Data Mining: A Gentle Introduction

Frans Coenen
Department of Computer Science
The University of Liverpool
coenen@liverpool.ac.uk

http://www.csc.liv.ac.uk/~frans/Seminars/dataMiningLiverpool2010-8-18.pdf

Presentation Overview

- What is Data Mining?
- Techniques:
  - Clustering,
  - Classification,
  - Frequent itemset mining.
- Example applications.
- Summary and conclusions.

What is Data Mining?

What is Data Mining?

- Data mining is the non-trivial discovery of interesting knowledge from data, beyond that which can be obtained by simple database querying.

What is Data Mining?

- Data mining is the **non-trivial discovery** of **interesting knowledge** from data, beyond that which can be obtained by simple database querying.

What is Data Mining?

- Data mining is the non-trivial discovery of interesting knowledge from data, beyond that which can be obtained by simple database querying.
- It can be thought of as "in-depth" data analysis --- something many students need to do as part of their study at The University of Liverpool.

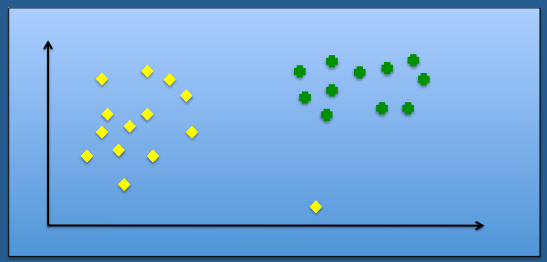
What do I mean by data?

- In the context of data mining the data we wish to mine is often in the form of a single table where: (i) the rows represent *records,* and (ii) the columns *fields* (or *attributes*).
- For the purpose of this presentation we will assume that this is the case.

Techniques (clustering, classification and frequent itemset mining)

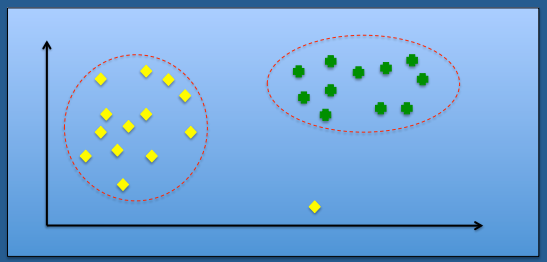
Clustering

- Clustering is the process of grouping records into k categories (where k is often predefined).



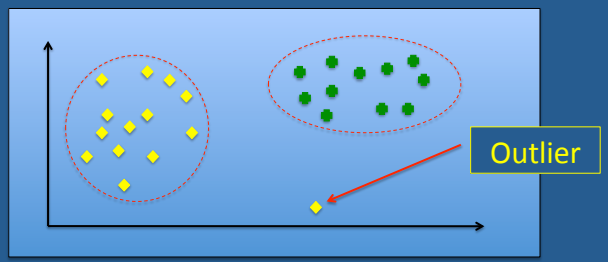
Clustering

- Clustering is the process of grouping records into k categories (where k is often predefined).



Clustering

- Clustering is the process of grouping records into k categories (where k is often predefined).

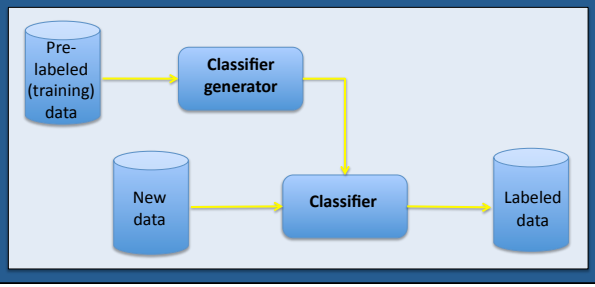


A Clustering Algorithm (Kmeans)

1. Choose first k records to represent prototypes for k clusters.
2. Process remaining records and add each to cluster according to a simple distance measure (nearest neighbour).
3. For each cluster identify most central record as new prototype.
4. If prototypes have changed repeat from (2), otherwise end.

Classification

- Classification is the process of building a classifier from a pre-labeled training set that can be used to "classify" unlabelled data.
- Note that the definition of a set of clusters can be used as a classifier (this rather blurs the distinction!).

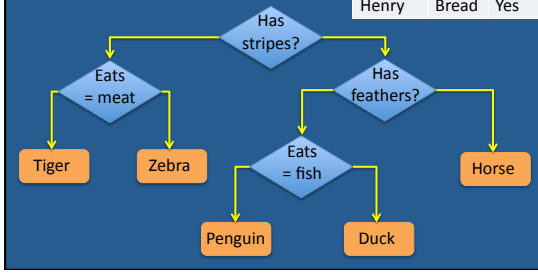


A Classifier Generator Algorithm (Decision Tree)

Name	Eats	Has Feathers?	Has stripes	Class
Anna	Grass	No	Yes	Zebra
Barry	Meat	No	Yes	Tiger
Charles	Meat	No	Yes	Tiger
Dawn	Fish	Yes	No	Penguin
Emma	Bread	Yes	No	Duck
Frank	Grass	No	Yes	Zebra
Gemma	Grass	No	No	Horse
Henry	Bread	Yes	No	Duck

A Classifier Generator Algorithm (Decision Tree)

Name | Eats | Has Feathers? | Has stripes | Class
Anna | Grass | No | Yes | Zebra
Barry | Meat | No | Yes | Tiger
Charles | Meat | No | Yes | Tiger
Dawn | Fish | Yes | No | Penguin
Emma | Bread | Yes | No | Duck
Frank | Grass | No | Yes | Zebra
Gemma | Grass | No | No | Horse
Henry | Bread | Yes | No | Duck



Frequent Itemset Mining

- A frequent itemset is a group (set) of field values (items) that occur frequently together, where frequently is defined according to some pre-specified *support threshold*.
- Frequent itemset mining typically operates on "binary valued data".

