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

ANZSCO - Australian and New Zealand Standard Classification of Occupations

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 [19]) 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 [38]). 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 [44]) questionnaires on the other hand allow participants to answer a series of questions using multiple short answer types. Marshall (2005 [38]) 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, [1]) 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 [38]) should be handled together and also the possibility of multi-value answers needs to be addressed. Chen et al. (2009 [10]) 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 be 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 [36]) 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.

The proposal is organised as follows. Section 1.2 covers a systematic literature review including any discovered gaps. The stakeholders that are likely to gain from this work are described in Section 1.3 . Section 1.4 introduces the main aim of the research and any questions that arise from that aim. The objectives that are born from these questions are explained in Section 1.5. Section 1.6 highlights the way in which this research will be undertaken to ensure it maintains sufficient rigour. Section 1.7 offers the research plan for the work including initial designs of how any proposed objectives will be attained along with a general architectural description of how these objectives fit in with the system as a whole. Section 1.8 talks about the impact this research will have both on the university's industry partner and also the research community in general. Finally Section 1.9 gives a broad timeline for the research.

Thesis statement: *Selecting a good job candidate, that matches a role, to progress to an actual medical is possible through association rule mining using only their closed health questionnaire responses.*

1.2. Systematic Literature Review

In order to confirm that this research is both novel and of relevance the methodology that it follows will be one of a systematic literature review (SLR).

This approach was formalised with a set of guidelines produced by Kitchenham and Charters in 2007 and is very useful in confirming that no relevant work is excluded. In order to conduct a review the guidelines suggest the following stages

- Identification of research
- Selection of primary studies
- Study quality assessment
- Data extraction and monitoring
- Data synthesis

The following section's take each of these stages and describe how they have been applied for this research. Through using this approach we firstly highlight key research in the field and then show how this research benefits our understanding of the field and leads to our choice of research topic.

1.2.1 Identification of research

Identification of research is essential in confirming amongst other things that the researcher has not been biased in what he/she wants to find. It is this rigour that delineates SLR from other approaches.

The following databases have been identified as representing a broad enough source for all relevant literature searches so as to nullify any question surrounding researcher bias.

1. Scopus (https://www.scopus.com/home.uri)
2. Science & Technology (https://www.proquest.com/libraries/government/science-technology/)
3. ACM Digital Library (https://dl.acm.org/)
4. Google Scholar (https://scholar.google.com/)
5. IEEE Explore (https://ieeexplore.ieee.org/Xplore/home.jsp)

These databases were selected as they cover a wide corpus of the literature associated with Information Technology and in particular that of machine learning and artificial intelligence.

Table 1.1 : Search items and corresponding keywords

Search item	**Keyword**
Classification	data mining, classification
Association rules	association rules
Questionnaire	questionnaire, poll, census, canvass, survey
Fuzzy	fuzzy, non crisp
Closed data	*initially it was hoped to search on only closed data but researchers do not categorise work on whether it is open or closed so results will be manually filtered*

This leads our search string to the following form *("data mining" OR "classification") AND "association rules" AND ("questionnaire" OR "poll" OR "census" OR "canvass" OR "survey") AND ("fuzzy" OR "non crisp")*

The large number of references produced from such searches will be recorded and managed through Mendeley Desktop.

Also any relevant information about the searches themselves will be recorded along with the search. This may include information such as date of search, years covered or any specific conditions relating to the search.

1.2.2 Inclusions and Exclusions

Inclusion and exclusion criteria will initially be set to the following

Inclusion:

1. Available online
2. Article is peer reviewed
3. Full text is available in English
4. Article on or after 2005
5. Can be an academic or commercial project

Exclusion:

1. Non English papers
2. Duplicate studies
3. Magazines, newspapers, websites, podcasts, blogs

The excluded resources will however be maintained so that if during the process of inclusion/exclusion too few works result are presented then the criteria maybe adjusted in order to include other relevant work.

1.2.3 Selection of primary studies

The query options for the *1st Filter* for each of the selected databases will be included in this section along with the numbers of relevant papers discovered shown

in Table 1.2. After the initial selection process *2nd Filter* shows the number of relevant papers after a title review to see if the papers are clearly not relevant. *3rd Filter* is after a review of the abstract to decide whether the paper is still suitable for the study.

Table 1.2 : Search count for SLR results from chosen databases

Database	**1st Filter**	**2nd Filter**	**3rd Filter**
Scopus	56	18	12
Proquest	30	14	9
ACM Digital	3	2	1
Google Scholar	13	7	4
IEEE Explore	9	5	3

1.2.3.1 Scopus

TITLE-ABS-KEY(”mining” OR ”classifi*”) and TITLE-ABS-KEY(”association rules”) and TITLE-ABS-KEY(”questionnaire” or ”poll” or ”census” OR ”canvass” OR ”survey”) and TITLE-ABS-KEY(”fuzzy” or ”non crisp”) AND (PUBYEAR > 2004)

1.2.3.2 Proquest

(ti(”mining” OR ”classifi*”) OR abs(”mining” OR ”classifi*”)) AND (ti(”association rules”) OR abs(”association rules”)) AND (ti(”questionnaire” OR ”poll” OR ”census” OR ”canvass” OR ”survey”)) OR abs(”questionnaire” OR ”poll” OR ”census” OR ”canvass” OR ”survey”)) AND (”fuzzy” OR ”non crisp”)

with a filter of on or after 1/1/2005

1.2.3.3 ACM Digital

Edit Query

(Title:("mining" OR "classifi*") OR Abstract:("mining" OR "classifi*")) AND (Title:("association rules") OR Abstract:("association rules")) AND (Title:("questionnaire" OR "poll" OR "census" OR "canvass" OR "survey") OR Abstract:("questionnaire" OR "poll" OR "census" OR "canvass" OR "survey")) AND AllField:("fuzzy" OR "non crisp")

Full Query Syntax

"query": (Title:("mining" OR "classifi*") OR Abstract:("mining" OR "classifi*")) AND (Title:("association rules") OR Abstract:("association rules")) AND (Title:("questionnaire" OR "poll" OR "census" OR "canvass" OR "survey") OR Abstract:("questionnaire" OR "poll" OR "census" OR "canvass" OR "survey")) AND AllField:("fuzzy" OR "non crisp") "filter": Publication Date: (01/01/2005 TO 12/31/2020), ACM Content: DL, NOT VirtualContent: true

1.2.3.4 Google Scholar

The search options in scholar are limited compared to other databases and the closest search to previous databases can be performed by sorting by date and selecting 'Abstract' rather than 'Everything'. This unfortunately only shows papers in the current year. The following two searches were performed to look for "mining" or "classification"

1. mining "association rules" questionnaire
2. classification "association rules" questionnaire

1.2.3.5 IEEE Explore

("Document Title":"mining" OR "Document Title":"classifi*" OR "Abstract":"mining" OR "Document Title":"classifi*") AND ("Document Title":"association rules" OR "Abstract":"association rules") AND (("Document Title":"questionnaire" OR "Document Title":"poll" OR "Document Title":"census" OR "Document Title":"canvas" OR "Document Title":"survey") OR ("Abstract":"questionnaire" OR "Abstract":"poll" OR "Abstract":"census" OR "Abstract":"canvas" OR "Abstract":"survey")) AND ("Full Text .AND. Metadata":"fuzzy" OR "Full Text .AND. Metadata":"non crisp")

with a filter of on or after 1/1/2005

1.2.4 Study quality assessment

Along with excluding and including based on criteria the guidelines suggest a more thorough determination of inclusion based on the quality of research papers. This may often be based on the quality of abstract of a paper but as pointed out in the guidelines the quality of abstract in Information Technology often varies widely compared to other fields.

1.2.5 Secondary Review

Another aspect of our study will attempt to verify whether using anomaly detection to discover outliers in our dataset will produce a better overall predictor for an assessment. It is therefore appropriate to gain insight into how such detection is handled by the research community. Firstly the field of class imbalance will be introduced and from that other concepts will be reviewed culminating in anomaly detection.

