such data (typically terabytes in size) is often stored in data warehouses and data marts.

## 1.2. What is Data Mining?

Data mining is a rapidly growing field that is concerned with developing techniques to assist managers and decision makers to make intelligent use of these repositories. The goal is to discover meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using techniques developed in pattern recognition, machine learning, artificial intelligence, statistics and mathematics.

Data mining involves a sequence of important steps.

1. The first step is called data pre-processing. It includes

- loading and integrating data from various data sources,
- cleansing data, dealing with missing values
- scaling and normalizing data,
- transforming and reducing variables
- partitioning the data into training, validation and test data sets,
- carrying out exploratory data analysis using graphical and statistical techniques

2. The second step involves modeling the data using techniques developed in the area of statistics and artificial intelligence. Examples of such techniques are
   - regression
   - decision-trees
   - association rules
   - k-nearest neighbors clustering methods
   - neural networks and
   - genetic and evolution algorithms.
3. The third step is for interpreting the model and drawing conclusions.

## 1.3. Applications in Finance and Risk Engineering

In finance, data mining techniques are used for determining financial indicators and future predictions from financial time series data. Forecasting stock market, currency exchange rate, bank bankruptcies, understanding and managing financial risk, trading futures, credit rating, loan management, bank customer profiling, and money laundering analyzes are core financial tasks for data mining. Some of these tasks such as bank customer profiling have many similarities with data mining for customer profiling in other fields. Stock market forecasting includes uncovering market trends, planning investment strategies, identifying the best time to purchase the stocks and what stocks to purchase.

## 1.4. More Applications

1. Target marketing
    - Business problem: Use list of prospects for direct mailing campaign
    - Solution: Use Data Mining to identify most promising

respondents combining demographic and geographic data with data on past purchase behavior
- Benefit: Better response rate, savings in campaign cost

2. Example: Fleet Financial Group
   - Redesign of customer service infrastructure, including $38 million investment in data warehouse and marketing automation
   - Used logistic regression to predict response probabilities to home-equity product for sample of 20,000 customer profiles from 15 million customer base
   - To predict profitable customers and customers who would be unprofitable even if they respond
3. Churn Analysis: Telcos
   - Business Problem: Prevent loss of customers, avoid adding churn-prone customers
   - Solution: Use neural nets, time series analysis to identify typical patterns of telephone usage of likely-to-defect and likely-to-churn customers

- Benefit: Retention of customers, more effective promotions

4. Fraud Detection
   - Business problem: Fraud increases costs or reduces revenue
   - Solution: Use logistic regression, neural nets to identify characteristics of fraudulent cases to prevent in future or prosecute more vigorously
   - Benefit: Increased profits by reducing undesirable customers
5. Example: Automobile Insurance Bureau of Massachusetts
   - Past reports on claims adjustors scrutinized by experts to identify cases of fraud
   - Several characteristics (over 60) of claimant, type of accident, type of injury/treatment coded into database
   - Dimension Reduction methods used to obtain weighted variables. Multiple Regression Step-wise Subset selection methods used to identify characteristics strong correlated with fraud
6. Risk Analysis

- Business problem: Reduce risk of loans to delinquent customers
- Solution: Use credit scoring models using discriminant analysis to create score functions that separate out risky customers
- Benefit: Decrease in cost of bad debts

7. Finance
   - Business problem: Pricing of corporate bonds depends on several factors, risk profile of company , seniority of debt, dividends, prior history, etc.
   - Solution Approach: Through DM, develop more accurate models of predicting prices.
8. Clicks to Customers
   - Business problem: 50% of Dells clients order their computer through the web. However, the retention rate is 0.5%, i.e. of visitors of Dells web page become customers.
   - Solution Approach: Through the sequence of their clicks, cluster customers and design website, interventions to maximize the number of customers who eventually buy.

- Benefit: Increase revenues

Emerging Major Data Mining applications are:

- Spam
- Bioinformatics/Genomics
- Medical History Data  Insurance Claims
- Personalization of services in e-commerce
- RF Tags : Gillette
- Security : Container Shipments  Network Intrusion Detection

## 2. Large Data Sets

It is possible to work with a few hundred records, each having tens of attributes using manual or semiautomatic computer-based analyses. However, effectively mining millions of data points, each described with tens or hundreds of characteristics is not a trivial task.

