UNIVERSITY OF TECHNOLOGY, SYDNEY
Faculty of Engineering and Information Technology

Association Rule Mining and Decision Making for Closed-Ended Hierarchical Questionnaire Data

by

Mark Caple

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Masters by Research

Sydney, Australia

2020

Contents

Abbreviation

ARM - Association Rule Mining

AUC - Area Under Curve

DSR - Design Science Research

ROC - Receiver Operating Characteristic

Chapter 1

Research Description

1.1. Introduction and Background

Since humans began to communicate they have asked questions about a multitude of diverse topics. Indeed studies show that children can ask tens to hundreds of questions of their parents per day. These questions would probably have a constantly changing theme dependant upon what was puzzling them at the time.

There are however many times when a theme would be beneficial when wanting to draw some sort of conclusion to answers given. Local councils may ask a series of questions to gauge public appetite for new works that they plan to introduce or a company may check their clients acceptance of staged branch closures before they happen.

When such scenarios arise the humble questionnaire is front and centre as a means to gather information. The questionnaire, although common place today, has a history of less than 200 years. It's origins have been attributed to the Statistical Society of London in 1838 (Gault, 1979 [**?**]) whose main goal was "procuring, arranging and publishing facts to illustrate the condition and prospects of society".

Distilling information from the answers given to these questionnaire's has been a goal of machine learning for quite some time and much research has covered various approaches. It is important however before mentioning the machine learning work to categorise the individual questions of a questionnaire into two broad types being either closed-ended or open-ended (Marshall, 2005 [**?**]). A closed-ended question

would have one correct answer or a limited number of options. An open-ended question on the other hand would not have a correct answer but rather would allow the participant to enter exactly what they believed appropriate. Open-ended can be considered to promote long responses whereas closed-ended short responses.

So when discussing machine learning research, open-ended questionnaires allow the participant to use free or open text to answer the question and typically the research then incorporates the use of techniques such as Natural Language Processing to infer a conclusion. Closed-ended (Howard & Presser, 1979 [**?**]) questionnaires on the other hand allow participants to answer a series of questions using multiple short answer types. Marshall (2005 [**?**]) defines five data types:

- category - represents a set of mutually exclusive categories (e.g male, female)
- list - multiple category choice is possible as the answers are not exclusive, e.g "what services have you used from your GP in the last year?".
- quantity/numeric - such as "how many times have you broken your leg?"
- ranking/scale - such as "how would you rate your doctor [1-7]"
- linguistic ranking/scale - such as "would you describe yourself as: very tall, tall,short,very short?"

The research community is not so united in their approach when it comes to closed-ended questionnaire data with no clear technique winning out over another. One of the characteristics of a closed-ended questionnaire that adds to its complexity is the fact that it is able to contain so many different types of answer.

One approach that has been considered worthy of handling questionnaire responses is Association Rule Mining (ARM). Agrawal et al. (1993, [**?**]) represents a seminal piece of research in the field defining not only the problem but also the

mechanism to handle it. The mechanism involves an easily understood procedure wherein frequent itemsets are included that obey some minimum support and then association rules are created that satisfy some minimum confidence.

ARM has been adopted widely to determine the purchasing habits of consumers but over time it has been applied to a diverse problem landscape including product recommendation, web page caching mechanisms, medical diagnosis, census data analysis and protein sequencing.

My literature review has identified that there has to date been very little research that is able to mine association rules from closed-ended questionnaire data. The approaches that have been adopted have predominantly used crisp or boolean values where a section of the questionnaire is analysed in isolation as the data is of the same data type. For the approach of this research all of the five data types mentioned by Marshal (2005 [**?**]) should be handled together and also the possibility of multi-value answers needs to be addressed. Chen et al. (2009 [**?**]) was the first to use fuzzy association rules on questionnaire data and in so doing was able to handle all of the data types simultaneously. They achieved this through no longer considering answers as true/false but as partial truths thus any answer that is of a linguistic type can considered alongside a non fuzzy type. Although the research had some success the authors do concede some shortcomings in the approach. The first being the use of static membership functions that are defined ahead of time that create roadblocks in the process.This work will consider a dynamic membership function which will derive the function from the data. Another being a mechanism for analysing associations between questionnaire's over time.

One further research direction that to the best of my knowledge has not previously been investigated in the context of a questionnaire, is to take the fuzzy association rules produced and apply some neural network algorithms to improve

the findings. Mamuda et al. (2017 [**?**]) show that it is possible to tune the parameters of fuzzy rules using traditional gradient descent. The added advantage of this was that it "allowed a membership function of the rule to be used more than one time in the fuzzy rule base".

