Detecting Adverse Events in an Active Theater of War Using Data Mining Techniques

Jozef Zurada
University of Louisville
USA
WSB Gdansk
Poland
jozef.zurada@louisville.edu

Waldemar Karwowski
University of Central Florida
USA
wkar@ucf.edu

Donghui Shi
Anhui Jianzhu University
China
sdonghui@gmail.com

Jian Guan
University of Louisville
USA
jeff.guan@louisville.edu

Erman Cakit
Aksaray University
Turkey
ermancakit@aksaray.edu.tr

***Abstract* – This study investigates the effectiveness of data mining techniques in detecting adverse events based on infrastructure development spending, the number of project types, and other variables in an active theater of war in Afghanistan using data sets provided by the Human Social Culture Behavior program management (2002-2010) of the U.S. Department of Defense. The study first applies feature reduction techniques to identify significant variables, then uses five cost-sensitive classification methods and reports the resulting classification accuracy rates and areas under the receiver operating characteristics charts for adverse events for each method for the entire country and its seven regions. The results show that when analysis is performed for the entire country, there is little correlation between adverse events and project types and the number of projects. However, the same type of analysis performed for each of its seven regions shows a connection between adverse events and the infrastructure budget and the number of projects allocated for the specific regions and time periods. Among the five classifiers, the C4.5 decision tree and k-nearest neighbor provided the best global performance.**

Keywords: active war theater; data mining; adverse events; prediction; classification

I. INTRODUCTION

The U.S. Department of Defense (DoD) uses the following definition for irregular warfare: "*a violent struggle among state and non-state actors for legitimacy and influence over the relevant population(s).*" Irregular warfare is a non-conventional warfare which includes non-proportional force to subdue and coerce the civilian population in the regions in which opposite forces are not large and effective. The success of irregular warfare operations depends heavily on protecting the civilian population by the military as the civilian population is the primary target of irregular warfare [1]. Recognizing the challenges of the dynamic of irregular warfare among various actors, the U.S. military has made some changes and accommodations to its force structure. Also, the DoD initiated and developed the Human Social Culture Behavior (HSCB) modeling program. The main goal of the program was to guide and help the U.S. military in understanding different cultures while operating in overseas countries and to better organize and control the human terrain during irregular warfare. The military uses HSCB models to understand the behavior and structure of organizational units at the macro level (i.e., health, politics, energy, economics, security, water and sanitation, and social and cultural aspects) and at the micro level (i.e., terrorist networks, tribes, customs, and military units). These HSCB models are important and attract a great deal of attention with regard to current and future operational military and non-military requirements. These models are also very complex as they exhibit non-linear and fuzzy behavior and are often ill-defined with respect to their socio-economic-cultural factors.

II. PREVIOUS WORK

Several studies have attempted to develop models of human behavior from patterns identified in the data in order to predict the effects of actions aimed at disrupting terrorist networks [2]. Since terrorist attacks are not random in space and time, it is possible to discover representative patterns and trends in adverse activity or behavior over time and space by analyzing the geospatial intelligence on reported incidents. The studies concluded that these patterns and trends could be used for prediction future attacks and that they might help decision-makers to allocate more resources and personnel to the places which are more likely to be attacked and also to try reduce the number of such attacks. These studies used

 ISBN: 978-1-61208-614-9

fuzzy inference systems (FIS), adaptive neuro-fuzzy inference system (ANFIS) and wavelet neural networks to analyze terrorist attacks time series.

Other studies built models based on the input variables such as infrastructure development spending projects, the number of projects, and population density. The models applied multiple linear regression, data mining and soft computing techniques such as neural networks, ANFIS, and FIS as well as fuzzy C-means and subtractive clustering for predicting four categories of adverse events, i.e., the number of killed, the number of wounded, the number of hijacked, and the number of events at month *t* in an active theater of war in Afghanistan [3]. These four categories of events are collectively called "adverse events". The studies performed analysis for the entire country and its seven regions and used variable reduction techniques to eliminate redundant attributes as well as implemented sensitivity analysis for the neural network to determine the cause and effect relationship between the input and output variables. However, due to the sparse nature of the input and output data (between 87% and 98% of values for the four adverse events are 0′s, with a 0 representing lack of events), the prediction errors generated by the models for the four adverse events were significantly high. Thus due to the unbalanced nature of the data precise prediction of the number of four adverse events was an extremely challenging and difficult task.