Apples (a)	Bananas (b)	Cherries (c)	Damsons (d)
Yes	Yes	Yes	No
Yes	Yes	No	No
Yes	No	Yes	No
Yes	No	Yes	Yes
Yes	Yes	No	No

a b c
a b
a c
a b d
a b

Downward Closure Property of Itemsets

a b c
a b
a c
a b d
a b

- There are four items (*attributes*), hence there are 2^4-1 (15) potential frequent item sets.
- In practice we can not generate all potential frequent item sets!

- Consequently we use something called the downward closure property of itemsets that states that "an item set cannot be frequent unless all its subsets are frequent".

A Frequent Itemset Mining Algorithm (Apriori)

1. `K=1` (Set itemset size counter to `1`).
2. Count occurrences in data set for all `K`-itemsets.
3. Throw away `k`-itemsets below threshold.
4. If no itemsets left end, else generate `k+1` candidate itemsets from `k`-itemsets.
5. `k=k+1` (incremnet `k`).
6. If no candidates end; else go to (2).

Some Example Data Mining Applications

Data Preparation

- To use data mining techniques, of the form described, the data invariably requires some form of data preparation.
- Often we need to combine several tables.
- We may wish to ignore certain columns.
- We may need to range certain fields.
- We need to be aware of data privacy and protection issues.

Customer DB (Transglobal Project)

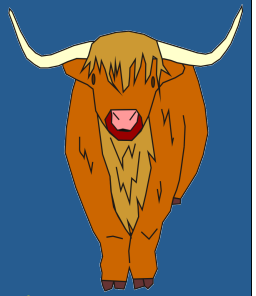
- Mining of customer databases is a standard commercial data mining activity (e.g. super market basket analysis).
- Doing this with a freight forwarder.
- Issues:
  - customer privacy.
  - Continuous numeric fields.
  - Address fields.
- We wish to identify patterns and trends in the data.

Text Classification

- Categorisation of documents such as news articles.
- Define documents in terms of key words/phrases (the bag of words/ phrases representation).
- We can then apply a variety of classification rule mining techniques.
- We can also find frequent item sets or attempt to cluster documents.

Trend Mining in Social Networks

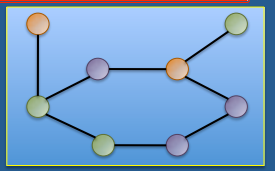
- Particular case is the UK's cattle movement DB, a large DB recording all cattle movements between locations in the UK.
- Represents a time stamped social network.
- Using *trend mining* techniques to identify changes in behaviour.



- Aim is to determine the effect that changes in government policy and working practices might have (or not have).

Frequent Sub-graph Mining

- Given a set of graphs, representing (say) chemical compounds, we can define the node-link-node triples as attributes.
- We can then find frequently occurring sub-graphs (frequent itemsets).
- Or we can build a classifier to categorise new graphs.



- Alternatively given one (very) large sub-graph we can discover frequently occurring sub-graphs in this single graph.

Image Mining (1): AMD (Age related Macular Degeneration)

- We wish to provide screening support for the early diagnosis of AMD.
- A common (standard) mechanism for doing this is by identifying "drusen" in retina scans.
- We represent each images as a series of graphs and build a classifier.

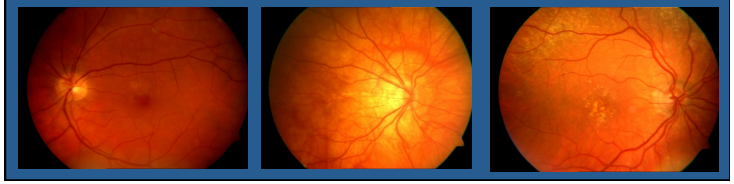
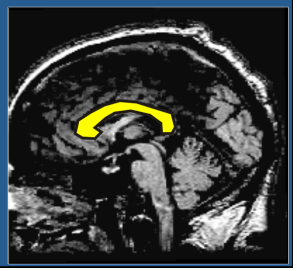


Image Mining (2): MRI scan data

- Classification of MRI scan data, for medical research purposes, according to a particular feature in these scans called the Corpus Callosum (CC)
- The conjecture is that the shape and size of the CC serves to distinguish, for example, musicians and non-musicians.
- It is also suspected that the CC shape and size plays a role in the identification of epilepsy, schitsophrenia, autism, etc.
- We are using a graph and time series representations.



Summary

- We have defined what we mean by data mining?
- We have considered a number of common data mining techniques: clustering, classification, frequent item set mining.
- We have looked at some example applications.

Conclusions

- Students undertaking research either as part of a taught programme, or in the context of a research degree, often need to perform some form of analysis on data (what ever the precise format of that data may be).
- Data mining provides some good techniques for conducting such analysis.