Data Mining Assignment 1: Data Mining Survey

# Part 1: Survey

## Introduction to Data Mining

Data Mining is a set of computational techniques that involves pattern discovery in large sets of data. Specific techniques classed as data mining include combinations of machine learning, statistics and databases. The primary objective of data mining is to extract useful information from datasets in the form of a useful pattern or structure. Examples of these useful results are identified clusters from cluster analysis, anomalous data within a dataset, association rules that tie together elements in a system, statistical patterns in sequences and predictive analytics.

## Modelling Methods

### Anomaly Detection

Anomaly detection involves the identification of outlying or anomalous data points within a dataset. Such data may be an interesting element that needs extra attention or an error in methodology or data collection that needs addressing. Such anomalies may also need to be identified for elimination from the dataset as certain techniques such as K-Means clustering need to be clear of outliers for accurate results.

### Association

Association involves simple correlation between several data items, usually of the same type or category. Such correlation is used to identify patterns and relationships between items, with application in the retail market. Customer buying habits can be tracked and used to advertise familiar and usually bought products, such a customer who typically buys milk being informed that they may have forgotten to buy milk on checkout.

### Classification

Classification involves the establishment of descriptive attributes that can be applied to a class of customer, item or object. This is useful when introducing a novel piece of information, finding closely similar information and effectively classifying said information into the correct class. As one example, a newly manufactured car may have its attributes analysed to properly categorise what type it is. This is particularly useful for vague information or situations in which there are blurred lines between categories of item. Classification can also be used in conjunction with other modelling methods, ranging from pre-processing data or as the result of a decision tree.

### Clustering

Clustering or cluster analysis is used to obtain a structure from individual pieces of data. Said clusters are excellent for visual presentation of similarities in data, with different items being identified from different clusters. There are different types of clustering such as hierarchical or connectivity based clustering, centroid based clustering, distribution based clustering and density based clustering.

### Decision Trees

Decision trees are typically used as part of the selection criteria for data selection. A decision tree takes the form of answering a question that leads then to another based on the answer for however long the tree functions. This helps classify data into separate, specific areas based on attributes queried.

### Prediction

Prediction is often used in conjunction with other methods in order to predict future events or trends using past and/or present data.

### Regression

Regression or regression analysis is a very common statistical technique used to estimate relationships between data points. It is used for prediction, particularly in machine learning. Specific methods of regression include linear regression, ordinary least squares, nonparametric regression and metric regression.

### Sequential Patterns

Sequential patterns are a method used for trend identification or tracking the regular frequency of related events.

### Summarisation

Summarisation within data mining is an attempt to give a user an overview of the structure of the data. A simple summary can be the objective of a data mining project, for example the total sales for a given month for a business. Typically, summarisation is an early, lesser part of the overall process used to provide insight into the nature of the initially gathered data or to present the final results.

## Data Mining Tools

### Rapidminer (YALE)

A java based program, Rapidminer (previously known as YALE) is a software as a service (SaaS) analytic tool that allows pre-processing, visualization of data, predictive analytics and statistical modelling. It is open source under the AGPL open source license and available free for download, although a commercial version is available for business. An online poll conducted (KDnuggets, 2013) found that Rapidminer was the most popular data analytics/big data/data mining tool in use.

A GUI interface allows the design and execution of analytical workflows called processes, consisting of multiple operators. An operator is responsible for a single task in the process and the output of an operator feeds into the next operator.

The engine can be used as an API or called directly from another program in use. Command line functionality can be used for single use functions and the program can be extended with R and python scripts.

### WEKA

Another java based tool, WEKA (Waikato Environment for Knowledge Analysis) is capable of pre-processing, clustering, classification, regression, visualisation, data analysis and predictive modelling. It notably lacks sequence modelling. It is entirely free and covered by the GNU General Public License allowing full customisation.

All of the techniques in WEKA assume that data is in the form of one flat file or relation, with every data point detailed by a set number of attributes. WEKA can access SQL databases through Java in order to process information obtained. Using Deeplearning4j, an open source deep learning programming library written in Java, WEKA can use deep learning.