1.2.5.1 Introduction

He and Garcia (2008 [24]) refer to the idea of intrinsic-based and extrinsic-based class imbalances. Intrinsic being the occurrence of a class skew due to the data itself and extrinsic being a possible skew due to reasons beyond the data. Examples of intrinsic imbalance include picking out the hopefully small number of spam emails from the multitude of daily emails received so that the recipient is not inconvenienced but also does not miss any email. Another would be in predicting "churn" within an insurer's member base, the vast majority of members would usually roll their policy over around its renewal time but a small percentage would switch. The insurer needs to target these potential churners before they make their decision but in targeting too many of the members they run the risk of alienating more than they can hope to retain. An example of extrinsic imbalance being a break in the recording of the data brought about through some type of interruption. Another being the data is collected at set times within which the minority class is most prevalent. Physical failures such as power supply or storage space may also create artificial skews in the data. A final example of extrinsic imbalance being manifested due to a change in the process of gathering of the actual data.

Although this work makes no distinction between intrinsic and extrinsic imbalance, it is the authors belief that extrinsic imbalance issues will prove to be very relevant to this work's existing industry dataset. The data collection process has changed over a number of years and is reliant on a repeatable standard line of enquiry from a human practitioner. As time moves on, practitioners change and if the standards are not strictly adhered to then over time the data becomes skewed through no other reason than change in interpretation. That being said intrinsic imbalance is also present within the dataset through the simple fact that some candidate's attributes will be underrepresented within the total pool.

1.2.5.2 Traditional Methods

Possible solutions to the traditional class imbalance issue are typically split into data level and algorithm level solutions as described in the work of Ali et al. (2015 [2]). The data level encompassing both data sampling and feature selection methods whilst the algorithm level includes cost sensitive and hybrid or ensemble applications.

Typical data sampling, cost sensitive techniques include Random Over-Sampling (ROS) and Random-Under Sampling (RUS). Whereas typical algorithm level methods include amongst others fuzzy rule-based classification (Chi et al., 1996 [11]).

Over sampling described in "subsubsection 1.2.5.3. Over sampling" and under sampling in "subsubsection 1.2.5.4. Under sampling" can be thought of as two sides of the same coin in that over-sampling involves expanding the minority class through repetition and thus increasing the occurrence of the class whereas under-sampling involves removing samples from the majority class and hence reducing its dominance on the overall classification. Whereas entries of the minority class maybe added either randomly or through some computed method the entries from the majority class tend to be removed randomly. In fact the process of adjusting the majority/minority class has been the subject of a lot of work in its own right (Khoshgoftaar et al., 2007 [31]; Van Hulse et al., 2007 [48]).

Finally it should be remembered that although til now the reader may have assumed that it is possible to improve classification results at either the data or algorithm level the hybrid scenario is one in which some combination of both of these may help. One example of this would be the introduction of the Random Forest (RF) classifier which although based on the earlier Random Decision Forest also has the inclusion of Bagging techniques (Gislason et al., 2006 [20]).

1.2.5.3 Over sampling

Two prevalent over sampling techniques are Random Over Sampling (ROS) (Zhang & Li, 2014 [54]) and Synthetic Minority Over-Sampling Technique (SMOTE) (Chawla et al., 2002 [7]; Chawla et al., 2003 [8]; Han Wang & Mao [23]). One common complaint of over sampling techniques is their propensity to create an "over-fitting" (Fig 1.1) issue whereby the training data is too heavily mimicked leading to a poor model predictor. This over-fitting concern however can be mostly overcome through the use of SMOTE (Fernández et al., 2017 [18]). Another complaint is an even larger training dataset than initially considered useful. In looking at the amount to which a minority class should be inflated it may appear reasonable that obtaining a 50/50 split would be ideal but Japkowicz and Stephen (2002 [27]), Fawcett and Provost (1997 [17]) along with Estabrooks and Japkowicz (2001 [15]) offer a counter argument and show that such a split would not produce the most favourable results.

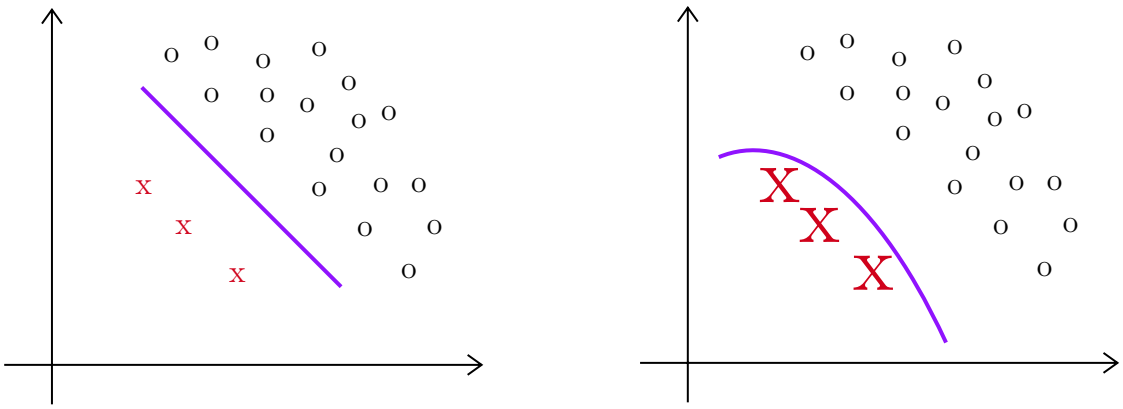


Figure 1.1 : Oversampling

1.2.5.4 Under sampling

One of the most used under sampling (Fig 1.2) techniques is Random Under Sampling (RUS) (Seiffert et al., 2009 [46]). Good practice in the removal process

from the majority class is that the data removed lies as far away from any boundary with the minority class as possible. Practitioners point out that one of the disadvantages of under sampling is the possible removal of valuable information from the core set and a possible change in the majority class's distribution if care is not taken in the extraction. For instance we could remove a large portion of young people from an otherwise well distributed cohort of data.

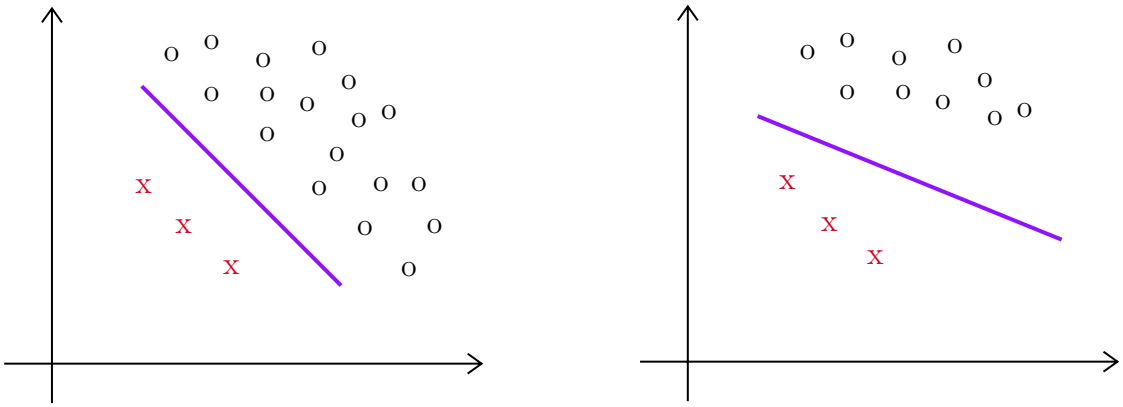


Figure 1.2 : Undersampling

1.2.5.5 Feature selection

Feature selection is the process of selecting the most influential features for a given dataset to produce an improved classification process. While not solely related to imbalance conditions a consequence of better feature selection usually leads the betterment of the effects of such imbalances (Yin et al., 2013 [52]; Mladenic & Grobelnik, 1999 [40]). A downside of using feature selection to reduce imbalance is the extra computational load that is required. The work undertaken by Mladenic and Grobelnik (1999 [40]) rates a number of feature selection ideas and puts the "odds ratio" ahead of the rest. This is confirmed in the work by Zhang et al (2018 [55]). This rating however is not confirmed in some works and the general belief is that feature selection is very much a case by case situation.