Our university industry partner has a core service that is pre-employment assessments. Currently they offer third party organisations an efficient means to bring candidates on-board which can involve interview(s), medical questionnaires and medical assessments. It is the intention of this research that through adopting association rule mining on those medical questionnaires and fine tuning with machine learning techniques the number of actual medical reviews and time to selection would be reduced.

Thesis statement should come at the end of this

A Thesis Statement is"A statement or theory that is put forward as a premise to be maintained or proved."

1.2. Stakeholders

A critical stakeholder in this research would be the industry partner.

The take up of machine learning within industry has been been somewhat stymied by the misunderstanding that machine learning cannot appropriately be deployed within the workplace as the results, even when accurate, cannot be explained. The term "black box" is often used which itself would suggest some kind of magic. However the individual components used within the field of machine learning are very well understood and hence this description is inappropriate. It should thus be apparent that by validating this research for the benefit of our industry partner that any industry group thinking about deploying a machine learning project would themselves be valid stakeholders.

The last stakeholder would be any researcher needing to make a prediction from closed-ended questionnaires.

1.3. Aims and questions

The main aim of the project is closely related to the most critical stakeholder and industry partner.

How can we apply machine learning techniques to a questionnaire to replace the role of high cost medical assessments used in selecting a candidate for a specific job role and yet still avoid the liability risk of an incorrect choice?

From the above aim we are able to produce the following research questions:

Question 1: Is it possible, in a timely manner, to reduce the need for a physical medical assessment for a job role by introducing a suitability predictor using only responses given in a medical questionnaire?

Question 2: Is it possible to improve upon the suitability predictor by allowing actual medical assessment results to be fed back into the live system?

Question 3: Would removing rare or anomalous candidates from the pool of candidates create a better suitability predictor?

Question 4: How to analyse and compare the results of repeat medical assessments from the same candidate for different job roles over time?

Question 5: How to verify and validate the above aims?

1.4. Objectives

This study will have 5 objectives. The first objective will use a current production system that has a lot of data but lacks some integrity constraints to ensure that data

is always correct. The next 3 objectives will use a system that is being developed in parallel with this study but suffers from not having real data but does address many of the concerns of the current system.

Objective 1. To classify a candidate into a small number of groups that give a sliding suitability score.

This objective corresponds to research question 1. To the best of my knowledge, through my literature review, no research exists that addresses the classification of closed-ended medical questionnaire's using fuzzy association rule mining. Furthermore, my literature review shows that no work has taken the predicted rule parameters from such mining and applied neural networks to fine tune the results.

Objective 2. To define a mechanism whereby results of physical medical assessments are fed back into the system for a better predictor.

This objective corresponds to research question 2. The physical medical assessment is outside the core architectural boundaries of the current system. There are many factors to this including geographic remoteness or the use of third party independent assessors. In order for any selected candidate to be both accurately determined and done so in a timely fashion it is vitally important that results of physical assessments are fed back into the system promptly. In fulfilling this the research will not only complete a design criteria but will also help to fill another literature gap. Chen et al. (2009 [**?**]) concluded that their design proved to be slow as the fuzzy membership routines were static. In feeding back this assessment information into the system the research will use the data to deduce dynamic membership routines.

Objective 3. To build an anomaly detection routine to predict a list of candidates of concern.

This objective corresponds to research question 3. The literature review has

demonstrated very little research that has catered for questionnaire responses to all of Marshall's (2005 [**?**]) five data types. The main objective of this research is to use fuzzy association rule mining to improve on this situation. This objective however will endeavour to apply a number of the state of the art algorithms used in machine learning to discover rare conditions within the candidates. In completing this objective the very small pool of anomaly detection that has been applied to closed-ended questionnaire data will be enhanced. Once these anomalies are detected the research will verify if removal of such candidates leads to a better predictor of the system as a whole.

Objective 4. To build a model whereby assessments maybe compared along a timeline so that assessments taken multiple times maybe analysed.

This objective corresponds to research question 4. Chen et al. (2009 [**?**]) had demonstrated the inability of their research to analyse associations between questionnaire's over time. By addressing this gap this research will also benefit the industry partner's goal of comparing candidate's assessments over time. This will answer a broader question of whether a given candidate may prove unsuitable for any role or simply an individual role. One further goal of the industry partner is to allow unsuccessful candidates to be given suggestions of possible alternate rolls that they would be suitable for. This objective goes a long way to achieving this goal.

