Data Management

University of Brighton
Richard Goodman
13842540

March 2, 2017

Abstract

This report will explore various data mining methods to highlight the importance of data mining. The methods discussed will be compared to see their advantages and disadvantages between one another. Finally, a data warehouse design (appendix A) will be created that is compliant with OLAP applications.

1. Introduction

In any company over time, the data grows as more data is fed into the company. 'The world's data volume is expected to grow 40 per cent per year, and 50 times by 2020.' (de Waal-Montgomery 2015). Companies are using this to their advantage to understand not only more information on their consumers, but the environment the companies domain revolves around.

One can argue that there is 2 main stages for business intelligence: *collecting* and *analysing*. In the scenario of the case study, the main source of information comes through a web service. By utilizing this to its full potential, a data warehouse can extract and clean any incoming data to be a more consistent product for further services and external sources, such as: OLAP servers and clients.

2. Data Warehouse

'A major component of Business Intelligence is the use of a data warehouses for making better decisions based on the information it provides' (Leonard 2009). A data warehouse *(DW)* is a fundamental piece of tooling that provides a huge benefit for any company collecting data and processing it into workable data. But what is a Data Warehouse? what does it do? and what benefits does it provide?

2.1. What

A DW is considered the "middle-man" between OLTP *(**O**n-**l**ine **T**ransactional **P**rocess)* and OLAP *(**O**n-**l**ine **A**nalytical **P**rocess)* systems. OLTP systems and DW's are closely related despite having different qualities, figure 1 *(source (Connolly & Begg 2015))* shows the differences in an OLTP system and a DW.

CHARACTERISTIC	OLTP SYSTEMS	DATA WAREHOUSING SYSTEMS
Main purpose	Support operational processing	Support analytical processing
Data age	Current	Historic (but trend is toward also including current data)
Data latency	Real-time	Depends on length of cycle for data supplements to warehouse (but trend is toward real-time supplements)
Data granularity	Detailed data	Detailed data, lightly and highly summarized data
Data processing	Predictable pattern of data insertions, deletions, updates, and queries. High level of transaction throughput.	Less predictable pattern of data queries; medium to low level of transaction throughput
Reporting	Predictable, one-dimensional, relatively static fixed reporting	Unpredictable, multidimensional, dynamic reporting
Users	Serves large number of operational users	Serves lower number of managerial users (but trend is also toward supporting analytical requirements of operational users)

Figure 1: *Differences in an OLTP system and a DW*

2.1.1. OLTP & OLAP

Before any further discussion, it's important to briefly discuss the difference between OLTP and OLAP. 'OLTP systems are optimized for a high number of transactions that are predictable, repetitive, and update intensive.' (Connolly & Begg 2015), these kind of transactions follow the typical procedure of CRUD *(Create Read Update Delete)*. Data processed in

an OLTP systems would be kept in it's rawest form, to preserve the originality of the incoming data.

On the other hand, OLAP systems are multi-dimensional systems that are generally created using existing data that is cleaned. A big feature of OLAP systems is the use of derived fields which can be used to make analytical querying easier as they are a field themselves rather than a calculation being executed at query run time. Another significant difference is OLAP having historical *(as well as current)* data, which provides powerful advantages for analytical statistics.

In an ideal scenario, a business would have an OLTP system where data is stored in it's original form. After data cleansing, the data would then be added to the large pool of data in an OLAP system where optional derived fields would be instantiated as well as additional tables if applicable. Generally analysts would use OLAP systems in a read only manner to extra valuable data for their needs.

2.2. Benefits

There are common risks in incoming data such as: data integrity *(data type mismatch, semantic differences etc.)* and inconsistency are a few. By implementing a DW an important procedure is the cleansing of data, making sure it is compliant with the schema's in the business domain.

