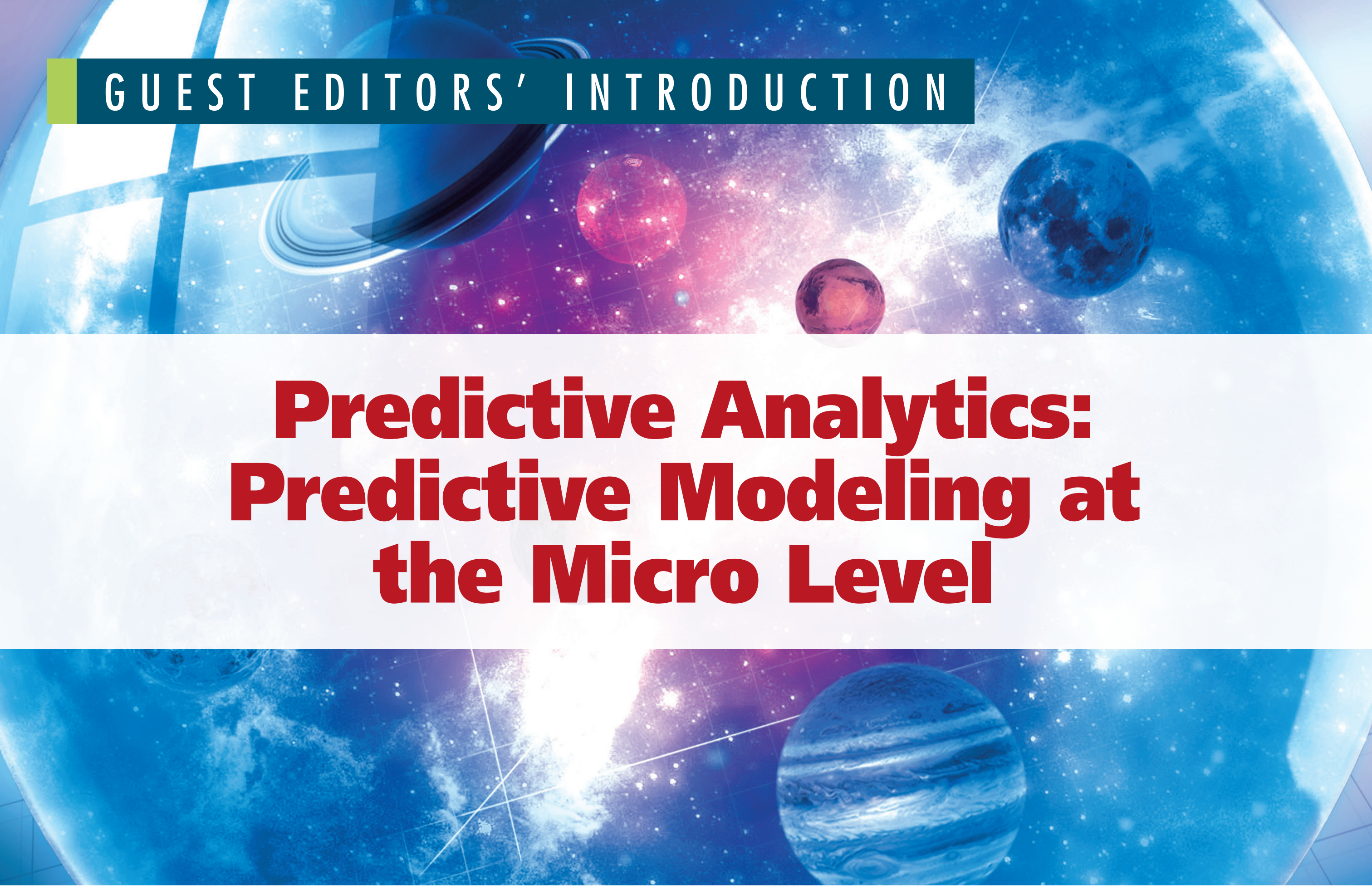

Predictive Analytics: Predictive Modeling at the Micro Level

Donald E. Brown and Ahmed Abbasi, *University of Virginia*

Raymond Y.K. Lau, *City University of Hong Kong*

In the last issue,[1] we presented three articles that were nice examples of analytics at the macro level, an area that has benefited dramatically from the volume and velocity of Big Data to leverage ever-growing datasets for real-time analysis. Micro-level predictive analytics, on the other hand, involves making inferences about future or unknown outcomes pertaining to individual firms, people, or instances. Emphasis on individual or small groups of people, organizations, and other types of granular entities has become the focus of an area referred to as behavior informatics.[2]

Micro versus Macro

Table 1 gives examples of micro- versus macro-level prediction tasks in various application areas, including marketing, security, sales, fraud, politics, security, and health. While the obvious difference is the level of granularity for specific "instances" in the data matrices, the shift in granularity also introduces some nuances with respect to the variety of predictors that can be utilized.

