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.

## 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 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.

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

### Business Understanding

The establishment of objectives and requirements from the business’s perspective, followed by the creation of a problem definition and the plan for delivering the project. This phase consists of first determining the business objectives, assessing the situation, determining the data mining goals and producing a project plan.

### Data Understanding

This involves initial data collection, gaining familiarity with the dataset and resolution of quality issues. 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 may be achieved here for the business. Data Understanding consists of collecting the initial data, describing the data, exploring the data and then verifying the data quality.

### Data Preparation

Data is pre-processed ready for the modelling phase. This processing can involve the selection of relevant data with the culling of unneeded data, sorting into the necessary order, cleaning of data to fit the modelling techniques and rationalisation of data to fit heterogeneous groups. Data preparation consists of selecting the data, cleaning the data, constructing the data, integrating the data and formatting the data.

### Modelling

A variety of modelling techniques are applied with their use properly calibrated. Techniques are implemented and 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. Modelling consists of selecting the modelling technique, generating a test design, building a model and assessing a model.

### Evaluation

With an apparently high quality model built, the model is evaluated and reviewed to ensure that the business objectives have been reached. The primary point of this stage 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. Evaluation consists of evaluating the results of the data mining approach, reviewing the process implemented and determining the next steps to take.

### Deployment

The results of the model being created are 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. Deployment consists of planning the deployment, plan monitoring and maintenance, producing a final report and reviewing the project. Documenting the accumulated experience of the project is the final step.

## 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

## 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 particular 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 particular 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 in order 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 or automatic 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 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.

## 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.

## Applications and Problem Types of Data Mining

# Part 2: Scenario

## Introduction and Assumptions of Scenario

## Choice of Strategy

## Possible Implementation Methods

# References

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].

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

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.

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].