

© <2020>. This manuscript version is made available under the CC-BY-NC-ND 4.0 license
<http://creativecommons.org/licenses/by-nc-nd/4.0/>
The definitive publisher version is available online at [https://doi.org/
10.1016/j.techfore.2020.120486](https://doi.org/10.1016/j.techfore.2020.120486)

Discovering dynamic adverse behavior of policyholders in the life insurance industry

Md Rafiqul Islam^a, Shaowu Liu^a, Rhys Biddle^a, Imran Razzak^b, Xianzhi Wang^c, Peter Tilocca^d and Guandong Xu^{a,*}

^aAdvanced Analytics Institute (AAI), University of Technology Sydney (UTS), Australia

^bSchool of Information Technology, Deakin University, Geelong, Australia

^cSchool of Computer Science, University of Technology Sydney (UTS), Australia

^dOnePath Life, Australia

ARTICLE INFO

Keywords:

Adverse behavior
Life insurance
Frequent pattern
High-risk users
Decision making

Abstract


Adverse selection (AS) is one of the significant causes of market failure worldwide. Analysis and deep insights into the Australian life insurance market show the existence of adverse activities to gain financial benefits, resulting in loss to insurance companies. Understanding the behavior of policyholders is essential to improve business strategies and overcome fraudulent claims. However, policyholders' behavior analysis is a complex process, usually involving several factors depending on their preferences and the nature of data such as data which is missing useful private information, the presence of asymmetric information of policyholders, the existence of anomalous information at the cell level rather than the data instance level and a lack of quantitative research. This study aims to analyze the life insurance policyholder's behavior to identify adverse behavior (AB). In this study, we present a novel association rule learning-based approach 'ARLAS' to detect the AS behavior of policyholders. In addition to the original data, we further created a synthetic AS dataset by randomly flipping the attribute values of 10% of the records in the test set. The experiment results on 31,800 Australian life insurance users show that the proposed approach achieves significant gains in performance comparatively.

1. Introduction

Life insurance plays a vital role in society and provides financial protection to policyholders in need. With the increase in life expectancy (increased from 80.3 to 83.9 years from 2000 to 2020) and increasing pressure on government budgets, life insurance companies must play a greater role in society. This industry, worth \$8 billion, is facing unprecedented challenges. Analysis and deep insights into the Australian life insurance market show the existence of adverse activities to gain financial benefits, resulting in loss to insurance companies (Leach et al., 2012; Butler, 2007). For example, an AS occurs when a policyholder obtains a policy at a much lower premium than the insurance company would charge if they were aware of the actual risk regarding the applicant, usually because the applicant withholds relevant information or provides false information that thwarts the effectiveness of the insurance company's risk evaluation system (Spears and Barki, 2010). Thus, understanding policyholders' behavior is necessary to reduce adverse claims and increase business profit and marketing planning (Lin et al., 2009). It also determines the intent of policyholders to purchase or not to purchase products or services. However, the behavior analysis of policyholders is a challenging task, usually involving several factors depending on their preferences and the nature of data such as the absence of useful private information, the presence of asymmetric information on policyholders, etc. Furthermore, anomalous information exists at the cell level, making it difficult to identify an adverse user.

Adverse selection behavior refers to a situation where sellers have information that buyers do not have, and vice versa (Bajari et al., 2014). In the life insurance business, the AS behavior of policyholders is typical and presents a risk to the integrity of the insurance market (Polyakova, 2016). One of the significant causes of market failure is when a high-risk policyholder intentionally hides or provides misleading information to the insurer to avoid paying high premiums, and to obtain greater benefits (Cohen and Siegelman, 2010; Butler, 2007). For example, a race car driver may acquire a life insurance policy without providing his correct occupational information, even though hiding one's

*Corresponding author

 Guandong.Xu@uts.edu.au (G. Xu)

ORCID(s):

occupation could be considered a criminal activity. As another example, a vehicle owner can obtain a lower policy rate by providing a false garage address in a suburb which has a lower policy premium. Similarly, a policyholder may obtain insurance coverage by providing a residential address that falls within an area with a very low crime rate despite living in an area with a very high crime rate. Insurance companies often bear the loss of these misleading practices due to shortfalls in covering the risk. Therefore, the ability to detect AS in the insurance market is critical to reduce company losses, increase service quality and improve risk adjustments when assessing AB, allowing insurers to focus on complications of the greatest concern and allowing for the improvement of insurance premium policies.

Insurance companies are often uncertain because of the actions imposed by the unregulated movement of high-risk policyholders (Riddell and Hales, 2018). As a result, life insurance companies are keen to understand policyholders' behavior to develop appropriate business strategies. Insurance managers (IMs) have already started carrying out detective analytics to manage and promote their business efficacy (Boodhun and Jayabalan, 2018). However, it is very challenging to identify them with some hidden characters related to different data of the insurance policy. There is still a shortage of considerable research regarding detective analytics for the enrichment of the life insurance domain. Existing research has pointed out that traditional techniques are rather time-consuming, taking up to several months, and it is costly to capture comprehensive information on policyholder behaviors. Therefore, it is important for IMs to remain alert for changes, demands, and necessary actions to manage and work with local industries (Bolhaar et al., 2012; Islam et al., 2017; Keane and Stavrunova, 2016). However, the major challenge for IMs is to keep track of the behavioral patterns of policyholders. Keeping track of policyholder behavior over time is difficult because of its dynamic nature. Behavioral patterns can help IMs make smart decisions that optimize business quality, increase profit, and improve policyholder feedback (Grewal et al., 2019; Nahar et al., 2013). Therefore, IMs need to have access to all the critical aspects of the related information, detailing which packages are the best to promote a product, how people prefer different premiums over time, what changes will make a premium more attractive, what actions should be taken to tackle future problems such as the sudden increment of policyholders mental illness claim, a natural disaster and so on.

Advanced data analytics approaches have attracted immense attention from the research community, business decision makers and companies to improve the gain in net profit and have shown considerably better performance via predictive and analytic capabilities (Viswanathan et al., 2020; Boodhun and Jayabalan, 2018; Olakanmi and Dada, 2019; Kaushik and Gandhi, 2019; Hutagaol and Mauritsius, 2020). In the insurance industry, while the existing methods explore the hidden behavior of dishonest policyholders, there is still potential to more accurately discover their hidden behaviors. Focusing on these issues, we propose a novel association rule learning-based approach 'ARLAS' to identify the behavior of policyholders. The rationale for taking this approach is as follows: in general, the adverse selection (AS) problem in life insurance does not fit the supervised learning paradigm since there are no labels. Still, life insurers need a method that can identify potential AS behaviors. After consulting with domain experts, we made the assumption that AS behaviors exist but are rare. We recognize that this assumption corresponds to the infrequent patterns in the data set, and such patterns can be extracted using association rule learning reversely, that is, looking at patterns with low confidence but high support. Thus, this approach provides a workaround to make predictions without labels, and the predictions can significantly narrow down the list of suspected AS behaviors to be further verified by insurers.

