E-Governance in Elections: Implementation of Efficient Decision Tree Algorithm to predict percentage of e-voting

S. Meenakshi,

Research Scholar, Manonmanim Sundaranar University,
Asst. Professor, PG Dept of IT,
Bhaktavatsalam Memorial College for Women,
Korattur, Chennai - 80.

Dr A. Murugan,

Associate Professor and Head,
PG and Research Dept of Comp. Science,
Dr.Ambedkar Govt. Arts College, Vyasarpadi,
Chennai- 39.

Abstract

One of the important strengths of Indian democracy is elections. All the eligible citizens are expected to come to the polling stations to elect their representatives to form the government. But the process is not that much simple in a big country like India and we are still unable to achieve hundred percent polling in our elections. Various measures have been taken by the government to achieve this without causing any damage to the fairness in the procedures involved in the electoral process because public confidence is the back bone of this grand system. In the recent past, many researchers and officials consider online voting as a valid and reliable choice to improve polling percentage. It is time to enter into the magical world of technology and apply its powerful tools to implement the change quickly and safely. The growth in the number of home computers and internet access is the ray of light that shows the direction of implementing online voting. This paper tries to find out the ways of improving poll percentage by an application of efficient decision tree algorithm to predict percentage of online voting. The proposed algorithm is compared with the already existing classifying algorithms and the accuracy value is predicted. Also the dataset is applied in the WEKA tool and the values are compared. The proposed algorithm suggests that polling percentage in elections in India can be definitely improved if online voting is introduced with all necessary factors of proper application.

Keywords: *Election, Polling percentage, online voting, Internet access, Decision tree algorithm, WEKA tool.*

1. Introduction

At first general elections in our country was conducted using ballot paper method. Later as a technological improvement, the Election Commission has introduced Electronic Voting Machines(EVMs) in elections from 1998[1]. In 2003 all state elections were conducted using EVMs. When EVMs were introduced, there were counter opinions about using them but now it is accepted. Many changes have been introduced to avoid the fall in number of the voters in each election. In this context, when we are living in the technology explosion period, it is natural to think about electing our political representative from our home or office over internet. It is true that people with home computers and internet access are increasing in our country. Hence it is the right time for us to consider and switch over to remote voting.

1.1 Why e-voting ?

Online voting refers to an election process in which people can cast their votes over the internet[2]. The major idea behind it is helping all the citizens of the country to participate in the election process irrespective of any difficulty regarding their place of residence, nature of job etc.,

The reasons for e-voting are many and to say a few of them:

- Time is saved
- Cost of the electoral process is lessened
- Creation of interest to cast vote is possible
- Faster counting of votes and declaration of results
- Accuracy in results since human error is removed
- Increased participation of voters
- Timely response to the taste of the technological power
- Audio ballot paper for blind voters
- Saving time of poll workers
- Reduced malpractice regarding vote selling

The concept that remote voting increases polling percentage is not simply hypothetical but scientific. To prove this, classification techniques are used to classify the dataset and finally decision tree algorithm is used to predict the online voting percentage. Let us discuss the classification techniques used in this research work in the following sections.

2. Classification Techniques

Classification is a data mining technique used to predict group membership for data instances[3]. It is also a form of data analysis that can be used to extract models to describe important data classes or to predict future data trends. They are widely used in data mining to classify data among various classes. They are used in different industry to identify the type and group to which a particular tuple belongs. There are many algorithms used for classification in data mining. The algorithms taken here for comparison are Naïve Bayes and KNN.

2.1 Naïve Bayes

The Naïve Bayes is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach we will come up with if we want to model a predictive modeling problem probabilistically. Naïve Bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value which is independent of all other attributes[4]. It is a strong assumption but results in a fast and effective method. The probability of a class value given for a value of a attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class. To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.

2.2 K– Nearest Neighbors

K-Nearest Neighbor is also known as Lazy learning classifier. K- Nearest neighbors is the simple algorithm that shows all available cases and classifies new cases based on a similarity measure. It is used in statistical estimation and pattern recognition[5]. It is a general technique to learn classification based on instance and do not have to develop an abstract model from the training data set. K-Nearest neighbor does not construct a classification model from data. It performs classification by matching the test instance with K training examples and decides its class based on the similarity to K nearest neighbors.

2.3 Decision Tree induction

Decision tree classification is the learning of decision trees from class labelled training tuples. A

decision tree is a flowchart like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. A decision tree represents a procedure for classifying categorical data based on their attributes[6]. It is also efficient for processing large amount of data and so it is often used in data mining application.

