Data Stream Models for Predicting Adverse Events in a War Theater

Donghui Shi
Department of Computer Engineering
School of Electronics and Information Engineering
Anhui Jianzhu University
Hefei, China 230601
sdonghui@gmail.com

Jozef Zurada
Department of Computer
Information Systems
College of Business
University of Louisville
Louisville, KY 40292
jozef.zurada@louisville.edu
WSB Gdansk, Gdansk, Poland

Waldemar Karwowski
Department of Industrial Engineering and
Management Systems
University of Central Florida, Orlando,
FL 32816-2993, USA
wkar@ucf.edu

Jian Guan
Department of Computer Information Systems
College of Business
University of Louisville
Louisville, KY 40292
jeff.guan@louisville.edu

Abstract

Predicting adverse events in a war theater has been an active area of research. Recent studies used machine learning methods to predict adverse events utilizing infrastructure development spending data as input variables. The goals of these studies were to find correlation and disclose the main factors between adverse events and human-social-infrastructure development projects, and reduce the occurrence of the adverse events. The predictions still have large errors compared with the real values using the existing methods. The reason could be that some significant variables are removed to comply with constraints in a soft computing model such as neural networks, fuzzy inference systems (FIS) and adaptive neuro-fuzzy inference systems (ANFIS) that work well with a smaller number of variables. In this paper, a data stream approach using three data stream regression algorithms, AMRules, TargetMean and FIMTDD, is proposed to predict the adverse events so that much more input variables could be included. The results show that the data stream methods generate better results than machine learning methods used in the previous studies, thus helping us better understand the relationship between infrastructure development and adverse events. In addition the data stream methods also outperform the traditional linear regression model. An important advantage in using data stream methods is the ability to create and apply predictive models with a relatively small amount of memory and time. Finally, the use of data stream methods provides an additional advantage by allowing the user to observe error distribution over time for more accurate assessment of the performance of the resulting models.

1. Introduction

Adverse events are caused by terrorist activities in a war theater in countries such as Afghanistan. The Human Social Culture Behavior (HSCB) modeling program [2, 14] was developed by the U.S. Department of Defense (DoD) to help the military to undertake infrastructure development efforts to stabilize the country, and consequently to decrease the number of terrorist events that mainly affect the civilian population.

Recently many methods such as linear regression, neural networks, FIS, ANFIS, fuzzy overlay models were applied in various studies to predict adverse events (the number of killed, the number of wounded, the number of hijacked, and the number of events) using infrastructure development spending as input variables in an active war theater in Afghanistan [6, 7, 8, 9, 10, 11]. Infrastructure development included areas such as Education, Community Development, Governance, Transport, and Agriculture. These studies used the data sets provided by the HSCB program management of the U.S. DoD. The mean absolute error (MAE) and the mean absolute percentage error (MAPE) were used to evaluate the prediction results.

Although machine learning methods are applied to predict the adverse events, the MAE and MAPE values in former studies were still large. The possible reason

URI: https://hdl.handle.net/10125/59556
ISBN: 978-0-9981331-2-6

could be that some significant input variables were not included when FIS and ANFIS were used in the studies. On the other hand, if too many input variables are retained, these models might not work normally due to the limitation of the memory for data sets including over 30000 instances and 100 variables. For example, in the study [6], the exhaustive search function was used for selecting input variables in ANFIS modeling using MATLAB. After an exhaustive search, only 1–4 input variables from a large set of inputs were picked. For a large data set with high dimensionality, the use of these traditional machine learning algorithms to process these types of data can present challenges and fail to produce desirable results. If some significant variables are removed as input variables, the prediction performance would suffer. In this study, feature selection techniques retained between 6 and 20 variables, depending on the scenario used.

Since data stream methods can run in a limited amount of memory and a limited time for a large data [4], the study [15] proposed the use of data stream methods to classify incidents in the aviation safety from incident reports. The data sets in the research include over twenty attributes which were extracted from a narrative field in the incident reports and over 168,227 instances. The results show that data stream methods can improve the classification accuracy for a larger data set with a high dimension.

Up to now, most data stream studies mainly focus on classification algorithms, and few studies have closely examined data stream regression methods. In the paper, we will investigate the use of the data stream regression models for the large data sets for the adverse events in an active war theater. We compare the performance of three data stream algorithms: AMRules, TargetMean and FIMTDD to the traditional linear regression and, due to space constraints, to only one of the previous studies [8]. The paper is organized as follows. Section 2 describes the error measures for data streams and the three data stream regression algorithms. Section 3 introduces the data set used in the simulations. Section 4 discusses the experiment results for the traditional linear regression, the three data stream algorithms, and the previous study. Finally, section 5 draws a conclusion.

2. Methods

2.1. The Measure Methods for Data Streams

The data stream environment is different from the traditional batch setting. The main significant features of the data stream methods are the following: (1) process an instance at one time, (2) use a limited amount of memory and a limited time, and (3) classify or predict the instance at any time [3, 4]. Thus, a data stream approach allows one to analyze the data continuously in real time. In a data stream setting Prequential method was used to test the model using each instance before the instance is used for training, and incrementally update the accuracy of the model. The data stream approach allows one to capture the accuracy profile of the model over time. In a real application, a sliding window or a fading factor forgetting mechanism is used for evaluating a classifier or a regression model by testing then training with each instance in order. In the study, the data stream regression algorithms are used for the adverse events data sets. MAE and RMSE are used for measuring the performance of the data stream regression algorithms.