### R-Programming

Project R is a GNU project primarily written in C and FORTRAN that uses considerable amounts of R language programming for its modules. Due to its ease of use and its ability to be easily extensible to accommodate a large variety of projects it has seen more use in recent years.

### Orange

Python-based and open source, Orange has components for machine learning, bioinformatics and text mining. One of the major selling points of the program is that it has extensive visual programming and visual display of information. Widgets, either pre-defined or user created are used to link together workflows. Open source under the GNU General Public License, core components are written in C++ with python wrappers. Orange takes advantage of python open source libraries such as numpy, scipy and scikit-learn. The program runs within a Qt framework, an open source application framework that is able to be run over many software and hardware platforms.

### KNIME

KNIME, written in Java, is an open source data analytics, reporting and integration tool. It uses a GUI to help the user in the assembly of nodes during pre-processing consisting of three phases; Extraction, Transformation and Loading. It does this without or with very little programming from the user. It has been used for pharmaceutical research (Tiwari and Sekhar, 2007), business intelligence and financial data analysis.

## CRISP-DM

The de-facto standard for developing a data mining process is a data mining process model, CRISP-DM, also known as the “Cross-industry standard process for data mining”.

The development of CRISP-DM began in late 1996 (CRISP-DM 1.0, 2000) through the work of three businesses; Daimler-Benz, SPSS (Statistical Package for the Social Sciences) and NCR (National Cash Register). Each of the companies had developed data mining technologies but agreed to work together to create a standardised process model in the hopes of marketing data mining to prospective customers in an emerging market. The three companies formed a consortium called CRISP-DM, obtained funding from the European Commission and began work on drafting the process model. To gain input from practitioners of data mining, the consortium created a special interest group and hosted a day in Amsterdam to gather views and ideas on how to create a standardised process model. The response was much better than expected and over the next two and a half years, the consortium worked to develop and refine CRISP-DM. Trials were run at Mercedes-Benz and CRISP-DM was integrated into commercial data mining tools. With the end of the European Community funded part of the project in 1999, the consortium had produced a good initial draft and approximately one year later the CRISP-DM 1.0 was published.

The methodology of CRISP-DM is described as a hierarchical process model, consisting of six phases. These phases are divided into generic tasks which are in turn split into a set of specialised tasks. The specialised tasks are split into process instances which are specific instances or records of the actions, decisions and results of data mining. Ultimately this division into sub levels with tasks to be performed in a specific order is an idealised approach, it can be necessary to repeat certain tasks, which is an approach taken by the more refined ASUM-DM model which was designed in part to account for this.

CRISP-DM separates the data mining process into six distinct phases:

### Business Understanding

This phase consists of four generic tasks, with this phase centred around a project management perspective that works to setup the project fully. Project managers with oversight from senior stakeholders are the primary resource involved with this phase.

The first generic task is determining the business objectives that consists of:

* Documenting the background of the project
* Determining the actual objectives of the business undertaking the project
* Defining the success criteria for those objectives having been completed

The second generic task, assessing the situation involves:

* Taking an inventory of project resources
* Recording the project requirements, assumptions made and the constraints on the project
* Conducting risk management to assess the risks and contingencies for potential issues
* Defining terminology within the project
* Detailing the costs and benefits of the project.

The third generic task, determining the data mining goals has the following outputs:

* Establishing the data mining goals for the project
* Defining the success criteria for the data mining work

The fourth generic task for business understanding, produce a project plan is composed of:

* Producing a project plan
* Making an initial assessment of tools and techniques to be used during the project

While not mentioned within the final generic task within the list of outputs in the CRISP-DM guide (CRISP-DM 1.0, 2000), it is mentioned elsewhere that a problem definition is made at this point.

### Data Understanding

This phase is based on the collection of data for the project. Initial work on the data may lead to the first insights into patterns or subsets of the data that be prove useful in the future. Quick wins taken from this insight may be achieved here for the business if suitable. The primary resource in this phase should be data scientists.