Doing this insures the data being processed will be valid for any OLAP system implemented. But it isn't just about the purity of data, as this general system is automatic, it provides several benefits *(source (Kimball & Ross 2002))*:

- Makes an organization's information easily accessible.
- Presents the organization's information consistently.
- If changes are made to any schema, it's easy to append to a DW requiring less work than if one isn't implemented.
- A level of security of information.

The last being an interesting point, one benefit of this automatic process is the privacy of valuable data such as personal information. Compilers or third party users can input sensitive data in, however you can specify certain data to go into OLAP system schemas so it can't be produced for any analytical or client procedures. This removes a strain from a business view as there is a significant lower chance of publishing sensitive information.

2.3. Conclusion & Design

It's evident that a DW is needed in any business which processes many data transactions, but also wishes to use data for analytical purposes. Appendix A presents a DW design suitable for the case study provided.

3. Data Mining

Data Mining is a powerful tool for data analysis. As a whole it is the concept of complex algorithms that are used to extract information from data into multi-dimensional subsets of data to identify trends and patterns.

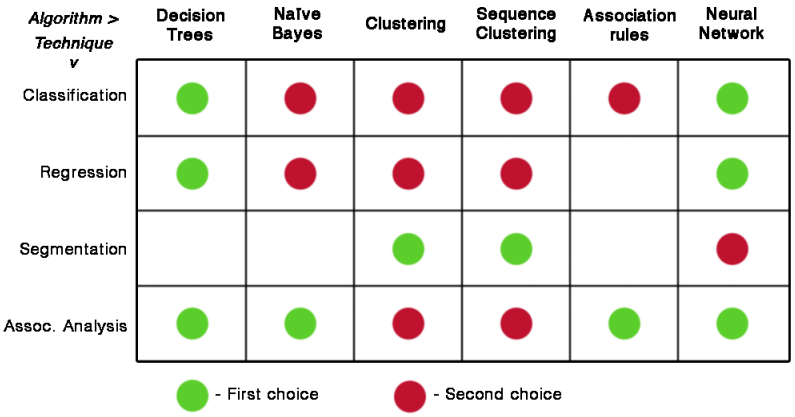
Data Mining is becoming common for business to use these tools on their data because simply just storing data isn't enough to identify patterns upon data to predict forecasts as well as see historical data in a more compelling fashion as well as provide that competitive edge.

It is important to note that there is a difference between data mining and general querying of data:

- Querying:
  - Find all backers who have backed project X
  - Find all projects backed in a specific region
- Data Mining:
  - Find all projects that are often backed when project X has been backed
  - Identify project categories that are common in certain locations *(Arts and Culture projects more frequently backed in Brighton)*

3.1. Techniques

The Data Mining algorithms that will be discussed are capable of executing various problems, figure 2 show what each Data Mining algorithm is capable of.



Original source: Data Minning Lecture, Fallakhair, S. University of Brighton

Figure 2: *Different Data Mining Algorithms*

3.1.1. Classification

Classification falls under the prediction category, 'leaning a function that maps *(classifies)* a data item into one of several predefined classes.' (Weiss & Kulikowski 1991).

What this means is, it acts as an IF THEN ELSE type algorithm, classic examples of this are mapped out into decision trees to predict an outcome of a defined class as seen in figure 3 *(source (Han et al. 2012))*.

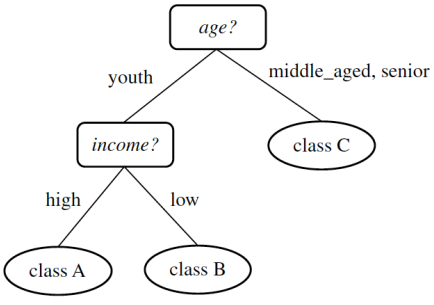


Figure 3: *A decision tree using the classification technique*

Example
Is this project likely to get successful funding?

3.1.2. Regression

Also a technique that is under the prediction category, 'Regression is a data mining function that predicts numeric values along a continuum' (Taylor 2013).