2.2 AMRules and TargetMean

Adaptive Model Rules (AMRules) algorithm developed by [1, 12] is an incremental algorithm for rules-based learning and is a popular data stream regression algorithm. AMRules can add and remove the rules as the data stream evolves. The form of the rule is the following [5]:

$$C \rightarrow M$$

In the above rule C represents the antecedent which is a conjunction of literals and M represents a model that can predict value a. The literal is a condition such as $A = a$, or $A \leqslant v$ or $A \geqslant v$, where A is a discrete attribute and a is one of its values, and A can also be continuous and v is a numerical value. M is a regression model. The AMRules algorithm has three types of regression models: (1) the mean values of the target attribute; (2) a linear combination of the attributes; and (3) a choice between (1) and (2), resulting in a regression model with a lower mean absolute error according to the recent instances.

AMRules has some different features from decision trees. For example, a decision tree model includes a set of exclusive and complete rules, whereas AMRules uses a set of rules that are neither exclusive nor complete. The rules need not cover all instances and that an instance may be covered by a set of rules. AMRules supports a set of ordered or unordered rules. If the rules are ordered rules, the prediction result of an instance is that of the first rule. If the rules are unordered, all rules that cover an instance are used and the algorithm averages their predicting results. A critical feature of AMRules is to create new rules, extend existing rules, and remove useless rules.

TargetMean is also a rules-based learning algorithm derived from AMRules. It uses the mean of the target variable calculated from the instances covered by the rule as the decision strategy. It is a special form of

AMRules. TargetMean is more robust as it can work with the nominal and numeric input variables. However, AMRules can work only with numeric input variables.

2.3 FIMTDD

FIMTDD [13] is a decision tree for streaming regression from data streams with drift detection. It is an extension of the Hoeffding Tree algorithm. FIMTDD has some features similar to Hoeffding Trees for classification, but it is used for data stream regression. It has some interesting features [5]: (1) variance reduction is used; (2) numeric attributes are processed using an exhaustive binary tree algorithm; (3) perceptrons are used at the leaves to adapt to drifts; (4) the Page-Hinkley method is applied to detect changes in the error rate at the inner nodes of the decision tree; (5) if a subtree is underperforming, a new tree is grown with new incoming instances; it replaces the subtree with the new tree that has better performance; and finally (6) it uses some pruning rules to avoid storing too many values of the outcome. One of the limitations of FIMTDD is that it does not work well with sparse data.

2.4. The framework of data stream methods

Figure 1 shows the framework for detecting adverse events using data stream methods. In the study, the three data stream algorithms are used to predict the adverse events for the whole dataset: Dead, Wounded, Hijacked and Events and for the sub datasets, one for each of the seven regions. The framework includes the two main steps: input variables selection and Prequential measurement for the data stream regression algorithms.

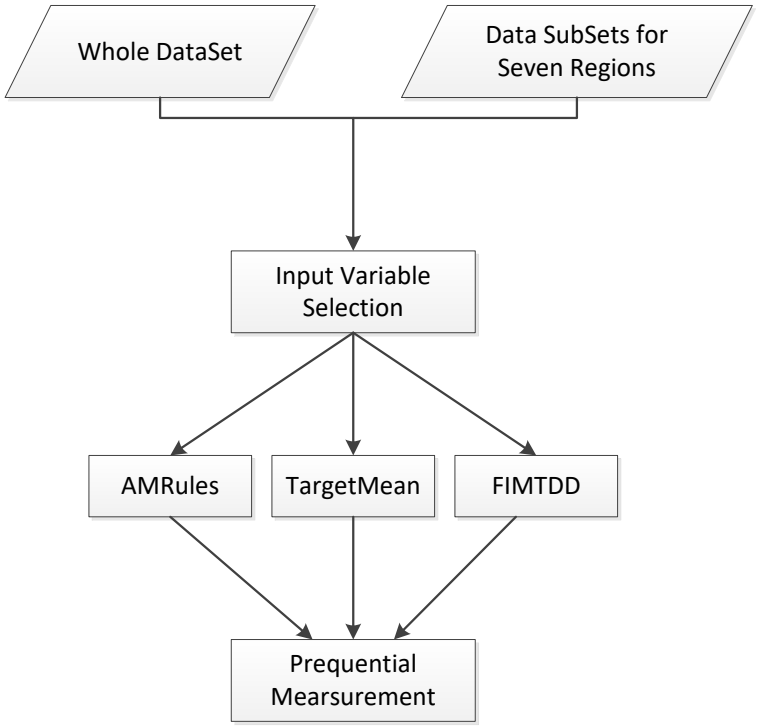

Figure 1. The Framework of Data Stream Methods for Detecting Adverse Events

3. Data Set

In this study, the data sets about Afghanistan provided by the HSCB program management are applied. Some infrastructure development variables are used as input variables, and the number of killed, the number of wounded, the number of hijacked, and the number of events are used as the four output variables. They are organized as the four data sets, each with one dependent variable representing the number of Dead, Wounded, Hijacked and the number of Events (Event_Nu(t)) at time t. The input variables in the four datasets also include the population density, province,

city, district, project types and their number, and as Education, Community Development, Governance, Transport, and Agriculture over the period of three years. There are 101 attributes and 33,600 records collected between 2002 and to 2010.