1.2.5.6 Cost sensitive methods

The general premise of cost sensitive methods is to assign a higher weight or prominence to a given occurrence of a minority classification than any majority one so as to boost the worth of the minority classifier and reduce the effects of any imbalance. Domingos (1999 [12]) and Elkan (2001 [14]) are early studies on this important topic. They rely upon something called a cost matrix (Table 1.3) which encapsulates all possible outcomes of a two class problem.

	actual negative	actual positive
predict negative	C(0, 0)	C(0, 1)
predict positive	C(1, 0)	C(1, 1)

Table 1.3 : Cost Matrix

An example given by Elkan (2001 [14]) of such a cost matrix would be in attempting to find a fraudulent credit card transaction (Table 1.4). In this scenario approval of a fraudulent transaction would cost the provider the total amount or *"-x"* whereas refusing the legitimate transaction would damage the provider's reputation with the client. Here Elkan suggests an arbitrary $20 loss to be sufficient. Refusing a fraudulent transaction he attributes a $20 benefit and finally a legitimate approval will cost the institution 2% of the cost of the total amount or *"x"*.

	fraudulent	legitimate
refuse	$20	-$20
approve	-x	-0.02x

Table 1.4 : Credit Card Fraud

1.2.5.7 Hybrid/ensemble methods

Although these methods may also be considered a type of cost sensitive method the approach is quite different in that the result of the classification is some amalgam of multiple classifiers (Seiffert et al., 2009 [46]). Two very popular ensemble methods are Bagging and Boosting (Graczyk et al., 2010 [21]). Bagging achieves its goals by producing more than a single training set with each set being uniquely classified. Each classification is then used to produce the final result. Boosting takes a similar divide and conquer approach as that of Bagging where multiple training sets are once again produced. The difference to Bagging is that the classification of each individual training set is weighted by the degree of error within that set so that the final combined result is induced by only the more favourable interim classifiers

A first look at the collaborating company dataset suggests a very high class imbalance and so the work on limiting the effects through over or under sampling is particularly relevant to this research.

1.2.5.8 Deep Learning Methods

Now that the reader is aware of how it is possible to mitigate the affects of class imbalance using traditional methods we turn our attention to the use of deep learning within such imbalances. It should be reassuring to know that many of the sampling and cost methods already mentioned can be applied within a deep learning framework. The work of Johnson and Khoshgoftaar (2019 [28]) gives a thorough road map of influential papers in the field of deep learning with imbalanced classification. Back in the 1990's Anand et al. (1993 [3]) researched ways in which the class imbalance issue could be resolved using shallow neural networks. The work describes how the majority class is responsible for swamping the net gradient of the dataset and hence has an overpowering affect on the final weights of the model.

As in traditional methods we can split deep learning into data level, algorithm

level and hybrid methods.

1.2.5.9 Data level

Hensman and Masko (2015 [39]) use a CNN to balance image data using ROS. The result of this is then in some ways contradicted by the work of Lee et al. (2016 [34]) which suggests a CNN using RUS solves the imbalance issue more favourably. In the work of Pouyanfar et al. (2018 [43]) a completely new sampling method is introduced and once again a CNN using image data is proposed. The basic idea of this method is to over sample the minority while at the same time under sample the majority class. Further work by Buda et al. (2018 [6]) compares RUS with ROS as well as two-phase learning again using image datasets. Their findings suggest that ROS is the most universal method for handling class imbalance whereas RUS is rated poor and two-phase learning with ROS and RUS being inferior to them being used individually. Two-phase learning has also been undertaken in the work of Lee et al. and shown to improve minority classification without affecting that of the majority class. It achieves this by only taking instances of the majority class during the pre-training phase so that the minority class is more influential at this phase but when the final training occurs the model is allowed to see all data.

1.2.5.10 Algorithm level

Custom loss functions are widely considered to be the easiest way to address class imbalance as unlike data level methods they do not increase the dataset size or training times, another advantage is that they do not require pre-processing steps. Much research has focused on their use within deep learning and many approaches exist. The work of Wang et al. (2016 [50]) and Lin et al. (2017 [35]) created new loss functions, allowing minority classes to become larger contributors to the overall loss. Wang et al. created a custom loss function using deep MLPs. They firstly demonstrate why a mean square error (MSE) loss function is unsuitable because of the

dominance of the majority class to affect the function. After this they propose two new loss functions being the mean false error (MFE) and mean squared false error (MSFE) which they then go on to show give a better balance between the majority and minority classes. In fact, most notably they outperform MSE more when the imbalance is more pronounced. Lin et al. (2017 [35]) on the other hand, introduce a focal loss (FL) in their custom implementation which is designed purely to help in classifying imbalances between the foreground and background objects in an image. Wang et al. (2018 [49]), Khan et al. (2017 [30]) and Zhang et al. (2016 [53]) all studied cost sensitive DNNs in various guises with those proposed by Khan et al. with Zhang et al. having the extra bonus of obtaining cost matrices through training. Zhang et al. (2018 [55]) introduce a truly hybrid approach with transfer learning, CNN feature extraction and a nearest neighbour idea to improve imbalanced classification. Wang et al. (2018 [49]) proposes something called a cost sensitive deep neural network wherein traditional one-hot encoding is superseded with "categorical feature embedding". The embedding along with extracted features through the use of a CNN are then used as input to a DNN for the final classification. Khan et al. (2017 [30]) introduce a custom method CoSen CNN, which learns both weighted parameters and misclassification costs during training. One major advantage of this is that the transfer learning approach within the CoSen method negates the requirement to choose an appropriate domain specific cost matrix. Consequently, no potentially expensive domain expert is required.

1.2.5.11 Hybrid methods

A number of key works have presented their findings where they combine data and algorithm level methods. Each of these concentrate their efforts solely on the use of image datasets. Huang et al. (2016 [26]) introduce something called a "Large Margin Local Embedding" (LMLE) method which is claimed to generate fairer class

sampling. This is achieved through the rationale that minority classes are typically sparse and can be populated easily with samples from another denser class. The downside of this approach is that it is both complicated and expensive to implement and as such would probably deter its future use. Ando and Huang (2017 [4]) propose the first Deep Over Sampling (DOS) method. The method maintains two parallel learning procedures wherein a lower layer is used to deduce an embedding function and the upper layer then uses this function to classify the imbalanced data. Dong et al. (2018 [13]) uses a novel loss function combined with hard sample mining. The authors themselves suggest that this approach may only be appropriate within large datasets. It is a progressive classifier whereby members from the minority class that are deemed to have more affect on each mini batch are selected. This means each mini batch requires smaller amounts of data to train compared to using the whole dataset. As each mini batch is trained in turn the authors attempt to rectify the class imbalance incrementally giving the minority class more say in the final result.

1.2.5.12 Related research topics and challenges

It is possible for classification imbalance to contain very high imbalance ratios. Indeed when the ratio does become excessively high it is sometimes more appropriate to think of the issue as not so much an imbalance one but as an anomaly detection issue. Within anomaly detection we assume that our dataset has an expected distribution and that anything that deviates sufficiently is an anomaly. The methods for such detection may differ from that of classification imbalance and so will not be discussed further in this review. They include clustering methods, one class SVM's and Isolation Forests. Our industry dataset has not to date been considered to contain an anomalous minority condition.

In a landscape where we harvest more and more data in the hope of either gaining hindsight now or sometime in the future from that data the idea of "big data" has

arisen. This big data phenomena has caused issues for traditional machine learning. One such issue is that a number of accepted algorithms used in the field are either not capable of dealing with big data or run too slowly. They may have been designed to see all the data at the same time or cannot be sped up through parallelism because of the way they have been implemented.

