DMV302 Data Mining & Visualisation
Assessment 1 - Data Mining Applications & Processes

David Lawler - A00075945
David.Lawler@mds.torrens.edu.au

Word Count: 1018(1460 total)

Executive Summary

The purpose of this report is to show an understanding of data, data mining processes, and data mining techniques and apply this knowledge in order to assist online customer-to-customer retailer DDMall in optimising their three main services; Targeted Advertising, Regulation and Demand Prediction.

The problem becomes firstly understanding the DDMall business, and then the data mining process. We can assume that any data we want is available according to the brief, so what data types and with what attributes will best serve our data mining techniques? What techniques should DDMall use to accomplish their goals? And what are the challenges associated with their use?

Throughout my study I found that you can use a lot of different models for similar purposes, but they all have strengths and weaknesses. I came to a decision to use a hybrid clustering and classification for demand prediction based on the Parikh 2003 study that suggested it was better than using a straight classification technique. For targeted advertising, Association analysis was chosen to use in analysing shopping carts. For regulation of frauds and bad actors, Clustering can be used to find copycat products and descriptions, it can also be used to group their accounts by information like bank accounts and username.

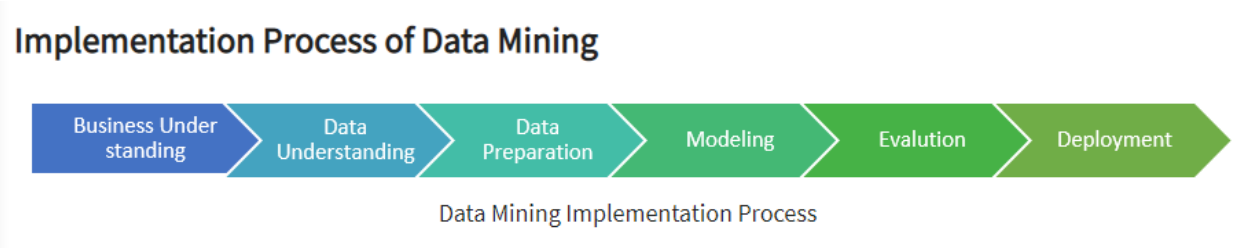
Introducing DDMall

DDMall is an online consumer-to-consumer retail business and this report will suggest data mining solutions to improve the three services it provides to its store owners; targeted advertising, regulation, and demand prediction. For the purpose of effective data mining DDMall should focus on acquiring as much data from all available sources relevant to stores, products, and customers. The brief has stated that we can assume any data we want is available to us, so this may include customer data, purchase histories, product description data, and time series based sale totals.

This report will look at how DDMall's data should be selected, cleaned, transformed, formatted, anonymised and constructed for appropriate data mining applications. After, three data mining applications will be recommended and explained, one for each service DDMall provides.

Data Mining Applications for DDMall's Services

Implementation Process of Data Mining(Taylor 2022)



Classification or Hybrid Classification & Clustering(HCC) Model for Demand Prediction

DDMall could implement a Classification or a HCC model based on the Parikh 2003 paper for demand forecasting to help retailers identify underachieving stores where potential sales exceed actual sales and product allocation to assist manufacturers to allocate products to stores and accounts.

The business objective is to predict potential sales of products considering various store, product and shopper's demographic attributes. This objective drives the entire data mining process.

Data Mining Process Overview(Parikh 2003)

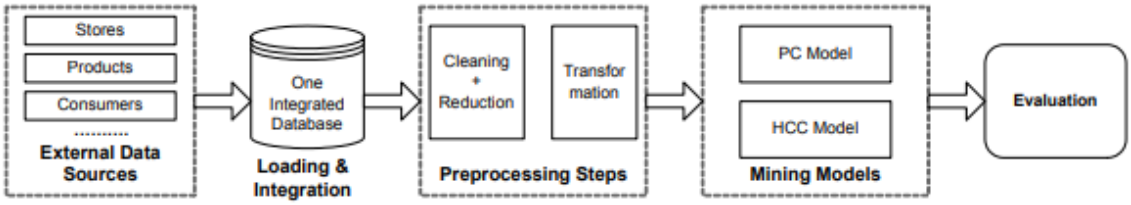


Figure 1: Data Mining Process Overview

First we identify the required external data sources required; Store, Product and Consumer Data. For example, store data might include Store Name nominal data, Location ordinal data, Products which would include measurements and other quantitative data. The loading and Integration process loads data from the external sources and combines them in one database instance.

