氏　　名	NGUYEN, Ba Hung		
学位の種類	博士(知識科学)		
学位記番号	博知第 276 号		
学位授与年月日	令和 2 年 9 月 24 日		
論文題目	Textual-based Ensemble Learning Models for Corporate Bankruptcy Prediction		
論文審査委員	主査	HUYNH, Nam Van	北陸先端科学技術大学院大学　教授
		藤波　努	同　教授
		DAM　HIEU CHI	同　准教授
		郷右近　英臣	同　准教授
		工藤　康生	室蘭工業大学　教授

論文の内容の要旨

A credit score is an estimation of the likelihood that a borrower will show some undesirable behaviors in the future and supports decision making in credit risk modelling. Nevertheless, the majority of studies were usually based on a snapshot of financial-related data at a specific time point in the past, excluded the trend in business performance over years, and ignored up-to-date business/social activity information that might suggest an early warning of changing in credit worthiness. In addition, advances in data mining for social media and machine learning in application for text mining can be applied for the identification of key features for credit scoring models in term of timeliness, to improve the trade-off between cost and accuracy. Hence, the research that utilises both time series data and textual data can help not only to address the shortage of data types and sources, but also to introduce a new approach in credit scoring.

My research tackle these crucial issues with (i) examining more recent and time-series based financial data with a trendy approach adapted from epidemiology and (ii) the development of new ensemble learning approaches that combine tradition statistical models and machine learning models in credit risk modelling capable of handling corporate rich-featured data, including both numeric and textual data. First, this study employs a large longitudinal data for the UK SMEs to examine their time-to-liquidation using survival analysis, a well-known technique from clinical research. Despite of severely lacking financial data, this study shows the significant effects of SME's demographic characteristics and also further stresses on improvement both in causal interpretation and in model discrimination power when utilising the extended hazard models using the time-varying nature of SMEs financial variables. Another crucial finding in the implication of using some traditional statistic models is the bias in decision-making, where we show that excluding the gender feature eventually reduce the acceptance rates of the better credit worthiness class in both traditional statistical and machine learning-based models. Which questions on the

current inconsistencies of existing regulations for the automated decision-making tools.

With two recent, imbalanced corporate credit datasets, this study then sheds more light on the comparison of corporate credit risk models with different balancing strategies and performance measurements. This study shows that the AUC is not a sufficient measure for the imbalanced dataset as the classifiers tend to overfitted toward the majority class with extremely low value of precision and recall, and second, sampling methods provide significant improvement toward the correctness of classifiers in problems that minority class play an important role as in credit risk management. As any single model has its drawbacks and advantages in a specific domain, combining several models might result in improvement in classification accuracy. In the light of reducing the risk of overfitting as well as underfitting, my research combine models using three approaches to build meta-algorithm including bagging, boosting, and stacking. This study shows that homogeneous and simple heterogeneous ensemble classifiers show better performance compared with the traditional individual classifiers. These findings based on two recent loan portfolios of Vietnamese and US corporate data provide more insights to the practice of corporate credit risk modelling.

Finally, to the utilisation of textual data in credit risk modelling, this study employs topic model on textual data to (i) explore the aspects that defines creditworthiness, (ii) learn the distributed representation of textual data, and (iii) combine it with traditional industry standard to improve the credit risk prediction. I uncover 30 topics embedded in the financial reports which reflect important business aspects and the evolution of words in many topics are in line with crucial economics events. More importantly, the topical features alone provide comparable performance with industrial standard using z-score. And by concatenating the topical features and zscore features, the classifier demonstrates the state-of-the-art performance in corporate bankruptcy prediction. In addition, I proposed novel models that learn from both numeric and textual data from financial reports to examine the predictability of models built from dictionary-based count vectorisation of financial report and dictionary-based sentiment classifier using a financial dictionary. The approach provides comparable and consistent predictive results, yet with more simple and intuitive features compared with the deep learning　model.

Keywords: bankruptcy prediction · ensemble model · textual analysis · topic modelling · sentiment analysis

論文審査の結果の要旨

Credit scoring aims to estimate the likelihood that a borrower will show some undesirable behaviors in the future and supports decision-making in credit risk modelling. Most of

previous studies were usually based on financial-related historical data but had overlooked the trend in business performance over years and up-to-date business/social activity information that might suggest an early warning of changing in credit worthiness. Utilizing both time series data and textual data would help not only to address the shortage of data sources, but also to introduce a new approach in credit scoring. The main objective of this research is to tackle these crucial issues with (i) examining more recent and time-series based financial data with a trendy approach adapted from epidemiology and (ii) the development of new ensemble learning approaches that combine traditional statistical models and ML models in credit risk modelling capable of handling corporate rich-featured data of both numeric and textual types.

Firstly, using two recent imbalanced datasets of loan portfolios of Vietnamese and US corporate data, this research conducted a comparative study of corporate credit risk models with different balancing strategies and performance measurements. It is shown that employing class balancing strategies can mitigate classifier errors, and both homogeneous and heterogeneous ensemble approaches can yield significant improvement on credit scoring. Secondly, it applied topic modeling to textual data to (i) explore the aspects that defines creditworthiness, (ii) learn the distributed representation of textual data, and (iii) combine it with traditional industry standard to improve the credit risk prediction. Thirdly, this dissertation proposed novel models that learn from both numeric and textual data from financial reports to examine the predictability of models built from dictionary-based count vectorization of financial reports and dictionary-based sentiment classifier using a financial dictionary. The approach provides comparable and consistent predictive results, yet with more simple and intuitive features compared with the state-of-the-art deep learning based model.

This dissertation has made significant contributions to methodological and experimental development within the area of credit risk modelling. The research work presented in the dissertation has resulted in 2 journal papers, and several refereed conference papers.

In summary, Mr. NGUYEN Ba Hung has successfully completed all the requirements in the doctoral program of the School of Knowledge Science, JAIST and finished the examination on August 7, 2020, all committee members approved awarding him a doctoral degree in Knowledge Science.