In the data sets, the input variables are the sum of budget allocated to 14 project types represented in this study by symbol B and their number represented by symbol A at years t (i.e., the year of event), t−1 (one year before), and t−2 (two years before). The 14 project types include: 1. Commerce and Industry; 2. Community Development; 3. Education; 4. Emergency Assistance; 5. Energy; 6. Environment; 7. Gender; 8. Governance; 9. Health; 10. Security; 11. Transport; 12. Water and Sanitation; 13. Agriculture; and 14. Capacity Building. Apart from these project types, other input variables are Urban male population density, Urban female population density, Rural male population density, Rural female population density, Number of $killed_{t-1}$, Number of $wounded_{t-1}$, Number of $hijacked_{t-1}$ and Number of $events_{t-1}$, where subscript t-1 represents the previous month. For example, A1(t-2) means the number of projects regarding Commerce and Industry at two years before. B14(t-1) means the sum of budget of the project type regarding Capacity Building at one year before. Number of $killed_{t-1}$ means the number of killed at one month before.

4. Simulation and Discussion of the Results

A forward stepwise least squares regression in SAS Enterprise Miner (SAS EM) was applied to the whole data set to select a subset of variables from all the variables according to R-square values. For the Dead, Wounded and Events data sets, minimum R-square was set to 0.005. However, for the Hijacked data set, allocated budget information for different projects such which is a sparse data including target variable with many 0's, minimum R-square was set to 0.0005. After computing the square correlation coefficient, between 6 and 14 variables were retained depending on the category of the adverse event (Table 1).

Massive Online Analysis (MOA) [4], used in this study, is an open source platform for data stream machine learning, which includes a lot of classification and regression algorithms. In the study, AMRules, TargetMean and FIMTDD were selected as the three data stream regression algorithms and the traditional linear regression was chosen as a benchmark. We also ran computer simulation for more advanced machine learning algorithms such as support vector machines (SVM), bagging, and boosting. However, SVM could not run in a reasonable time, whereas bagging and boosting produced results comparable to linear regression. Due to space constraints, those results for bagging and boosting are not presented in this study. In the simulations using the whole data set, the sample frequency was set to 200, and the window size was set to 100, 500 and 1000 respectively.

Table 1 lists the output variables and input variables for the four data sets. Among others, the input variables always include project number representing Education (A3). Region, a nominal variable representing region, is included in the data sets. Except for the Region variable, other input variables are numeric. Because the AMRules model does not support the input variables with nominal values, Region is removed from the data set when AMRules model is used. When we use the linear regression, TargetMean and FIMTDD models, the Region variable is retained.

Table 1. Output and Input Variables

Output	Input
Dead	Region, A3(t-2), A3(t-1), Dead(t-1), Wounded(t-1), Event_Nu(t-1)
Wounded	Region, A3(t-2), A3(t-1), Urban Male Population Density, Urban Female Population Density, Wounded(t-1), Event_Nu(t-1)
Hijacked	Region, B5(t-2), A2(t-2), A3(t-2), A6(t-2), A12(t-2), B6(t-1), B14(t-1), A3(t-1), A9(t-1), Rural Male Population Density, Wounded(t-1), Hijacked(t-1), Event_Nu(t-1)
Events	Data_year, Region, A3(t-1), A5(t), Urban Male Population Density, Urban Female Population Density, Event_Nu(t-1)

Table 2. MAE and RMSE Results of Linear Regression for 10 folds and MAE Reported in [8]

	Linear Regression		Previous Study [8]
Output	**MAE**	**RMSE**	**MAE**
Dead	0.5633	1.9296	2.0177
Wounded	0.8934	3.7623	4.3022
Hijacked	0.1594	1.2239	0.5051
Events	0.3215	0.7845	0.9352

The traditional linear regression is used for the four data sets to establish a benchmark. In the simulation, the ten-fold cross validation technique is applied. Table 2 shows the MAE and RMSE results, which are taken as the baseline to be compared with the three data stream algorithms. The maximum values of MAE and RMSE are 0.8934 and 3.7623 for Wounded. The minimum values of MAE and RMSE are 0.1594 for Hijacked and 0.7845 for Events. The MAE values reported in [8] are several times larger than the MAE values depicted in Tables 2 and 3.

Table 3 shows the MAE and RMSE results using AMRules, TargetMean and FIMTDD for the Dead, Wounded, Hijacked and Events. Different window sizes 100, 500, and 1000 are set when the data stream methods are used. For example, for Dead, the MAE and RMSE of TargetMean are 0.1277 and 0.5457 when window size is 100, the MAE and RMSE are 0.1338 and 0.7571 when window size is 500, and the MAE and RMSE are 0.1346 and 0.8132 when window size is 1000. When the window size is larger, the values of MAE and RMSE are slightly worse. But the AMRules model is different. When the window size is larger, its MAE values are slightly better, and RMSE values are worse. The AMRules algorithm uses the regression models by selecting a lower mean absolute error between the mean values of the target attribute and a linear combination of the attributes, and TargetMean uses only the model with the mean values of the target attribute. When window size is larger, the mean values of the target attribute could increase, the MAE values for TargetMean will be worse.