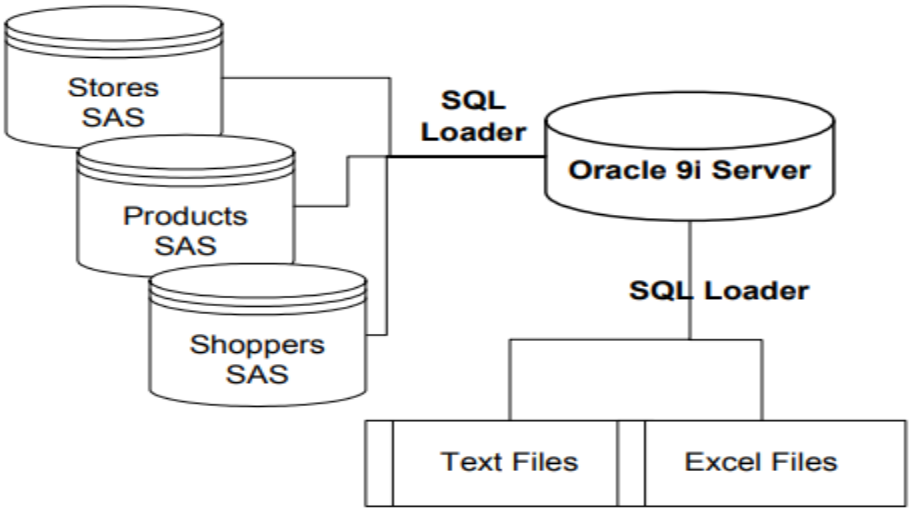


Figure 3: Integration from External Data Sources

The data preparation steps are performed to ensure the quality of the selected data sets.
The first issue is handling inconsistent data types, for example user name might be string from one source and varchar from another and will need to be converted to the most appropriate data type. Handling missing attributes by removing them, or removing whole stores if they are missing a lot of them. Another challenge involves deleting duplicates and correcting outliers.

After the data is cleaned, it needs to be transformed to appropriate forms. For numerical attributes, log transformation is applied to correct problems with skewed data, and then min-max normalisation is used to transform the data into a defined range.

The proposed data mining models, Pure Classification (PC) and Hybrid Clustering Classification (HCC), implement different mining techniques on the processed data. Both models support multi-relation data mining with efficient search and indexing techniques, combining with existing data mining techniques. The Evaluation step analyzes test experiments of each model, and then compares both models in terms of accuracy and performance. Experimental results show that the Hybrid Clustering Classification model provides better accuracy with significant efficiency improvement over the Pure Classification model*(Parikh 2003)*.

Association Analysis for Targeted Advertising

Targeted Advertising will assist DDMall in attracting new customers and retaining existing customers. In order to reduce cost per conversion(CPA) and increase Return on Ad Spend(ROAS)(Nguyen 2019), DDMall chould use Association Analysis to conduct market basket analysis based on identifying elements(sex, advertisement number, date etc.) that have a position relationship with conversion.

(Example mode for Association Analysis(Nguyen 2019)

	antecedents	consequents	support	confidence	lift
0	(Male)	(Conversion)	0.092	0.191268	1.045181
24	(ad1)	(Conversion)	0.060	0.194805	1.064509
26	(ad2)	(Conversion)	0.070	0.191257	1.045119
28	(lp2)	(Conversion)	0.052	0.191882	1.048535
30	(lp3)	(Conversion)	0.045	0.192308	1.050862
32	(lp4)	(Conversion)	0.052	0.201550	1.101368
35	(Friday)	(Conversion)	0.031	0.200000	1.092896
36	(Age 25-34)	(Conversion)	0.083	0.201456	1.100854
38	(Low Income)	(Conversion)	0.045	0.207373	1.133187
234	(ad1, Male)	(Conversion)	0.032	0.209150	1.142898
240	(ad2, Male)	(Conversion)	0.034	0.188889	1.032180
246	(Age 25-34, Male)	(Conversion)	0.045	0.220588	1.205400
556	(ad2, Age 25-34)	(Conversion)	0.036	0.214286	1.170960

Support in this instance is the percentage of the dataset that has the antecedents and the consequent. Confidence is how likely the consequent is of happening if the antecedent happens, Lift is the chance of consequent happening increases, given the antecedent happens.