It is pertinent however to distinguish between big data and non big data so that we may confirm whether any methods to improve upon any classification imbalance are suitable in both cases. Katal et al. (2013 [29]) explains that big data does not as its name suggests only refer to the size of the data but variety (multi-typed), volume (size), velocity (capture rate), variability (inconsistent load payloads), complexity (multi- sourced) and value (worth to organisation).

For our collaborating company, as it expands its global reach the need for staying abreast of the best practices for dealing with "big data" will become paramount. The expectation to become a "source of truth" to other entities will also create a maze of connected services each having to deal with very diverse data sources.

1.2.6 Comparative analysis

Based on the reviews previously mentioned Table 1.5 attempts to articulate the gaps found during our review

Table 1.5 : Comparative analysis

Author	**Closed-ended Questionnaire**	**Mining Method**	**Medical Data**	**Dynamic Membership**	**Anomaly Detection**	**Timeline Analysis**
Chen et al. (2008 [9])	Yes	ARM	No	No	No	No
Chen et al. (2009 [10])	Yes	ARM	No	No	No	No
450	PL 450	TO-38	single	osram	50–90	120
638	ML520G54	TO-56	single	mitsubishi	90–100	150
Our Contribution	Yes	ARM	Yes	Yes	Yes	Yes

1.3. 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.

1.4. 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?

1.5. Objectives

This study will have 5 objectives:

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

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

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

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

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

These objectives are further elaborated upon in Section 1.7

1.6. Methodology

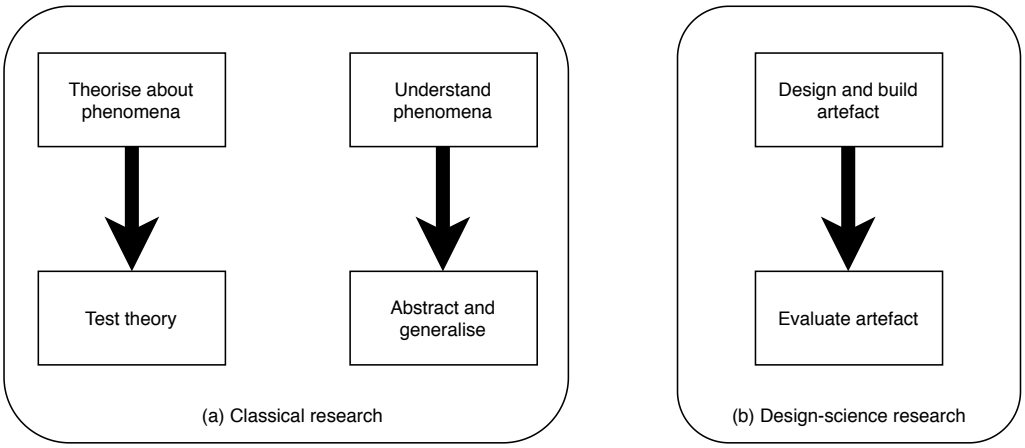
Babbie (2015 [5]) states that research is a *"systematic and orderly approach taken towards the collection and analysis of data so that information can be obtained from*

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 [51]).

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 [51]). Figure 1.3 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.

Figure 1.3 : Classic v Design-Science Research within Information Systems



Hevner et al. (2004 [25]) argued that DSR artefacts can take one of the four forms

shown in Figure 1.4. 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 [22]) disagree with this principle and instead suggest a framework where design-research theory is paramount.

Figure 1.4 : 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.

Indeed one of the issues of DSR is in proving any such research has 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 [25]) which suggests a set of 7 non mandatory guidelines (Table 1.6) that should be "addressed in some manner for design-science research to be complete".

Table 1.6 : 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

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 [22]) does not suffer from this generic criticism. They 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 1.7).

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 1.4 that defines DSR.

Peffers et al. (2007 [42]) suggest a research methodology using a six step process for correct implementation of DSR (Table 1.8). 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

Table 1.7 : 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)*

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 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.

Table 1.8 : Peffers et al. 6 Step Iterative Process for Conducting DSR

No.	**Step**
1	Identify, define, and motivate the focal problem
2	Define objectives that a solution (possibly partial) to the focal problem must achieve
3	Design and develop the artefact
4	Demonstrate the artefact can be used to help solve the focal problem
5	Evaluate how well the artefact solves the focal problem
6	Communicate the outcomes of the research

1.7. Research Plan

The research plan of this study will follow the work of Peffers et al. described in Section 1.6. Each required step of Peffers et al. will be considered in its own section and how that step applies to the current research will be discussed.

1.7.1 Identify, define, and motivate the focal problem

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. Through experience they have realised that sometimes a candidate that appears perfect for a role fails at the last hurdle of the selection process, being a medical assessment. It is this late assessment failure that brings rise to the core problem of this research. *How can a potential candidate be assessed on some medical criteria without involving an actual medical assessment?* A curious reader may wonder why the assessment occurs so late in the process? This

is because the selection process starts with many multiples of the final number of candidates actually being assessed and so would be prohibitively expensive to offer the assessments any earlier.

1.7.2 Define objectives that a solution (possibly partial) to the focal problem must achieve

In order to address the focal problem any potential solution should garner useful information from the candidate's answers to a preselection questionnaire that they are required to complete. The questions contained within any such questionnaire should take into account the specific role for which the candidate is applying and any typical risks or needs that are associated with that role.

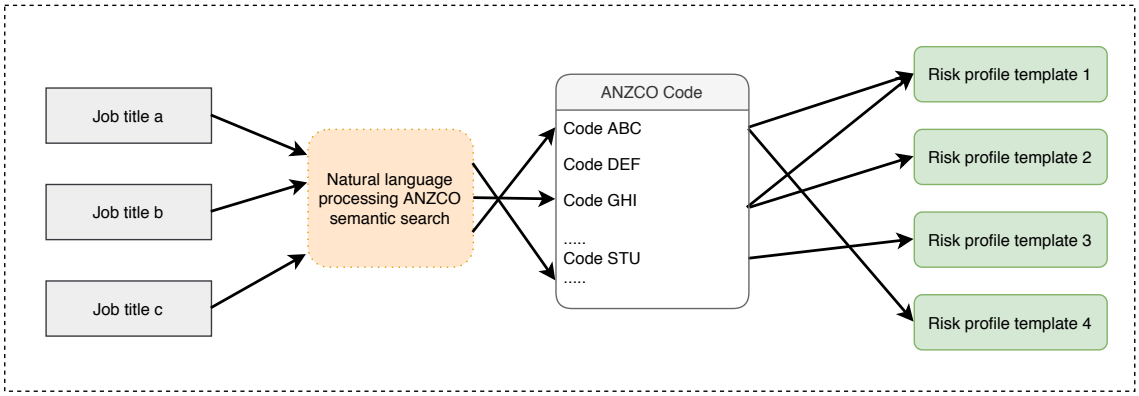
1.7.3 Design and develop the artefact

Before detailing the artefacts to be developed that correspond to the objectives of this work it would be prudent to give an overview of the proposed system.

1.7.3.1 System Overview

Being able to definitively label a role or job title is critical to the operation of any system being developed. For instance a bus driver could also be referred to as a bus operator or omnibus driver and so a standard convention is required. This convention within Australia and New Zealand is the Australian and New Zealand Standard Classification of Occupations (ANZSCO). This ANZSCO standard will form the "bridge" between the description of a role given by a third party and the definition of the role within the system. This "bridge" is described in Figure 1.5. Within the bridge exists an ANZSCO semantic search which represents another research opportunity currently being written within the university. The bridge also contains multiple "risk profile templates" which represent a predefined "risk profile" enabling third party's to have a building block for their specific role.