Compared with the MAE and RMSE of the linear regression model, the MAE and RMSE of the TargetMean model are better for all the data sets. For example, for Dead, the MAE and RMSE of the linear regression are 0.5633 and 1.9296, and the MAE and RMSE of the TargetMean model with window size 1000 are 0.1346 and 0.8132. For Hijacked, the MAE and RMSE of the linear regression are 0.1594 and 1.2239, and the MAE and RMSE of the TargetMean model with window size 1000 are 0.0464 and 0.4852. For AMRules, some results are better than those of the linear regression, and some are worse than the linear regression. For FIMTDD, most results are better than the results of the linear regression, except for the Hijacked data set. Hijacked is a very sparse dataset, in which the vast majority of values of the output variable are 0's. Among the three data stream algorithms, the TargetMean model is the best.

Table 3. MAE and RMSE Results for the Three Data Stream Methods

Output	Window Size	Measurement	AMRules	TargetMean	FIMTDD
Dead	100	MAE	0.9242	0.1277	0.2353
		RMSE	1.9545	0.5457	0.8116
	500	MAE	0.8885	0.1338	0.2490
		RMSE	2.0595	0.7571	1.0475
	1000	MAE	0.8770	0.1346	0.2524
		RMSE	2.0612	0.8132	1.1158
Wounded	100	MAE	0.7227	0.2131	0.4575
		RMSE	2.5740	0.9301	1.6633
	500	MAE	0.7174	0.2608	0.5909
		RMSE	3.2696	1.9282	3.5496
	1000	MAE	0.7122	0.2716	0.6357
		RMSE	3.4409	2.2506	4.2512
Hijacked	100	MAE	0.1381	0.0620	0.2639
		RMSE	0.7116	0.4005	0.9079
	500	MAE	0.1198	0.0504	9.1203
		RMSE	0.8142	0.4647	197.7790
	1000	MAE	0.1127	0.0464	11.3323
		RMSE	0.8392	0.4852	348.6402
Events	100	MAE	0.3088	0.1137	0.1203
		RMSE	0.7230	0.3542	0.3549
	500	MAE	0.2991	0.1243	0.1351
		RMSE	0.7188	0.4478	0.4156
	1000	MAE	0.2933	0.1253	0.1378
		RMSE	0.7091	0.4617	0.4316

Figure 2 shows the MAE values of Dead, Wounded, Hijacked and Events using the linear regression, AMRules, TargetMean and FIMTDD when window size is set to 100. In the figure, we can see that the performance of TargetMean model is the best for all datasets. The performance of FIMTDD is the second. The figure is consistent with the results in Table 2. TargetMean and FIMTDD are better than the linear regression and AMRules for the four data sets.

Figure 3 shows the MAE values of Dead, Figure 3 shows the MAE values of Dead, Wounded, Hijacked and Events using TargetMean when window size is set to 100. The MAE values for Dead and Events are lower than those for Wounded. For Hijacked, in most points, the MAE values are very low, but in some observations between 20000 and 25000, the MAE values are very high. The results could be caused by the sparse data.

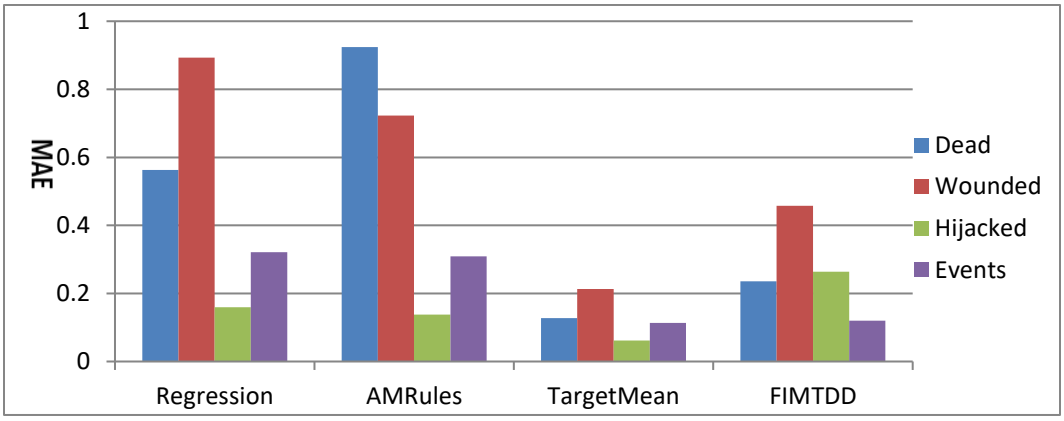

Figure 2. MAE Results of Dead, Wounded, Hijacked and Events Using the Four Methods for Window Size 100

