Towards Simplified Insurance Application via Sparse Questionnaire Optimization

Shaowu Liu*, Guandong Xu*, Xiao Zhu*, Zili Zhou†*
*Advanced Analytics Institute, University of Technology Sydney.
†School of Computer Engineering and Science, Shanghai University.
{shaowu.liu, guandong.xu@uts.edu.au}, chooxxxmail@gmail.com, zhouzili@i.shu.edu.cn

***Abstract*—Life insurance application requires in-person meetings with underwriters, tedious paperwork, and an average waiting period of six weeks before an offer can be made. This outdated process has become a barrier for broader consumer adoption, resulting large coverage gap. In this work, we aim to closing this gap by leveraging data mining techniques to optimize the insurance questionnaire form. Our experiment on 10 years of insurance application data has identified that only $\sim 2\%$ of all questions have shown high relevancy to determining the risks of applicants, resulting a significantly simplified questionnaire.**

I. Introduction

Applying life insurance often requires in-person meetings with underwriters, tedious paperwork, and an average waiting time of six weeks to get policy active. This outdated process has become a barrier for broader consumer adoption, resulting large coverage gap. One of the major issues within this process is that applicants have to complete an overcomplicated questionnaire in order to get assessed by underwriters. The questionnaire, which contains both useful and useless questions, can be as lengthy as 100 pages with more than 2.5K possible questions. Filling such questionnaire is tedious even though some nested questions can be skipped. Therefore, optimizing the questionnaire has become a very first step towards simplified insurance application process.

Advances in machine learning has made it possible to evaluate the importance of each question in a data-driven manner. Specifically, the questionnaire optimization can be considered as a feature selection problem [3] in machine learning. Essentially, each question is considered as a feature and the claims are considered as the labels. The feature selection task is then to select the subset of questions that can efficiently determine how likely an applicant will have a claim in the future.

The rest of the paper is organized as follows. Section II introduces the basic concepts of insurance application process. Section III is devoted to describe the proposed method. In Section IV, the proposed method is applied to an insurance dataset to achieve questionnaire simplification. Finally, conclusions are drawn in Section V.

II. Preliminaries

Life insurance application takes five typically steps [1]:

Compare Quotes The first step is to compare quotes from different companies, even though the final price will vary depending on applicant's situation.

The Application The second step is then to lodge an application by filling some basic personal information, which normally takes half an hour to complete.

The Medical Exam The third step is to take a quick medical exam for blood pressure, weight, etc. This normally takes half an hour to complete.

Underwriting The longest part of the whole application is this underwriting process, in which applicants and underwriters need to complete a lengthy application form. This process takes six months on average.

Decision Once the underwriting is complete, a decision will be made and an offer is given to the applicant. However, if the offer does not meet applicant's expectation, it might be rejected, and the whole process is wasted.

This six-month process may lead to an rejection of offer, therefore it is crucial to shorten the application process, particularly, the underwriting process.

III. Sparse Questionnaire Optimization

Given the questionnaire with more than 2.5K questions, the underwriter will try to estimate the risk of an applicant. However, the questionnaire is extremely sparse as some questions are only enabled if certain answers are provided to some parent questions. To identify which subset of questions have strong impact on claims (risks), we adopted the Minimum-redundancy-maximum-relevance (mRMR) [4] feature selection method.

The goal of mRMR is to measure the relevance of each question to claim reasons. Meanwhile, the redundancy of each question should also be minimized, as some questions may be highly correlated. Specifically, the relevance of a question set Q for the claim reason c is defined by the mean of mutual information between each question q_i and the claim reason c:

$$D(Q, c) = \frac{1}{|Q|} \sum_{q_i \in Q} I(q_i; c) \tag{1}$$

On the other hand, the redundancy of question set Q is calculated as the mean of mutual information between each