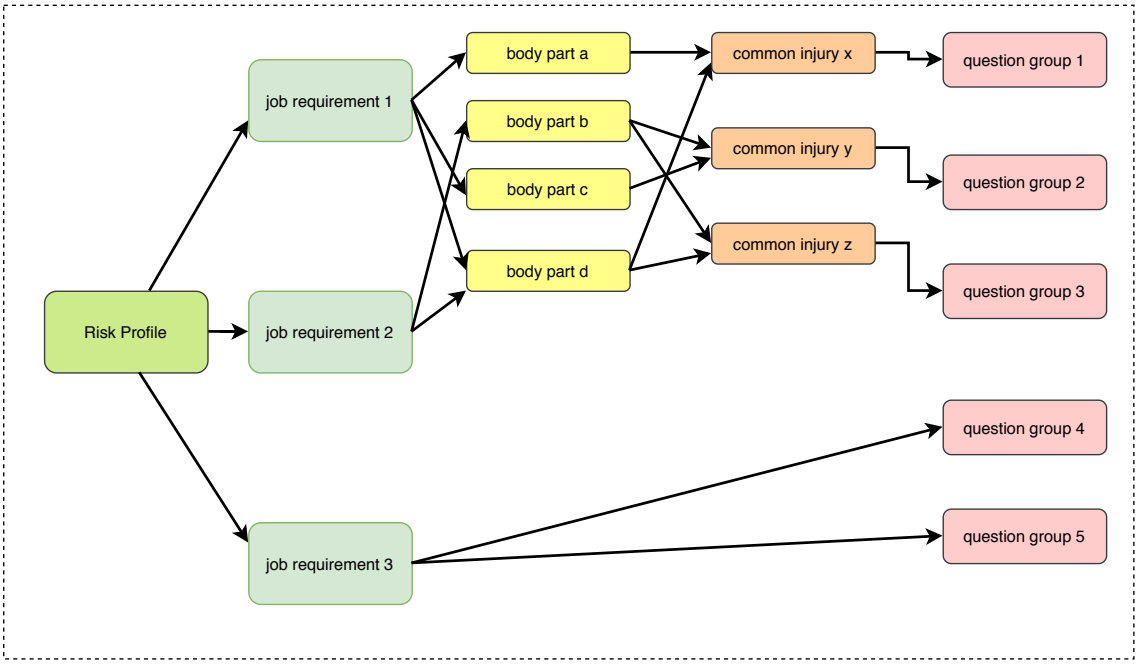
Figure 1.5 : ANZSCO Bridge



The artefacts to be developed should be able to categorise candidates for specific roles where each role is associated with one or more "risk profiles". The purpose of a risk profile is to form the association between particular job requirements such as "lifting heavy weight from floor to waist" or "sitting for extended periods" through to the body parts that are affected due to that requirement, such as back, arm, leg, or shoulder. From these body parts a number of common injuries are linked and from those injuries ultimately a series of questions are added to the questionnaire for that role. This risk profile association for an individual role is shown in Figure 1.6. In this figure we see two job requirements 1 & 2 that follow the classic association just described whereas job requirement 3 links directly to multiple question groups. An example of such a requirement would be a bus driver where a particular licence is a requirement which obviously has no reliance on a particular body part.

There are a few complications of the questionnaire that a model must adapt to. The first being its hierarchical or cascading nature which is best described through an example. A question could be framed asking "have you ever broken your leg?" the answer a candidate gives could be either yes/no. If the candidate enters "yes" however this could trigger further questions about frequency, time to heal etc. and

Figure 1.6 : Risk Profile Association



each successive answer could themselves create a hierarchical set of questions. The other complication is the fact that answers to questions can be of various types, for instance a simple quantity like 4, a category like 'male', a list of possible answers, a ranking like [0-7] or a linguistic ranking like 'tall'.

Initially categorisation of a candidate will involve training an individual classification model for each role from the answers collected through the associated questionnaire. This classification will be a binary classifier of "suitable" or "not suitable". It is envisaged that some fuzzy categorisation of the candidate will also be possible which will allow for a human to override a decision on the candidate who is borderline. The opportunity to use some form of transfer or cluster learning would remove the need to train all roles separately but the initial design will not cater for this scenario.

1.7.3.2 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.

Fuzzy classification can be considered as two distinct major steps:

Step 1: Create our distinct linguistic definitions to be able to describe a given domain feature in a fuzzy or non crisp manner.

In this step we introduce similarity matrix which allow us to compare two linguistic terms in a fuzzy manner. For instance suppose we had a health ranking feature (Table 1.9) By using this table we could say that a health ranking value of *Very Poor* had a similarity of 0.5 when compared to *Poor*.

Table 1.9 : Health Ranking Similarity Matrix

	Very Poor	*Poor*	*Average*	*Good*	*Very Good*
Very Poor	1	0.5	0	0	0
Poor	0.5	1	0.5	0	0
Average	0	0.5	1	0.5	0
Good	0	0	0.5	1	0.5
Very Good	0	0	0	0.5	1

Membership functions are also defined in this step. A membership function best describes the degree of truth in fuzzy logic. As an example if we were considering *di-*

astolic blood pressure as a feature then this could be described with the membership functions shown in Figure 1.7.

Figure 1.7 : Diastolic Blood Pressure Membership

$$DBP_{normal}(q) = \begin{cases} 1 & \text{if } q < 60, \\ \frac{60-q}{80-60} & \text{if } 60 \geq q < 80. \end{cases}$$

$$DBP_{high}(q) = \begin{cases} \frac{117-q}{85} & \text{if } 90 \geq q < 100, \\ 1 & \text{if } \geq 100. \end{cases}$$

Step 2: Select the most noteworthy classification rules using the linguistic definitions from step 1 along with any crisp values. This step needs to include a mechanism for handling exponential growth in noteworthy rules as the features in a domain increase.

For traditional crisp association rule mining our feature values maybe boolean or quantitative. The selection process involves finding rules that satisfy a minimum support and confidence value. So if we suppose we have a set of items

$I = \{i_1, i_2, ..., i_m\}$

and D represents the set of all transactions where each transaction t is a set of items and $t \subseteq I$. Let A be a set of arbitrary items in I. A transaction t contains $A \iff A \subseteq I$. Each association rule is of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cap B = \phi$. We can now define support (Sup) as the percentage of transactions that contain both A and B

$$Sup(A \Rightarrow B) = \frac{\mid t_A \cap t_B \mid}{\mid D \mid} \tag{1.1}$$

and confidence *Conf* as the percentage of transactions in D containing A that also contain B.

$$Conf = \frac{Sup(A \Rightarrow B)}{Sup(A)} = \frac{\mid t_A \cap t_B \mid}{\mid t_B \mid} \tag{1.2}$$

Now typically in a non fuzzy approach we would introduce the Apriori algorithm which uses the definitions for *Sup* and *Conf* to perform a two step procedure where itemsets are generated and tested against *Sup*. The steps are:

Step 2.1: set $k = 1$; *largeItemSet* = *empty*

Step 2.2: find all *candidateItemsets* of size k from original dataset

Step 2.3: if a *candidateItemset* transaction count is greater than *Sup* then it is a frequent itemset and added to the *largeItemSet*.

Step 2.4: set $k = k + 1$

Step 2.5: find all *candidateItemsets* of size k from *largeItemSet*

Step 2.6: if a *candidateItemset* transaction count is greater than *Sup* then it is a frequent itemset and added to the *largeItemSet*.

Step 2.7: if $k \leq$ *numberOfFeatures* go to Step 2.4

Step 2.8: output all *largeItemSet* combinations that have a confidence value larger than *Conf*

Our questionnaire data however does not only contain quantitative data and so we need to define support and confidence formula for our linguistic answers.

To date we have implemented a proof of concept of this technique that replicates the findings of Chen et al.(2008 [9], 2009 [10]). The solution is written in Python and uses Neo4j as a datastore, through using a more dynamic approach the implementation of our research goals should be faster to achieve.

1.7.3.3 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 [10]) concluded that their design proved to be slow as the fuzzy membership routines were static. In feeding back this assessment information from clinical experts into the system the membership functions themselves will be altered over time and in doing so become dynamic.