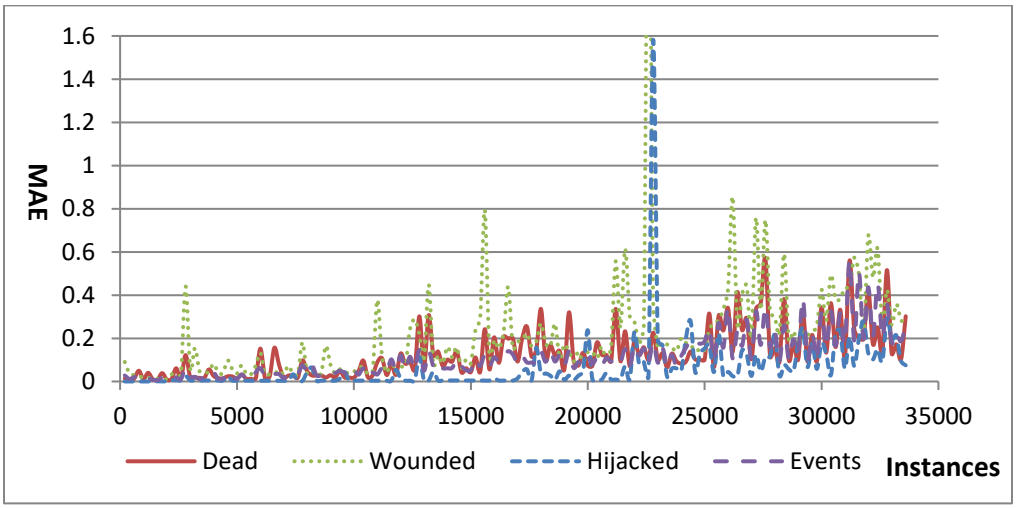

Figure 3. MAE Results of Dead, Wounded, Hijacked and Events Using TargetMean for Window Size 100

Figure 4 shows the MAE results of Dead, Wounded, Hijacked and Events using TargetMean when window size is set to 1000. The curves are smoother than those in Figure 3. The MAE results of Dead, Hijacked and Events are better than those of Wounded. The MAE values of Hijacked are lowest. When window size is set to 1000, the big fluctuations of errors disappeared. In the simulations in the seven regions, we only show the MAE and RMSE when window size is set to 1000. In addition, Figures 3 and 4 all show that the MAE values increase with the instances. The reason could be that with instances, the percentage of adverse events occurrence increases, the MAE values also increase. In Figure 4, for Dead, Hijacked, and Events, the MAE values are lower for before 12000 observation, which is under 0.1. The MAE values are over 0.1 after 15000 observation for Dead, Wounded, and Events. The MAE values are highest for Wounded and the MAE values are lowest for Hijacked.

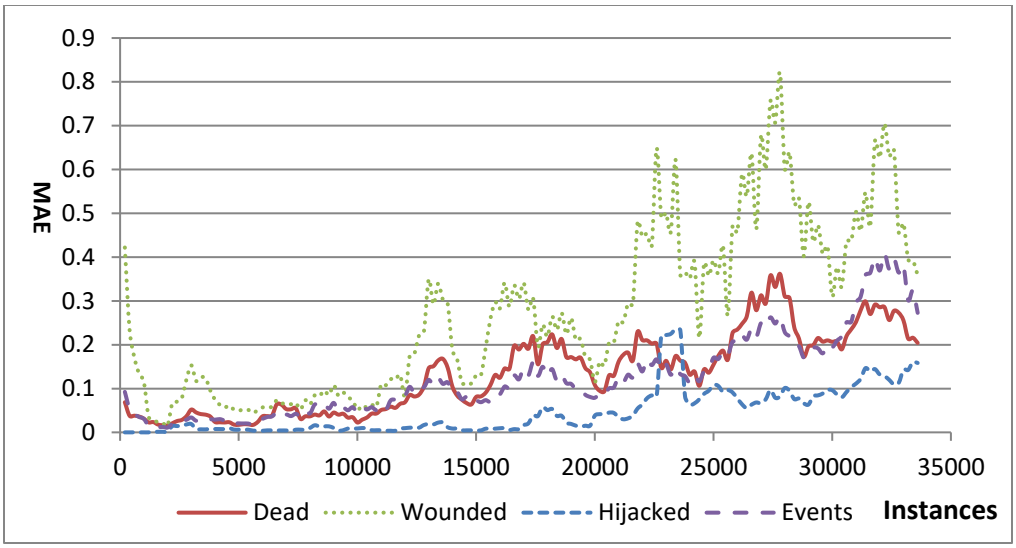

Figure 4. MAE Results of Dead, Wounded, Hijacked and Events Using TargetMean for Window Size 1000

The same variable selection method, a forward stepwise least squares regression in SAS EM, was used for rejecting insignificant variables for the seven regions: Central, Eastern, Northeastern, Northwestern, Southeastern, Southwestern and Western. Depending on the region and the category of the adverse event, between 5 and 20 variables were retained. Every region has the four data sets: Dead, Wounded, Hijacked and Events. The input variables are the ones selected by evaluating the R-square. After the selection of the significant variables, AMRules, TargetMean and FIMTDD are used for the data set of the seven regions. In the simulations, all the window size values are set to 1000, and the sample frequency is set to 30. The reason for that is that the size of every data set by region is about 1/7 of the whole data set.

Table 4 shows the MAE and RMSE values using the linear regression, AMRules, TargetMean, FIMTDD, and MAE from [8] for Dead, Wounded, Hijacked, and Events in the seven regions. The MAE values reported in [8] are very high. The MAE and RMSE values of AMRules and TargetMean are better than those of linear regression and FIMTDD for almost all the four data sets in the seven regions. For example, in Central, for Dead, the MAE and RMSE values for AMRules are 0.2984 and 1.5783, the MAE and RMSE values for TargetMean are 0.3146 and 1.5828, the MAE and RMSE values for Linear Regression are 0.4764 and 1.6207, and the MAE and RMSE values for FIMTDD are 0.7542 and 2.3625. The performance of AMRules model and that of TargetMean are very close, and in some region AMRules has the best performance, and in some region TargetMean is the best.

