Feature Engineering for Credit Risk Evaluation in Online P2P Lending

Shuxia Wang, Beijing Institute of Petrochemical Technology, Beijing, China
Bin Fu, Peking University, Beijing, China
Hongzhi Liu, Peking University, Beijing, China
Zhengshen Jiang, Peking University, Beijing, China
Zhonghai Wu, Peking University, Beijing, China
D. Frank Hsu, Fordham University, New York City, NY, USA

ABSTRACT

The rise of online P2P lending, as a novel economic lending model, brings new opportunities and challenges for the research of credit risk evaluation. This paper aims to mine information from different data sources to improve the performance of credit risk evaluation models. Besides the personal financial and demographic data used in traditional models, the authors collect information from (1) text description, (2) social network and (3) macro-economic data. They design methods to extract features from unstructured data. To avoid the curse of dimensionality caused by too many features and identify the key factors in credit risk, the authors remove the irrelevant and redundant features by feature selection. Using the data provided by Prosper.com, one of the biggest P2P lending platforms in the world, they show that: (1) it can achieve better performance, measured by both AUC (area under the receiver operating characteristic curve) and classification accuracy, by fusion of information from different data sources; (2) it requires only ten features from different data sources to get better performance.

KEYWORDS
Credit Risk Evaluation, Feature Engineering, Knowledge Engineering, Online P2P Lending, Social Features, Textual Features

1. INTRODUCTION

A credit risk is the risk of default on a debt that may arise from a borrower failing to make required payments (Basel Committee on Banking Supervision, 2000). The goal of credit risk evaluation is to estimate the probability that a borrower will default on a debt. It is a crucial step in relative economic activities. However, it is a difficult
problem because of the information asymmetry between lenders and borrowers. That is, lenders have less information about borrowers’ capability and willingness to pay back than borrowers do. The rise of online P2P lending, as a novel economic lending model, brings new opportunities and challenges for the research of credit risk evaluation. Besides the personal financial and demographic information used in conventional models, there are various new sources of information from the Internet that could be used to evaluate the credit risk of a loan (Lin et al., 2013; Wang et al., 2016). However, too much information will cause information overload. The objective of this paper to study how to extract and select the most useful features from different data sources.

The key of credit risk evaluation is to evaluate borrowers’ capability and willingness of repayment. Repayment capability refers to the borrowers’ ability to repay the loan on time. It depends on the assets, debts, future incomes and expenditures of the borrower. Repayment willingness is a subjective factor, which depends on the thoughts and ideas of the borrower. We mine information from different data sources to evaluate the repayment capability and willingness of borrowers from different aspects to reduce the information asymmetry in online P2P lending.

Most of related works focus on evaluating the effects of various information on the funding success of a loan in online P2P lending, including personal financial information (Puro et al., 2010), social information (Lin et al., 2013; Brandes et al., 2011; Greiner & Wang, 2009; Herrero-Lopez, 2009), textual information (Michaels, 2011; Larrimore et al., 2011; Herzenstein et al., 2011), and so on. Our current work aims to: (1) mine information from different data sources to improve the performance of credit risk evaluation models, i.e. the accuracy of default prediction of loans, and (2) identify the key factors that affect the credit risk.

The rest of this paper is organized as follows. In Section 2, we present some related works. In Section 3, we give the definition of the problem that we plan to solve. In Section 4, we present our assumptions and describe how to extract credit risk features from unstructured data sources. In Section 5, we describe how to avoid the curse of dimensionality caused by too many features. Empirical evaluation and comparison are shown in Section 6. Section 7 concludes the paper.

2. RELATED WORK

Personal financial data is the main information source of traditional credit risk evaluation models. Puro et al. (2010) studied the relationship between loan amount, interest rate and the funding success. Their experimental results showed that lower interest rates decrease the chances of getting the loan funded, while lower loan amounts increase the chance of funded. Emekter et al. (2015) studied the relation between various financial factors and the default rate. Their results showed that credit grade, debt-to-income ratio, FICO score and revolving line utilization play an important role in loan defaults. Loans with lower credit grade and longer duration as associated with high default rate.
Simulation and Visualization of Concept Algebra in MATLAB

An Interactive Visualization of Genetic Algorithm on 2-D Graph
[www.igi-global.com/article/interactive-visualization-genetic-algorithm-graph/67997?camid=4v1a](www.igi-global.com/article/interactive-visualization-genetic-algorithm-graph/67997?camid=4v1a)