To better explain this, imagine we are dealing with a crisp set, a membership function is either *True* or *False*, for example someone is either born in Australia or not, there is no middle ground. What about however if we wanted to specify if a person is *Healthy* this is far less rigid. It could be decided on a number of factors such as they (i) can perform a set number of exercises without becoming breathless; (ii) have not visited a doctor in X years; (iii) do not consume cigarettes, alcohol or other stimulants. It is for this type of attribute that membership functions are required, they will assign a value in the range of 0 (*False*) & 1 (*True*) to establish in this case how *Healthy* someone is from assigning weightings to the factors (i),(ii) & (iii) previously mentioned.

Marasini et al. (2016 [37]) state that there are four main approaches that can be used to define a membership function:

- a "best" mathematical function
- a probabilistic point of view
- decision-theoretic approach
- axiomatic measurement theory

The idea of intuitionistic fuzzy sets in questionnaire analysis has been covered extensively within the work of Marasini et al. (2016 [37]).

Kostikova et al. (2016 [32]) cover how dynamic membership functions maybe achieved.

1.7.3.4 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 [38]) 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.

We have to date implemented an anomaly detection technique originally in Octave and later ported into MATLAB. The implemented work uses a current production system that has a lot of data but lacks some integrity constraints to ensure

that data is always correct. A similar approach will be undertaken with the system currently being developed once sufficient data becomes available.

The approach can be broken down into the following steps:

Step 1: Select features that will be used to base our decision on whether a candidate maybe considered anomalous. Our industry partner has initially chosen BMI Banding (Table 1.10), daily consumption of cigarettes, age and sex. These features should be open to change as the project evolves.

Table 1.10 : BMI Banding

BMI	**Banding**
0	Unknown BMI
> 0 *and* < 0.15	Very Severely Underweight
≥ 0.15 *and* < 0.16	Severely Underweight
≥ 0.16 *and* < 0.185	Underweight
≥ 0.185 *and* < 0.25	Normal
≥ 0.25 *and* < 0.30	Overweight
≥ 0.30 *and* < 0.35	Obese Class I
≥ 0.35 *and* < 0.40	Obese Class II
≥ 0.40 *and* < 0.45	Obese Class III
≥ 0.45 *and* < 0.50	Obese Class IV
≥ 0.50 *and* < 0.60	Obese Class V
≥ 0.60	Obese Class VI

Step 2: Load, explore and clean data. The data cleansing part of this step is particularly important due to the integrity issues of the current system. The implementation removes data with poor integrity. It is at this step the distribution

of a feature will be examined and all features will have their values normalised. Normalisation prevents features with a large range swamping those that do not.

Step 3: Define the network architecture. Whilst using the data solely from the original system it was decided not to use any type of Deep Neural Network. At this point in the research the decision was made to use both the Linear regression (Seber & Lee, 2012 [45]) and SVM (Noble, 2006 [41]) algorithms as having the ability to exclude rare occurrences was considered more important than the accuracy of those decisions. Moving forward it is the intention of the research to introduce a Deep Neural Network architecture.

Step 4: Specify training options.

Step 5: Train the network.

The current system although lacking integrity does contain sufficient data for a classic split of the data into training, cross validation and test data. When the new system comes on line a less rigid split will need to be used involving some kind of k-fold cross validation to cater for the initial lack of data. To date some work has been done using a leave-one-out technique to gain familiarity with the approach.

Step 6: Predict the labels of new data and calculate the classification accuracy. Both linear regression and SVM struggled to correctly classify anomalous candidates. In total from a test population of 2834 there were actually 20 candidates rejected however our solution predicted none of these. Upon closer examination of these 20 however it became apparent that these candidates had been rejected for reasons outside of the questionnaire. For example one candidate was rejected after the assessor discovered that they had only one arm which was a question not asked through the particular job questionnaire but the job had required both arms. Other candidates were rejected for reasons that the company gave to the assessor directly and not through the questionnaire process. These findings will be fed back into

giving the next iteration of the system more data integrity.

1.7.3.5 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 [10]) 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.

As different job roles will require the use of varying sets of questions it is paramount that we are able to in some way compare the answers for different roles in order that we may recommend an unsuccessful candidate for a more suitable role. Krusaa et al. ([33]) propose a *Questionnaire-Based Efficient Adaptive Transfer Neural Network*, QEATNN that learns to transfer knowledge between a social media and questionnaire domain. We will adapt this approach of transferring knowledge between disparate domains and instead use differing questionnaire domains.

The basic steps are:

Step 1: Select representative questions from each questionnaire that are able to be transferred

Step 2: Apply the work of Shi et al. (2017, [47]) to create a model using Local Representative-Based Matrix Factorisation, LRMF that splits the representative questions into global (global attributes) and local (personal attributes) representa-

tives.

Step 3: Deduce whether sufficient local representatives of an unsuccessful candidate from one questionnaire domain match those of a successful candidate from another domain and if so recommend this candidate for the alternate role.

1.7.4 Demonstrate the artefact can be used to help solve the focal problem

Figure 1.8 demonstrates a more specific job requirement for lifting heavy weights from floor to waist. It has been partially completed for the purpose of explanation only and does not represent a genuine scenario. Of particular note are the edges for heavy weight, hip and flexor. It is on these edges that properties will initially be attached such as the value for the actual weight considered "heavy" by an expert in the field. As our classifier begins to acquire data it is these properties that will dynamically be altered in order to solve our core problem of classifying a candidate. An added bonus of the classifier will be the ability to suggest alternate candidate roles to an unsuccessful candidate that they may be suitable for.

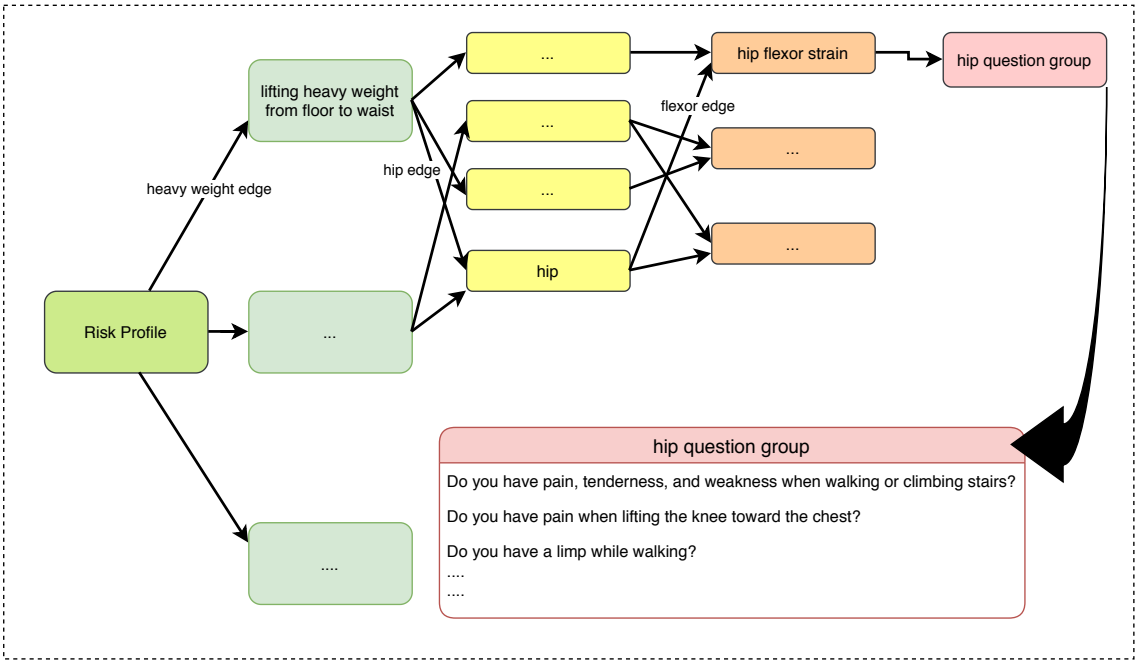
1.7.5 Evaluate how well the artefact solves the focal problem

Objective 5 of Section 1.5 discusses evaluation of the developed artefacts and in so doing satisfies this requirement of Peffers et al. described in the Methodology Section 1.6.

1.7.5.1 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.