This first generic task is collecting the initial data which consists of performing the initial data collection and then producing a report on the results of this initial gathering.

The second task, describing the data involves formally documenting a description of the data in an associated report.

The third task is exploring the data, taking a deep dive into the data gathered to in part, look around at the quality of the data, gain familiarity with the dataset and to find anomalous data. The results of this are recorded in a formal report.

The final generic task is verification of the data quality and then documenting the quality of the data in a data quality report. Issues with the quality of data should be resolved at this point.

### Data Preparation

Data Preparation consists of five generic tasks based on getting pre-processing data for the modelling phase. Data scientists are the primary resource.

The first task, data selection consists of the data being selected based on defined rationale for the inclusion and exclusion of given data from the dataset used.

The second task, clean data involves preparing data from its raw form into a form that can be accurately read into the modelling techniques without issue. Corrupt and inaccurate records are removed at this point in addition to typographical errors. A data cleaning report is created once this task is complete.

The third task is constructing data, the putting together of data into its useable project specific form using derived attributes and generated records.

Fourth is the integration of data, where data is merged into a larger dataset. Data rationalisation, a process involving fitting together heterogeneous groups such as files with similar but different column headers occurs here.

Fifth, formatting of data is where data is formatted correctly ready for use in the modelling process. Prefixes and suffixes may be added or taken away, datatypes may be changed into another form such as integers into strings and the data is transformed from a rawer form into a form ready to be directly read into a modelling technique.

Once the tasks are complete, the dataset to be used is explicitly defined in terms of scope with a dataset description noted.

### Modelling

The modelling phase is where the actual implementation of data mining takes place using the data that has been prepared in the previous phase. As with the previous two phases, data scientists should be the main resource used.

The first generic task is the selecting of the appropriate modelling techniques, documenting the techniques to be used and then stating the assumptions to be made in using said techniques.

The second task is the generation of a test design or designs for recording results later.

The third task is building the model. The model first needs to have its parameters set, modelling techniques implemented within the grander model and then descriptions of the model noted. A variety of modelling techniques are applied with their use properly calibrated. If necessary, further data preparation is used to tailor the dataset for certain techniques. This part of the process can be iterative, until the model is of high enough quality.

The final task is assessing the model. The model is given a full assessment with parameters revised as necessary.

### Evaluation

Evaluation is where the data mining process is evaluated fully. The primary point of this phase is to determine if there are any business issues that have not been properly addressed. With the end of this phase, a decision on how to use the results of the data mining should be finalised. Resource wise, data scientists should present the results to project managers who then fully document the outcome of the data mining process.

The first task is to undertake a full evaluation of results with the assessment of data mining results with regards to the business success criteria and the recording of the approved models used.

The second task is, reviewing the data mining process.

The third task is to determine the next steps. A list of possible actions and decisions to take the project forward is drawn up.

### Deployment

The final phase of CRISP-DM, deployment consists of the results of the model being created, being transformed into useful, organised and presented in a customer focussed manner. This phase may range from a simple report to iterating through a big data, data mining process through a full business cycle. The customer is often the one to deploy the model and so they need to be aware of the actions needed to implement it. The primary resource for this phase is project managers due to the reporting of the project’s results and project closure.

The first generic task is planning the deployment with a deployment plan.

The second task, plan monitoring and maintenance, consists of developing a monitoring and maintenance plan to equip future users of the data mining model to properly maintain the model if it is to be in use into the future. Much of the time, a successful model is retained for possible future use.

The third task is the production of a final report and a final presentation which fully summarise results and end the project. The presentation should be aimed towards senior stakeholders and owners of the project.

The final task is the reviewing of the project with the experience gained from the project fully documented.

## ASUM-DM

A recently developed process model for implementing a data mining or predictive analytics project, ASUM-DM (Analytics Solutions Unified Method-Data Mining) is according to Haffar:

the “Analytical” activities and tasks of CRISP-DM but the method was augmented with missing activities and tasks as well as templates and guidelines. In other words ASUM-DM is nothing more but an extended and refined CRISP-DM. (Haffar, 2018).

