Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 19 Issue 2 Version 1.0 Year 2019
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Analyzing Political Opinions and Prediction of Voting Patterns in the US Election with Data Mining Approaches

By Md. Sohel Ahammed, Md. Nahid Newaz & Arunavo Dey

Bangladesh University of Business & Technology (BUBT)

Abstract- Data is the precious resources. Data contains the useful patterns which provide the crucial information about the prediction of what is going to be happened in the next. In this paper, we aim to identify the political preferences and tendency of the US populations using classification and data mining techniques. To provide the usefulness of proposed model we analyze the electoral data sets in US election obtained from the official website which contains the information about 1984 United States Congressional voting records. This paper shows the classification techniques that can be used to predicting voting patterns in the US House of Representatives and shows the close correspondence between election results and extracted opinion. This paper also shows the political support of the voters and prediction the characteristics of the voter with their political tendency.

GJCST-C Classification: *H.2.8*

Strictly as per the compliance and regulations of:

Analyzing Political Opinions and Prediction of Voting Patterns in the US Election with Data Mining Approaches

Md. Sohel Ahammed[α], Md. Nahid Newaz[σ] & Arunavo Dey[ρ]

Abstract- Data is the precious resources. Data contains the useful patterns which provide the crucial information about the prediction of what is going to be happened in the next. In this paper, we aim to identify the political preferences and tendency of the US populations using classification and data mining techniques. To provide the usefulness of proposed model we analyze the electoral data sets in US election obtained from the official website which contains the information about 1984 United States Congressional voting records. This paper shows the classification techniques that can be used to predicting voting patterns in the US House of Representatives and shows the close correspondence between election results and extracted opinion. This paper also shows the political support of the voters and prediction the characteristics of the voter with their political tendency.

I. Introduction

Election is important because it allows the electorate to decide who's going to make decision for their country for the next couple of years. But this election can be forecasted with a reasonable accuracy. Forecasting election using small polling system is very common approach but this often do not produce reasonable accuracy.

Data mining is a process that examines large preexisting databases in order to generate new information. There are also various works that uses data mining approaches to predict various types of results such as weather forecasting, sports result prediction, future buying decision prediction, etc. But there are very few works that uses data mining approaches to predict voting patterns on election. In this work, we uses data mining approaches to predict voting patterns in USA election. For this study we uses data preprocessing for removing missing value, identifying best attributes and removing duplicate values. We split the dataset into training datasets and test datasets. Then we applied four algorithms Tree J48, Naïve Bayes Classifier, Trees Random Forest and Rules zero or Classifier for predicting voting patterns and also compares the results of those model and finds the best models from those models.

Author α σ ρ: Dept. of Computer Science & Engineering, Bangladesh University of Business & Technology (BUBT), Mirpur-2, Dhaka-1216, Bangladesh. e-mails: sohel.ruet10@gmail.com, md.nahidnewaz@gmail.com, arunavo071@gmail.com

II. Related Works

Gregg R. Murray and Anthony Scime uses data mining approaches to predict individual voting behavior including abstention with the intent of segmenting the electorate in useful and meaningful ways [1]. Gregg R. Murray, Chris Riley, and Anthony Scime, in another study, uses iterative expert data mining to build a likely voter model for presidential election in USA [2]. Bae, Jung-Hwan, Ji-Eun, Song, Min uses Twitter data for predicting trends in South Korea Presidential Election by Text Mining techniques [3]. Tariq Mahmood, TasmiyahIqbal, Farnaz Amin, WaheedaLohanna, Atika Mustafa uses Twitter data to predict 2013 Pakistan Election winner [4].

III. Data Preprocessing

I. *Handling with Missing Attributes:* In this section, we uses the technique of replacing missing values with mean, median or mode. We uses this approach because it is better approach when the dataset is small and it can prevent data loss.

II. *Removing Duplicates:* We used WEKA tools for removing duplicates from the datasets. We used *Remove Duplicates ()* function in WEKA for removing duplicates.

III. *Best Attributes Selection:* We used **Gain Ratio Attribute Eval** which evaluates the worth of an attribute by measuring the gain ratio with respect to the class and **Ranker** which Ranks attributes by their individual evaluations. The top 12 attributes from the whole dataset according to rank from the attributes are presented in Figure 1.

Table 1: Selection of 12 best Attribute from "**Ranker** and **Gain Ratio Attribute Eval**" method

Accuracy	Ranked Number	Attribute Name
0.7366801	4	physician-fee-freeze
0.436943	3	adoption-of-the-budget-resolution
0.3936309	5	el-Salvador-aid
0.3522504	8	aid-to-Nicaraguan-contras
0.3416423	12	education-spending
0.3104985	14	Crime
0.2963994	9	mx-missile
0.2213305	13	superfund-right-to-sue
0.2191455	15	Duty-free-exports
0.1929617	7	anti-satellite-test-ban
0.1510354	6	religious-groups-in-schools -Africa
0.1470019	1	Handicapped-infants

IV. Experimental Methodology

We used 4 algorithms and 8 models (2 models for each algorithm) to predict the voting pattern in the US election. We then analyse and compare the results of those models and finds the best models with most accuracy. The algorithms which are applied for generating models are given below.

i. Trees J48
ii. Naive Bayes classifier
iii. Trees RandomForest
iv. Rules ZeroOR Classifier

a) Trees J48

We used **Model 1** for training dataset and **Model 2** for test dataset evaluation.

