Prediction of Potential Customers with Data Streams

謝宜臻
Yi-Jen Shie

國立中山大學應用數學系
Department of Applied Mathematics
National Sun Yat-sen University

Abstract

Prediction of potential customers is an important issue for e-commerce companies. In this study, we analyze this problem for two data sets. The first dataset, based on the Kaggle competition called "Acquire Value Shopper Challenge", contains all the given transactions with size over 22GB. The objective is to make prediction for customers who remain loyal after a promotional period. The second dataset is from a local e-commerce company, which contains customer purchase history and product information. The objective is to make prediction for potential customers who will buy new product given their past purchase information. First we transfer the data stream to a new data format where columns represent aggregated features and rows represent customers. The aspect of feature engineering was based on the aggregation of customer transactions in the past. We consider the prediction models including XGBoost classifier and support vector machine. To evaluate the models, the area under the operating characteristic curve (AUC) and prediction accuracies are adopted. The results provide useful information for e-companies to keep their advertising budget in line with promotional and marketing goals.

Key words: e-commerce, data streams, aggregated features, XGBoost classification, support vector machine