For example, micro-level prediction contains greater intra-entity information, including perceptual constructs, individual behavior, and spatial-temporal indicators. When considering an individual's susceptibility to a particular cybersecurity threat, for instance, recent work suggests that users' security awareness and perceived vulnerability to the threat are powerful predictors.[3,4] Similarly, when contrasting macro-level health

1541-1672/15/$31.00 © 2015 IEEE
Published by the IEEE Computer Society

Table 1. Examples of macro- and micro-level prediction tasks.

Application area	Prediction task	Macro level	Micro level
Marketing	Churn	Predicting a firm's quarterly churn rate	Predicting individuals' likelihood of churn
Politics	Election outcome	Predicting overall election winner	Predicting how a particular person will vote
Security	Cybersecurity threats	Predicting attack volume over next year	Predicting an individual's susceptibility to a cyberattack
Sales	Sales forecasting	Forecasting sales volume over a period of time	Predicting when a given customer will make a purchase
Health	ER visits	Predicting annual patient volume in the ER for a hospital or region	Predicting whether a particular patient will be admitted to the ER the following year
Fraud	Financial statement fraud	Predicting fraud levels in a particular industry segment over a period of time	Predicting whether fraud occurred over a specific firm-period instance

outcome prediction with the micro level, the latter's veracity issues, such as an individual's health-related self-reporting and disclosure biases and inaccuracies, present unique challenges.[5,6] Likewise, when examining fraud at a specific firm-year level, anomalous reporting relative to industry peers provides powerful insights that aren't as effective for macro-level fraud detection.[7] Other examples include unemployment rate prediction[8] and crime detection.[9]

In summary, micro-level prediction tasks embody sufficient differences relative to macro-level tasks. Accordingly, in this special issue on the topic of predictive analytics, we feature four articles that apply analytics at the micro level to individual customers, users, and firms.

In This Issue

In the first article, entitled "Predicting Location-Based Sequential Purchasing Events by Using Spatial, Temporal, and Social Patterns," Yun Wang and Sudha Ram examine a micro-level predictive analytics task looks at individuals' sequential purchase patterns based on spatial, temporal, and social relationship features. Although several studies have exploited these features in isolation, the novelty of Wang and Ram's work lies in their combination and the application of sophisticated smoothing methods to regulate probability distributions related to space and time. Experiments based on 3 million transactions initiated by 13,000 customers from nearly 300 locations reveal that the proposed model outperforms other baseline methods in both spatial sequence and temporal preference prediction. The important takeaway from this work is that social relationship features are useful and complementary to the temporal preferences and spatial contexts for sequential pattern prediction. Given the omnipresent nature of spatial-temporal data—and its centrality to various forms of behavior prediction—the results have important implications that extend well beyond the important purchasing task examined in this study to include a plethora of behavior prediction tasks in domains as broad and diverse as health and security.

The second article, entitled "A Twitter Hashtag Recommendation Model that Accommodates for Temporal Clustering Effects" by Hsin-Min Lu and Chien-Hua Lee, looks at how to identify relevant hashtags to facilitate users' access to interesting tweets and thereby enhance social media-based event tracking. The key nuance of the authors' work is the development of a combined topic-over-time mixed membership model that can exploit the joint usage of tweets and hashtags along with the temporal clustering effect of tweets to predict hashtags (that is, topics) in which users might be interested. Experiments based on 1 million tweets reveal that the proposed TOT-MMM method outperforms other baselines such as a purely tweet similarity-based method and the classical latent Dirichlet allocation topic modeling method. The important takeaway from this work is that latent semantics (topics) embedded in tweets and hashtags can be mined to predict users' potential interests.

The third article is entitled "Does Summarization Help Stock Prediction? A News Impact Analysis," by Xiaodong Li and his colleagues. While previous studies have explored a variety of features for stock return prediction, the main novelty of this work is how it leverages a text summarization method to extract the most informative lexical features from financial news article for stock return prediction. More specifically, the proposed summarization model (self-present sentence relevance; SPSR) can iteratively exploit the semantic links among sentences to identify the most informative lexical units from news articles. Experiments based on six years of stock data from the Hong Kong stock market and financial news articles from the corresponding period reveal that lexical features extracted from summarized news

THE AUTHORS

Donald E. Brown is director of the Data Science Institute and W.S. Calcott Professor in the Department of Systems and Information Engineering at the University of Virginia. His research focuses on data fusion, statistical learning, and predictive modeling with applications to security and safety. Brown is a fellow of the IEEE, past president of the IEEE Systems, Man, and Cybernetics Society, and former EIC of *IEEE Transaction on Systems, Man, and Cybernetics, Part A: Systems and Humans*. He's the recipient of the Norbert Wiener Award and IEEE Millennium Medal. Contact him at deb@virginia.edu.