Figure 1.8 : Specific Job Requirement



Briefly if we are describing a binary classification such as the candidate is "suitable" or "not suitable" we can plot a confusion matrix of the form shown in Fig 1.9. 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 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.

Figure 1.9 : Confusion matrix

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

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

Fawcett (2006 [16]) explains the intricacies of an ROC graph, an example of which, is shown in Fig 1.10. 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

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 1.9

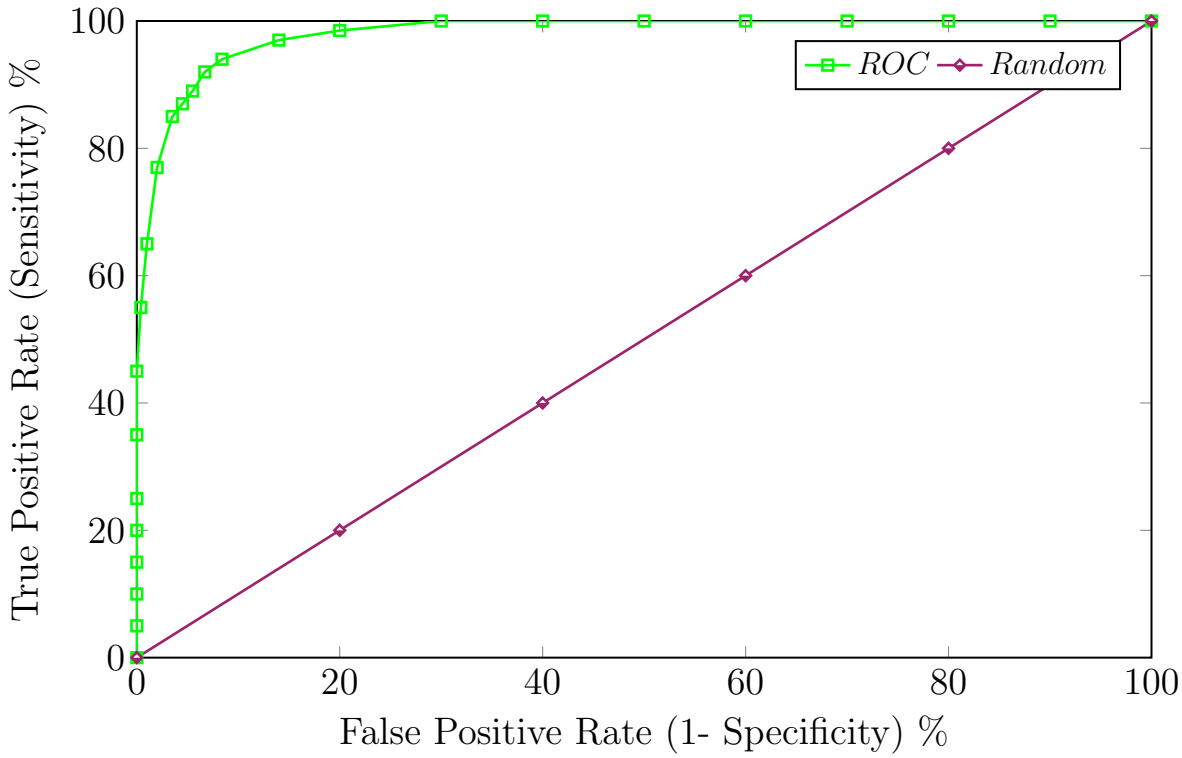
$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

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

1.7.6 Communicate the outcomes of the research

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 artefact 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

Figure 1.10 : ROC graph



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

problem of classifying the answers to closed survey/questionnaire data.

Through progressing the research to completion the communication of the outcomes will satisfy the industry partner. The wider research community will become aware of the outcomes through publishing one or two papers at recognised conferences.

1.8. Future Impact/Significance

The significance of this research will be described both from a theoretical and practical standpoint.

1.8.1 Theoretical significance:

The findings from this study contribute to the machine learning community by addressing some areas of misrepresentation of the closed-data questionnaire. Specifically this research sets out to address the problem of handling a range of different response data types all at the same time by introducing a fuzzy association rule mining approach. Although fuzzy ARM has been looked at in a very small number of papers they have neither found a solution to the time consuming task of creating membership functions statically nor attempted to improve upon the results by tuning the rule parameters using gradient descent. The research will also enrich the understanding of analysing questionnaire responses along a timeline.

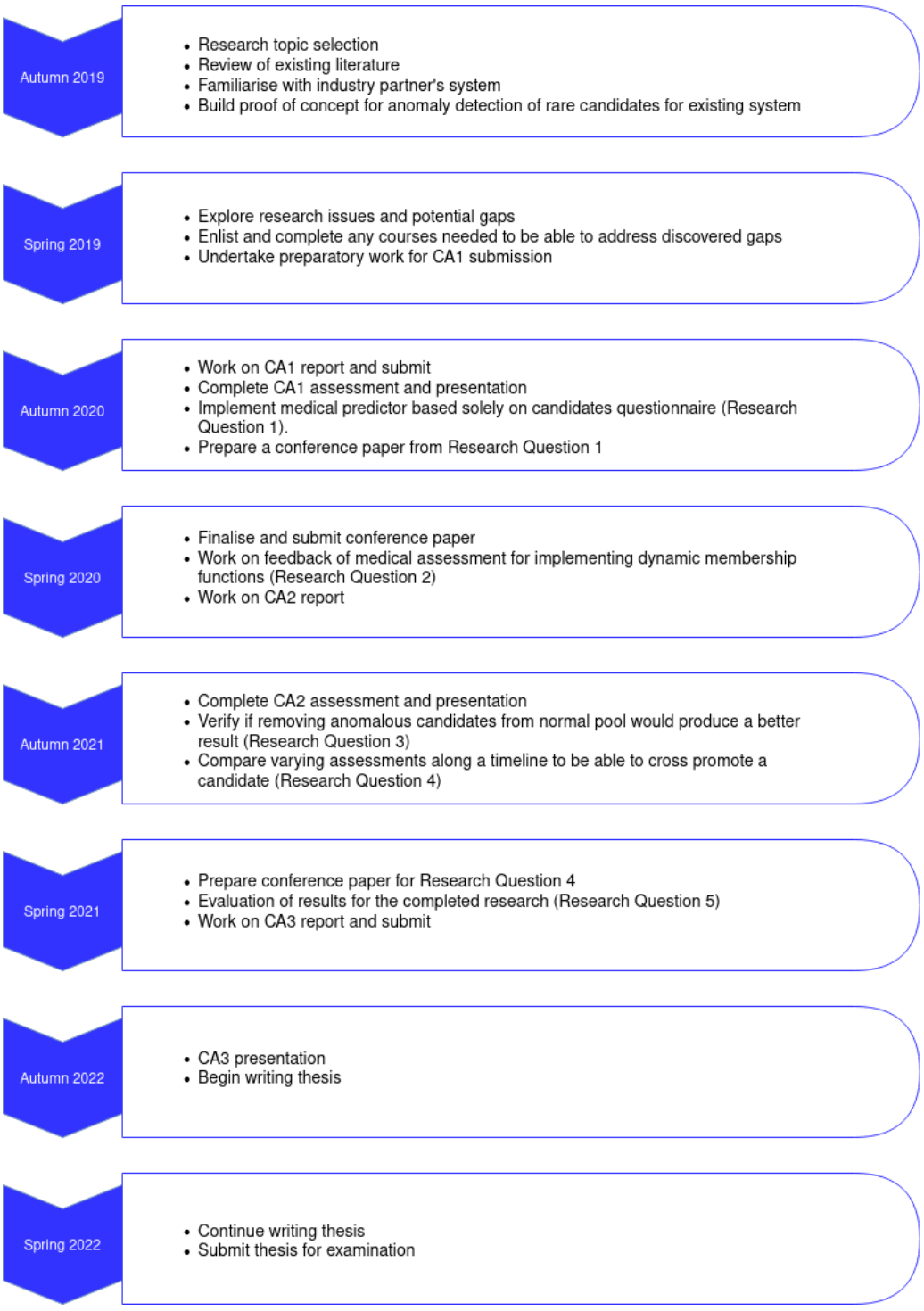
1.8.2 Practical significance:

Providing unnecessary medical assessments to candidates that are unsuitable for a role raises a number of issues to the selection process of our industry partner

- The cost of providing the assessment can be prohibitive.
- As the candidate put forward for an assessment is usually a long way through the selection process suitable secondary candidate's are not always available. This may be because the other candidate has already found a new role. This can also lead to the even larger problem of having to initiate the candidate recruitment cycle all over starting with the initial interview stage.
- The candidate that is unsuitable for the role may well be suitable for a different role but has now missed the opportunity to apply for that role.