The main contribution of this study is to propose the first unsupervised learning method to detect AS behaviors in relation to life insurance. This problem can also be viewed as an unsupervised outlier detection problem. Hence, for comparison purposes, we included a few outlier detection techniques such as Local Outlier Factor (LOF), Cluster-Based Local Outlier Factor (CBLOF), One-class SVM, and Isolation Forest (IF) to evaluate the performance of our proposed method. We conducted extensive experiments to study model performance and behavior on one of the largest life insurance data sets ever studied in the literature. The experiment results on the life insurance data of 31,800 policyholders suggest that association rules can identify AS behavior and assist the insurance authority to reduce loss and guide changes to insurance premium policy for further development management, and planning. The **key contributions** of this work are as follows:

- We present an end-to-end framework to analyze policyholders' adverse behavior that will help the insurance industry reduce the risk of adverse claims.
- We analyze the life insurance user status to identify adverse behavior using ARLAS along with LOF, CBLOF, IF, and One-class SVM.
- To evaluate the performance, we simulated adverse behaviors by randomly flipping the attribute values. We change a random set of 10% (i.e. 318) of the test set records to be adverse-selected and the attributes are reassigned

by drawing from the corresponding attribute.

- We analyze 10 years of data on 31,800 policyholders, and create novel association rules that show better performance compared to state-of-the-art methods.

The rest of the study is organized as follows. In Section 2, we discuss the relevant work on adverse user identification followed by the proposed framework and description of the method in Section 3. In Section 4, we discuss the empirical analysis, which is applied to a real-life insurance dataset to solve our research problem. In Section 5, we present a comprehensive analysis. Finally, conclusions and future directions are presented in Section 6.

2. Literature review

In the existing literature, machine learning is mainly used for the prediction and optimisation tasks (Liu et al., 2017; Yu et al., 2018; Biddle et al., 2018b,a) in life insurance. In this paper, we explore the task of detecting adverse behavior (AB). Adverse behavior from policyholders is typical and raises the risk of instability in the insurance market. High-risk policyholders deliberately provide false information to the insurer to escape higher premiums, or to avoid being excluded for eligibility (Riddell and Hales, 2018; Islam et al., 2020a). Existing studies on the AS of the policyholder demonstrate that AB policyholders are better informed about the market likelihood, and use information to select their insurance plans (Chu and Chau, 2014; Chau et al., 2013; Sengupta and Rooj, 2019). Additionally, the psychological disorder of the individual can have a deleterious effect on AS behavior. Thus, there is no ambiguity that AS issue has created significant challenges and controversy for insurance industries.

The existing studies by (Cohen and Siegelman, 2010; Bolhaar et al., 2012; Bates et al., 1995) present a clearer view of AS detection. Several studies have highlighted the potential effect of asymmetric information, the proposed methods and key ideas, and have detailed various causes of AS, as shown in Table 1. Thus, in the context of the life insurance industry, it has been shown that scrutiny for AS has not extended to the same extent as that for other issues. Cohen and Siegelman (2010) reviewed many empirical studies and found evidence of AS in insurance markets, that people in poorer health prefer policies that provide more generous coverage, and policyholders who buy more insurance coverage appear to be riskier (Pauly and Zeng, 2004; Lester et al., 2019; Ettner, 1997). Boodhun and Jayabalan (2018) used a supervised machine learning algorithm to assess risk and provide solutions to refine the underwriting process. Boxwala et al. (2011) used statistical and machine learning approaches to identify suspicious records. Although engagement classification is related to AS in life insurance markets, there is no analysis of engagement for observing or measuring which individual factors are more likely to cause AS, which individual policyholders are engaged, who are disengaged, and those who are in between (Wu and Wang, 2011; Angiulli and Pizzuti, 2005; He, 2008). As a result, it is not possible to identify real AS users, and many honest policyholders may suffer. It is worth mentioning that the AS detection method developed in this paper is different from the outlier and anomaly detection methods used in other applications (Yin et al., 2018, 2020; Razzak et al., 2020b; Tewari and Gupta, 2020; Li et al., 2019; Razzak et al., 2020a; Li et al., 2012; Li and Wang, 2017; Wang et al., 2019; Singh and Vardhan, 2019) since no explicit labels are provided; instead, we leverage the rule learning technique, which has a long history but still shines in recent works (Zhou et al., 2019).

From the above review of the existing studies, it is clear that most of the methods focus on limited aspects and were limited in their performance and capability. For instance, (McCarthy and Mitchell, 2010; Cutler and Zeckhauser, 1998; Song et al., 2014; Finkelstein, 2004) provide evidence for AS in insurance markets but they used limited information. However, there are many aspects associated with demographic and socio-economic information such as age, postcode, occupation, and gender, which have made insufficient research concern for AS purposes (Aquino and Douglas, 2003). They go on the AS hypothesis test using statistical models, which could cause bias in the results of the estimation. Yet importantly, there have been some attempts to used data mining and machine learning to analyze and propose solutions using policyholder data within the life insurance industry (Boodhun and Jayabalan, 2018; Huang and Meng, 2019; Islam et al., 2020b). To analyze and describe potential predictive factors, they use straightforward regression models. However, the predictive performance of the existing techniques is rather low. While an increasing body of research combines insurance data with machine learning techniques to observe policyholder behavior, it is challenging to do this with sensitive policyholder data and there is scope to apply advanced machine learning and data analytics techniques.

In summary, the existing research on AS behavior analysis is limited. Very few studies have considered advanced data analytics techniques to meet the practical requirement of the insurance industry. Furthermore, a very limited number of datasets have been used in literature. To deal with the aforementioned challenges, in this work, we analyzed

Table 1

Key studies: different methods for adverse-selection detection in the insurance market.

Source	Solution methods	Key Ideas	Purposes
(Sengupta and Rooj, 2019)	Instrument-free semi-parametric copula regression technique	Identification of AS in the healthcare market	To estimate the effect of health insurance status on healthcare utilization
(Lester et al., 2019)	Search-theoretic model	Identification of AS and imperfect competition	To explore the interaction between AS, screening, and imperfect competition in frictional markets.
(Riddell and Hales, 2018)	Baseline and control optimism classification model	Risk misperceptions and selection in the insurance market	To investigate the relative influence of baseline and control optimism on selection in an insurance market.
(Boodhun and Jayabalan, 2018)	Supervised learning algorithms	Risk prediction in the life insurance industry	To classify the risk level by applying a predictive model.
(Keane and Stavrunova, 2016)	Smooth Mixture of Tobits	Analyze AS and moral hazard in a unified economical framework	To estimate AS and moral hazard effects jointly in the Medigap market.
(Song et al., 2014)	Machine learning methods	Assess financial fraud risk	To identify the risks associated with financial fraud, and help to reduce enterprise risks.
(Meyer et al., 2014)	Data mining classification techniques	Improve the dynamic decision strategies.	To discover treatment strategies by predicting and eliminating treatment failures.
(Boxwala et al., 2011)	Statistical and machine-learning approach	Identify suspicious records	To help privacy officers detect suspicious access to EHRs.
(McCarthy and Mitchell, 2010)	A over E	Adverse selection in life insurance and annuities	To assess the extent to which life insurers can hedge mortality exposure by writing both life insurance and annuities.
(He, 2008)	Conditional correlation approach	Find the relationship between a high-risk and low-risk person	To examine the presence of AS in the life insurance market.

10 years of data on 31, 800 policyholders, and propose the first unsupervised learning method for detecting AS behaviors in the life insurance industry.

3. Methodology

In this section, we first describe the details of data collection and processing. We then present the proposed framework and method for the detection of AB in the life insurance industry.