Evaluation of **Model 1** Training dataset is given below:

Table 2: Evaluation on J48 Training Data (**Model 1**)

Correctly Classified Instances	421	96.7816%
Incorrectly Classified Instances	14	3.2184%
Kappa statistics	0.9324	
Mean Absolute Error	0.0582	
Root Mean Squared Error	0.1706	
Relative Absolute Error	12.2709%	
Root Relative Squared Error	35.0341%	
Total Number of Instances	435	

Table 3: **Model 1** Precision, Recall, F-measure rate according to Democrat class

TP Rate	FP Rate	Precision	Recal-l	F measures	MCC	Rock Area	PRC area	Class
0.966	0.030	0.981	0.966	0.974	0.933	0.975	0.973	democrat

Sensitivity & Specificity Calculation for Training Data (Model 1)

Formula of Sensitivity = TP/ (TP+FN)

Formula of Specificity = TN/ (TN+FP)

So **Sensitivity** = TP Rate = 0.966 & **Specificity** = 0.030

Evaluation of **Model 2** test dataset is given below

Table 4: Evaluation on J48 Test Data (**Model 2**)

Correctly Classified Instances	105	96.3303%
Incorrectly Classified Instances	4	3.6697%
Kappa statistics	0.921	
Mean Absolute Error	0.0619	
Root Mean Squared Error	0.1894	
Relative Absolute Error	13.2259%	
Root Relative Squared Error	39.4312%	
Total Number of Instances	109	

Table 5: **Model-2** Precision, Recall, F-measure rate according to Democrat class

TP Rate	FP Rate	Precision	Recall	F measures	MCC	Rock Area	PRC area	Class
0.957	0.026	0.985	0.957	0.971	0.922	0.969	0.972	democrat

Sensitivity & Specificity Calculation for Model 2

Sensitivity = TP Rate = 0.957 **& Specificity** = 0.026

b) *Naive Bayes classifier*

Evaluation on Training Data set: Naive Bayes classifier algorithm

Table 6: Evaluation of Naive Bayes Training Data (**Model 3**)

Correctly Classified Instances	395	90.8046%
Incorrectly Classified Instances	40	9.1954%
Kappa statistics	0.8094	
Mean Absolute Error	0.0965	
Root Mean Squared Error	0.2921	
Relative Absolute Error	20.34%	
Root Relative Squared Error	59.9863%	
Total Number of Instances	435	

Table 7: **MODEL-3** Precision, Recall, F-measure rate according to Democrat class

TP Rate	FP Rate	Precision	Recall	F measures	MCC	Rock Area	PRC area	Class
0.895	0.071	0.952	0.895	0.923	0.812	0.972	0.983	democrat

Sensitivity & Specificity Calculation for Training Data (Model 3)

So **Sensitivity** = TP Rate = 0.895 & **Specificity** = 0.071

Evaluation of **Model 4** test dataset is given below

Table 8: Evaluation on Naïve Bayes Test Data (**Model 4**)

Correctly Classified Instances	99	90.8257%
Incorrectly Classified Instances	10	9.1743%
Kappa statistics	0.8069	
Mean Absolute Error	0.0978	
Root Mean Squared Error	0.2934	
Relative Absolute Error	20.9083%	
Root Relative Squared Error	61.0861%	
Total Number of Instances	109	

Table 9: **Model-4** Precision, Recall, F-measure rate according to Democrat class

TP Rate	FP Rate	Precision	Recall	F measures	MCC	Rock Area	PRC area	Class
0.886	0.051	0.969	0.886	0.925	0.812	0.969	0.984	democrat

Sensitivity & Specificity Calculation for Model 4

Sensitivity = TP Rate = 0.886 **& Specificity** = 0.051

c) *Trees Random Forest*

Evaluation on Training Data set: Trees Random Forest algorithm

Table 10: Evaluation of Trees Random Forest Training Data (**Model 5**)

Correctly Classified Instances	427	98.1609%
Incorrectly Classified Instances	8	1.8391%
Kappa statistics	0.9613	
Mean Absolute Error	0.0376	
Root Mean Squared Error	0.1222	
Relative Absolute Error	7.9365%	
Root Relative Squared Error	25.0915%	
Total Number of Instances	435	

Table 11: **MODEL-5** Precision, Recall, F-measure rate according to Democrat class

TP Rate	FP Rate	Precision	Recall	F measures	MCC	Rock Area	PRC area	Class
0.981	0.018	0.989	0.981	0.985	0.961	0.998	0.999	democrat

