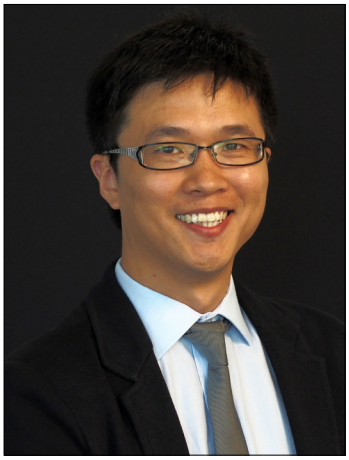
Capital Markets CRC Limited

Computer science research produces new type of regression algorithm for stock picking



Tony Zhao
PhD candidate, Macquarie U.
CMCRC Researcher

Zhao is a PhD student at Macquarie University. He has worked on a wide range of topics including embedded systems, digital signal processing, machine learning, large-scale data processing and natural language processing but his main research area is analysing the sentiment of financial announcements and evaluating topical collocation models.

Study looks at combining text data and financial quantitative data to produce a model for predicting a stocks daily return.

Predicting the daily return of a stock accurately is of critical interest to many companies and individuals who invest in the markets. To predict the daily returns almost all academic work has focused on either pure text data or financial quantitative data. A study by Zhendong (Tony) Zhao, Nataliya Sokolovska and Professor Mark Johnson looks at combining quantitative and text data rather than treating them separately. The resulting regression algorithm produced from the research has come up with some interesting and positive results.

The combination of these two different types of data gives the research far more variables to analyse, which seems to have led to more accurate predictions. Pure text data are things like company announcements, media news (e.g. Australian Financial Review, Sydney Morning Herald, The Australian) and social media (e.g. Twitter) while financial quantitative data include factors like past daily returns, capital size, volatility and the stock price to name but a few. Very little previous research in this area has looked at combining these two types of data which makes Zhendong's research rather unique.

The study compares the predictive performance of four different combinations of features including text data, quantitative data, quantitative and text data together and quantitative and text data with unequal penalty factors (this is where there is a weighting on the quantitative and text variables calculated by using advanced statistical and mathematical techniques). To examine the performance of these combinations the research uses 19,282 ASX announcements from the first half of 2010. The research uses 80% of the announcements to train the algorithm and 20% to test the different combinations. The best performance was from the quantitative and text data algorithm with unequal penalty factors which significantly decreased the mean square error and showed a 2.6% improvement compared to the quantitative data only method.

This study is important as very little academic work has looked at combining quantitative and text data to predict daily stock returns. The results clearly show that this combination outperforms the other methods and suggest that this combination using unequal penalty factors performs best of all. Further work needs to be done to back test Zhendong's research results but if this back testing proves to be successful his research will play a very important role in putting new science behind the theories of predicting daily stock returns.

The *Capital Markets Cooperative Research Centre* is a world-leading research organisation that provides thought leadership and break-through technology solutions for capital and insurance markets (www.cmcrc.com).