CS513-Data Mining

Lecture 1: Introduction

Waheed Noor

Computer Science and Information Technology,
University of Balochistan,
Quetta, Pakistan

Outline

Outline

CS513-Data Mining (3-0)

Objective

To equip students with the understandings of basic and advancement in data mining, and its applications in real world. After this course students will be able to design and deploy data mining solutions using different tools.

In this course:

- Concentrate on theoretical aspects of data mining
- Learn & use different data mining implementation tools
- Learn the design of your experiments and validation process of your results (important aspect of scientific research, specifically in the field of computer science)
- Assignments & problem sets after each unit
- A mini project or a research paper (depends on student's choice)

Tools & Resources

Tools

There are plenty of commercial and open source tools available, check `http://www.kdnuggets.com/software/suites.html`

- *Weka* *www.cs.waikato.ac.nz/ml/weka*
- *Shogun* *www.shogun-toolbox.org*

Books

- Principles of Data Mining, MIT Press by Hand et al. [1]
- Data Mining: Practical Machine Learning Tools and Techniques, Second Edition by Witten and Frank [2]

Data Mining Competitions & Conferences

Competitions

- KDD Cup: [3]
- Netflix: Online Video Recommendation

Conferences

- ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
- SIAM International Conference on Data Mining
- IEEE International Conference on Data Mining

Problems and dataset repository

UCI Machine Learning Repository:
`http://archive.ics.uci.edu/ml/`

Outline

Data Mining

Definition (Hand et al. [1])

Data Mining is the analysis of very large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Definition (Witten and Frank [2])

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data.

Data Mining

Definition (Hand et al. [1])

Data Mining is the analysis of very large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Definition (Witten and Frank [2])

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data.

Data Mining

Objective

- Building intelligent computer programs that sift through databases automatically, seeking regularities or patterns.
- Strong patterns, if found, will likely generalize to make accurate predictions on future data.

Challenges

- Many patterns will be uninteresting.
- Others will be accidental coincidences in the particular dataset.
- Real data is imperfect as some parts will be disjoint and some will be missing.

How to Achieve the Objective

Algorithms need to be robust enough to cope with imperfect data and to extract regularities that are inexact but useful.

Data Mining

Objective

- Building intelligent computer programs that sift through databases automatically, seeking regularities or patterns.
- Strong patterns, if found, will likely generalize to make accurate predictions on future data.

Challenges

- Many patterns will be uninteresting.
- Others will be accidental coincidences in the particular dataset.
- Real data is imperfect as some parts will be disjoint and some will be missing.

How to Achieve the Objective

Algorithms need to be robust enough to cope with imperfect data and to extract regularities that are inexact but useful.

Data Mining

Objective

- Building intelligent computer programs that sift through databases automatically, seeking regularities or patterns.
- Strong patterns, if found, will likely generalize to make accurate predictions on future data.

Challenges

- Many patterns will be uninteresting.
- Others will be accidental coincidences in the particular dataset.
- Real data is imperfect as some parts will be disjoint and some will be missing.

How to Achieve the Objective

Algorithms need to be robust enough to cope with imperfect data and to extract regularities that are inexact but useful.

Data mining Vs non-data mining tasks

Data Mining

- Grouping books based on the interest of readers.
- Defining rules for co-occurring events in the data, e.g., frequency of purchase of pair of products by a certain age group of customers.
- Identifying key words said on a subject in a blog.

Non-Data Mining tasks

- Searching for a phone number or address in the directory
- Query search engines or data centers for specific information.

Activity: Identify any data collection or processing task you have recently done or heard, which can be used for data mining.

Data mining Vs non-data mining tasks

Data Mining

- Grouping books based on the interest of readers.
- Defining rules for co-occurring events in the data, e.g., frequency of purchase of pair of products by a certain age group of customers.
- Identifying key words said on a subject in a blog.

Non-Data Mining tasks

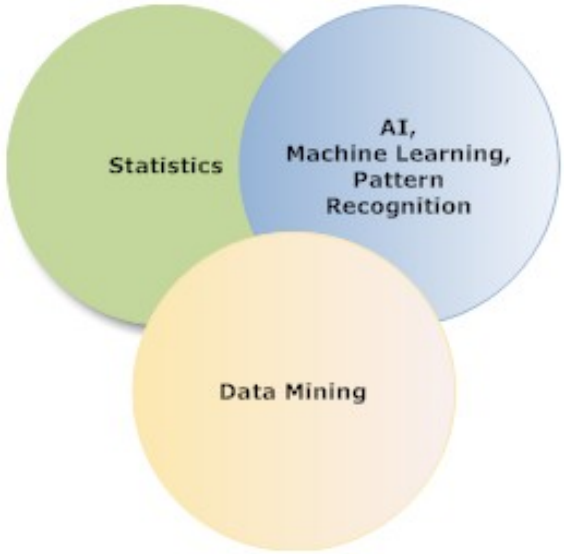
- Searching for a phone number or address in the directory
- Query search engines or data centers for specific information.