Objective 5. To evaluate the developed artefacts from the previous objectives.

This objective corresponds to research question 5. For our classification objective there are a number of typical tests to indicate the correctness of a classification and these include amongst other measures confusion matrix, area under the ROC curve and F1 score.

Briefly if we are describing a binary classification such as the candidate is "suit-

able" or "not suitable" we can plot a confusion matrix of the form shown in Fig **??**. From this matrix we can calculate attributes such as accuracy, precision and recall to decide whether our results have been noteworthy. The reader may at this point believe that to be noteworthy we should simply strive for the highest accuracy in our classification but that may not always be the case. For instance we may gain an accuracy of over 98% but if we ultimately select a candidate who would have been rejected if given a physical assessment then this could present an unforeseen cost to the client and also a loss of faith in the predictor. This then brings us to the other two attributes precision and recall which represent the ratio of true positives in the model to the predicted positives and actual positives respectively. Precision should be closely watched when the cost of a false positive is high and recall when a the cost of a false negative is high.

Initially all of our candidates will have a medical assessment and so the cost of a false positive will not truly be of concern. That does not however mean that our success of a classification for this research should only look at accuracy as eventually the medical assessor should not need to be called upon for every situation. It does however offer some flexibility in deciding the success criteria at this stage of the research.

Fawcett (2006 [**?**]) explains the intricacies of an ROC graph, an example of which, is shown in Fig **??**. It is a very visual means by which the correctness of a classifier can be judged by varying the threshold. The threshold being a value between 0% and 100%, that is used to set the limit to decide upon which class an instance belongs to. In our case, we will classify a candidate as either "suitable" or "not suitable". The graph plots the true positive rate against the false positive rate. For us the true positive rate is the rate in which candidates are correctly identified as "suitable" for the job role in question. A classifier that approaches the top left of the graph is considered a better classifier than one further away. The closer the curve

Figure 1.1 : Confusion matrix

	Suitable (Actual)	**Not Suitable (Actual)**
Suitable (Predicted)	TP (True Positive)	FP (False Positive)
Not Suitable (Predicted)	FN (False Negative)	TN (True Negative)

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$
$$precision = \frac{TP}{(TP + FP)}$$
$$recall = \frac{TP}{(TP + FN)}$$

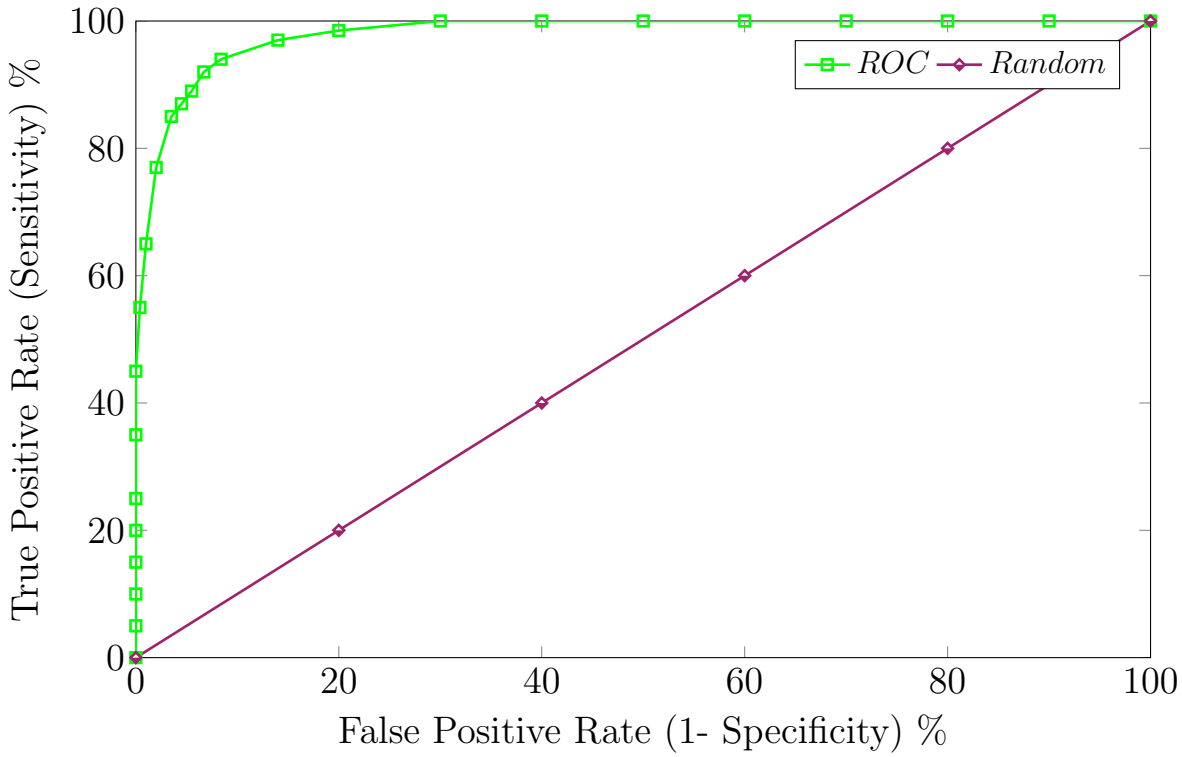
comes to the "random" 45 degree line the less accurate the classifier and this would equate to using a "coin toss" to decide upon the candidates suitability. Although an ROC graph is a very visual tool, to evaluate multiple classifiers the approach used is to take the area under the ROC curve (AUC). Generally, although not always true, a high AUC score is a better predictor than one that is lower. One advantage that an ROC graph has over a confusion matrix is that it does not depend on class distribution and hence is still suitable for evaluating classifiers that contain rare or anomalous values.

