Data Mining summary

Medjkoune Sofiane 201400007781
Hadjeres Yasmine 201400007136

Chapiter 1 :

Through decades, information technology evolved from the basic file processing, to relational databases, warehouse thus handling the huge amount of information coming from different sources, its necessity became even more obvious with the avenuing of the internet in the early 90's. Data became present in every prospect of our lives and information, it became a need to extract knowledge from this data and use it to improve productivity in general, and this is what data mining does, extracting knowledge from raw data and creating value out of the mining.

To achieve this task, it goes through several steps :

- Data Cleaning
- Data integration (where multiple data sources may be combined)
- Data selection (Getting the relevant data)
- Data transformation
- Data mining (Extracting data patterns)
- Pattern evaluation.
- Knowledge presentation

One of data mining's major pros is that it can be on different data structures as Databases, data warehouses, transactions and some more advanced structures. Regardless to the structure used, the main purpose is to mine different types of patterns such as :

- Characterisation and discrimination where characterisation consists on summarising the data of targeted class and discrimination on comparing the targeted classed with contrasting classes. It's not rare to use both techniques.
- Mining Frequent Patterns, Associations, and Correlations : Consists on mining the frequent pattern appearances which lead to an association analysis that help us evaluate two information : the support and the confidence.
- Classification and regression : where classification is the method that allow us to find a model that describes a certain class and thus help classifying the data that's not labeled. Whereas classification is used on discret data, regression is used to predict continuous values.
- Cluster Analysis : Consists in analysing and gathering data that not is labeled. Data is grouped regarding the similarity of the different attributes of the data.
- Outlier analysis : Analysing objects that do not behave like the general model of the data, it may be useful for fraud detection for example, it is a kind of anomaly detection.

Different technologies are used depending on the mining method, for instance statistical models are widely used for characterisation and discrimination, verifying the results of the mining. Machine learning learning methods such as supervised and unsupervised trainings are used for classification/regression and cluster analysis respectively. Some database systems have built dat analysis capabilities permitting, real time fast streaming of large datasets. Information retrieval is also used when it comes to searching informations in different types of documents based on probabilistic models.

Data mining is used in multiple fields. Business intelligence is a great example of its importance when it comes to extract knowledge from the informations about their customers and competitors, in this field, data mining is essential. Web search engines are also a great source of applications for data mining, to finds the most accurate informations as fast as possible.

Chapter 2

It is now clear that data is the key for the data mining, before diving into this process it is important to have some informations about the data, for instance it is useful to know the types of the different attributes, some statistical descriptions to learn more about the values or even visualisation.

When we talk about data, it is more precise to speak of data objects that represent an entity which is represented by attributes thus describing that data object. Each attribute has a type, we can
summarise them in the following categories :

- Nominal attributes (That generally refer to names)
- Binary attributes (As we know them to have only two possible values : true of false)
- Ordinal attributes (That have a limited number of meaningful values unlike the nominal attributes, e.g take an attribute eye_color it can only take values : blue, green, brown, grey and black)
- Numeric attributes (It is clear that it is the attributes with a numerical value)
- Data attributes can either be discrete if it takes a finite number of values. On the contrary, it is called continuous attributes.

Another type of information we can get is statistical properties of the data. Many techniques can be used to get such intel, we can talk about :
Measuring the Central Tendency:

- Mean : The mean is the average of the numbers, calculated by summing different values of an attributes from different objects and dividing the result by the number of objects tested. However it presents a major issue, when the values are extreme, the results are easily corrupted and that's where the Median technique comes in.
- Median : To find the Median, we sort the values of the attribute and find the middle number thus solving the precedent issue. e.g: for the Median of {13, 23, 11, 16, 15, 10, 26}we sort the values{10, 11, 13, 15, 16, 23, 26} The middle number is 15, so the median is 15. (If there are two middle numbers, we average them)
- Mode : calculated as follows : mean − mode ≈ 3 × (mean − median).
  Another used technique would be the Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range (Will be detailed if time permits it)then we can use different graphical displays of these statistical informations.

Chapter 3

Now that we know more about our data, the next step is the preprocessing. It is important to measure the quality of the data that stands on three principles : accuracy, completeness and consistency and in real life, it is common to find low quality data in large scale databases and warehouses. Pretending the data quality depends directly on the intended use of the data. Some other factors also affect data like timeliness, believability of the data and interpretability or the ease to understand data.
The preprocessing goes through several steps :

- Data Cleaning : We start with the missing values, actually, there's different methods to fill in the void :
  - Ignore the tuple (Usually done when the class label is missing), not the best method.
  - Fill in the missing value manually, time consuming and small scale operation.
  - Use a global constant to fill in the missing value.
  - Use the most probable value to fill in the missing value : may be found using regression or a decision tree using other related informations of the customer for example.
  - Use the attribute mean or median for all samples belonging to the same class as the given tuple.

    It is to note that sometimes a missing value is not an error nor a misleading attribute. Now that we've gone through missing attributes, we face another problem which is noisy data.
- Noisy data : To simply put it, a noise is an incoherent value of an attribute depending on the context, and to get rid of that noise many techniques are possible :
  - Binning: smoothes the value of the attribute by consulting the values around it, statistical methods like the mean, median and mode can be used for that.
  - Regression: involves finding the "best" line to fit two attributes so that one attribute can be used to predict the other.
  - Outlier analysis : usually using clustering, this is a technique that is described in an upcoming chapter of the textbook.
- Data integration : used to reduces redundancies and inconstancies. Before starting the integration, we need to make sure that when it comes to merges from different databases, the structure of the data is coherent to ensure that any attribute functional dependencies and referential constraints in the source system match those in the target system. There's also the redundancy problem that can be detected with a correlation analysis. One other common issue is the data value conflict. To put it simply, it is the values of the same attribute recorded in different places (common issue with data warehouses)that doesn't have the same value. e.g : in Algeria we record weight with kg and in the U.K with pounds.
- Data Reduction : Generally we face huge amounts of data making the data analysis and this is where data reduction comes in to reduce the size of the dataset while maintaining its integrity. Reduction strategies are divided into :
  - Dimentionality reduction : it's the process of reducing the number of random variables or attributes under consideration, it includes wavelet transforms and principal component analysis methods.
  - Numerosity reduction : techniques replace the original data volume by alternative, smaller forms of data representation
  - Data compression : Consists on having a compressed version of the data, If the original data can be reconstructed from the compressed data without any information loss, the data reduction is called lossless. If, instead, we can reconstruct only an approximation of the original data, then the data reduction is called lossy.