Sensitivity & Specificity Calculation for Training Data (Model 5)

So **Sensitivity** = TP Rate = 0.981& **Specificity** = 0.018

Evaluation of **Model 6** test dataset is given below

Table 12: Evaluation on Trees Random Forest Test Data (**Model 6**)

Correctly Classified Instances	106	97.2477%
Incorrectly Classified Instances	03	2.7523%
Kappa statistics	0.9404	
Mean Absolute Error	0.0432	
Root Mean Squared Error	0.1508	
Relative Absolute Error	9.2437%	
Root Relative Squared Error	31.408%	
Total Number of Instances	109	

Table 13: **Model-6** Precision, Recall, F-measure rate according to Democrat class

TP Rate	FP Rate	Precision	Recall	F measures	MCC	Rock Area	PRC area	Class
0.971	0.026	0.986	0.971	0.978	0.941	0.996	0.997	democrat

Sensitivity & Specificity Calculation for Model 6

Sensitivity = TP Rate = 0.971 **& Specificity** = 0.026

d) Rules ZeroOR Classifier

Evaluation on Training Data set: Rules ZeroOR Classifier algorithm

Table 14: Evaluation of Rules ZeroOR Classifier Training Data (**Model 7**)

Correctly Classified Instances	267	61.3793%
Incorrectly Classified Instances	168	38.6207%
Kappa statistics	0	
Mean Absolute Error	0.4742	
Root Mean Squared Error	0.4869	
Relative Absolute Error	100%	
Root Relative Squared Error	100%	
Total Number of Instances	435	

Table 15: **MODEL-7** Precision, Recall, F-measure rate according to Democrat class

TP Rate	FP Rate	Precision	Recall	F measures	MCC	Rock Area	PRC area	Class
1.0	1.0	0.614	1.0	0.761	-	0.500	0.614	democrat

Sensitivity & Specificity Calculation for Training Data (Model 7)

So **Sensitivity** = TP Rate = 1.0 & **Specificity** = 1.0

Evaluation of **Model 8** test dataset is given below

Table 16: Evaluation on Rules ZeroOR Classifier Test Data (**Model 8**)

Correctly Classified Instances	70	64.2202%
Incorrectly Classified Instances	39	35.7798%
Kappa statistics	0	
Mean Absolute Error	0.4678	
Root Mean Squared Error	0.4802	
Relative Absolute Error	100%	
Root Relative Squared Error	100%	
Total Number of Instances	109	

Table 17: **Model-8** Precision, Recall, F-measure rate according to Democrat class

TP Rate	FP Rate	Precision	Recall	F measures	MCC	Rock Area	PRC area	Class
1.0	1.0	0.642	1.0	0.782	-	0.500	0.642	democrat

Sensitivity & Specificity Calculation for Model 8

Sensitivity = TP Rate = 1.0**& Specificity** = 1.0

V. Revaluation of the Best, Second Best and Third Best Model

Table 18: Comparison among Models to select best, 2nd best and 3rd best Model

Model	Accuracy	precision	recall	sensitivity	specificity	Rank
Model 1	96.7816%	0.981	0.966	0.966	0.030	2nd best
Model 2	96.3303%	0.985	0.957	0.957	0.026	3rd best
Model 3	90.8046%	0.952	0.895	0.895	0.071	
Model 4	90.8257%	0.969	0.886	0.886	0.051	
Model 5	98.1609%	0.989	0.985	0.981	0.018	Best
Model 6	97.2477%	0.986	0.978	0.971	0.026	
Model 7	61.3793%	0.614	1.00	1.00	1.00	
Model 8	64.2202%	0.642	1.00	1.00	1.00	

From the above table, the best model was identified based on the value of the parameters accuracy, precision, recall, sensitivity, and specificity. The higher the value of accuracy, precision, recall and (sensitivity> specificity), the higher the rank.

VI. Conclusion

Though there are lot of techniques and methods for predicting voting patterns, data mining is the most efficient and effective methods in this fields. In our study, we clearly found that among various data mining algorithms Trees Random Forest performs the best with 98.17% accuracy. In future, we will expand our research in most recent dataset for validating our findings with recent ones.

References Références Referencias

1. Gregg R. Murray and Anthony Scime, "Micro targeting and Electorate Segmentation: Data Mining the American National Election Studies", Journal of political marketing, volume 9, issue 3, 2010.
2. Greg R. Murray, Chris Riley, and Anthony Scime, "Pre-Election Polling: Identifying Likely Voters Using Iterative Expert Data Mining", Public opinion Quarterly, volume 73, issue 1, 2009.
3. Bae, Jung-Hwan, Ji-Eun, Song, Min, "Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques", Journal of intelligence and Information Systems, Volume 19, issue 3, 2013.
4. Tariq Mahmood, Tasmiyahlqbal, Farnaz Amin, WaheedaLohanna, Atika Mustafa, "Mining Twitter big data to predict 2013 Pakistan election winner", INMIC, IEEE, 2013.