The main reason for the creation of ASUM-DM is given by Haffar as CRISP-DM not covering the infrastructure and operations side of implementing a project. Project management tasks you’d expect to see in a data mining project are notably absent by and large within CRISP-DM. While the external version of ASUM-DM is available publicly, it is obvious from a cursory read through of the documentation that the process model is geared towards IBM internal processes, it is not written in a generalist way that CRISP-DM is. The “Set up Environments” stage is particularly geared towards IBM internal processes with almost the entire stage focussed on the installation of IBM software and the setting up of QA and production teams of IBM staff. Even so, if the documentation is suitably adapted to an organisation, ASUM-DM would be a better view of a data mining project plan from a project manager’s or senior stakeholder’s perspective. Going further, the style of the documentation is written in a manner reminiscent of Microsoft Project.

Unlike the six phases of CRISP-DM, ASUM-DM has three phases in part because the initial phase is repeatable:

### Analyse, Design, Configure and Build

This phase is repeatable as data mining/predictive analytics projects are iterative in nature, according to IBM (IBM Analytics Solutions Unified Method (ASUM) - External, 2015).

There are fourteen stages within this phase:

* Prepare for Implementation
* Conduct Readiness Assessment
* Conduct Project Kick-off
* Understand Business
* Understand Data
* Design and Validate Infrastructure
* Set up Environments
* Prepare Data
* Build Model
* Evaluate Model
* Conduct Analytical Knowledge Transfer
* Define Deployment Approach
* Design Operational Testing Strategy
* Validate and Test in QA Environment

It is easy to see that compared with CRISP-DM, ASUM-DM combines the first five phases of CRISP-DM into its first phase. Everything except the deployment of the model is done. The biggest differences are the addition of project-management oriented tasks which make up a large part of this phase.

### Deploy

Deployment consists of six stages:

* Conduct Operational Knowledge Transfer
* Prepare for Ongoing Maintenance
* Deploy Solution
* Transit to IBM Support
* Launch
* Prepare for Project Closure

### Operate and Optimise

Operate and Optimise consists of five stages:

* Monitor Model
* Operate, Optimise and Improve System
* Support User Community
* Manage Infrastructure
* Govern System Lifecycle Program

## Applications and Problem Types of Data Mining

### Automatic Credit Approval

Automatic credit approval has been identified as an application where data mining is used within the context of the banking sector (Chitra and Subashini, 2013). Fraud is prevented during the approval process using classification models based on decision trees, support vector machines and logistic regression techniques.

The two decision trees used within this study were C5.0 and CART, which both serve to act as classification models when selecting someone for credit approval.

C5.0 creates a decision tree from training data, splitting sets of data into subsets at each node of the tree based on the criteria of the normalised information gain measured as a difference in information entropy. An attribute with the highest information gain is used to make decisions within the context of a decision tree. While a little difficult to understand initially, this type of decision tree is used because small decision trees result from this method, meaning few questions that need to be asked.

CART (classification and regression tree) in comparison uses a binary decision tree based on splitting data into two child nodes at each level, starting with the root node that contains the whole set of the training data. A calculation used within the algorithm to generate this tree is Gini impurity, the measure of how often a randomly chosen element is correctly labelled if it is randomly labelled, relative to the distribution of labels in the subset. It is calculated by summing the probability of each item chosen times the probability of a mistake being made in categorising it.

Another technique used, a support vector machine (SVM) is used in machine learning as a polynomial kernel, a kernel function. The paper (Chitra and Subashini, 2013) is of questionable value in describing this technique as a standalone paper due to the Wikipedia page describing polynomial kernels being identical and almost confirming direct plagiarism by the authors in a simple copy and paste operation. In any case, a SVM represents the similarity of vectors in feature space over polynomials of the original variables which allows the machine to learn non-linear models. The paper is unclear on how the results are gained from these calculations.

A final technique described is logistic regression which is a type of regression analysis use for predicting the outcome of a categorical dependent variable based on one or multiple predictor variables. Logistic regression is described as easy to implement with good performance from a computational perspective.