Among the seven regions, the MAE and RMSE values of the four methods in Northwestern for Dead, Wounded, Hijacked and Events are lowest. For example, in Northwestern, for Dead, the MAE and RMSE values for Linear Regression are 0.1355 and 0.5378, the MAE and RMSE values for AMRules are 0.0789 and 0.3516, the MAE and RMSE values for TargetMean are 0.0703 and 0.3510, and the MAE and RMSE values for FIMTDD 0.1103 and 0.4312. The MAE and RMSE values of the four methods in Southwestern for Dead, Wounded, Hijacked and Events are highest. In the Southwestern region, for Dead, the MAE and RMSE values for Linear Regression are 1.4060 and 3.6975, the MAE and RMSE values for AMRules are 1.2925 and 3.4602, the MAE and RMSE values for TargetMean are 1.3013 and 3.4988, and the MAE and RMSE values for FIMTDD 1.7959 and 5.7678.

Figure 5 shows the histogram of MAE values of Dead, Wounded, Hijacked and Events for the seven regions using AMRules when window size is set to 1000. One can see that in the Northwestern region, the MAE values are lowest. In the Southwestern region, those are highest. These results are consistent with those of Table 4.

Figure 6 shows the MAE results of Dead using Linear Regression, AMRules, TargetMean and FIMTDD in the seven regions. One can see that AMRules and TargetMean are very close in some regions and AMRules models have better performance than TargetMean. In some regions, TargetMean models are better. They both have better performance than Linear Regression and FIMTDD.

Table 4. MAE and RMSE Results Using the Four Methods and MAE Reported in [8] for Seven Regions

Region	Output Variable	Error Measures	Regression	AMRules	TargetMean	FIMTDD	Previous Study [8]
Central	Dead	MAE	0.4764	0.2984	0.3146	0.7542	1.1566
		RMSE	1.6207	1.5783	1.5828	2.3625	
	Wounded	MAE	1.4620	0.2345	0.2569	0.4313	4.9301
		RMSE	5.4126	0.7153	0.7367	2.7501	
	Hijacked	MAE	0.1170	0.0554	0.0560	0.1453	0.3982
		RMSE	0.5323	0.3690	0.3689	0.5603	
	Events	MAE	0.2526	0.2345	0.2569	0.4313	0.9763
		RMSE	0.6645	0.7153	0.7367	2.7501	
Eastern	Dead	MAE	0.3567	0.2370	0.2321	0.3770	0.7458
		RMSE	1.0726	0.8951	0.8927	1.0390	
	Wounded	MAE	0.9408	0.5622	0.5602	1.2136	2.6807
		RMSE	4.8827	3.5716	3.5724	6.0860	
	Hijacked	MAE	0.1633	0.0771	0.0772	0.1566	0.4412
		RMSE	0.7468	0.5453	0.5453	1.2132	
	Events	MAE	0.3306	0.2400	0.2393	0.3547	0.7168
		RMSE	0.6210	0.5165	0.5141	1.1407	
North Eastern	Dead	MAE	0.2342	0.1162	0.1181	0.3322	0.6238
		RMSE	1.3341	0.8917	0.8916	1.2776	
	Wounded	MAE	0.3870	0.1701	0.1801	0.4443	1.0443
		RMSE	2.0839	1.3371	1.3362	2.2677	
	Hijacked	MAE	0.0740	0.0223	0.0228	0.0448	0.2356
		RMSE	0.5965	0.2390	0.2390	0.2411	
	Events	MAE	0.1745	0.1061	0.1043	0.1412	0.4827
		RMSE	0.4507	0.3308	0.3293	0.3422	
South Eastern	Dead	MAE	0.5583	0.5068	0.5565	0.7566	1.5004
		RMSE	1.7053	1.7870	1.7799	2.7599	
	Wounded	MAE	0.8941	0.7153	0.7653	1.1047	2.3699
		RMSE	3.2313	3.0627	3.0366	5.0387	
	Hijacked	MAE	0.2005	0.1352	0.1404	0.2421	0.6444
		RMSE	1.2567	0.9351	0.9377	1.2836	
	Events	MAE	0.3961	0.4283	0.4602	0.4442	1.1770
		RMSE	1.0503	1.3462	1.2337	2.1692	
Western	Dead	MAE	0.4013	0.2819	0.2890	0.4362	1.3051
		RMSE	1.3484	1.1286	1.1286	1.6562	
	Wounded	MAE	0.4457	0.3352	0.3355	0.4757	1.4825
		RMSE	1.8326	1.6970	1.6947	1.9034	
	Hijacked	MAE	0.2932	0.1331	0.1372	0.3192	0.5161
		RMSE	2.5531	1.5024	1.5025	1.5984	
	Events	MAE	0.2873	0.2134	0.2138	0.2457	0.9506
		RMSE	0.6389	0.5532	0.5649	0.5740	
North Western	Dead	MAE	0.1355	0.0789	0.0703	0.1103	0.4419
		RMSE	0.5378	0.3516	0.3510	0.4312	
	Wounded	MAE	0.1980	0.1025	0.1039	0.2560	0.5694
		RMSE	1.1131	0.6438	0.6437	0.8962	
	Hijacked	MAE	0.0825	0.0159	0.0158	0.0516	0.3600
		RMSE	0.4368	0.1252	0.1252	0.3532	
	Events	MAE	0.1637	0.0951	0.0914	0.1696	0.5362
		RMSE	0.4430	0.2834	0.2833	0.8877	
South Western	Dead	MAE	1.4060	1.2925	1.3013	1.7959	2.0278
		RMSE	3.6975	3.4602	3.4988	5.7678	
	Wounded	MAE	1.6289	1.4242	1.4728	2.8054	2.0806
		RMSE	4.7087	4.7739	4.7720	23.9269	
	Hijacked	MAE	0.2164	0.1397	0.1477	0.5222	0.5926
		RMSE	1.2355	0.9379	0.9382	7.4381	
	Events	MAE	0.5509	0.4750	0.5701	0.5731	1.2946
		RMSE	1.1489	1.0200	1.1318	1.4677	