3.1. Data collection and processing

In this paper, we use two types of datasets, namely 1) questionnaire based behavioral data, and 2) demographic data. We collect the data from one of the most popular insurance companies in Australia, where users are required to answer various questions. The behavioral dataset contains 31, 870 data records related to one of the insurance applicants and includes 834 columns, each pertaining to a yes or no question. On the other hand, the demographic dataset contains information on the policyholders' ID, gender, postcode, age, and occupation. When our dataset was ready, we started processing the data. Before any data analysis process can begin, the dataset requires cleaning and pre-processing to remove ambiguity. Any ambiguity or confusion in the dataset can lead to an incorrect analysis. Therefore, we wrote Python scripts to start the data cleaning process and cleaned our dataset. Finally, we resolved missing and invalid data, and all data was subjected to a quality test.

3.2. Proposed model

In this section, we present a model 'ARLAS' to detect the AS behavior of users in the life insurance market (see Figure 1). Our approach is similar to the method proposed by (Grewal et al., 2019), which has been applied to smart home

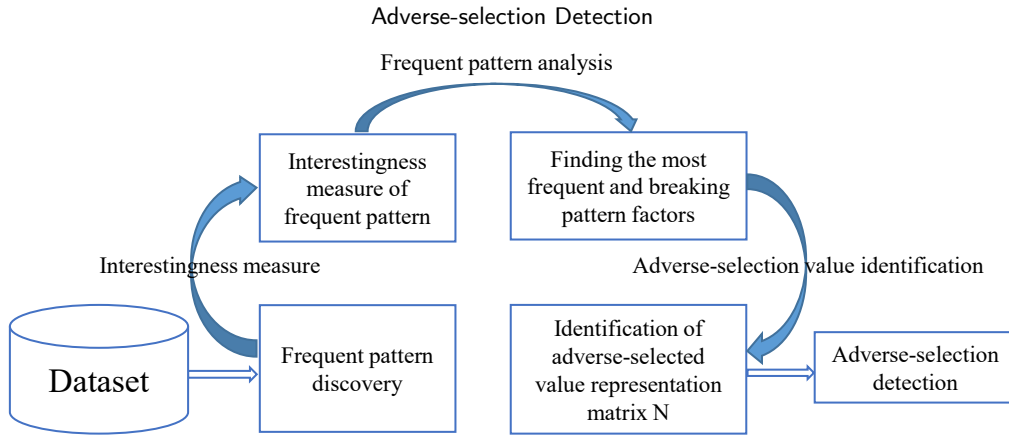


Figure 1: The structure of the ARLAS Framework.

data for behavior monitoring and abnormality detection. We use a frequent pattern mining algorithm which has the ability to locate repeating relationships between unique items in a data set and represent them in the form of association rules. To analyze insurance data to identify the AS behavior of policyholders, we carry out the following steps:

3.2.1. Frequent pattern discovery

The Apriori-based frequent itemset algorithm is used to mine frequent itemsets to generate patterns (Agrawal et al., 1993). It uses an iterative level-wise search technique to discover the $(k + 1)$ item sets from k -item sets, for example, a sample of the questionnaire database that comprises the various questions answered by different users. First, it scans the database to identify all the frequent itemsets by counting each of them and capturing those which satisfy the minimum support threshold. The identification of each frequent itemset set requires scanning the entire database until no more frequent k -question sets are identified.

3.2.2. The interestingness measure of the frequent pattern

To illustrate, we assume that the formal description of a frequent pattern is as follows:

$$(A \rightarrow B) \tag{1}$$

In this description, $A = \{a_1, a_2, a_3, \dots, a_n\} \in I$ and $B = \{b_1, b_2, b_3, \dots, b_n\} \in I$. I show itemsets and $A \cap B = \phi$. The patterns should meet a certain support threshold s . Therefore, according to (Ju et al., 2015), the standard five measures are used to characterize our frequent patterns.

Support: For a transaction set D , the support of an itemset X is given by

$$\text{supp}(X) = \frac{|t \in D; X \subseteq t|}{|D|} \tag{2}$$

Confidence: Confidence is the conditional probability of subsequent occurrence as a result of the previous data. The rule $(A \rightarrow B)$ has confidence c in the transaction set D , where c is the percentage of transactions in D containing A that also contain B , i.e.,

$$\text{conf}(A \rightarrow B) = \frac{\text{supp}(A \cup B)}{\text{supp}(A)} \tag{3}$$

Lift: Lift refers to the ratio of the occurrence probability of B under condition A to that without considering condition A , which reflects the relationship between A and B . The interest of the rule $(A \rightarrow B)$, also known as lift, is:

$$\text{lift}(A \rightarrow B) = \frac{\text{supp}(A \cup B)}{\text{supp}(A) \times \text{supp}(B)} \tag{4}$$

Leverage : A new interestingness measurement method of ARL is based on the description of the defects of the traditional interestingness measurement method. The leverage of the rule $(A \rightarrow B)$ is defined as:

$$\text{leverage}(A \rightarrow B) = \text{supp}(A \cup B) - \text{supp}(A) \times \text{supp}(B) \tag{5}$$

Conviction: The conviction of $(A \rightarrow B)$ is defined as:

$$\text{conviction}(A \rightarrow B) = \frac{1 - \text{supp}(B)}{1 - \text{conf}(A \rightarrow B)} \quad (6)$$

3.2.3. Finding the most frequent patterns and frequent pattern-breaking factors

A dataset contains many factors used to create distinct patterns. However, not all factors can create patterns all the time. The factors are more informative when they play an important role in creating the pattern. Therefore, the factors that are used to create a pattern frequently are the correct factors. In contrast, when it breaks, the most frequent patterns are the adverse-selected factors (ASF). For example, suppose $D = \{t_1, t_2, t_3, \dots, t_n\}$ is a database containing a set of n items $I = \{i_1, i_2, i_3, \dots, i_n\}$. An itemset X is a non-empty subset of I . Given a minimum support threshold, minisupp , find all itemsets when they break the rules with supports greater or equal to minisupp .

We created a list of the frequent patterns (see Section 4.2). We extracted the most frequent pattern through a user-specified minisupp threshold value. In this step, the user-specified support threshold is set to 0.015. We test the support threshold with different sizes ranging between 0.001 and 0.030 where minisupp 0.001 extracts many patterns but affects the execution time and minisupp 0.030 extracts very few patterns. Therefore, by setting the minisupp threshold to 0.015, we decrease the execution time and obtain a reasonable number of patterns. Additionally, to decrease execution time, we omit patterns with lengths greater than four. We then determine how often the factors used to create the most frequent pattern fail. The initial assumption of the breaking frequent patterns (BFP) is that when the factors used to create the most frequent pattern fail, we identify them as the breaking frequent patterns of factors. The set of all breaking frequent patterns is denoted by $\text{BFP}(D, \text{minisupp})$, i.e.,

$$\text{BFP}(D, \text{minisupp}) = X \not\subseteq I \mid \text{supp}(X) \geq \text{minisupp}. \quad (7)$$

For each transaction t , the frequent pattern ASF of t is defined as:

$$\text{ASF}(t) = \frac{\sum_{X \subseteq U, X \in \text{BFP}(D, \text{minisupp})} \text{supp}(X)}{|\text{BFP}(D, \text{minisupp})|} \quad (8)$$

The interpretation of Equation (8) is: if I contains more breaking frequent patterns, its ASF value will be large, which shows that it is more likely to be an AS factor. In contrast, the factors with small ASF values are unlikely to be an AS factor. Obviously, the ASF value is between 0 and 1.