Ahmed Abbasi is associate professor of information technology, director of the Center for Business Analytics, and a member of the Predictive Analytics Lab at the University of Virginia. His work on security, health, social media, and big data has been funded through multiple NSF grants, an IBM Faculty Award, and an AWS Research Grant. Abbasi is a senior member of IEEE and an associate editor for *ACM Transactions on MIS, IEEE Intelligent Systems, Information Systems Research, and Decision Sciences*. Contact him at abbasi@comm.virginia.edu.

Raymond Y.K. Lau is an associate professor in the Department of Information Systems and associate director of the Center for Intelligent Information Systems at City University of Hong Kong. His research on information retrieval, social media, and big data analytics has been funded by multiple Hong Kong SAR and China regional grants. Lau is a senior member of IEEE and a life member of the Hong Kong Computer Society. Contact him at raylau@cityu.edu.hk.

articles rather than full-length news articles can more accurately predict stock returns at individual stock, business sector, and whole stock market levels. Whereas a fair amount of work has examined stock price movements for a portfolio of firms, individual firm-level prediction is generally considered more challenging due to the volatility of individual firm's financial performance.[10] This further underscores the impressiveness of the results attained by Li and his coauthors.

The final article in this special issue is our own: "Predicting Behavior" (handled in a separate peer review process). Traditional behavior modeling approaches have their shortcomings: a heavy reliance on objective, observed data, and a failure to consider the granular, micro-level decisions and actions that collectively drive macro-level behavior. To address these shortcomings, our framework advocates the integration of objective and perceptual information and decomposes behavior into a series of closely interrelated stages to facilitate enhanced behavior prediction performance. Our framework's strength lies in its simplicity and generalizability: each instantiation of the framework leverages different specific categories of predictors and varying classifiers. We also discuss how the framework could be applied to essentially any application area and present examples pertaining to cybersecurity and health behavior prediction tasks.

We believe the four articles highlighted here are good exemplars of predictive analytics, encompassing novel insights, key nuances such as the importance of leveraging an appropriate variety of predictors, rigorous analytical methods, large-scale experimentation, and interesting findings. All four provide important takeaways for researchers and practitioners working in these respective areas.

References

1. D.E. Brown, A. Abbasi, and R.Y.K. Lau, "Predictive Analytics," *IEEE Intelligent Systems*, vol. 30, no. 2, pp. 6–8.
2. L. Cao, "Behavior Informatics: A New Perspective," *IEEE Intelligent Systems*, vol. 29, no. 4, 2014, pp. 62–80.
3. A. Abbasi et al., "Enhancing Predictive Analytics for Anti-Phishing by Exploiting Website Genre Information," *J. MIS*, vol. 31, no. 4, 2015.
4. F.M. Zahedi, A. Abbasi, and Y. Chen, "Fake-Website Detection Tools: Identifying Elements that Promote Individuals' Use and Enhance Their Performance," to be published in *J. AIS*, 2015.
5. D. Adjeroh et al., "Signal Fusion for Social Media Analysis of Adverse Drug Events," *IEEE Intelligent Systems*, vol. 29, no. 2, 2014, pp. 74–80.
6. I. Bardhan et al., "Predictive Analytics for Readmission of Patients with Congestive Heart Failure: Analysis across Multiple Hospitals," to be published in *Information Systems Research*, 2015.
7. A. Abbasi et al., "MetaFraud: A Meta-Learning Framework for Detecting Financial Fraud," *MIS Quarterly*, vol. 36, no. 4, 2012, pp. 1293–1327.
8. L. Ziang et al., "An Ontology-Based Web Mining Method for Unemployment Rate Prediction," *Decision Support Systems*, vol. 66, 2014, pp. 114–122.
9. R.Y.K. Lau, Y. Xia, and Y. Ye, "A Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media," *IEEE Computational Intelligence*, vol. 9, no. 1, 2014, pp. 31–43.
10. H. Chen, "Business and Market Intelligence 2.0, Part 2," *IEEE Intelligent Systems*, vol. 25, no. 2, 2010, pp. 74–82.