ANALYSIS OF SUPERVISED AND UNSUPERVISED LEARNING ALGORITHM TO PREDICT ELECTION RESULTS

ADIT KUMAR

ABSTRACT

Three supervised (k-nearest, perceptron, ID3 decision tree) and two unsupervised (k-means and agglomerative nesting) learning algorithms are analyzed on the dataset of election results and their accuracy and performance is compared

1. DESCRIPTION

In supervised learning algorithm the system is first trained with dataset and tested while in unsupervised learning algorithm, the dataset are clustered into different classes. The following procedures are adopted in each algorithms-

K-means

Data is first normalized by recalculating the attribute values of county relative to total US population. The mean of all attribute value is taken, so that a county can be represented by this single value. The clusters are formed by dividing the county into 2 dataset according to their election result. For instance if a county voted for republican then it will go into republican cluster. The centroids of cluster are calculated by taking average of above county mean value. The data is then tested, for each county in test data again the mean of attribute value is taken and then result is predicted by selecting minimum Euclidean distance to centroid of both clusters. An accuracy of around 86% is obtained.

Agglomerative Nesting

In this algorithm, during training phase each county starts with its own cluster, the cluster value is represented by mean of all normalized attribute value. The cluster sets of republican counties is separated from democrat counties in order to differentiate between them. Each cluster set are stored in two different priority queue. In each priority queue first two clusters are taken , merged and then put back into queue, this process is repeated until 1 cluster is obtained in both queues. The resulting means of these clusters are their respective centroid. During test phase, the same procedure as that of k-means is followed and a county result is predicted by taking the nearest distance to each centroids of clusters. An accuracy of 83% is obtained.

K-Nearest Neighbor

This is a lazy learning algorithm, so the processing directly starts with testing phase. A county in test set is compared with all the counties in training set by calculating the sum of Euclidean distance of each features of both counties and then storing it in the list along with training county value so that it can be traced later. The values of this list is sorted in increasing order and then first K (=5) of them are selected, the training county of these are traced and their results are determined. The result of test county is decided by majority of type of counties in the set (k=5) i.e. if 3 out of 5 county in set is supporter of republican then the test county will be predicted as supporter of republican. A prediction accuracy of 86% is obtained

Perceptron Learning

In this algorithm each attribute of a county are assigned a weight. The prediction is done by taking sum of product of each attribute value by its corresponding weight and then compared with expected result. The algorithm starts by finding the best optimum value for each weights by modifying them according to expected results while traversing through each county in training phase. After the weights are trained, these are then tested on counties in test set. A prediction accuracy of 76% is obtained. This accuracy can be improved if the learning rate is decreased at regular intervals for increasing the convergence of result.

ID3 Decision Tree Learning

The algorithm works by first constructing the decision tree in which each node is associated with a decision and then predicting the output by traversing the tree from root to leave. The algorithm tries to represent the best attribute value on the decision node. During tree generation phase for each columns (attribute) is traversed and checked if it can be used to divide the rows into homogeneous groups i.e. the rows with homogeneous output result. This is measured by entropy, more is entropy, and the less is the homogeneity in the rows. The attribute value with least entropy is selected and put into decision tree node. At leaf node the final decision result is stored. The tree is generated using training data and then classification is done on testing data. The overall accuracy of 81% is obtained

3. CONCLUSION

The best predication accuracy was obtained using K-means and K-nearest Neighbor algorithm. The best running time performance is observed in K-means algorithm. However the worst running time performance is observed in K-Nearest Neighbor algorithm. It is also important to note that accuracy of some supervised algorithms like perceptron learning can be improved by using better model like multi-layer models.