Predicting Financial Literacy via Semi supervised Learnoing

David Hason Rudd, Huan Huo, Guandong Xu

UNIVERSITY OF TECHNOLOGY SYDNEY

02/02/2022

Abstract

It is essential to accurately estimate FL to allocate specific intervention programs to less financially literate groups. The current research work investigated mechanisms to learn customer FL levels from financial data. We propose the SMOGN-COREG model for semi-supervised regression to deal with unbalanced unlabelled data. We compared the SMOGN-COREG model with six well-known regressors on five datasets to evaluate the model's effectiveness on unbalanced and unlabelled financial data.

Introduction

Current SSL methods in FL, most were applied as supervised methods for classification, since real-valued target variables raise practical difficulties for SSL in regression. We applied SMOGN to oversample values to predict rare or uncommon data in the skewed dataset. Empirical results confirmed that the proposed SMOGN-COREG model outperformed all current models. Thus, including unlabelled examples via SSR improves prediction accuracy more than using only labelled data in supervised methods.

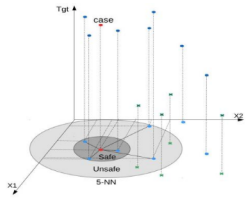

Fig. 1: A synthetic example in SMOGN

Approach

Depending on the model application, SSL can be categorised into inductive and transductive frameworks. Inductive semi-supervised learning can handle unseen data, whereas transductive learning only works on labelled. A transductive learning approach via COREG can successfully use unlabelled data to boost regression predictions. The SMOGN uses a synthetic minority over-sampling technique for regression, with the additional step of using Gaussian noise to perturb interpolated values. After post-processing, SMOGN returns an updated data frame with under and oversampled (synthetic) observations.

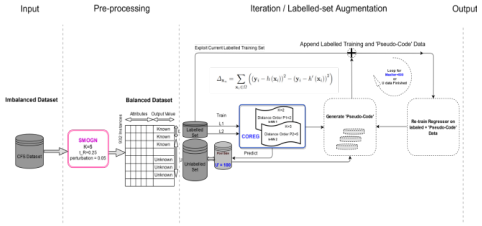

Fig. 2: Proposed SMOGN-COREG model workflow

The well-known SMOTER is employed in SMOGN. It selects K-NN or Gaussian noise based on the distance between the data points. If the neighbour is within a safe distance, it is suitable to conduct interpolation via the SMOTER method. On the other hand, if the selected neighbour is located in an unsafe zone introducing Gaussian Noise is a better selection to generate a new instance.

Experiment Analysis

We initially employed cross-validation with 10 folds of the datasets, one-fold for the test set and the remainder for learning. Unlabelled ratio UR = 80% was used to split the training set in each fold, i.e., only 20% labelled data were involved in learning. COREG maximum iterations = 100, U 0 pool size = 100, and always Δxu > 0 in each iteration, hence maximum labeling capacity = 50000 iterations.

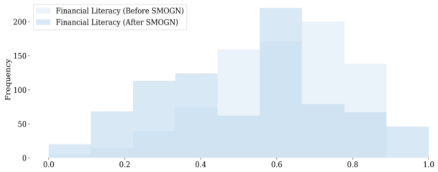

Fig. 3: The results presented in Figure 3 show the skewed data distribution was modified after applying SMOGN. The dark blue histogram shows that after applying SMOGN, fewer sample data points were extracted between the values of 0.6 and 0.9. In contrast, some extra samples were generated from lower than 0.5 values.

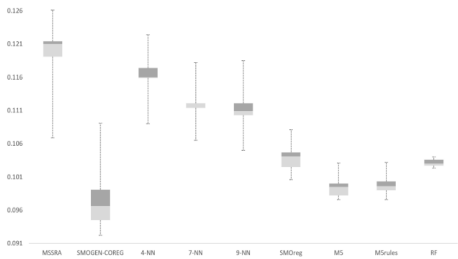

Fig.4: The SMOGN-COREG model performance were compared with Linear Regression, k-NN , SMOreg, M5 Rules model tree, Random Forest, and Meta multi-scheme SSR algorithm (MSSRA).

Conclusion

Irrational financial decisions can have irreversible impacts on quality of life. To prevent this, it is essential to estimate FL and hence allocate specific intervention programs and financial advice to less financially literate groups. Empirical results confirmed that combining SMOGN and COREG algorithms on unlabelled data reduced cost and model runtime, and improved prediction accuracy beyond current supervised regression methods. This study results represent a further step towards applying SSR techniques to assist FinTech companies in narrowing their consumer financial behavior and targeted marketing campaigns.

The proposed solution was based on an offline learning process because the proposed statistical predictive method was applied to previously collected data. Future study will investigate implementing online learning on streaming data.

Acknowledgment

This work is partially supported by the Australian Research Council (ARC) under Grant No. DP200101374 and LP170100891.

Contact Info

david.hasonrudd@student.uts.edu.au