### Automatic Text Summarisation

Automatic text summarisation is the summarisation of a text document to create a shorter, more condensed version that contains the main points in the original document. There are two approaches to this summarisation, extraction and abstraction. Extraction works by picking out existing elements from the text to create the summary. Abstraction works by making a semantic representation of the document and using natural language generation to create a summary in a human-like style.

One primary example of text summarisation is key phrase extraction. According to Turney (Turney, 2000), journals ask authors to provide a list of keywords for their articles. Key-phrase extraction is supposed to serve the goals of:

* Enabling the reader to quickly determine whether the given article is in the reader’s field of interest
* To be used in indexing to enable the reader to find a relevant article
* For search engines that use keywords, giving more precise results

Automatic key phrase extraction is a more general task of automatic key phrase generation where the key phrases generated do not necessarily even appear within the document used but 75% of generated key phrases do appear somewhere in said document. Key phrase extraction algorithms have been in use for decades, notably in Microsoft Word 97 and what are now defunct software products such as Metabot and Verity Search 97 in use at the time of writing. Within Word 97, keywords are automatically selected and set as the key words for the document within the document metadata.

## Customer Relationship Management

Customer Relationship Management, also known as CRM involves management of the relationship between current customers, potential customer and a company. Using data mining techniques, a customer’s history is used to drive retention efforts and feed sales growth.

CRM began in the 1970’s with customer surveys on the front line or through annual surveys (Financesonline.com, 2018). It later evolved into using databases of customer information and statistical methods of analysis starting in 1982 with Kate and Robert Kestnbaum. By 1995, CRM as a phrase first entered the public domain with credit going to either Tom Siebel of Siebel Systems who first developed a CRM-like program or IBM. In 2003, Microsoft rolled out Dynamics CRM and in 2004 the first open source CRM was developed by SugarCRM. In the late 2000’s connections were made to cloud computing and social media. In the years since CRM has used “business intelligence” to perform data driven decisions within companies, culminating in industry specific implementations of CRM.

A typical CRM is composed of data warehouses technology to provide the data used and software as a service (SaaS) usage of CRM software. The specific role being determined by its application in operational use, analytical use, collaborative use or as a customer data platform. Operational use of CRM includes sales force automation, marketing automation and service automation. Analytical CRM systems analyse customer data and present it for business managers to make data-driven decisions. Data mining features prominently in selecting, extracting, processing and then presenting the data to power these decisions. Collaborative CRM systems are used to integrate external entities such as suppliers and distributors into a larger network to share customer information. A customer data platform is a system that collects data on individual people and builds up a large database which can be accessed from external software.

## Fraud Detection

According to a BBC news article (BBC, 2016), fraud costs nearly £3000 per head of population with total fraud coming in a cost of £192 billion. The study cited by BBC (PKF, 2016), states that UK fraud by sector works out to proportions of 74% (144bn) in the private sector, 20% (37.5bn) in the public sector, 5% (10bn) by individuals and 1% (1.9bn) in the charity sector. The report goes further to say that in the private sector:

It is estimated private sector fraud could cost the UK economy up to £143.6 billion. But further analysis suggests that may be a conservative figure, given the general sentiment among our biggest businesses against releasing commercially sensitive, or potentially damaging, financial fraud data. Right now no comprehensive data exists in the public domain in the UK. (PDF, 2016)

With the use of AI techniques as an emerging field, data mining is a major component of detecting where fraud occurs. Classification, clustering, segmentation and the finding of associations and patterns within the data where fraud is present is performed. Such patterns and rules found with data mining can then be used within machine learning techniques and/or neural networks to more automatically find cases of fraud in a set of data.

## Healthcare

Data mining has potential to identify inefficiencies and best practices that improve care and reduce costs according to health catalyst (Health Catalyst, 2018). Some estimates give as high as 30% of overall healthcare spending being reduced consequently. Due to the complexity of healthcare and a slow rate of technological adaption, healthcare has been slow to take up data mining practices that have been used elsewhere in business and finance.