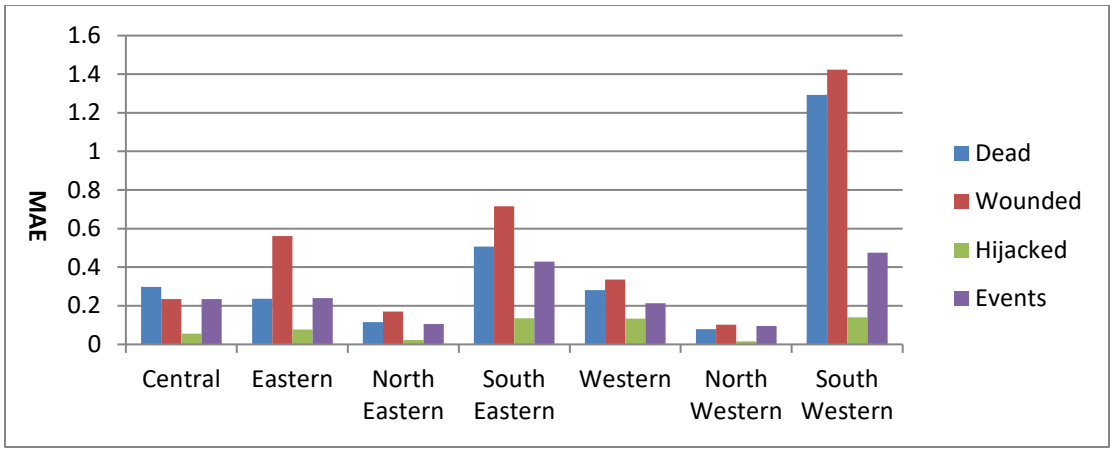

Figure 5. MAE Results of Dead, Wounded, Hijacked and Events for Seven Regions Using AMRules for Window Size 1000

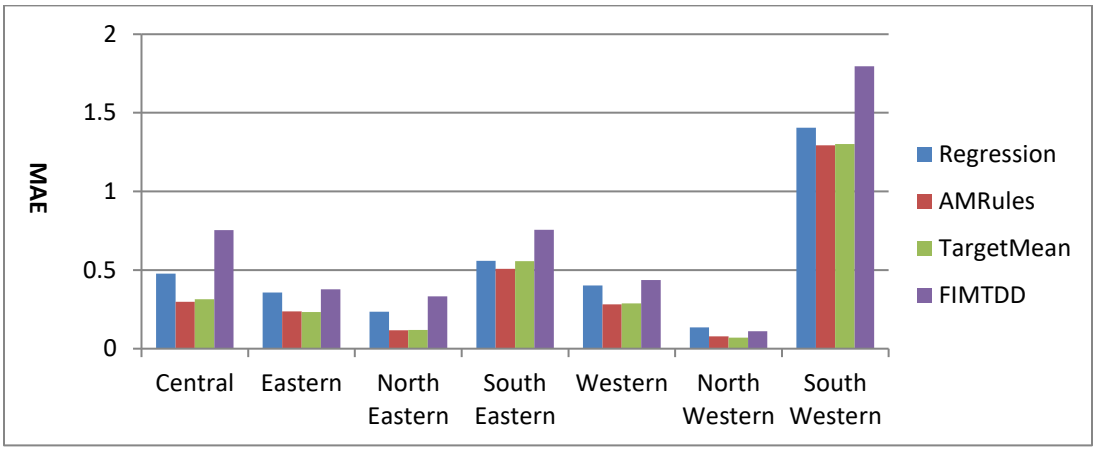

Figure 6. MAE Results of Dead Using the Four Methods for Window Size 1000

5. Conclusions

Currently most studies use traditional linear regression models or machine learning models for predicting the adverse events in an active war theater. The performance of these models is rather poor and can be improved. In our study, we use a new approach, based on data stream methods, to improve the prediction results. First, a forward stepwise least squares regression was applied to select the significant variables from over 100 input attributes for the four data sets: Dead, Wounded, Hijacked and Events. Then traditional linear regression, and three data stream regression algorithms, AMRules, TargetMean and FIMTDD were used on the four data sets. The results show that the data stream algorithm TargetMean has the best performance in the four data sets and its MAE values are the lowest. FIMTDD also has a fairly good performance in most scenarios, but for Hijacked, a sparse dataset, it has the worse results. This shows that FIMTDD may not be suitable for sparse data. The AMRules method does not show a good performance. It may be caused by the fact that we removed a nominal attribute: Region, since AMRules cannot support nominal attributes. When window size is set to 100, 500 and 1000, the MAE values curves become smoother for window size 1000.