The final success indicator that we will employ is the F1 score. This score is based on the precision and recall values mentioned in Fig **??**

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

It is used when a balance is required between precision and recall. The reason

Figure 1.2 : ROC graph



$$TruePositiveRate = \frac{TP}{(TP + FN)}$$
$$FalsePositiveRate = \frac{FP}{(TP + FN)}$$

for the F1 score over using straight accuracy is that accuracy is affected highly by true negatives which are not often a focus in business problems. False negatives and positives usually claim the majority of the focus as they are responsible for most of the cost involved in incorrect classification. Thus the F1 score seeks to find balance between precision and recall when there is an uneven class distribution.

Each objective maybe linked to the publication of a research paper

1.5. Methodology

Babbie (2015 [**?**]) states that research is a *"systematic and orderly approach taken towards the collection and analysis of data so that information can be obtained from those data"* A research methodology establishes the framework for research, amongst other things it defines strategy, approach and components for the research.

Methodologies can be categorised as either qualitative or quantitative and further broken down into approaches such as survey designs, case study, action research, constructivist grounded theory, bibliometric research, design science research, researching history, ethnographic research and experimental research (Williamson & Johanson (2017 [**?**]).

Over time certain methodologies have been put forward that suite the field of Information Systems which is the chosen field of this work. It is Design Study Research (DSR) which has been selected as the most suitable of these for this particular work. In DSR *"researchers focus on building some kind of artefact they believe will be useful to a particular stakeholder community. They then evaluate the merits of the artefact in various ways"* (Williamson & Johanson (2017 [**?**]). Figure **??** shows the differences between the classic approach, which creates artefacts to attempt to build or test a theory and the DSR approach which builds artefacts that are useful to certain stakeholders.

Hevner et al. (2004 [**?**]) argued that DSR artefacts can take one of the four forms shown in Figure **??**. This definition has not however gained widespread approval and other researchers have put forward differing claims to what defines a DSR approach. Authors such as Gregor and Jones (2007 [**?**]) disagree with this principle and instead suggest a framework where design-research theory is paramount.

Indeed one of the issues of DSR is in proving any such research has indeed

Figure 1.3 : DSR artefact forms

Form 1: **Constructs** represent conceptual objects which describe real world "things" such as businesses, employees, levels of debt, sale of products, state of liquidity...

Form 2: **Models** represent a subset of real world "things", it is a way by which we can reduce complexity. An example maybe the way we break database systems down into their three sub levels of internal schema, conceptual schema and external schema.

Form 3: **Methods** are a set of actions that used together achieve an outcome. The outcome could be a product or a service. One example being test based development which is a method that improves upon software development where the development represents a product.

Form 4: **Instantiations** are hardware or software systems that we use produce either a construct, model or method.

been done in a rigorous manner as no single approach has been adopted as the gold standard. The approaches that have been put forward include Hevner et al. (2004 [**?**]) which suggests a set of 7 non mandatory guidelines (Table **??**) that should be "addressed in some manner for design-science research to be complete".

A number of concerns of this approach have been pointed out by academics, including its generic applicability to other types of research apart from DSR and the difficulty in gauging some of the guideline's aims. For instance what makes an artefact viable or how do we know when an artefact has been produced rigorously?

