Applications of Decision Tree Analytics on Semi-Structured North Atlantic Tropical Cyclone Forecasts

Michael Kevin Hernandez, Boeing, Charleston, USA
Caroline Howard, Colorado Technical University, Colorado Springs, USA
Richard Livingood, Colorado Technical University, Colorado Springs, USA
Cynthia Calongne, Colorado Technical University, Colorado Springs, USA

ABSTRACT

This interdisciplinary quantitative study examines how a text mining technique that is widely used to understand financial market forecasts could also help in understanding North Atlantic Tropical Cyclone (TC) forecasts. TCs are a destructive circulation of thunderstorms over a surface low-pressure center. The C4.5 decision tree algorithm has been used successfully to aid in the understanding of financial market forecasts with accuracy rates greater than 55%. This study has examined the use of the C4.5 decision tree algorithm on a 15-year period of the National Hurricane Centers five-day TC forecasts to see if the algorithm could provide a statistically significant value to improving the overall TC forecast accuracy. Improvements in the overall TC forecast accuracy can aid in providing those impacted by a TC adequate early, relevant, and lifesaving TC watches and warnings. This study has helped identify key weather pattern components that have significant information gain, which can help both researchers and practitioners prioritize projects that could help improve TC forecasts.

KEYWORDS

Big Data, Decision Trees, Diffusion of Innovation, Explicit Knowledge, Financial Models, Interdisciplinary Study, Knowledge Management, Text Mining, Tropical Cyclones

INTRODUCTION

Tropical cyclones (TC) are severe thunderstorm system that rotates over a closed surface level low-pressure center that can vary in strength and potentially destructive power based on the TC’s maximum sustained wind speeds (Nakamura, Lall, Kushnir, & Rajagopalan, 2015; NHC, n.d.a; Simpson & Saffir 1974). This potential destructive power of TCs threatens to make landfall on world's coastline yearly. Thus, the more intense the strength of the TC, the more intense is the potential describe power, which could lead to extensive fatalities and property damage (McAdie & Lawrence, 2000; Rappaport et al., 2009; Sheets, 1990; Zhao, Lin, Lee, Sun, & Zhang, 2016). On average $1 billion
in damages accrue from a landfalling TC (National Centers for Environmental Information, 2016). To mitigate this extensive amount of fatalities and property damage the goal of TC forecasters is to provide early and relevant warnings on potential landfalling TCs (Comes et al., 2015; Gall, Franklin, Marks, Rappaport, and Toepfer, 2013; Wang et al., 2015). Early and relevant warnings give people time for preparation and evacuation.

The Hurricane Forecast Improvement Project (HFIP) a primary goal is to improve the forecast accuracy by 50% by 2019 to provide better earlier and relevant warnings (Gall et al., 2013). There are multiple ways to improve forecast accuracy and Gall et al. primarily focused on the use of dynamical and ensemble forecasting models to improve TC forecasts. But there was no mention of the use of novel methods like predictive data analytics. Therefore, this quantitative study focused on improving forecast accuracy by using the C4.5 decision tree. The C4.5 algorithm is a predictive data analytics algorithm that used the National Hurricane Center’s (NHC) tropical discussions from 2001-2015 to help improve forecast accuracy.

The NHC’s tropical discussion contains the explicitly recorded TC forecaster’s logic and knowledge behind each of the forecaster’s TC forecasts (Cangialosi, 2016; Rappaport et al., 2009; Williamson et al., 2014). Since 2001, the NHC has been creating five-day TC forecasts (Cangialosi, 2016). Therefore, from 2001 to 2015, the NHC has accrued 5,131 forecasts in the form of tropical discussions which contains over 1.35 million words (NHC, n.d.a). Thus, this dataset helped categorized this study as a study in big text analytics.

The application of big text analytics on meteorological data deepens the body of knowledge in big data analytics while furthering the field of meteorology by introducing new techniques and procedures (Corrales et al., 2015). This study will evaluate the results from both a meteorological and data analytics perspective to verify the importance and accuracy of the results (Garcia, Ferraz, & Vivaqua, 2009).

Thus, this study posed the following research question: Using the C4.5 algorithm on the five-day tropical discussions from 2001 to 2015, which weather pattern components can improve the Atlantic TC forecast accuracy? For this study, the null hypothesis (H0) is non-directional, whereas the alternative hypothesis (H1) is directional:

**H0:** There are no significant differences in the C4.5 algorithm derived weather pattern components, which can decipher the difference between a successful and unsuccessful TC forecast.

**H1:** There are significant differences in the C4.5 algorithm derived weather pattern components, which can decipher the difference between a successful and unsuccessful TC forecast.

This study used the Cohen’s kappa statistic value (for interrater agreement) of 0.81 to 1.00 as the proof criteria.

The independent variables were the individual weather pattern components that interact with the TC. The dependent variable is the classified TC forecast. This study analyzed each TC forecast as either scoring above (unsuccessful forecasts) or below (successful forecast) the 15-year average forecast error to classify TC forecasts. Subsequently, the results could help provide insights to TC forecasters to improve the forecasters’ TC forecasts and to TC researchers to focus on improving an understanding of certain weather pattern components (Gall et al., 2013; Rappaport, Jiing, Landsea, Murillo, & Franklin, 2012).

Section two presents the foundation for this study, which describes the reasoning as to the use of the C4.5 algorithm’s use of tropical discussions. In Section three, there is a further exploration of the implementation of the C4.5 algorithm on this 15-year dataset. Section four presents the results of the data collection, data processing, and data analysis. Finally, section five provides an overview of the results, the limitations of the study, and identifies future research.
Use of Data Reduction Process to Bankruptcy Prediction: Evidence from an Emerging Market
www.igi-global.com/article/use-of-data-reduction-process-to-bankruptcy-prediction/148637?camid=4v1a

Realising Virtual Reality: A Reflection on the Continuing Evolution of New Media
Allan McLay (2012). *Technological Change and Societal Growth: Analyzing the Future* (pp. 144-161).
www.igi-global.com/chapter/realising-virtual-reality/62782?camid=4v1a