DATA MINING

PART 1

by Yannick Giovanakis

March 19, 2018

Contents

1 Introduction

Digital data storing is a task that started in the early '60s and became since then an important field of computer science. With the passing of time technologies became more powerful and storing large amounts of data became easier.
Analysing bigger and bigger amounts of data became a difficult task over time so **automation** was required . In the late '80s / start of '90s **data mining** became an important task to make sense and interpret massive amounts of data and has been a crucial since then in many fields (customer attrition,credit assessment ,customer segmentation, community detection and many more) as part of a larger subject called **data science**
Data mining is the non-trivial process of identifying

- **valid**
- **novel/interesting**
- **useful**
- **understandable**

patterns in data that results in some **worthy** information.
The aim of data mining is to build programs that run automatically on large databases to seek for regularities or patterns. The main problem is that most patterns are **uninteresting, spurious,inexact** and based on real **imperfect** data. This is why data mining algorithms need to be **robust** to cope with imperfections and to extract regularities that are inexact but **useful**.
The informations found can then be used for:

- **predictive** tasks → create predictions based on patterns
- **descriptive** tasks → create insights based on patterns
- **prescriptive** tasks → combination of both

1.1 Data mining process

1. Interesting question
   What is the goal? What must be predicted?
2. Get the data
   How was data sampled ? Which data is relevant?
3. Explore the data
   Plot data,compute statistics , search for anomalies
4. Build model
   Build,fit ,validate
5. Communicate result
   What did we learn? Is there a result?

The data part is the most important :

- **Selection**
  What are the data we actually need?
- **Cleaning**
  Are there errors / inconsistencies that need to be eliminated?
- **Transformation**
  Some data can be eliminated because equivalent to other data or used to get new data
- **Mining**
  Select mining approach : **classification ,regression ,clustering...** and apply algorithms
- **Validation**
  Are the patterns found **sound** ? According to which criteria? Can the results be explained?

1.2 Data mining tasks

- Prediction & Regression
- Classification
- Clustering
- Association rules
- Trend & evolution analysis
- Outlier analysis
- Text mining ,topic modelling, graph mining

1.3 Issues

Data mining generates many patterns, but typically only **few** are interesting. It is important to find an **interestingness** measure : a pattern is interesting if it is understood, valid on new data/test data with some degree of certainty, potentially useful,novel or validates some hypothesis that needed confirmation.
Interestingness measures can be **objective** (based on **statistics** and **patterns**) or **subjective** (based on **belief** in data) .
Can **all** interesting patterns be found?

Completeness problem

Can a data mining system find **all** the interesting data within a dataset? This depends also on the data mining approach that has been chosen.

Optimization problem

The data mining system should **only** find useful and interesting patterns , by either filtering **all possible patterns** or by using **mining query optimization**

2 Data representation

Inside databases and datasets we can identify:

- **instances** → observations/cases/records that represent atomic elements of information
- **Attributes** → variables/features that measure aspects of an instance. Each instance has a certain number of attributes
- **Concepts** → things that can be learned inside the data

2.1 Attribute types

Numeric attributes

- Real-valued or integer-valued domain
- Interval-scaled when only differences are useful → temperature
- Ratio-scaled when only rations are meaningful → age

Numerical attributes are **ordered** and measured in fixed units.
Zero point is only defied for **ratio attributes**.
Can be divided into *discrete or continuous*.

Categorical attributes

- Set-valued domain composed of a set of symbols
- *Nominal*
  When only equality is meaningful Values are distinct symbols that serve as labels. No relation is implied among nominal values and only equality test can be performed.
- *Ordinal*
  When both equality and inequality are meaningful. As in the nominal case, also here talking about **difference** doesn't make sense.

Binary attributes

Represented by either 0/1

Sometimes the same attribute can be either considered **nominal** or **ordinal** : `if age == young AND ...` is nominal whereas `if age < presbyopic AND...` is ordinal.