To put this into context, this technique predicts the result of a continuous attribute as long as it is supplied with a subset of data corresponding to the problem. Figure 4 shows an example of where regression is used to predict a loan amount with the two variables, debt and income (Fayyad et al. 1996).

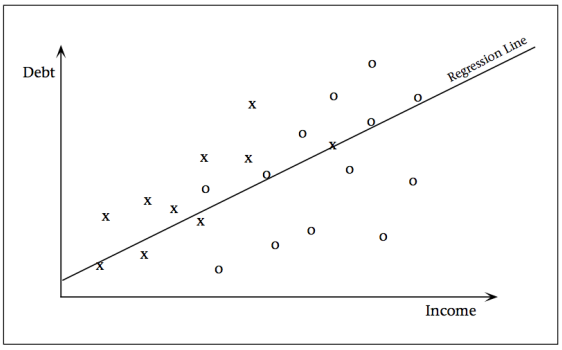


Figure 4: *Linear regression*

Example
What will the registration count be next month?

3.1.3. Segmentation

Being in the clustering category, segmentation is a function where 'They partition the objects into groups, or clusters, so that objects within a cluster are âĂIJsimilarâĂİ to one another.' (Han et al. 2012).

What this allows is the concept of clustering data into groups which have common attributes or behaviours that aren't necessarily obvious when looking at the data. Figure 5 shows an example taken from Liekens article where they have used clustering to group music in terms of a users taste (Liekens 2007).

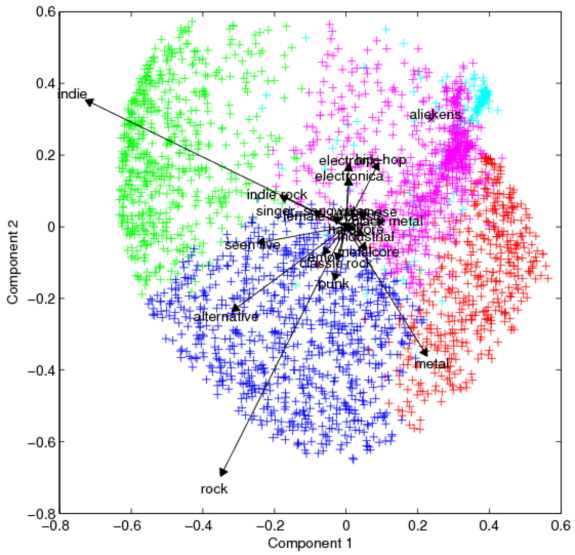


Figure 5: *Clustering segmentation*

Example
Identify groups of projects *(3D Printers, Digital Cameras etc.)*

3.1.4. Association Analysis

Association Analysis is under the recommendation category that is used in major e-commerce business such as Amazon. By imagining all objects *(in this case, projects)* as boolean values, 'The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together' (Han et al. 2012).

As mentioned, major e-commerce companies use this to promote other items to attempt the customer purchasing more than originally intended.

Example
If user backs project X then they are 60% likely to back project Y.

3.1.5. Misc

There are other techniques out there, such as: Time Series, Sequence analysis and Anomaly Detection. However, are irrelevant for the problem domain. This is because of the following reasons:

- Time Series - Although the company can potentially use time series for historical statistics, basic time series can easily be created without the need for data mining, but by doing historical queries.
- Sequence Analysis - This technique is mostly used on complex sequential datasets such as 'symbolic sequences, and biological sequences' (Han et al. 2012).
- Anomaly Detection - Although the domain deals with online banking transactions, using a technique such as this would be a waste of time and resources, as this is aimed at companies that specialise with fraudulent services.