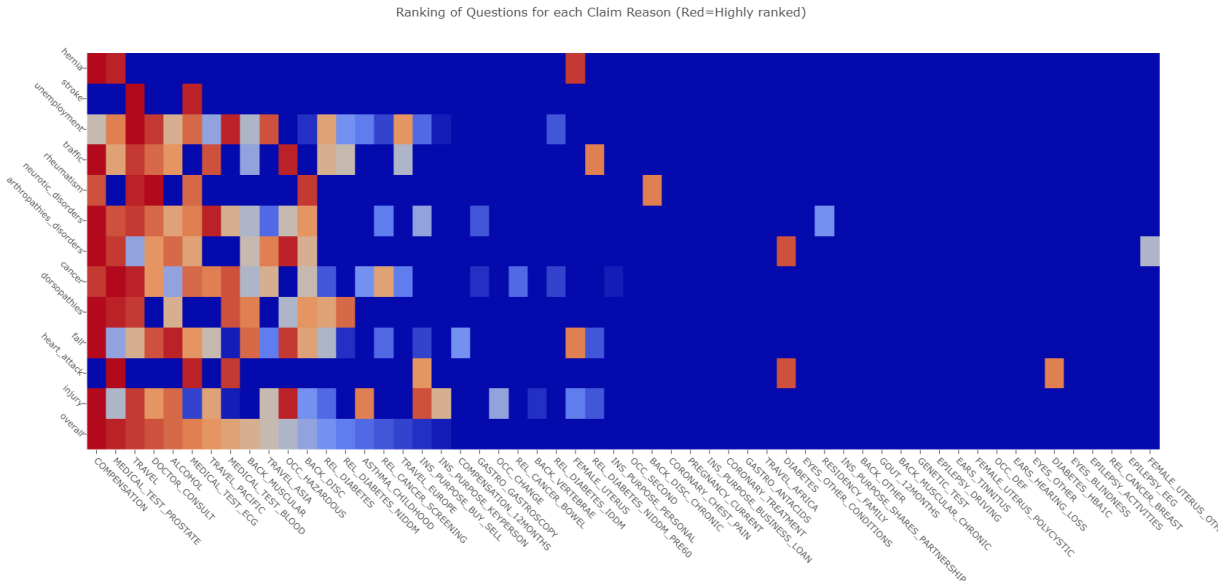

Fig. 1. Impact of Questions

pair of questions q_i and q_j within question set Q:

$$R(Q) = \frac{1}{|Q|^2} \sum_{q_i, q_j \in Q} I(q_i; q_j) \tag{2}$$

Put relevance and redundancy together gives the mRMR measure:

$$\text{mRMR} = \max_{Q} \left[\frac{1}{|Q|} \sum_{q_i \in Q} I(q_i; c) - \frac{1}{|Q|^2} \sum_{q_i, q_j \in Q} I(q_i; q_j) \right] \tag{3}$$

The mRMR algorithm approximates the optimal maximum-dependency feature selection algorithm. The algorithm considers the pairwise interactions of two questions in the question set, however, it doesn't take higher order interactions into consideration. This may lead to potential issues when some "useless" questions become informative when combined with several other questions [2]. Besides, the difference between relevance and redundancy is calculated, while the quotient scheme can also be used. However, due to page constraint, we limit our discussions to the difference scheme only.

IV. Experiment

The experiment was conducted on a dataset provided by one of the largest insurance companies in Australia. The dataset contains roughly 136K life insurance applications spanning between 2006 and 2016. The applications are associated to a claim dataset, such that we know whether a claim has been made or not.

The mRMR feature selection method described in Section III was applied to this dataset to select a subset of questions. The importance of each question is then plotted on the heatmap in Fig. 1, where x-axis shows the selected questions and y-axis shows the reasons of claim. A warmer color indicates higher relevance. For example, the question *OCC_HAZARDOUS* has shown strong impact on claim *injury*. By examining the relevance of each question with respect to the claim reasons, the insurance company is now urge to refine their questionnaire.

V. Conclusions

In this paper we proposed to use mRMR to simplify redundant questionnaire by minimizing redundancy of questions as well as maximizing relevance of questions to insurance claims. By applying this method to six years of insurance data, we managed to identify a subset of $\sim 2\%$ of questions that show strong impact on insurance claims. These findings have helped one of the largest insurance companies in Australia to make transformation of their questionnaire design.

Acknowledgment

The authors would like to thank the Australian insurance company for providing this invaluable dataset.

References

[1] Kenneth Black and Harold D Skipper. *Life and health insurance*. Pearson, 2000.

[2] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(Jan):27–66, 2012.

[3] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[4] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.