Gregor and Jones (2007 [**?**]) does not suffer from this generic criticism. They

Table 1.1 : Hevner et al. 7 Guidelines for Design-Science Research

No.	**Guideline**
1	Produce a viable artefact
2	Ensure that the artefact produced is relevant and important
3	Rigorously evaluate the artefact produced
4	Produce an artefact that makes a research contribution
5	Follow rigorous construction methods
6	Show the artefact is the outcome of a search process
7	Clearly communicate the research process and outcome

put forward an approach which looks at design-science theory and states that design-science theory has 6 obligatory components and potentially a further 2 optional ones (Table **??**).

Again this approach has suffered criticism in giving only minimum guidance in the pursuit of research rigour. Another criticism is it only considers 'method' and 'product' artefacts from the 4 possibilities mentioned in Figure **??** that defines DSR.

Peffers et al. (2007 [**?**]) suggest a research methodology using a six step process for correct implementation of DSR (Table **??**). Through these steps Peffers et al. claim they are able to confirm design-science research which is "valuable, rigorous, publishable". Others have made similar claims with many frameworks having some comparable ideas along with their own strengths and weaknesses. Many practitioners advise that you simply choose one that fits your research and then simply adhere to it rigorously.

Objectives that a researcher must conform to demonstrate well structured DSR should show clearly that they understand the problem at hand and not jump to

Table 1.2 : Gregor and Jones's 8 Guidelines for Design-Science Theory

No.	**Guideline**
1	Purpose and scope
2	Constructs
3	Form and function
4	Mutability
5	Testable propositions
6	Justifactory knowledge
7	Implementation principles *(Optional)*
8	Instantiation *(Optional)*

a solution first approach. Do not create a 'solution looking for a problem'. In describing the problem they must specify who is experiencing the problem, what is the nature of the problem and success criteria. Why this problem can't be solved with existing means. When the problem arises, where it occurs and stakeholders affected by it. Through covering all these questions the researcher is explaining they understand the nature and boundary of the problem. The boundary of a problem is important as too narrow a scope would be considered uninteresting in the research community and too wide a scope, impractical.

Figure 1.4 : Peffers et al. 6 Step Iterative Process for Conducting DSR

Amongst the outcomes of this research will be the development of a number of novel algorithms to be incorporated into a commercial software product. It is the algorithms that are developed during the design phase that will satisfy the artifact requirement of DSR. The stakeholder community will initially involve the industry

partner of the university but will ultimately be useful to anyone dealing with the problem of classifying the answers to closed survey/questionnaire data.

article [demo]graphicx [font=small,labelfont=bf]caption

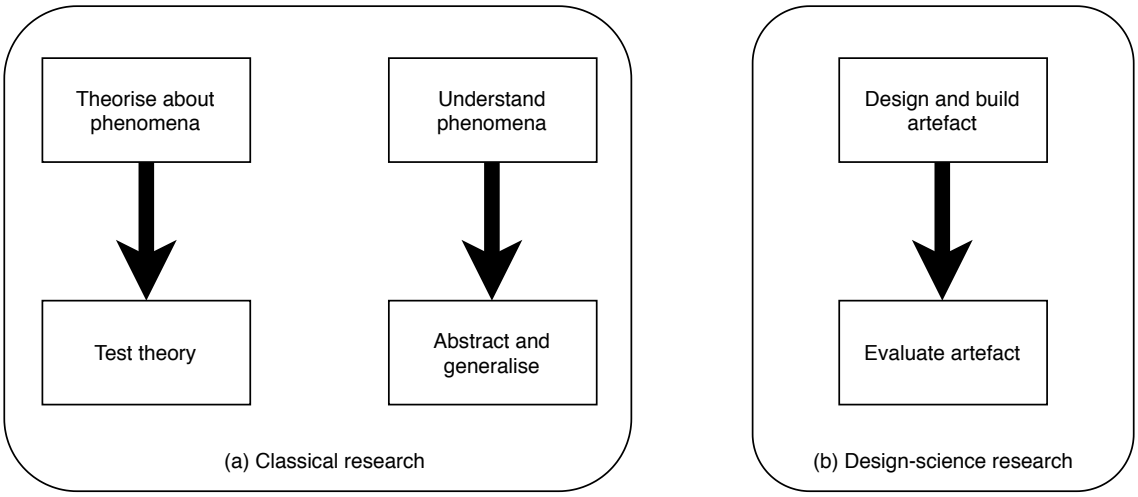


Figure 1.5 : Types of research