3.2. Conclusion

The Main techniques discussed all provide exclusive benefits for the problem domain:

- Classification - Allows to make predictive decisions with a level of certainty. A feature the company could offer from statistics is to see if a project is likely to get funded based on the parameters given.
- Regression - Can be used for more internal statistics and forecasts which can be used to measure various needs, such as predicting future numerical statistics, and comparing them when they become present.
- Segmentation - Can deal with noisy and high-dimensional data. Allows the idea of reducing data by creating clusters to represent large sets of data. Can help benefit the navigation of users by grouping projects into sub categories.
- Assoc. Analysis - Provides a competitive edge to the company. Provides an extra level of commitment to the customers by using recommendation to their actions.

Because of these reasons, and from the information collected, the most relevant techniques to use would be classification, segmentation and association analysis.

3.3. Algorithms

Before a final conclusion is made, details of each algorithm implementing these techniques *(if possible)* is presented.

3.3.1. Decision Tree

Decision trees have been briefly looked upon earlier, it is a flow-like structure *(can be represented vertically or horizontally)*, where each node represents an attribute. This allows an easy visual reference to large sets of data as seen in figure **??**.

'In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand.' (Han et al. 2012). This method is generally used for production as they have that level of certainty.

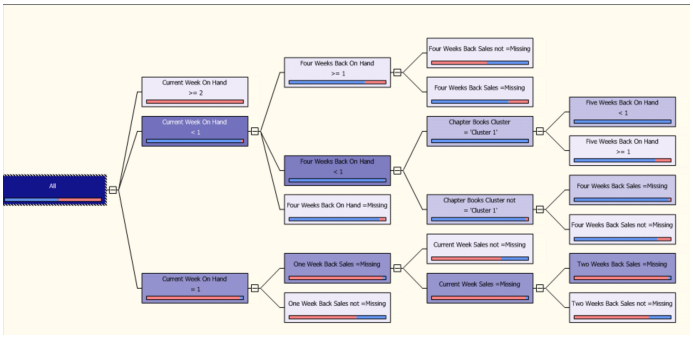


Figure 6: *MS SQL Server Decision Trees* [1]

3.3.2. Naïve Bayes

Naïve Bayes is one of the most known robust algorithms around. 'The induction of these classifiers is extremely fast, requiring only a single pass through the data if all attributes are discrete.' (Kohavi 2011).

This algorithm can be represented visually in a numerous amount of ways, a common method is found in figure 7, in this example its predicting the likelihood of a new object's color.

[1] source: `https://technet.microsoft.com/en-us/library/cc917727.aspx`

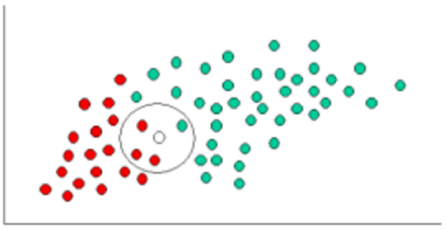


Figure 7: *Naïve Bayes Classification on prediction of color* [2]

3.3.3. Clustering

Clustering shares a similar description to segmentation, this is one tool where its speciality is segmentation, as can be seen in figure 5. There are slight variations which can be adapted to make it more visually appealing and easier to read. It's a reliable tool to use for grouping of data, and can make representation of data quick and easy to understand.

3.3.4. Sequence Clustering

Sequence clustering is some what a merge of clustering and decision trees. It can become a complex visual representation to look at, however it offers that edge of identifying routes that simply clustering can not provide.

A common case where this is used is in Microsoft SQL Server, as seen in figure 8. Depending on the data set used, the diagram can easily become cluttered to the extent of no clear context, however, it is a great tool to use for visualising links between attributes.