2.4 WEKA as a Data Mining Tool

In this research work, WEKA, a data mining tool is also used for classification techniques. WEKA is developed at the University of Waikato in New Zealand. WEKA stands for Waikato Environment of Knowledge Analysis. The software is written in Java. Data mining algorithms and tools used in WEKA are Associations, Attribute selection, Classifiers, Clusters and Preprocessing filters. The Lok Sabha election 2014 dataset is applied in WEKA tool for different algorithms and the results are found out.

3. Related Work

Alan Bermingham and Alan F.Smeaton[2011] from School of Computing, Dublin City University have worked on prediction of election results using twitter responses[7]. To investigate the potential to model political sentiment through mining of social media, they have taken Irish general election. This approach combines sentiment analysis using supervised learning and volume based measures.

Rohan Sampath[2014] worked on classification and regression approaches to predict United States elections[8]. He used LMS algorithm to predict margin of victory in Senate elections. He also used SVM classifier and Random Forest to predict the outcome of senate elections.

Barbara Ondrisek[2009] from Vienna University of Technology published a paper on e-voting system security optimization[9]. This paper presents the e-voting system security optimization method. It points out security flaws, shows security optimization potential and can be used to compare different election systems.

Vishnuprasad Nagadevara[2005] published a paper on building predictive models for election results in India– an application of classification trees and Neural Networks. Two techniques namely Classification trees and Artificial Neural Networks are used to build the predictive models for the Karnataka assembly elections[10]. Oversampling technique is used to eliminate the predictive biases. The overall accuracy of the predictive models varies from 90 to 98 percent.

4. Proposed Algorithm Description

The proposed algorithm is evaluated with Indian Lok sabha elections, 2014 datasets state wise with attributes such as name of the state, total voters, male voters, female voters, total male turnout, total female turnout, % of male turnout, % of female turnout, total % of turnout, total literacy, male literacy, female literacy, literacy growth state wise, total internet user etc., The proposed algorithm represents mathematical design to improve the classification accuracy. This approach can classify internet user growth and literacy growth because increase in the internet user and literacy is a strong point to adopt e-voting. It expresses the flowing performance matrix separately namely as Mean Absolute Error(MAE), Root Mean Square Error(RMSE) for given attribute and instances.

4.1 Mean Absolute Error (MAE)

In this section, proposed approach explains mathematical model in Equation (1) to measure how close predictions are to the eventual outcomes. The mean absolute error (MAE) is calculated as correctly classified malicious data with respect to overall data.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|f_i - y_i| \quad \text{-------} \quad (1)$$

The mean absolute error is an average of the absolute errors where f_i is the prediction and y_i the true value.

4.2 Root Mean Square Error (RMSE)

In this section, proposed method describes mathematical model to calculate the differences between value predicted by a model and the values actually observed from the environment that is being modeled. These individual differences are also called residuals. The root mean squared error (RMSE) is calculated as

$$RMSE = \sqrt{\frac{1}{n}\sum\left(\hat{y}_i - y_i\right)^2} \quad \text{----------} \quad (2)$$

$\hat{y}$ is a vector of n predictions and y is the vector of observed values corresponding to the inputs to the function which generated the predictions.

The main objective of the proposed algorithm is to reduce classification error and minimize retrieval process in comparison with available dataset. This system assists to reduce misclassification error pruning. For each leaf in a tree, it is possible to make a number of instances which are misclassified on a training set by propagating errors from. It can be compared with error rate if the leaf was changed by the most common class resulting from. To evaluate error reduction, sub tree leaf can be considered for pruning. The measurement of error reduction is done for all leaf in tree and one which have higher reduction error value. This process will continue with newly cutting tree until they are unable to find error reduction rate at any leaf. The error is compared with total instance and training set instances. The proposed classifier algorithm evaluates normalized information gain that results from selecting an attribute for splitting the data. The splitting process interrupts if all instances in a subset belong to the similar class. Here leaf nodes represent to select class. In this scenario, proposed classifier algorithm creates a decision node higher in the tree by using a excepted class value. The value is regenerated for every iteration with minimal tree. Ranges are measured based on value and the minimal attributes. This system reduced retrieval time and data classification times.

The classification is performed on the instances of training set and tree is formed. The main work of pruning is to decrease the classification error. It creates a binary tree. After tree formation, proposed classifier is applied on every tuple in dataset and predicts classification result for every tuple. After building of classifier, proposed algorithm does not consider missing value.