In theory, large data sets have the potential to yield more valuable information and lead to stronger conclusions. There are problems,

however, due to the increase computational complexity and the risk of finding some low-probability solutions that evaluates well for the given data set, but may not meet future expectations.

Data can be classified into three types: structured data, semi-structured data, and unstructured data. The first type is often referred to as traditional data. The latter two types are called nontraditional data.

Most business databases contain structured data consisting of well-defined fields with numeric or alphanumeric values. Examples of semi-structured data are electronic images of business documents, medical reports, executive summaries, and repair manuals. The majority of web documents also fall in this category. Examples of unstructured data include a video recorded by a surveillance camera in a department store and multimedia recordings of events and processes of interest.

Most current data-mining methods are applied to traditional data. Development of data-mining methods for nontraditional data is progressing at a rapid rate.

The standard model of structured data for data mining is a collection of cases or samples. Associated with each case are attributes or

features. The data is typically represented in a tabular form, where the attributes for each case are stored in a given row, and each column contains the attributes for each case in the entire collection. It is not uncommon to have million of samples and hundreds of features in the collection.

The success of any data mining method may depend on the quality of the data. There are a number of indication of data quality:

1. The data should be accurate. Names should be spelled correctly, numbers should lie in a given range, values should be complete, etc.
2. The data should be stored according to their data type. Numeric values should not be in character form, and integer should not be in the form of real numbers.
3. The data should have integrity. A Data Base Management System should be in place to ensure proper updates, backups and recovery of data.
4. The data should be consistent. The form and the content should be the same after integration of large data sets from different

sources.

5. The data should not be redundant. Redundant and duplicated data should be eliminated.
6. The data should be timely. The time component of data should be made clear from the data.
7. The data should be well understood. Naming standards must be established and followed.
8. The data should be complete. Missing data should be minimized. Some data mining methods are robust enough to work with data having missing values.

Processes to ensure data quality are performed very often using data warehousing technologies. In addition, many data mining preprocessing methods are employed.

## 3. Data Warehouses

The task of data mining is made a lot easier by having access to a data warehouse. The data warehouse is a collection of integrated,

subject-oriented databases designed to support the decision-support functions. It is an organization's repository of data, set up to support strategic decision-making. Data warehouses typically store billions of records.

There are 4 main transformations that are performed on raw data to obtain the data stored in a data warehouse:

1. *Simple transformations* – They are the building blocks of all other more complex transformations. Data in each field are manipulated individually without taking into account its values in other related fields. Examples include changing the data type of a field or replacing an encoded field value with a decoded value.

2. *Cleansing and scrubbing* – These transformation ensure consistent formatting and usage of a field. This can include a proper formatting of address information. This class of transformation includes checks for valid values in a particular field, usually checking the range from an enumerated list.

3. *Integration* – This is a process of taking operational data from

one or more sources and mapping it, field by filed, onto a new data structure in the data warehouse. The most difficult integration issue in building a data warehouse is the identifier problem. This situation occurs when there are multiple system sources for the same entities and there is no clear way to identify those entities as the same. Another data-integration problem arises when there are multiple sources for the same data element. It is not uncommon to find some of these values contradictory.

4. *Aggregation and summarization* – These are methods of condensing instances of data found in the operation environment into fewer instances in the warehouse environment.

Data in a data warehouse are not directly suitable for data mining. The data need to undergo preprocessing.

## References

[1] M. Kantardzic, *Data Mining - Concepts, Models, Methods, and Algorithms*, IEEE Press, Wiley-Interscience, 2003, ISBN 0-471-22852-4. This is our adopted textbook, from which this set of lecture notes are derived primarily.

[2] A. Berson, A. S. Smith, and K. Thearling, *Building Data Mining Applications for CRM*, McGraw Hill, New York, 2000.

[3] Michael W Berry and Murray Browne, *Lecture Notes in Data Mining*, 2006. ISBN-13 978-981-256-802-1, ISBN-10 981-256-802-6.

[4] Daniel T. Larose, *Data Mining Methods and Models*, Wiley-IEEE, 2006, ISBN 0471756474.

[5] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*, Elsevier 2006, ISBN 1558609016.

[6] D. Hand, H, Mannila, and P. Smith, *Principles of Data Mining*, MIT Press, Cambridge, MA, 2001.

[7] C.Westphal, and T. Baxton, *Data Mining SOlutions: Methods and Tools for Solving Real-World Problems*, John-Wiley, New

York, 1998.