3.2.4. Analysis of AS detection

In this step, AS factors and high-risk policyholders are detected. We first construct a frequent pattern value matrix and further transfer it to the AS value matrix by breaking rules that are interrupted to generate the frequent pattern. We define the AS value matrix N as follows.

$$N = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \quad (9)$$

where $a_{m,n} = \text{ASF}$ value which is mentioned in Equation (8).

We extracted the breaking pattern value of the factors and constructed the $n \times m$ matrix where the row shows the user and the columns show the AS factors. But the different factors in the matrix have different values. We then transferred the values of the factor to a common scale by normalization. We did this to change the values of numeric columns to a common scale, without distorting differences in the ranges of values using the following equation:

$$X_{\text{new}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \quad (10)$$

where X_{new} is a set of the re-scaled values present in X that will now be $0 \leq X_{\text{new}} \leq 1$, X_{min} is the minimum values in X and X_{max} is the maximum values in X . The greater the breaking frequent pattern values of the factor, the higher the probability of it being an AS factor. Therefore, when a user has many AS factor values, they are more likely to engage in AS.

3.2.5. Computation complexity

The proposed association rule learning-based approach is based on frequent pattern mining which is an NP-hard problem (Yang, 2004). Thus, the complexity of the proposed framework can be determined by the frequent itemset mining algorithm. Therefore, the complexity of the proposed method is $O(nN^2)$, where n represents the data records and N represent the number of items.

4. Empirical evaluation

In this section, the data description, interesting pattern list construction, the results and the analysis of our proposed model for solving the AS problem are addressed, respectively.

4.1. Data description

We utilized the following datasets for the empirical works in this study. We applied these datasets to provide a broader and more comprehensive analysis of AS behavioral modeling in the insurance industry. We processed all the datasets to remove any personal identification, including anonymizing names and personal contact details used in the datasets.

Questionnaire dataset: We acquire the dataset from a screening questionnaire provided by a local Australian life insurance company. The questionnaire was large and detailed, comprising data on 31,800 users and each user answered 834 questions ranging from personal details, lifestyle, and family history to occupational details. The data is binary data, where if the applicant answered ‘yes’ to the question, the cell contained a ‘1’, and if the user answered ‘no’, it contained a ‘0’. For example, if a user drinks alcohol, the ‘alcohol’ attribute must contain a ‘1’ in the dataset, and if the user does not drink alcohol, the attribute must contain a ‘0’.

Demographic datasets: There are five different variables in our demographic dataset: policyholder life insurance ID, gender, age, occupation, and policyholder postcode for the 31,800 policyholders. The ‘Gender’ attribute is denoted as either ‘M’ or ‘F’ for ‘Male’ and ‘Female’ respectively. The ‘Postcode’ attribute contains the Australian postcode of the applicant’s residence. The ‘Age’ attribute contains the age of the applicant in whole years, where the youngest applicant is 3 years old and the oldest is 78 years old. The ‘Occupation’ attribute contains 18 different categories. Examples of these include ‘T-Trades’, ‘S-Supervisor of Trades’, ‘R-Special Risk’, ‘Q-Qualified Professional’, ‘OR-Ordinary Rates’, ‘L-Light Trades’, ‘H-Heavy Trades’, ‘F-Financial Professional’, ‘D-Medical/ Dental’, ‘I-Indoor Sedentary’, and ‘C-Community Professional’. As part of the demographic information analysis, we use the Socio Economic Indexes for Areas (SEIFA) data set which consists of four indexes: the Index of Relative Socio-economic Advantage and Disadvantage (IRSAD); the Index of Relative Socio-economic Disadvantage (IRSD); the Index of Economic Resources (IER); and the Index of Education and Occupation (IEO) to rank areas in Australia by relative socio-economic advantage and disadvantage (Australian Bureau of Statistics, 2018).

4.2. Interesting pattern list construction

In this section, we apply the proposed frequent itemset mining algorithm to our collected data. In our proposed framework, we exploit the user-defined minimum support threshold *minisupp* to imply the ‘minimum frequency’ for ‘ARLAS’ model construction and to determine whether there are valid relationships in the provided data. The support value dictates how frequently a particular itemset appears within a dataset where an itemset with higher support justifies greater commonality or popularity. Therefore, we explore the effect of setting various support thresholds, ranging between 0.001 and 0.030, with respect to the number of features identified as candidates.

In Figure 2 we see that the algorithm identifies 66,000 patterns with *minisupp* 0.001, which is all the patterns in the stemmed list. The number of patterns gradually decreases to 4,000 when *minisupp* is set to 0.015, then continues to decrease slightly with an increase support thresholds. When *minisupp* is 0.030, only 147 patterns are returned. Notably, pattern generation is automated; thus, users may consider the output to choose their pieces of interest. This method is suitable in process because the pattern number is mostly small. The return of this work is that user behavior patterns are analyzed from the given data rather than from a pre-decided set. Hence, this condition provides a better extensive and consistent list of patterns to be formed.

Several other measurement methods such as ‘confidence’, ‘lift’, ‘leverage’ and ‘conviction’ are applied where various factors of interest to users are also identified. These detailed aspects include ‘family history’, ‘life style’, ‘insurance history’, ‘employment information’, ‘medical history’, ‘health and risk factors’, and ‘socio-economic factors’ such as ‘remote area’, ‘family type’, ‘gender’, and ‘age’ which are significant to insurance managers. Table 2 shows some interesting patterns’. These results are impressive because these terms are related to policyholders’ behavior information.

Table 2
Different interestingness measure of the different patterns.

Patterns	Supp.	Conf.	Lift	Lev.	Conv.
1. Screening off work 7days minor → Screening off work 7days minor recovered	0.001	0.98	68.41	0.001	50.76
2. Screening Musculo Skeletal Back, Screening Skin Lesion → Screening Consult Test Prescription	0.002	0.98	17.30	0.002	77.79
3. Screening Neurological, Screening Musculo Skeletal Joint → Screening Consult Test Prescription	0.001	0.98	17.27	0.009	66.94
4. Screening Respiratory Asthma, Screening Sensory Eyes → Screening Consult Test Prescription	0.001	0.98	17.32	0.001	85.80
5. High BP Medication, Screening Sensory → Screening Cardio High BP	0.003	0.98	40.12	0.003	49.55
6. Screening consult test , High BP medication → Screening cardio high BP	0.002	0.99	40.54	0.002	147.30
7. Screening consult test prescription, High BP medication → Screening cardio high BP	0.002	0.99	40.54	0.002	145.34
8. Joints area arm, Mental health specialist referral, Screening consult test prescription → Mental health medication	0.001	0.98	20.21	0.001	82.75

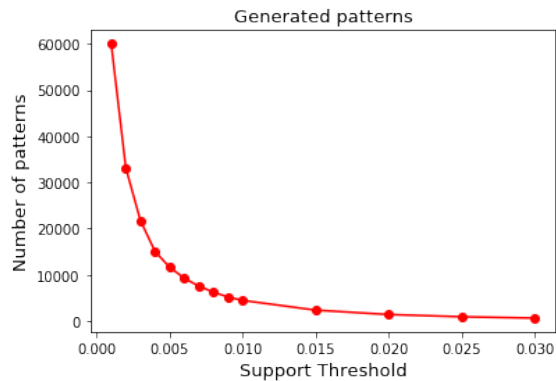


Figure 2: Identifying different rules with different support threshold.