4.3 Pseudo code for proposed Classifier algorithm

Input : ARFF dataset

Output : Predict(P) the Precision, Recall(R), F-measure(F), Mean Absolute Error MAE,

Root Mean Square Error (RMSE)

Procedure : Load Datasets D

Verify the instance and attribute;

Check the base cases;

Find the features

Check the class label of dataset;

If class have multiple category

Perform tree formation;

Let us consider attribute A

Find normalized data for every attribute from splitting attribute;

Let A have highest normalized data

At every leaf, left tree should contain low value and right tree leaf high value;

Allow design the branch;

Remove unwanted branch (not helping to reach leaf node);

Predict the accuracy;

Predict classification error;

Display the P, R, F, MAE & RMSE;

Else

Display exception unable to predict the class

5. Results and Discussion

The following results tabulate the comparison of accuracy, MAE and RMSE values:

Table 1: Comparison of accuracy, MAE and RMSE values

Algorithm	Accuracy	MAE	RMSE
Naïve Bayes	31.43	0.61	0.73
KNN	48.57	0.51	0.70
Proposed Algorithm	77.14	0.35	0.42

Graphical representation of error values and accuracy for the same is shown below:

Fig : 1 Graph showing comparison of the error values of algorithms

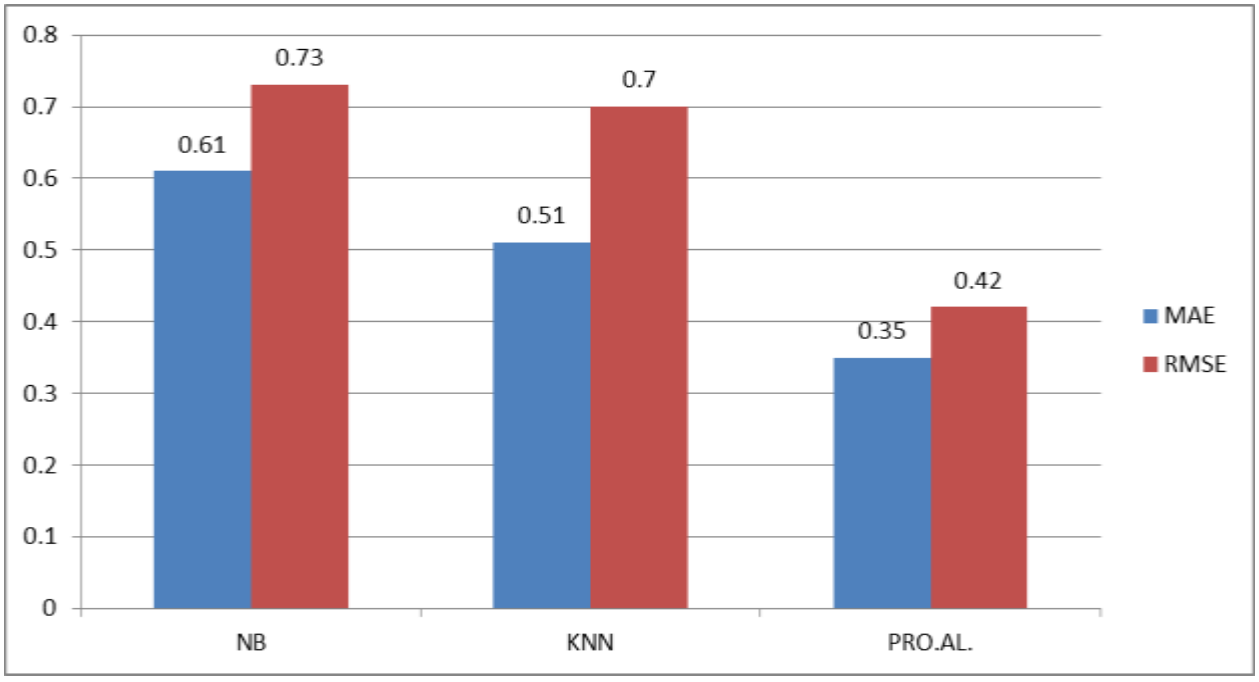
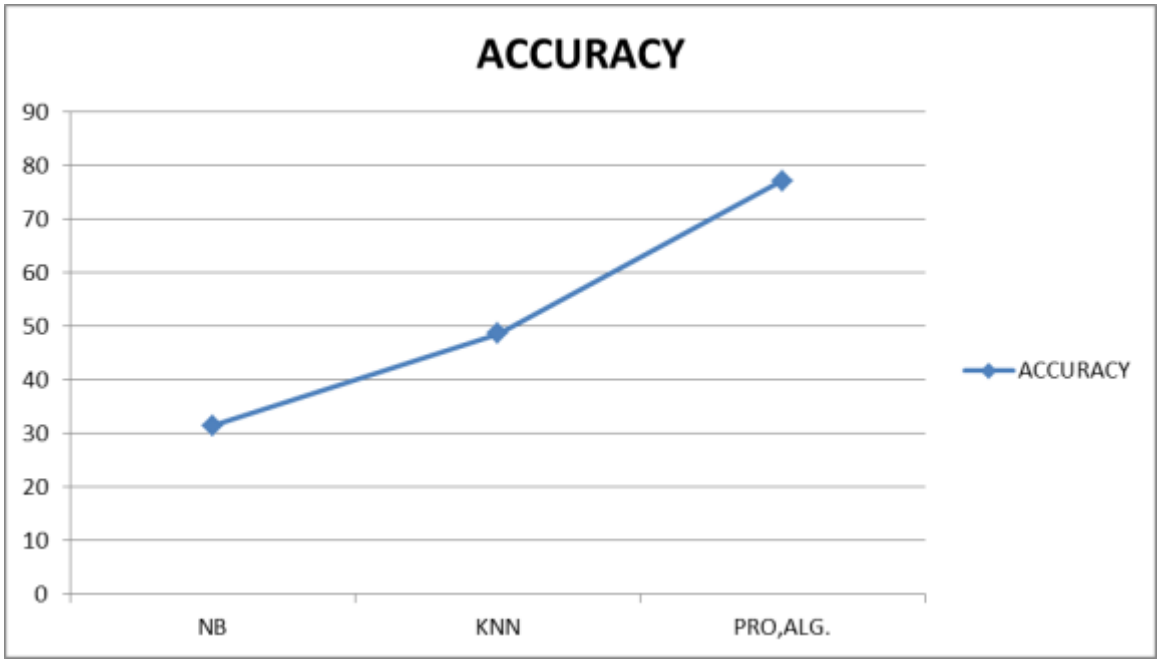