[2] source: `http://www.statsoft.com/textbook/naive-bayes-classifier`

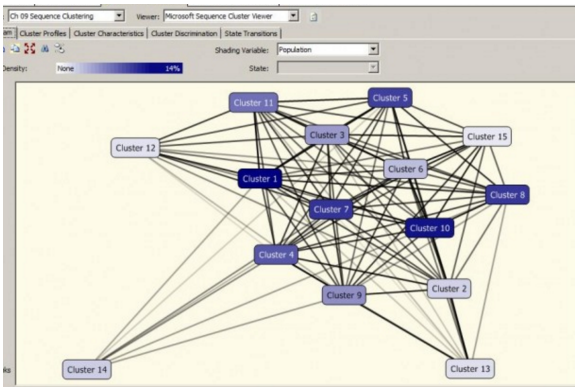


Figure 8: *Sequence Clustering in MS SQL Server* [3]

3.3.5. Neural Network

'Neural network is a set of connected input/output units in which each connection has a weight associated with it.' (Han et al. 2012). The only drawback this offers is that computation times can be lengthy, however, if this is feasible it isn't a limitation to using it.

The way it works is, it takes in parameters *(such as attributes or data)*, and processes them in an "invisible layer" in which it finds patterns and processes them into an output as seen in figure 9 (Analytics 2015).

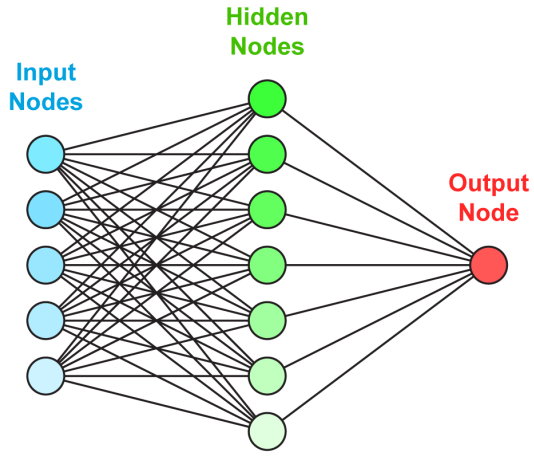


Figure 9: *Neuron Network*

[3] source `https://rlsqldev.wordpress.com/category/ssas-data-mining/`

3.3.6. Misc

As seen in figure 2, some algorithms have been excluded from a detailed description, this is because as a whole they do not support the techniques recommended.

3.4. Conclusion

When looking at the case study, and from the findings above, the most suitable data mining algorithm to implement would be **OMG**, this is because...

References

Analytics, B. (2015), 'Neural network data mining explained.'. `http://www.butleranalytics.com/neural-network-data-mining-explained/` [Accessed: 02 March 2017].

Connolly, T. M. & Begg, C. (2015), *Database Systems: practical approach to design, implementation, and management*, Pearson Education Limited.

de Waal-Montgomery, M. (2015), 'World's data volume to grow 40% per year & 50 times by 2020: Aureus.'. `https://e27.co/worlds-data-volume-to-grow-40-per-year-50-times-by-2020-aureus-20150115-2/` [Accessed: 22 February 2017].

Fayyad, U., Piatestsky-Shapiro, G. & Smyth, P. (1996), 'From data mining to knowledge discovery in databases', *AI Magazine* **17**(3).

Han, J., Kamber, M. & Pei, J. (2012), *Data mining: concepts and techniques*, Elsevier.

Kimball, R. & Ross, M. (2002), *The data warehouse Toolkit: The complete guide to dimensional modeling.*, 2 edn, Wiley, John & Sons., New York.

Kohavi, R. (2011), Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid, Master's thesis. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=`

10.1.1.462.9093&rep=rep1&type=pdf [Accessed: 02 March 2017].

Leonard, E. (2009), Design and implementation of an enterprise data warehouse., Master's thesis. http://epublications.marquette.edu/cgi/viewcontent.cgi?article=1118&context=theses_open [Accessed: 22 February 2017].

Liekens, A. (2007), 'Data mining musical profiles'. http://anthony.liekens.net/index.php/Computers/DataMining [Accessed: 02 March 2017].

Taylor, K. L. (2013), *Oracle Data Mining Concepts*. 11g Release 2 (11.2).

Weiss, S. I. & Kulikowski, C. (1991), *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems.*, Morgan Kaufmann., San Francisco, California.