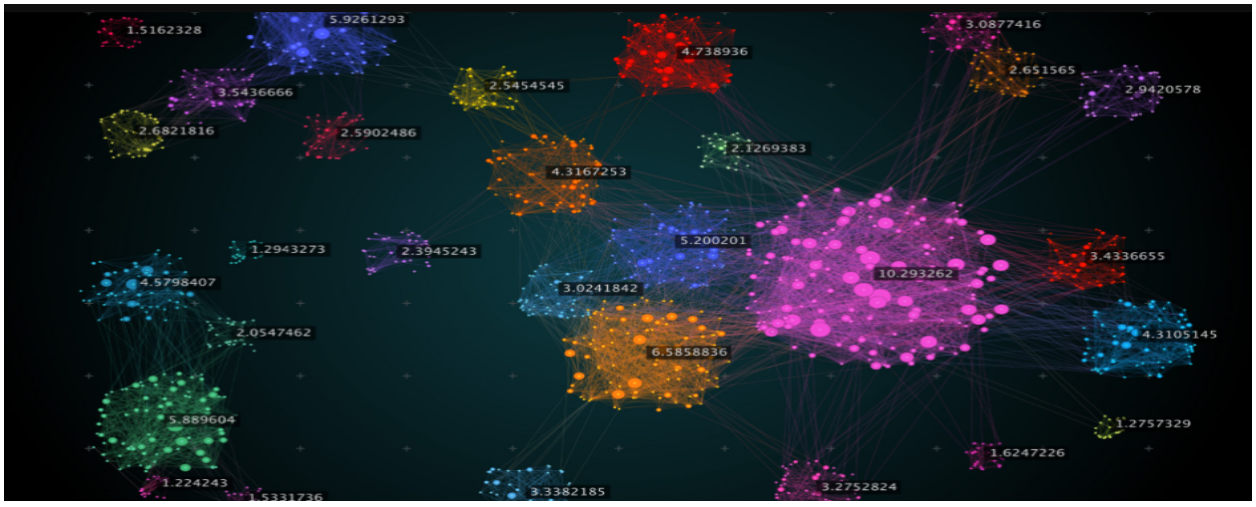
Association rules are a quick and easy way to get insights from an advertising dataset. Data preprocessing for Association Analysis involves the creation of variables for antecedents, and turning continuous features into nominal features(which might lose some meaning in the process).

Clustering for Regulation

To improve DDMall's ability to detect copyright or counterfeit products, fraudulent or inaccurate descriptions and fake consumer reviews, Clustering could be used to group products of similar descriptions, and to detect fraudsters - who will typically reuse information and attributes on their accounts, like date of birth, email or bank account data. Clustering is an application of machine learning that takes data and groups it into clusters of similarity or likeness(Tausz, 2020), to accomplish a purpose.

For say K-means clustering, data will need to be sorted where each row is a member of the group we want to turn into K subgroups or centroids. For example, to detect counterfeit products, we use the description of a known real product as a centroid and measure mean distances to others. Once a fraud is detected, all accounts associated with them by clustering can be invalidated.

Data needs to be collected and merged into a form where each row represents something we want clusters of. Data preparation for clustering depends on the algorithm which can group features by distance measure(i.e. euclidean), centroids, density of data or distribution models. While the distance models are easy to describe, they can be slow on big data sets as you compute a lot of distances and have trouble finding non convex clusters, as do centroid methods. If you wanted to detect non convex clusters, a better approach would be density clustering, though this can ignore outliers and can have low describability.



Software, Hardware and Analytical Tools for Data Mining

The hardware requirements for DDMall data mining are computers and databases etc.
The software requirements are a programming language for data science like Python and relevant libraries.There are open source software tools for the techniques used in this report including Orange, rapidminer, KNIME and BigML.

Conclusion & Recommendations

This report has demonstrated an understanding of the data, the data mining process, as well as the different data mining applications; clustering, classification and association analysis. DDMall's CTO can now leverage the data mining techniques throughout this report to optimise the three main services of the company; for attracting new and retaining existing customers, Association analysis is recommended; for regulation of fraudulent and bad faith behaviour, clustering is recommended; and for demand prediction a hybrid classification and clustering model should be used.

References

Taylor, D. (2022, August 6). Data Mining Tutorial: What is Data Mining? Techniques, Process Source. https://www.guru99.com/data-mining-tutorial.html

Nguyen, M. (2019, September 4). Association rules analysis applied in advertising optimization Source. https://towardsdatascience.com/association-rules-analysis-applied-in-advertising-optimization-d913845eea3e

Gayathri. [IT Miner]. (2015, July 19). Data Mining - Clustering [Video] YouTube. https://www.youtube.com/watch?v=2QTeuO0C-fY

Parikh, B. (2003). Applying Data Mining to Demand Forecasting and Product Allocations. PSU Graduate College. https://h3turing.cs.hbg.psu.edu/mspapers/sources/bhavin-parikh.pdf

Tausz, A. (2020, February 20). Similarity clustering to catch fraud rings. Source. https://stripe.com/blog/similarity-clustering

Poulson, B. (2016, September 6). Data Science Foundations. Data Mining. Source. https://www.linkedin.com/learning/data-science-foundations-data-mining/welcome?autoplay=true&resume=false&u=56744473. Linkedin Learning.

[Ekeeda]. (2021, April 3). Types of Attributes of Data - Data Exploration - Data Mining and Business Intelligence YouTube. https://www.youtube.com/watch?v=S_AosAc9_G8