With more instances, the percentage of adverse events occurrence increases, which could be the reason that the MAE values increases. Data stream methods show the MAE and RMSE values as new data arrive, thus allowing us to take snapshots for the prediction model at any time to see the changing performance of the model. This is in contrast to linear regression, where one can only see the final mean MAE and RMSE values.

In the analysis by region, significant variables are selected using the forward stepwise least squares regression. Similarly, linear regression and the three data stream methods are used on the four data sets, Dead, Wounded, Hijacked and Events in the seven regions, Central, Eastern, Northeastern, Northwestern,

Southeastern, Southwestern and Western. The MAE and RMSE results of data stream methods AMRules and TargetMean have better performance than traditional linear regression in all the four data sets in the seven regions. The MAE values are lowest in North Western, and those are highest in Southwestern. From the past studies, we can find that the percentage of adverse events occurrence is the lowest in the Northwestern and the percentage of adverse events occurrence is highest in the Southwestern. In the simulations by region, one can find that the performances of AMRules and TargetMean are very close. The improvement of performance of AMRules can be explained by the fact that in the analysis by region there is no longer a nominal attribute (i.e., Region). For the entire country and for seven regions MAE values reported in one of the previous studies [8] are much worse than those presented in this study.

Our results show that data stream methods demonstrate their advantages in improving the performance and providing a dynamic observing window for the models. In the future, it may be interesting to observe the performance of soft computing algorithms in the data stream setting, understand key factors of influence on the adverse events and find a general framework for adverse events not only used in an active war theater but also in other areas such as healthcare and aviation safety.

Acknowledgment

This study was supported in part by Grant no. 10523339, Complex Systems Engineering for Rapid Computational Socio-Cultural Network Analysis, from the Office of Naval Research (ONR) awarded to Waldemar Karwowski at University of Central Florida.

References

[1] Almeida E., Ferreira C., and Gama J. (2013). Adaptive Model Rules from Data Streams. *Lecture Notes in Computer Science*, 8188:480-492.

[2] Bhattacharjee, Y. (2007). Pentagon Asks Academics for Help in Understanding its Enemies. *Science*, 316 (5824), 534-535

[3] Bifet, A. and Kirkby, R. (2009) Data Stream Mining: A Practical Approach https://www.cs.waikato.ac.nz/~abifet/MOA/StreamMining.pdf

[4] Bifet, A., Kirkby, R., Kranen P., and Reutemann P. (2012). Massive Online Analysis Manual https://sourceforge.net/projects/moa-datastream/files/documentation/Manual.pdf

[5] Bifet A., Gavaldà R., and Pfahringer B. (2018). Machine Learning for Data Streams with Practical Examples in MOA.

[6] Çakıt, E., Karwowski, W., Bozkurt, H., Ahram, T., Thompson, W., Mikusinski. P., and Lee, G. (2014). Investigating the Relationship between Adverse Events and Infrastructure Development in an Active War Theater Using Soft Computing Techniques. *Applied Soft Computing*. 25, 204-214.

[7] Cakit, E., and Karwowski, W. (2015a). Assessing the Relationship between Economic Factors and Adverse Events in an Active War Theater Using Fuzzy Inference System Approach. *International Journal of Machine Learning and Computing*. 5(3), 252-257.

[8] Çakıt, E., and Karwowski, W. (2015b). Fuzzy Inference Modelling with the Help of Fuzzy Clustering for Predicting the Occurrence of Adverse Events in an Active Theater of War. *Applied Artificial Intelligence*, 29, 945-961.

[9] Çakıt, E., & Karwowski, W. (2017a). 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.

[10] Çakıt, E., and Karwowski, W. (2017b). Understanding the Social and Economic Factors Affecting Adverse Events in an Active Theater of War: A Neural Network Approach. *In International Conference on Applied Human Factors and Ergonomics* (pp. 215-223). Springer, Cham.

[11] Çakıt, E., and Karwowski, W. (2018). A Fuzzy Overlay Model for Mapping Adverse Event Risk in an Active War Theater. *Journal of Experimental & Theoretical Artificial Intelligence*, 1-10. https://doi.org/10.1080/0952813X.2018.1467494.

[12] Duarte, J., Gama, J., and Bifet, A. (2016). Adaptive Models Rules from High-Speed Data Streams. *ACM Transactions on Knowledge Discovery from Data.* 10(3): 30:1-30:22.

[13] Ikonomovska, E., Jo˜ao Gama, and Dzeroski, S. (2011). Learning Model Trees from Evolving Data Streams. Data Min. Knowl. Discov., 23(1):128–168.

[14] HSCB Modeling Program. (2009). Available via http://www.dtic.mil/biosys/docs/HSCB news-spring-2009.pdf.

[15] Shi, D., Guan, J., Zurada, J., and Manikas, A. (2017). A Data-mining Approach to Identification of Risk Factors in Safety Management Systems. *Journal of Management Information Systems*, 34(4), 1054-10 81.