According to Koh et al. (Koh and Tan, 2005), the vast potential for data mining in healthcare can be achieved through many fields:

* Evaluation of treatment effectiveness
* Management of healthcare
* Customer Relationship Management
* Detection of fraud and abuse
* Predictive Medicine

The effectiveness of medical treatments can be evaluated through the comparing and contrasting causes, symptoms and courses of treatments. United Healthcare, the largest healthcare company in the world by revenue has used data mining on its treatment record data to explore ways in which to cut costs and deliver better medicine.

To improve healthcare management, data mining can be used to identify and track both chronic disease states and high-risk patients and reduce the total number of hospital admissions. Blue Cross Blue Shield, an American health insurance federation has used data mining to reduce spending by having better disease management using emergency department, hospitalisation claims data, pharmaceutical records and physician interviews to help identify unknown asthmatics and intervene where necessary.

Customer relationship management (CRM) as mentioned previously as an application of data mining can be applied also to healthcare. By determining the preferences, usage patterns and current and future needs of patients to improve their satisfaction, customer relationships can be improved.

A notable case of fraud and abuse detection occurred in 1998 when the Texas Medicaid Fraud and Abuse Detection System recovered $2.2 million and identified 1400 suspects in less than one year’s operation. This data mining system in 2004 according to Secure Tech Alliance (Securetechalliance.org, 2006), used data from the Smart Card Program, a program aimed at combatting fraud. It utilised data taken from verification of patients attending appointments, time stamped data signatures with verification of duration of visits and helped combat identify fraud directly by using biometric cards.

# Part 2: Scenario

## Introduction and Assumptions of Scenario

## Choice of Strategy

## Possible Implementation Methods

# References

BBC News. (2016). *Fraud costing UK '£193bn a year'*. [online] Available at: http://www.bbc.co.uk/news/uk-36379546 [Accessed 9 Feb. 2018].

Chitra, K. and Subashini, B. (2013). Data Mining Techniques and its Applications in Banking Sector. *International Journal of Emerging Technology and Advanced Engineering*, 3(8).

CRISP-DM 1.0. (2000). 1st ed. [ebook] The CRISP-DM consortium. Available at: https://www.the-modeling-agency.com/crisp-dm.pdf [Accessed 7 Feb. 2018].

Financesonline.com. (2018). *Best CRM Software Reviews & Comparisons | 2018 List of Expert's Choices*. [online] Available at: https://crm.financesonline.com/#history-of [Accessed 9 Feb. 2018].

Health Catalyst. (2018). *What is Data Mining in Healthcare?*. [online] Available at: https://www.healthcatalyst.com/data-mining-in-healthcare [Accessed 9 Feb. 2018].

Kdnuggets.com. (2013). KDnuggets Annual Software Poll:RapidMiner and R vie for first place. [online] Available at: https://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html [Accessed 5 Feb. 2018].

Haffar, J. (2018). *Have you seen ASUM-DM? - SPSS Predictive Analytics*. [online] SPSS Predictive Analytics. Available at: https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/ [Accessed 7 Feb. 2018].

IBM Analytics Solutions Unified Method (ASUM) - External. (2015). IBM.

Koh, H. and Tan, G. (2005). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management*, 19(2).

PKF (2016). *Annual Fraud Indicator 2016*. Annual Fraud Indicator. [online] Portsmouth: Experian. Available at: http://www.port.ac.uk/media/contacts-and-departments/icjs/ccfs/Annual-Fraud-Indicator-2016.pdf [Accessed 9 Feb. 2018].

Securetechalliance.org. (2006). *Texas Medicaid*. [online] Available at: https://www.securetechalliance.org/resources/lib/Texas\_Medicaid.pdf [Accessed 9 Feb. 2018].

Tiwari, A. and Sekhar, A. (2007). Workflow based framework for life science informatics. *Computational Biology and Chemistry*, 31(5-6), pp.305-319.

Turney, P. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4), pp.303-336.