Activity: Identify any data collection or processing task you have recently done or heard, which can be used for data mining.

Why Data Mining

- Data collection and storage growing exponentially by sensors, microarrays, scientific simulations, Internet, etc.
- Infeasible to process such huge raw data for identifying or extracting useful information.
- Data mining makes it possible. Classification, segmentations, rule extraction etc.

Data mining origin



- Enormous data
- High dimensional data
- Heterogeneous data
- Distributed nature of data

Outline

Machine Learning I

Definition

Machine learning concerned with the designing and developing algorithms and techniques that implement various types of learning, mechanisms capable of inducing knowledge from examples of data.

Applications

Includes language processing, search engines, medical diagnosis, bio-informatics, speech and handwriting recognition, object recognition in computer vision, game playing and robot locomotion.

Types of learning

The main types of learning systems are supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning.

Machine Learning II

Example

Chess computer game. It doesn't require a huge database to play against you. It only needs the examples and the knowledge. Teach the system what are moves and rules and it will know how to respond and play in real time.

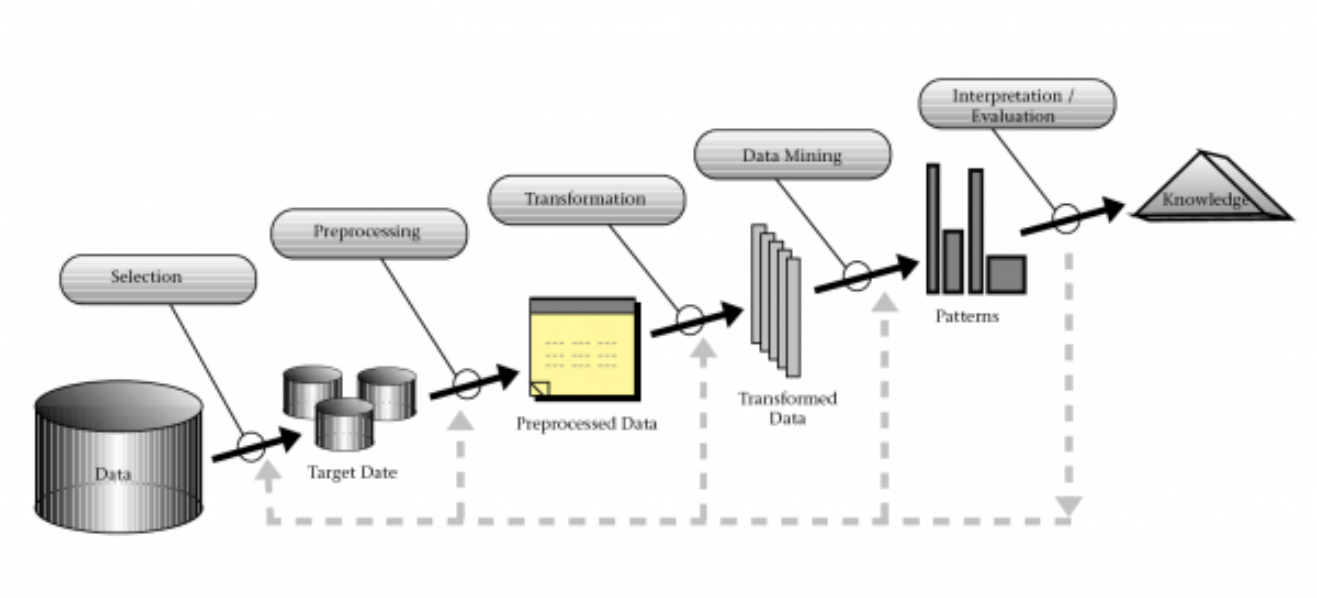
Recall "What is data minig", are they same?

Outline

KDD

Definition

KDD's concept first emerged in 1989 to refer to the broad process of finding knowledge in data. i.e, the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. KDD differs from machine learning in that the task is more general and is concerned with issues specific to databases.



Outline

DM Tasks

- Predictive: Predicting future based on historical data in the database
- Descriptive: human interpretable patterns describing data for decision and policy making

References I

[1] David Hand, Heikki Mannila, and P. Smith. *Principles of Data Mining*. MIT Press, USA, 2001.

[2] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, San Francisco, CA, 2005.

[3] Ismail Parsa. Kdd–cup 1998: Direct marketing for profit optimization, 1998. http://www.sigkdd.org/kddcup/index.php.