4.3. Result and analysis

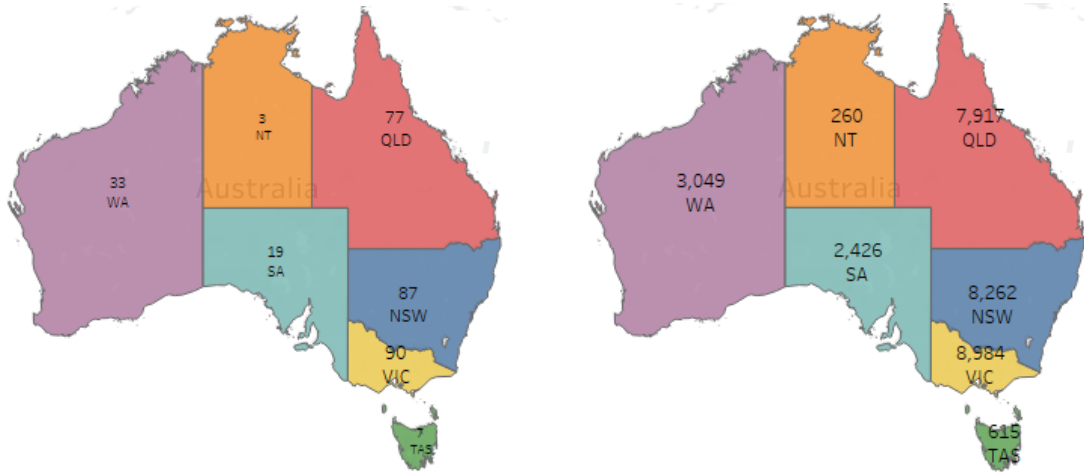
We propose a novel ARL based method to detect the AB of policyholders in the context of life insurance. As this is a new method, we compare our findings with other established unsupervised learning methods which are used for similar analysis such as LOF, CBLOF, IF, One-class SVM. By comparing and contrasting our results, we determine whether it produces similar results. However, without label data we cannot evaluate the performance of the models using simple measures such as accuracy, ROC, AUC, etc. Therefore, in this research, we have select unsupervised learning measures, such as the silhouette (SI) score to evaluate the performance of the baseline models. To evaluate the performance of our proposed method, we create synthetic AS results by randomly flipping the attribute (RFA) values. We begin by partitioning the dataset into training and testing sets. We sample 10% of the data for the test set, and the remaining 90% is chosen as the training set. The original dataset used to generate these results has 31,800 records, so after partitioning process the training set has 28,620 records and the test set has 3,180. We change a randomly sampled 10% (i.e. 318) of the test set records to be AS. For each record that is modified, we choose a random set of up to 1 attribute. The values for these attributes are reassigned by drawing from the corresponding attribute marginally. The higher the value of 1, the greater the degree of AS. However, these scores only evaluate the performance of the models according to the inter- and intra-distance measures and they cannot evaluate the models according to business requirements, thus it is the best option for evaluating the manual validation by insurance professionals. The results and findings of these methods are obtained and compared in Table 3 and we provide further details as follows.

Section 4.3, provides our results. Thus, we should apply the rules to the whole data set to analyze the AS patterns.

Table 3

The result of the experiment.

	LOF	CBLOF	Isolation Forest	One-class SVM	ARLAS
SI	0.50	0.49	0.58	0.58	–
RFA	–	–	–	–	0.63
Number of generated patterns	–	–	–	–	4000
Number of clusters	–	15	–	–	–
Total no. of policyholder	31,800	31,800	31,800	31,800	31,800
Number of adverse policyholders	296	301	307	308	319



(a) Number of adverse policyholders per state.

(b) Number of policyholders per state.

Figure 3: Mapped location of adverse policyholders.

We do not use the synthetic AS data in the rest of the paper. Therefore, through our extensive analysis, we visualize the distribution of AS policyholders. Using the list of AS policyholders derived from ARLAS, we look at the distribution of locations of individual users. The list derived from our approach gives us proportions of 29.15% (90 users) of risky users from Victoria, 26.96% (87 users) from New South Wales, 23.51% (77 users) from Queensland, 10.66% (33 users) from Western Australia, 5.02% (19 users) from South Australia and the remaining 4.7% are spread evenly between the Australian Capital Territory, the Northern Territory, and Tasmania. From this information, it is clear that these figures are correlated with state and territory populations, except for Victoria, which produce a higher proportion of risky users relative to its population of around 6.5 million compared to roughly 8 million in New South Wales. Looking further within New South Wales, we divide the risky users based on more precise locations and discover that a large portion of risky users come from the inner west and eastern suburbs, this being 23.3% and 19.8% respectively.

Similarly, Figure 4 highlights the different occupations of these AS users, and the number of people within each occupation pertaining to the results derived from our proposed approach. We found that the highest percentage of applicants worked in indoor sedentary occupations with 21.9% of AB users (70 individuals) making up this section. An indoor sedentary occupation is defined as any job where the employee spends most of their time sitting down. This covers most desk jobs, and jobs at call centers, professional drivers such as bus drivers, taxi drivers, truck drivers, train conductors and pilots, software developers, accountants, and designers. An article released by (Hoffmann et al., 2016) provided evidence that sitting down for extended periods of time has been linked to a variety of health risks and diseases such as “obesity, diabetes, hypertension, and heart disease.” With all these health risks linked to indoor sedentary occupations, life insurance companies consequently charge higher premiums for this type of occupation while additionally charging more if applicants encounter such health risks.

Next, we categorized our AB users by age. Figure 5 visualizes the AB users within the age ranges to see which age range contains the most problematic users. We discovered that a considerably large portion of applicants fall within the

Adverse-selection Detection

Adverse selection count by occupation and gender

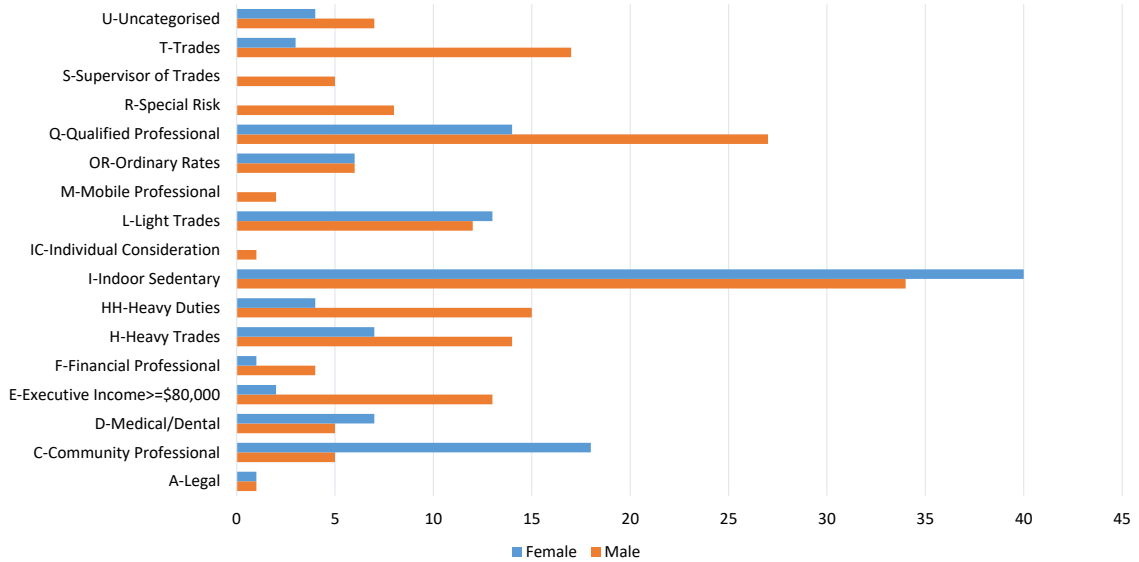
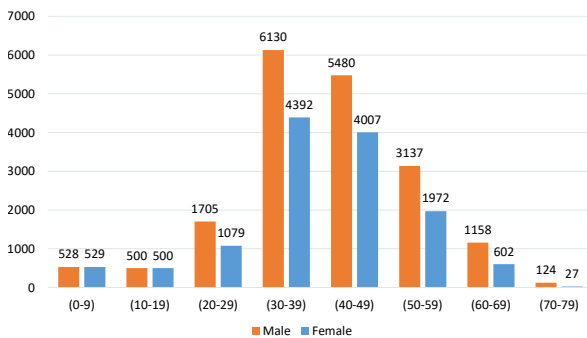


