Semi-Supervised Soft K-means Clustering of Life Insurance Questionnaire Responses

Rhys Biddle, Shaowu Liu, Guandong Xu
Advanced Analytics Institute, University of Technology Sydney.
Sydney, Australia
rhys.biddle@student.uts.edu.au, {shaowu.liu, guandong.xu@uts.edu.au}

***Abstract*—The life insurance questionnaire is a large document containing responses in a mixture of structured and unstructured data. The unstructured data poses issues for the user, in the form of extra input effort, and the insurance company, in the form of interpretation and analysis. In this work, we aim to address these problems by proposing a semi-supervised framework for clustering responses into categories using vector space embedding of responses and soft k-means clustering. Our experiments show that our method achieves adequate results. The resulting category clusters from our method can be used for analysis and to replace free text input questions with structured questions in the questionnaire.**

I. Introduction

One of the most important repositories of information in the life insurance application process is the questionnaire. In applying for life insurance an applicant must fill out a large questionnaire covering their medical history, occupation and general lifestyle details. This questionnaire contains a mixture of structured and unstructured data. The structured data can be easily analyzed however, the unstructured text data must be read and analyzed manually by an expert human underwriter in order to assess the client at application time. This process is time-consuming, costly, subjective and may be error-prone. The time to fill out the free text questions is a known burden for applicants.

Techniques from the fields of Natural Language Processing (NLP) and Data Mining (DM) can be used to assist in distilling the unstructured text data into structured data allowing for easier analysis and providing suggestions for replacing the free text questions with categorical input options, in the form of drop-down or check boxes. From a technical viewpoint this can be considered a document clustering problem.

The rest of this paper is as follows. Section II briefly introduces the techniques from NLP and DM used in our method. Section III outlines the proposed method. Section IV provides details on the experiments performed on real-world life insurance data. The paper is concluded with a discussion of the results in section V.

II. Preliminaries

A. Document Clustering

Clustering is a learning process where data points are grouped into clusters of similar data points and is one of the most popular learning paradigms in machine learning [1]. Document clustering is specific subset of the clustering from NLP where the task is to form clusters of similar text documents [1] . In this case study a free text response to a question within the application form is considered a document. Formally for each free text response r we want to create a vector $\vec{x_c}$ where each element $\in [0, 1]$ indicates whether the response belongs to cluster c.

B. Soft k-means

Soft k-means is a variation on the original k-means algorithm [2]. The softness refers to the assignment of membership degrees for each cluster rather than a hard binary assignment for a cluster [2]. This allows for a single data point to belong to more than one cluster to varying degrees [2]. Given a set of data points x and a user defined number, k, of clusters with centroids μ_i the membership degree of a data point j to cluster i is given by r_{ij} and is calculated by (1).

$$r_{ij} = \frac{e^{-\beta||x_j - \mu_i||^2}}{\sum_{l=1}^{k} e^{-\beta||x_j - \mu_l||^2}} \tag{1}$$

Once membership r_{ij} has been computed for all data points and clusters then the cluster centroids μ_i must be updated as per (2).

$$\mu_i = \frac{\sum_{j=1}^{n} r_{ij} x_j}{\sum_{j=1}^{n} r_{ij}} \tag{2}$$

The computation of memberships and then update of cluster centroids are continued until the cluster centroids begin to stabilize.

C. Vector-Space Model and TF-IDF

The vector-space model (VSM) is based off the idea of document vectors where each document $\vec{d}$ consists of a weight w_t for each term in the vocabulary in the entire document collection [3]. TF-IDF stands for Term Frequency-Inverse Document Frequency and is a common term weighting scheme in the VSM [3]. The weight w_{td} of term t in document d is calculated by (3)

$$tf_{td} * idf_t \tag{3}$$

Where tf_{td} is the count of term t in document d and idf_t is calculated by (4).

$$log \frac{N}{df_t} \tag{4}$$

Where df_t is the number of documents in the collect where term t appears and N is the number of documents. The

motivation for this weighting is that it diminishes the weight of terms that occur in many documents and as such lack discriminative power for that particular document [3].