1.9. Research Timeline

The following represents a guide to the expected duration of each activity in this research study.



Bibliography

[1] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 1993, pp. 207–216.

[2] A. Ali, S. M. Shamsuddin, A. L. Ralescu *et al.*, "Classification with class imbalance problem: a review," *Int. J. Advance Soft Compu. Appl*, vol. 7, no. 3, pp. 176–204, 2015.

[3] R. Anand, K. G. Mehrotra, C. K. Mohan, and S. Ranka, "An improved algorithm for neural network classification of imbalanced training sets," *IEEE Transactions on Neural Networks*, vol. 4, no. 6, pp. 962–969, 1993.

[4] S. Ando and C. Y. Huang, "Deep over-sampling framework for classifying imbalanced data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 770–785.

[5] E. R. Babbie, *The practice of social research*. Nelson Education, 2015.

[6] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[8] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "Smoteboost:

Improving prediction of the minority class in boosting," in *European conference on principles of data mining and knowledge discovery*. Springer, 2003, pp. 107–119.

[9] Y.-L. Chen and C.-H. Weng, "Mining association rules from imprecise ordinal data," *Fuzzy Sets and Systems*, vol. 159, no. 4, pp. 460–474, 2008.

[10] ——, "Mining fuzzy association rules from questionnaire data," *Knowledge-Based Systems*, vol. 22, no. 1, pp. 46–56, 2009.

[11] Z. Chi, H. Yan, and T. Pham, *Fuzzy algorithms: with applications to image processing and pattern recognition*. World Scientific, 1996, vol. 10.

[12] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *KDD*, vol. 99, 1999, pp. 155–164.

[13] Q. Dong, S. Gong, and X. Zhu, "Imbalanced deep learning by minority class incremental rectification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 6, pp. 1367–1381, 2018.

[14] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[15] A. Estabrooks and N. Japkowicz, "A mixture-of-experts framework for learning from imbalanced data sets," in *International Symposium on Intelligent Data Analysis*. Springer, 2001, pp. 34–43.

[16] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[17] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data mining and knowledge discovery*, vol. 1, no. 3, pp. 291–316, 1997.

[18] A. Fernández, S. del Río, N. V. Chawla, and F. Herrera, "An insight into imbalanced big data classification: outcomes and challenges," *Complex & Intelligent Systems*, vol. 3, no. 2, pp. 105–120, 2017.

[19] R. H. Gault, "A history of the questionnaire method of research in psychology," *The Pedagogical Seminary*, vol. 14, no. 3, pp. 366–383, 1907.

[20] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognition Letters*, vol. 27, no. 4, pp. 294–300, 2006.

[21] M. Graczyk, T. Lasota, B. Trawiński, and K. Trawiński, "Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal," in *Asian conference on intelligent information and database systems*. Springer, 2010, pp. 340–350.

[22] S. Gregor, D. Jones *et al.*, "The anatomy of a design theory." Association for Information Systems, 2007.

[23] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*. Springer, 2005, pp. 878–887.

[24] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge & Data Engineering*, no. 9, pp. 1263–1284, 2008.

[25] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS quarterly*, pp. 75–105, 2004.

[26] C. Huang, Y. Li, C. Change Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5375–5384.

[27] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[28] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, 2019.

[29] A. Katal, M. Wazid, and R. Goudar, "Big data: issues, challenges, tools and good practices," in *2013 Sixth international conference on contemporary computing (IC3)*. IEEE, 2013, pp. 404–409.

[30] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 8, pp. 3573–3587, 2017.

[31] T. M. Khoshgoftaar, C. Seiffert, J. Van Hulse, A. Napolitano, and A. Folleco, "Learning with limited minority class data," in *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*. IEEE, 2007, pp. 348–353.

[32] A. V. Kostikova, P. V. Tereliansky, A. V. Shuvaev, V. N. Parakhina, and P. N. Timoshenko, "Expert fuzzy modeling of dynamic properties of complex systems," *ARPN Journal of Engineering and Applied Sciences*, vol. 11, no. 17, pp. 10 601–10 608, 2016.

[33] R. Krusaa, C. Kloster, and V. Kamp, "Transfer learning for better cold-start recommendation using multiple domains."

[34] H. Lee, M. Park, and J. Kim, "Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3713–3717.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[36] M. Mamuda and S. Sathasivam, "On the fusion of tuning parameters of fuzzy rules and neural network," in *AIP Conference Proceedings*, vol. 1870, no. 1. AIP Publishing LLC, 2017, p. 040045.

[37] D. Marasini, P. Quatto, and E. Ripamonti, "Intuitionistic fuzzy sets in questionnaire analysis," *Quality & Quantity*, vol. 50, no. 2, pp. 767–790, 2016.

[38] G. Marshall, "The purpose, design and administration of a questionnaire for data collection," *Radiography*, vol. 11, no. 2, pp. 131–136, 2005.

[39] D. Masko and P. Hensman, "The impact of imbalanced training data for convolutional neural networks," 2015.

[40] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and naive bayes," in *ICML*, vol. 99, 1999, pp. 258–267.

[41] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[42] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.

[43] S. Pouyanfar, Y. Tao, A. Mohan, H. Tian, A. S. Kaseb, K. Gauen, R. Dailey, S. Aghajanzadeh, Y.-H. Lu, S.-C. Chen *et al.*, "Dynamic sampling in convolutional neural networks for imbalanced data classification," in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 2018, pp. 112–117.

[44] H. Schuman and S. Presser, "The open and closed question," *American Sociological Review*, vol. 44, no. 5, pp. 692–712, 1979. [Online]. Available: http://www.jstor.org/stable/2094521

[45] G. A. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012, vol. 329.

[46] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197, 2009.

[47] L. Shi, W. X. Zhao, and Y.-D. Shen, "Local representative-based matrix factorization for cold-start recommendation," *ACM Transactions on Information Systems (TOIS)*, vol. 36, no. 2, pp. 1–28, 2017.

[48] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 935–942.

[49] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. B. Abdallah, and A. Kronzer, "Predicting hospital readmission via cost-sensitive deep learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 15, no. 6, pp. 1968–1978, 2018.

[50] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in *2016 international joint conference on neural networks (IJCNN)*. IEEE, 2016, pp. 4368–4374.

[51] K. Williamson and G. Johanson, *Research methods: information, systems, and contexts*. Chandos Publishing, 2017.

[52] L. Yin, Y. Ge, K. Xiao, X. Wang, and X. Quan, "Feature selection for high-dimensional imbalanced data," *Neurocomputing*, vol. 105, pp. 3–11, 2013.

[53] C. Zhang, K. C. Tan, and R. Ren, "Training cost-sensitive deep belief networks on imbalance data problems," in *2016 international joint conference on neural networks (IJCNN)*. IEEE, 2016, pp. 4362–4367.

[54] H. Zhang and M. Li, "Rwo-sampling: A random walk over-sampling approach to imbalanced data classification," *Information Fusion*, vol. 20, pp. 99–116, 2014.

[55] Y. Zhang, L. Shuai, Y. Ren, and H. Chen, "Image classification with category centers in class imbalance situation," in *2018 33rd Youth Academic annual conference of Chinese Association of Automation (YAC)*. IEEE, 2018, pp. 359–363.