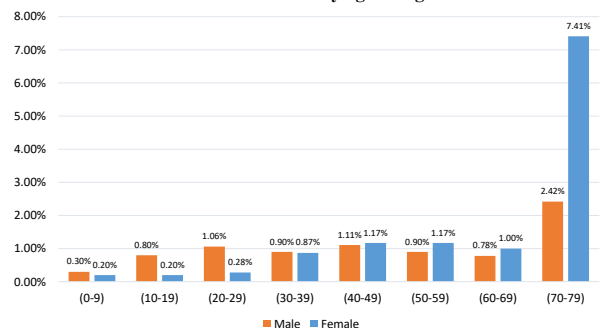
Figure 4: Number of adverse policyholder by occupation and gender.

Gender distribution by age group



(a) Number of policyholders by age group.

Adverse-selection % by age and gender



(b) Number of adverse policyholders by age group.

Figure 5: Gender distribution of adverse policyholder by age group.

31 – 40 range, with 130 people (40.75%) making up this category. Additionally, 63.23% of our AB users (202 people) are aged between 31 and 50. Studies have shown that both males and females in this age bracket have increased chances of diseases such as heart disease, obesity, cancer, and diabetes, which consequently increases insurance premiums if diagnosed. Because of health issues and disease, these being the main contributors to increased premiums, those who wish to avoid such an expense are more inclined to deny being at risk of these diseases and also have the incentive to lie in an application regarding such health issues. This results in the aforementioned consequences that affect not only the insurance company but also other applicants who are forced to pay more to account for the claims made by AB users. Another interesting finding that we made after the visualization process was that 23 AB users (7.2%) were in the < 20 age category. Young people in this bracket rarely have health issues that warrant life insurance. However, if these are special cases, it would make sense as to why our detection methods detected them as being adverse answers.

5. Discussion and implications

In this section, we provide a brief discussion and the implications of AS detection. To identify the AB of policyholders accurately, we consider both the questionnaire-based behavioral and demographic data of Australian life insurance

policyholders. Earlier studies on Australian life insurance mainly focused on statistical approaches. However, in this work, we use the ARL-based approach to explore the AB in depth to have a better understanding of what the data represents, the behaviour of the adverse policyholder, how the adverse policyholder differs from real users, etc. Through our research work, we found that the life insurance industry is at risk in Australia. To manage insurance data, the insurance authority needs to have a comprehensive understanding of normal and risky policyholder information and then be able to identify AB behavior. Therefore, we describe a model to obtain the details of policyholders who help to identify AB.

The analysis of the demographic information in Figure 3 suggests that IMs should pay more attention to NSW and VIC where policyholders stand to receive considerable benefit. Brisbane is also a high-potential area, where high-risk policyholders spend a long time. Therefore, IMs can investigate and develop business strategies among policyholders when they buy their insurance policy and reduce insurance loss.

The behavior analysis of policyholders provides an example of how different professional information can be extracted and analyzed for valuable insights. Prior studies often focused on the significant factors (Haddad and Anbaji, 2010; Lester et al., 2019). However, less significant factors should also be given attention because they can generate a substantial profit for high price ranges. Therefore, the occupation distribution in Figure 4 helps IMs to realize the fact that premiums may rise significantly based on the profession of a policyholder.

The analysis of the age distribution of males and females shown in Figure 5 is necessary for IMs in designing appropriate policy packages for the future. It shows that the average age of the adverse policyholder derived from our proposed method is 41 years old. The difference in the ages of males and females is higher in the age bracket (70 – 79) but in the age bracket (40 – 50), both are almost equal. Female policyholders are more adverse than males; they have less income but higher consumption expenditures than a male policyholder.

During our research, we encountered several limitations such as the availability of the required datasets, the implementation of some other more accurate machine learning algorithms or models such as logistic regression, CNN, ensemble deep learning, etc. There is no standard label data available, thus there is a strong need for a dataset for supervised learning aided by expert knowledge. The results will be more accurate with a huge dataset where all the policyholders information is confirmed. The behavior preference is more applicable and practical when the dataset is huge, which could be one of the limitations of the research that we found during the analysis phase. On the other hand, we only focused on the analysis of user behavior in Australia before COVID-19 since there was not much user information due to lockdown in Australia and many other countries. Additionally, our proposed approach has some limitations. First, Apriori-based frequent itemset mining generates large candidate sets and repeatedly scans the database, which requires a lot of run time and memory. Second, in frequent itemset mining, the order of items in the itemsets is unimportant. However, there are some situations in which the order of items inside the item is important. Third, if a pattern is frequent, its sub-patterns are also frequent. However, there are some cases where patterns and sub-pattern are not the same.

We make the following **key observations**

- The extensive study of 10 years of data of 31,800 policyholders showed that, for the age range 31 – 50, the number of adverse-selected female policyholders is considerably higher than male policyholders, thus IMs should pay more attention to female policyholders in NSW and VIC.
- Our study suggests that, premiums may rise significantly based on the profession of the policyholders.
- We note that the average age of the adverse-selected policyholders is between 35 and 45 years old. The risk of adverse claims for this population can be reduced by considering other factors.
- For a larger dataset, behavior preference could be used to improve the performance of AS.

6. Conclusion and future work

Understanding the behavior of policyholders is necessary to reduce AS, increase business profit, and improve marketing planning. Therefore, we proposed the first unsupervised learning method for detecting AS behaviors in life insurance. We conducted extensive experiments to study the proposed method's performance and behavior on one of the largest life insurance data set ever studied in the literature. A comparison and evaluation on real-world insurance data showed that the proposed approach showed a considerable gains in performance by identifying 319 adverse cases compared to 296, 301, 307 and 308 using LOF, CBLOF, IF, and One-class SVM. This research also lays

out a fundamental framework and structure to support further research on such topics. Being able to recognise a future trend in the insurance industry would help IMs in the decision-making process.

For future work, since there are different outlier detection methods (e.g., clustering), we aim to combine these with our method to develop a hybrid approach, e.g., using ensemble learning to combine and learn the weights of different AS detection methods. We will also investigate the effects on the insurance industry after the removal of the months of lockdown and the change in behaviors and plans for the post-COVID-19 period.

