World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:02, 2016

Predictive Analytics for Theory Building

Authors : Ho-Won Jung, Donghun Lee, Hyung-Jin Kim

Abstract : Predictive analytics (data analysis) uses a subset of measurements (the features, predictor, or independent variable) to predict another measurement (the outcome, target, or dependent variable) on a single person or unit. It applies empirical methods in statistics, operations research, and machine learning to predict the future, or otherwise unknown events or outcome on a single or person or unit, based on patterns in data. Most analyses of metabolic syndrome are not predictive analytics but statistical explanatory studies that build a proposed model (theory building) and then validate metabolic syndrome predictors hypothesized (theory testing). A proposed theoretical model forms with causal hypotheses that specify how and why certain empirical phenomena occur. Predictive analytics and explanatory modeling have their own territories in analysis. However, predictive analytics can perform vital roles in explanatory studies, i.e., scientific activities such as theory building, theory testing, and relevance assessment. In the context, this study is to demonstrate how to use our predictive analytics to support theory building (i.e., hypothesis generation). For the purpose, this study utilized a big data predictive analytics platform TM based on a co-occurrence graph. The co-occurrence graph is depicted with nodes (e.g., items in a basket) and arcs (direct connections between two nodes), where items in a basket are fully connected. A cluster is a collection of fully connected items, where the specific group of items has co-occurred in several rows in a data set. Clusters can be ranked using importance metrics, such as node size (number of items), frequency, surprise (observed frequency vs. expected), among others. The size of a graph can be represented by the numbers of nodes and arcs. Since the size of a co-occurrence graph does not depend directly on the number of observations (transactions), huge amounts of transactions can be represented and processed efficiently. For a demonstration, a total of 13,254 metabolic syndrome training data is plugged into the analytics platform to generate rules (potential hypotheses). Each observation includes 31 predictors, for example, associated with sociodemographic, habits, and activities. Some are intentionally included to get predictive analytics insights on variable selection such as cancer examination, house type, and vaccination. The platform automatically generates plausible hypotheses (rules) without statistical modeling. Then the rules are validated with an external testing dataset including 4,090 observations. Results as a kind of inductive reasoning show potential hypotheses extracted as a set of association rules. Most statistical models generate just one estimated equation. On the other hand, a set of rules (many estimated equations from a statistical perspective) in this study may imply heterogeneity in a population (i.e., different subpopulations with unique features are aggregated). Next step of theory development, i.e., theory testing, statistically tests whether a proposed theoretical model is a plausible explanation of a phenomenon interested in. If hypotheses generated are tested statistically with several thousand observations, most of the variables will become significant as the p-values approach zero. Thus, theory validation needs statistical methods utilizing a part of observations such as bootstrap resampling with an appropriate sample size.

Keywords : explanatory modeling, metabolic syndrome, predictive analytics, theory building
Conference Title : ICMLDA 2016 : International Conference on Machine Learning and Data Analysis
Conference Location : Paris, France
Conference Dates : February 22-23, 2016