An enhanced feature representation based on linear regression model for stock market prediction

ABSTRACT

Stock price prediction has been an attractive research domain for both investors and computer scientists for more than a decade. Reaction prediction to the stock market, especially based on released financial news articles and published stock prices, still poses a great challenge to researchers because the prediction accuracy is relatively low. For prediction purposes, linear regression is a popular method. Statistical metrics, such as the Document Frequency (DF), term frequency-invert document frequency (TF-IDF) and information gain (IG), are used for feature selection to extract the most expressive features to reduce the high dimensionality of the data. However, the effectivenesses of the available metrics have not been explored in identifying important financial feature representations that have dependable and strong relations with the stock price. The objective of this study are (i) to investigate the performance of five statistical metrics, namely, DF, TF-IDF, IG, Chi-square Statistics (Chi-Sqr) and occurrence in identifying important features that can represent the news and have a strong relationship with the stock price; (ii) to introduce feedback variables, namely, the prediction accuracy (PA), directional accuracy (DA) and closeness accuracy (CA), to capture the interaction between the released news and the published stock prices; and (iii) to introduce a prediction model that integrates features from financial news and a stock price value series based on a 20-minute time lag using linear regression. The experiment used the ELR-BoW method to build a number of 330 datasets with five statistical metrics to select different feature sizes of 50, 100, 150, 200, 250, 300, 400, 500, 600, 700 and 800. The performance of ELR-BoW is observed based on three parameters, namely, PA, DA and CA, and is compared against Naïve Bayes (NB) as the benchmark approach and the Support Vector Machine (SVM). The proposed ELR-BoW-SVM obtained a higher accuracy compared to ELR-BoW-NB, where the best feedback measure is PA, which has an F-measure value of 0.842. In addition, the best number of features is 300 features and using document frequency DF statistical metric. The identification of the top feature representations for financial news is highly promising for automatic news processing for stock prediction. This study demonstrates that the identification of the top feature representations for financial news is highly promising for news article processing in stock prediction.

Keyword: Financial news; Linear regression; Stock market prediction; Statistical metric; Feature representation