Acknowledgement

This work is partially supported by the Australian Research Council (ARC) under Grant No. DP200101374 and LP170100891. We would like to thank the Australian insurance company for providing us with the unique dataset.

References

- Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases, in: *Acm sigmod record*, ACM. pp. 207–216. doi:<https://doi.org/10.1145/170036.170072>.
- Angiulli, F., Pizzuti, C., 2005. Outlier mining in large high-dimensional data sets. *IEEE transactions on Knowledge and Data engineering* 17, 203–215. doi:<https://doi.org/10.1109/TKDE.2005.31>.
- Aquino, K., Douglas, S., 2003. Identity threat and antisocial behavior in organizations: The moderating effects of individual differences, aggressive modeling, and hierarchical status. *Organizational Behavior and Human Decision Processes* 90, 195–208. doi:[https://doi.org/10.1016/S0749-5978\(02\)00517-4](https://doi.org/10.1016/S0749-5978(02)00517-4).
- Australian Bureau of Statistics, 2018. Socio-economic indexes for areas (seifa) 2016 technical paper.
- Bajari, P., Dalton, C., Hong, H., Khwaja, A., 2014. Moral hazard, adverse selection, and health expenditures: A semiparametric analysis. *The RAND Journal of Economics* 45, 747–763. doi:<https://doi.org/10.1111/1756-2171.12069>.
- Bates, D.W., Cullen, D.J., Laird, N., Petersen, L.A., Small, S.D., Servi, D., Laffel, G., Sweitzer, B.J., Shea, B.F., Hallisey, R., et al., 1995. Incidence of adverse drug events and potential adverse drug events: implications for prevention. *Jama* 274, 29–34. doi:<https://doi.org/10.1001/jama.1995.03530010043033>.
- Biddle, R., Liu, S., Tilocca, P., Xu, G., 2018a. Automated underwriting in life insurance: Predictions and optimisation, in: *Australasian Database Conference*, Springer. pp. 135–146. doi:https://doi.org/10.1007/978-3-319-92013-9_11.
- Biddle, R., Liu, S., Xu, G., 2018b. Semi-supervised soft k-means clustering of life insurance questionnaire responses, in: *2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*, IEEE. pp. 30–31. doi:<https://doi.org/10.1109/BESC.2018.8697227>.
- Bolhaar, J., Lindeboom, M., Van Der Klaauw, B., 2012. A dynamic analysis of the demand for health insurance and health care. *European Economic Review* 56, 669–690. doi:<https://doi.org/10.1016/j.euroecorev.2012.03.002>.
- Boodhun, N., Jayabalan, M., 2018. Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems* 4, 145–154. doi:<https://doi.org/10.1007/s40747-018-0072-1>.
- Boxwala, A.A., Kim, J., Grillo, J.M., Ohno-Machado, L., 2011. Using statistical and machine learning to help institutions detect suspicious access to electronic health records. *Journal of the American Medical Informatics Association* 18, 498–505. doi:<https://doi.org/10.1136/amiaajnl-2011-000217>.
- Butler, J., 2007. Adverse selection in Australian private health insurance, in: *ACERH Policy Forum*.
- Chau, P.Y., Ho, S.Y., Ho, K.K., Yao, Y., 2013. Examining the effects of malfunctioning personalized services on online users' distrust and behaviors. *Decision Support Systems* 56, 180–191. doi:<https://doi.org/10.1016/j.dss.2013.05.023>.
- Chu, A.M., Chau, P.Y., 2014. Development and validation of instruments of information security deviant behavior. *Decision Support Systems* 66, 93–101. doi:<https://doi.org/10.1016/j.dss.2014.06.008>.
- Cohen, A., Siegelman, P., 2010. Testing for adverse selection in insurance markets. *Journal of Risk and Insurance* 77, 39–84. doi:<https://doi.org/10.1111/j.1539-6975.2009.01337.x>.
- Cutler, D.M., Zeckhauser, R.J., 1998. Adverse selection in health insurance, in: *Forum for Health Economics & Policy*, De Gruyter.
- Ettner, S.L., 1997. Adverse selection and the purchase of medigap insurance by the elderly. *Journal of health economics* 16, 543–562. doi:[https://doi.org/10.1016/S0167-6296\(97\)00011-8](https://doi.org/10.1016/S0167-6296(97)00011-8).
- Finkelstein, A., 2004. Minimum standards, insurance regulation and adverse selection: evidence from the medigap market. *Journal of Public Economics* 88, 2515–2547. doi:<https://doi.org/10.1016/j.jpubeco.2004.02.003>.
- Grewal, A., Kaur, M., Park, J.H., 2019. A unified framework for behaviour monitoring and abnormality detection for smart home. *Wireless Communications and Mobile Computing* 2019. doi:<https://doi.org/10.1155/2019/1734615>.
- Haddad, G.K., Anbaji, M.Z., 2010. Analysis of adverse selection and moral hazard in the health insurance market of Iran. *The Geneva Papers on Risk and Insurance-Issues and Practice* 35, 581–599. doi:<https://doi.org/10.1057/gpp.2010.20>.
- He, D., 2008. The life insurance market: adverse selection revisited. *Economics Department, Washington University in St. Louis Campus*.
- Hoffmann, J.C., Mittal, S., Hoffmann, C.H., Fadl, A., Baadh, A., Katz, D.S., Flug, J., 2016. Combating the health risks of sedentary behavior in the contemporary radiology reading room. *American Journal of Roentgenology* 206, 1135–1140. doi:<https://doi.org/10.2214/AJR.15.15496>.
- Huang, Y., Meng, S., 2019. Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems* 127, 113156.
- Hutagaol, B.J., Mauritsius, T., 2020. Risk level prediction of life insurance applicant using machine learning. *International Journal* 9. doi:<https://doi.org/10.30534/ijatcse/2020/199922020>.