Fig 2: Graph showing the accuracy



5.1 Classification in Weka

The classification is based on supervised algorithms. Algorithms are applicable for the input data. The classify tab is also supported which shows the list of machine learning tools.[11] These tools in general operate on a classification algorithm and run it multiple times to manipulate algorithm parameters or input data weight to increase the accuracy of the classifier. The performance evaluation of proposed approach is also tested in WEKA tool with default parameter setting. For classification, 10 cross fold validation test is conducted to measure the precision, recall and F1 score for following classifier namely NB(Navie Bayes), and KNN(K– Nearest Neighbor). The table shows precision, recall and F1 score and ROC for classification of Lok Sabha elections, 2014 dataset.

Table 2: Comparison of P, R, F1 and ROC values of different algorithms from WEKA

Algorithm	Precision	Recall	F1	ROC
NB	0.486	0.314	0.363	0.38
KNN	0.556	0.486	0.517	0.35
PRO.ALG.	0.595	0.771	0.672	0.354

6. Conclusion

When we look at the voters turn out in Polling place method from the year 1952, we come to know that it was not more than 65%. But the above results clearly suggest adopting online voting method which increases the voting turn out by a large margin. The benefits of it are already well known and widely discussed. Even small countries like Portugal are implementing both polling place e-voting and remote voting for the better execution of election process. Especially in Estonia, remote voting has been very successful. Estonia is using digital signature and smart ID cards for identification purpose. We can implement a remote voting in our country with the latest aspects involving sharp cyber technologies to avoid hacking, virus and stranger interference. Aadhar card as the unique identification document serving many purposes such as address proof and identity of a person

is given to all the citizens of our country. It involves iris recognition technology which will not give any space for malpractice because iris is unique for every human being. To strengthen the concept of online voting, Aadhar card can be attached to the process of elections in the country.

References

[1] Rubin, Avin. 'Security considerations for remote electronic Voting over the internet'. http://avirubin..com/e-voting-security.html

[2]. www.eods.eu/library/IDEA/

[3]. Gupta, Megha. 'Classification Techniques Analysis'. NCCI 2010– National Conference on Computational Instrumentation, CSIO Chandigarh, India, 19-20 March 2010.

[4]. http:/machinelearningmastery.com/naivebayes-users

[5]. http://www.saedsayed.com/k-nearest-neighbors.html

[6]. www.csun.edu/twang/595DM/slides/week.pdf

[7]. Bermingham, Adam. 'On using Twitter to monitor political sentiment and predict election results.' Proceedings of the workshop on Sentiment analysis where AI meets psychology(SAAIP), IJCNLP 2011, pp 2-10, Thailand, November 13,2011.

[8]. Sampath, Rohan. 'Classification and Regression approaches to predicting United States Elections'. http://c8229.stanford.edu/proj2014/Rohansampath.

[9]. Ondrisek, Barbara. 'E-voting system security optimization'. Proceedings of the 42nd Hawaii International Conference on System sciences– 2009.

[10]. Nagadevara, Vishnuprasad. 'Building predictive models for election results in India- an application of Classification trees and Neural networks'. *Journal of Academy of Business and Economics.* Vol 5, Issue 3, March 2005.

[11]. Shrivatsava, Manish Kumar. 'Exploring Data Mining Classification Techniques', *International Journal of Engineering Research and Technology(IJERT),* Vol 2, Issue 6, June 2013.