D. Semi-Supervised Learning

Semi-supervised refers to the problem where some data points have known labels while others do not [4]. In terms of the clustering problem this means that a clustering algorithm is initialized with the labelled data as the initial starting cluster centroids [5]. Optimal initialisation of clustering algorithms is known to have a strong impact on the resulting clusters[5].

III. Proposed Solution

Our proposed solution is a semi-supervised framework where the labelled portion of the dataset is used as seeds for soft k-means clustering algorithm similar to [5], see 1.

A. Framework

- Label a small proportion of the dataset
- Split the labelled dataset into two partitions, one for computing seeds for clustering another for validation of clustering
- Preprocess the text data through tokenization, stopword removal and spell check
- Generate seed super responses for each cluster c by concatenating all responses labelled with cluster c into a single response
- Embed all responses into VSM using TF-IDF
- Initialize soft k-means with VSM embedded super responses as cluster centroids
- Evaluate performance on holdout validation set using $f1$-score. Any cluster c with cluster membership greater than threshold t is considered as belonging to that cluster for evaluation purposes.

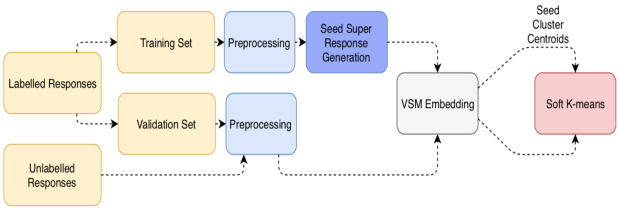

Fig. 1. Framework diagram

B. Motivation

The proposed solution is motivated by the nature of the free text responses contained in the personal disclosure statement. The responses are short answers to specific questions such as: *Please describe the details of your back pain?*, *Please describe the location of your chest pain?*. As a result of both these issues the text contained in the responses lack contextual information and are of a limited scope making the combination of VSM and TF-IDF suitable for capturing the important information, the keywords, in the responses. The choice of soft clustering as opposed to hard clustering was motivated by the fact that in numerous questions there is often multiple pieces of information in each response.

IV. Experiment

The experiment was conducted on a real-world life insurance dataset from a leading Australia life insurer containing nearly 100 thousand applications. The particular question in focus for these experiments were those relating to mental illness. The mean character count for responses to this question is only 49.8 before preprocessing with a drop to 33.1 after preprocessing.The proposed solution was applied to this question where we labelled 7 thousand responses and found 9 separate clusters. From the labelled data points, 2 thousand were used for creating the seed responses for each cluster and the remaining 5 thousand used for validation. The validation metric of choice was $f1$-score which is the weighted harmonic-mean of precision and recall [3].As soft k-means produces membership degrees $\in [0, 1]$ a threshold t was applied to convert to a binary indicator variable in order to compare with the labeled validation set. A mean weighted $f1$-score of 0.85 across all the clusters was achieved on the validation set showcasing that the method performed well in clustering the unseen responses to the correct clusters. Experiments on the same data using a multi-label classification framework with Random Forest as classifier achieved a mean-weighted $f1$-score of 0.70.

V. Conclusion

In this paper we have proposed a semi-supervised framework based on soft k-means clustering. The goals of this framework were to extract simple to understand features from unstructured text responses on life insurance application and to provide suggestions for replacement of free text questions with structured questions. Our proposed method has provided for each free text question, a human readable feature vector and suggestions for question adaptation, both of which are highly desirable by the insurance company. The semi-supervised approach has allowed this task to be performed in a fraction of the time it would take the insurance company to manually sift through hundreds of unstructured free text questions across the hundreds of thousands different applications.

Acknowledgment

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

References

[1] M. Steinbach, G. Karypis, V. Kumar *et al.*, "A comparison of document clustering techniques," in *KDD workshop on text mining*, vol. 400, no. 1. Boston, 2000, pp. 525–526.

[2] D. J. MacKay and D. J. Mac Kay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[3] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press, 2008, vol. 39.

[4] X. Zhu, "Semi-supervised learning," in *Encyclopedia of machine learning*. Springer, 2011, pp. 892–897.

[5] S. Basu, A. Banerjee, and R. Mooney, "Semi-supervised clustering by seeding," in *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*. Citeseer, 2002.