- Islam, M.R., Liu, S., Razzak, I., Kabir, M.A., Wang, X., Xu, G., 2020a. Mhivis: Visual analytics for exploring mental illness of policyholder's in life insurance industry, in: 2020 7th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC), IEEE.
- Islam, M.R., Liu, S., Wang, X., Xu, G., 2020b. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining* 10, 1–20. doi:<https://doi.org/10.1007/s13278-020-00696-x>.
- Islam, M.R., Sultana, N., Moni, M.A., Sarkar, P.C., Rahman, B., 2017. A comprehensive survey of time series anomaly detection in online social network data. *International Journal of Computer Applications* 180, 13–22. doi:<https://doi.org/10.5120/ijca2017915989>.
- Ju, C., Bao, F., Xu, C., Fu, X., 2015. A novel method of interestingness measures for association rules mining based on profit. *Discrete Dynamics in Nature and Society* 2015. doi:<https://doi.org/10.1155/2015/868634>.
- Kaushik, S., Gandhi, C., 2019. Ensure hierarchal identity based data security in cloud environment. *International Journal of Cloud Applications and Computing (IJCAC)* 9, 21–36. doi:<https://doi.org/10.4018/IJCAC.2019100102>.
- Keane, M., Stavrunova, O., 2016. Adverse selection, moral hazard and the demand for medigap insurance. *Journal of Econometrics* 190, 62–78. doi:<https://doi.org/10.1016/j.jeconom.2015.08.002>.
- Leach, L.S., Butterworth, P., Whiteford, H., 2012. Private health insurance, mental health and service use in australia. *Australian & New Zealand Journal of Psychiatry* 46, 468–475. doi:<https://doi.org/10.1177/0004867411434713>.
- Lester, B., Shourideh, A., Venkateswaran, V., Zetlin-Jones, A., 2019. Screening and adverse selection in frictional markets. *Journal of Political Economy* 127, 338–377. doi:<https://doi.org/10.1086/700730>.
- Li, D., Deng, L., Gupta, B.B., Wang, H., Choi, C., 2019. A novel cnn based security guaranteed image watermarking generation scenario for smart city applications. *Information Sciences* 479, 432–447. doi:<https://doi.org/10.1016/j.ins.2018.02.060>.
- Li, Q., Schaffer, P., Pang, J., Mauw, S., 2012. Comparative analysis of clustering protocols with probabilistic model checking, in: 2012 Sixth International Symposium on Theoretical Aspects of Software Engineering, IEEE. pp. 249–252. doi:<https://doi.org/10.1109/TASE.2012.28>.
- Li, Q., Wang, Z., 2017. Riemannian submanifold tracking on low-rank algebraic variety, in: Thirty-First AAAI Conference on Artificial Intelligence.
- Lin, C., Lin, C.M., Lin, B., Yang, M.C., 2009. A decision support system for improving doctors' prescribing behavior. *Expert Systems with Applications* 36, 7975–7984. doi:<https://doi.org/10.1108/17410401011052887>.
- Liu, S., Xu, G., Zhu, X., Zhou, Z., 2017. Towards simplified insurance application via sparse questionnaire optimization, in: 2017 International Conference on Behavioral, Economic, Socio-Cultural Computing (BESC), IEEE. pp. 1–2. doi:<https://doi.org/10.1109/BESC.2017.8256362>.
- McCarthy, D., Mitchell, O.S., 2010. International adverse selection in life insurance and annuities, in: *Ageing in advanced industrial states*. Springer, pp. 119–135. doi:<https://doi.org/10.3386/w9975>.
- Meyer, G., Adomavicius, G., Johnson, P.E., Elidrissi, M., Rush, W.A., Sperl-Hillen, J.M., O'Connor, P.J., 2014. A machine learning approach to improving dynamic decision making. *Information Systems Research* 25, 239–263. doi:<https://doi.org/10.1287/isre.2014.0513>.
- Nahar, J., Imam, T., Tickle, K.S., Chen, Y.P.P., 2013. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications* 40, 1086–1093. doi:<https://doi.org/10.1016/j.eswa.2012.08.028>.
- Olayanmi, O.O., Dada, A., 2019. An efficient privacy-preserving approach for secure verifiable outsourced computing on untrusted platforms. *International Journal of Cloud Applications and Computing (IJCAC)* 9, 79–98. doi:<https://doi.org/10.4018/IJCAC.2019040105>.
- Pauly, M.V., Zeng, Y., 2004. Adverse selection and the challenges to stand-alone prescription drug insurance, in: *Forum for Health Economics & Policy*, De Gruyter. doi:<https://doi.org/10.2202/1558-9544.1051>.
- Polyakova, M., 2016. Regulation of insurance with adverse selection and switching costs: Evidence from medicare part d. *American Economic Journal: Applied Economics* 8, 165–95. doi:<https://doi.org/10.1257/app.20150004>.
- Razzak, I., Saris, R.A., Blumenstein, M., Xu, G., 2020a. Integrating joint feature selection into subspace learning: A formulation of 2dpcra for outliers robust feature selection. *Neural Networks* 121, 441–451. doi:<https://doi.org/10.1016/j.neunet.2019.08.030>.
- Razzak, I., Zafar, K., Imran, M., Xu, G., 2020b. Randomized nonlinear one-class support vector machines with bounded loss function to detect of outliers for large scale iot data. *Future Generation Computer Systems* 112, 715–723. doi:<https://doi.org/10.1016/j.future.2020.05.045>.
- Riddell, M., Hales, D., 2018. Risk misperceptions and selection in insurance markets: An application to demand for cancer insurance. *Journal of Risk and Insurance* 85, 749–785. doi:<https://doi.org/10.34917/7645907>.
- Sengupta, R., Roj, D., 2019. The effect of health insurance on hospitalization: Identification of adverse selection, moral hazard and the vulnerable population in the indian healthcare market. *World Development* 122, 110–129. doi:<https://doi.org/10.1016/j.worlddev.2019.05.012>.
- Singh, N., Vardhan, M., 2019. Distributed ledger technology based property transaction system with support for iot devices. *International Journal of Cloud Applications and Computing (IJCAC)* 9, 60–78. doi:<https://doi.org/10.4018/IJCAC.2019040104>.
- Song, X.P., Hu, Z.H., Du, J.G., Sheng, Z.H., 2014. Application of machine learning methods to risk assessment of financial statement fraud: evidence from china. *Journal of Forecasting* 33, 611–626. doi:<https://doi.org/10.1002/for.2294>.
- Spears, J.L., Barki, H., 2010. User participation in information systems security risk management. *MIS quarterly* , 503–522doi:<https://doi.org/10.2307/25750689>.
- Tewari, A., Gupta, B., 2020. Security, privacy and trust of different layers in internet-of-things (iots) framework. *Future generation computer systems* 108, 909–920. doi:<https://doi.org/10.1016/j.future.2018.04.027>.
- Viswanathan, P., Srinivasan, S., Hariharan, N., 2020. Predicting financial health of banks for investor guidance using machine learning algorithms. *Journal of Emerging Market Finance* , 0972652720913478doi:<https://doi.org/10.1177/0972652720913478>.
- Wang, Z., Li, Q., Li, G., Xu, G., 2019. Polynomial representation for persistence diagram, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6123–6132. doi:<https://doi.org/10.1109/CVPR.2019.00628>.
- Wu, S., Wang, S., 2011. Information-theoretic outlier detection for large-scale categorical data. *IEEE transactions on knowledge and data engineering* 25, 589–602. doi:<https://doi.org/10.1109/TKDE.2011.261>.
- Yang, G., 2004. The complexity of mining maximal frequent itemsets and maximal frequent patterns, in: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 344–353. doi:<https://doi.org/10.1145/1014052.1014091>.

- Yin, J., Li, Q., Liu, S., Wu, Z., Xu, G., 2020. Leveraging multi-level dependency of relational sequences for social spammer detection. arXiv preprint arXiv:2009.06231 .
- Yin, J., Zhou, Z., Liu, S., Wu, Z., Xu, G., 2018. Social spammer detection: A multi-relational embedding approach, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer. pp. 615–627. doi:[https://10.1007/978-3-319-93034-3_49](https://doi.org/10.1007/978-3-319-93034-3_49).
- Yu, C., Li, J., Li, X., Ren, X., Gupta, B.B., 2018. Four-image encryption scheme based on quaternion fresnel transform, chaos and computer generated hologram. *Multimedia Tools and Applications* 77, 4585–4608. doi:<https://doi.org/10.1007/s11042-017-4637-6>.
- Zhou, Z., Liu, S., Xu, G., Zhang, W., 2019. On completing sparse knowledge base with transitive relation embedding, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3125–3132. doi:<https://doi.org/10.1609/aaai.v33i01.33013125>.