III. DATA SETS

The data sets for the five mentioned studies and this study were provided by the HSCB program management of the U.S. DoD. The time-dependent data were collected over the years 2002 through 2010 and represent more than 30,000 records and over 100 variables. Among other variables, the data sets included the following input variables: the budgeted amount [$US] for 14 categories of infrastructure investments in the areas such as Agriculture and Health, the number of 14 project types at years *t-2*, *t-1*, and *t*, as well as the mentioned four categories of adverse events at month *t-1*, seven regions, and the male and female urban and rural population densities. The output variables included the mentioned four categories of adverse events at month *t*.

IV. DESCRIPTION OF THE STUDY AND RESULTS

This study investigates the effectiveness of data mining techniques in detecting/classifying adverse events based on the infrastructure development spending in 14 project categories, the number of project types, and other variables in an active theater of war in Afghanistan using the same data sets that were used in the five mentioned studies. First, the study recodes the four output variables (the number of killed, wounded, hijacked, and events) representing adverse events into the binary representation, i.e. two classes. For example, killed (Yes or 1) or not killed (No or 0) or an event happened (Yes or 1) or did not happen (No or 0). Then it applies feature reduction techniques to identify significant variables. Next to compensate for class imbalances, the study uses five cost-sensitive classifiers such as neural networks (NN), k-nearest neighbors (*k*-NN), C4.5 decision trees (DT), support vector machines (SVM), and random forest (RF) to detect adverse events. Finally, the study reports the resulting classification accuracy rates and areas under the receiver operating characteristics (AUROC) charts for the four adverse events for each classifier for the entire country and its seven regions. The AUROC values, which testify to the global performance of the classifiers, are measured on the [0.5, 1] scale, where 0.5 and 1 indicates a bad classifier and a good classifier, respectively. For example, the AUROC values for the entire country for the four adverse events were within the [.688, .805] range. The results show that the AUROC values for events are generally higher than the AUROC values for dead, wounded and hijacked; and that the AUROC values for hijacked are generally lower than the AUROC values for dead, wounded and events. The hijacked category was the most highly underrepresented in the data sets.

The results show that when analysis is performed for the entire country, there is little correlation between adverse events and project types and the number of projects. However, the same type of analysis performed for each of its seven regions shows a connection between adverse events and the infrastructure budget and the number of projects types allocated for the specific regions and time periods. For example, for region Eastern the following variables (project categories) were identified as significant: Energy, Governance, Emergency Assistance, and Gender, as well as urban male and female population densities, rural female population density, killed at month *t-1*, and number of events at month *t-1*. Among the five classifiers, the DT and *k*-NN generated the best rates in terms of global performance.

V. CONCLUSION

The models presented in this study could support decision makers who analyze historical economic data on how regional funds allocation can best help minimize adverse events. Though the models used Afghanistan data, they may be applicable for other countries that are looking to build infrastructure while the threat of terrorist and military activities are present.

REFERENCES

[1]. J. Clancy, and C. Crossett. "Measuring effectiveness in irregular warfare", Parameters, 37(2), 88-100, 2007.

[2]. D. Schmorrow, and D. Nicholson. Advances in Cross-cultural Decision Making. Boca Raton: CRC Press, Chapter 38, 374-384, 2011.

[3]. E. Çakıt, and W. Karwowski. "Predicting the occurrence of adverse events using an adaptive neuro-fuzzy inference system (ANFIS) approach with the help of ANFIS input selection". Artificial Intelligence Review, 48(2), 139-155, 2017.

 ISBN: 978-1-61208-614-9