Comparison of Decision Trees, Boosted Trees, Random Forest, Support Vector Machines and Neural Network for Classification Problems

Atharva Bhagwat

Machine Learning Intern

AI Tech Systems (www.ai-techsystems.com)

atharva.bhagwat42@gmail.com

***Abstract*— The need of classification is rising day-by-day as automation is taking over. Be it in self driving cars, face recognition systems or just an app which classifies dogs and cats. This paper investigates Decision Tree, Boosted Tree, Random Forest, Support Vector Machines, and Neural Network classifier. The focus of this report is on comparing these classifiers by evaluating the accuracy, based on iris dataset.**

Keywords*— *Machine Learning, Decision Tree, Boosted Tree, Random Forest, Support Vector Machines, Neural Network

I. Introduction

As Machine Learning is one of the fastest growing areas in computer science, the problem of classification gained more importance as well. Machine Learning is being integrated in almost all of the fields, be it revealing a photograph of a black hole or classifying a tumor.

In this age, where petabytes of data is generated within hours, classification techniques can be used to label the data and help organizing it. In this report we will compare 5 classification algorithms.

II. Dataset Processing and Model Building

A. Dataset Processing

For the sake of comparison we shall use titanic dataset from kaggle (https://www.kaggle.com/c/titanic). The dataset contains various columns. Columns in the dataset are: PassengerId, Survived, PClass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked. Survived column is the target column. During the preprocessing of the dataset, we dropped 4 columns: Cabin, Name, Ticket and PassengerId. We filled the missing values for column Embarked using mode and for Age using median. We mapped the Gender of the passenger to 0 and 1(female:0, male:1), Embarked status to 0,1 and 2 (S:0, Q:1, C;2). We then standardized using StandardScalar. On the last step we split the dataset into training set (80%) and testing set (20%).

Fig 1.Correlation between parameters.

B. Model Descriptions

i) Decision Tree:

Decision Trees are supervised learning methods used for classification and regression. The Decision Tree Classifier is capable of both binary classification and multiclass classification.

The dataset is split into subsets such that they have the same value for an attribute. The main parameter is placed at the root, the values of the parameter are compared and corresponding branch is followed. Then the best attribute is kept at the root of the subtree and the same process is followed. Some advantages of Decision Trees are:

- Trees can be visualised.
- Easy to interpret. The entire process of coming to a conclusion can be tracked unlike in a neural net.
- Decision Tree are robust to errors, hence if the training data contains errors/missing values it can be used.

Some drawbacks of Decision Trees are:

- The more the number of decisions in a tree, less is the accuracy of any expected outcome.
- Decision Tree considers only one attribute at a time and might not be best suited for actual data in dataset.

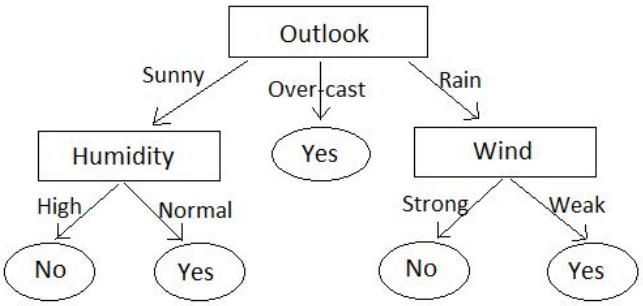

Fig 2.Example of working of a decision tree.

ii) Boosted Tree:

There are three boosting algorithms:

- AdaBoost (Adaptive Boosting): Training data is randomly sampled and decision stump algorithm is applied to classify the points. After classification, decision tree stump (1 level decision tree) is applied to the complete data. This process is repeated until the maximum number of estimators is reached. It identifies the shortcomings of the weak learners (decision stump) by misclassification rate.

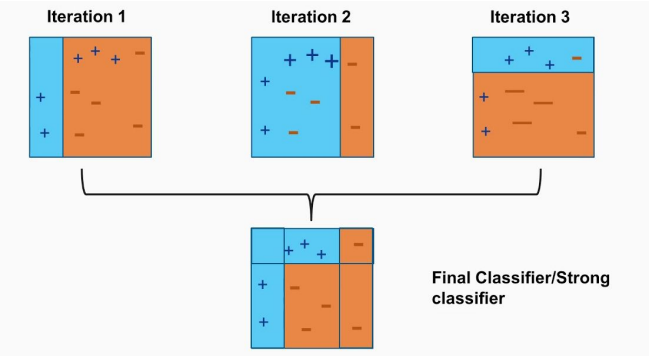

Fig 3. AdaBoost working.

- Gradient Boosting: It is the same as AdaBoosting, but the way it identifies shortcomings of the weak learners (decision trees) is different. It uses the gradients in the loss function as a measure for models performance.

- XGBoost: It is similar to gradient boosting algorithm but has few more features like:
  - Extra Randomisation Parameter.
  - Proportional shrinking of leaf nodes.

iii) Random Forest:

Random Forest consists of a large number of individual decision trees that act as an ensemble. Ensemble uses multiple hypotheses to form a better hypothesis. Each individual tree in random forest predicts a certain class, the class with the most number of predictions is the final classification of the model. Random Forest uses averaging to improve the predictive accuracy and control overfitting. The predictions made by the trees need to have low correlation with each other. Small changes to the training dataset changes the tree structure significantly. This is achieved by a process called Bagging. It allows each tree to randomly sample from the dataset with replacement, resulting in different trees.

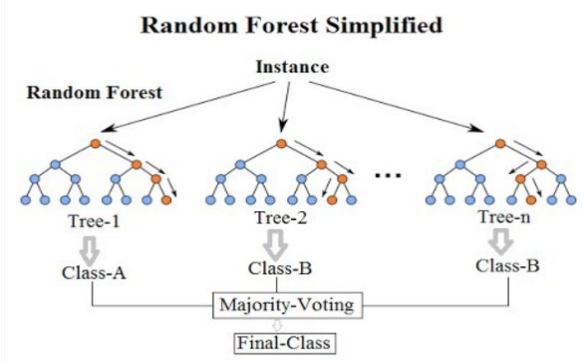

Fig 4. Random Forest working.

iv) Support Vector Machines:

Support Vector Machine classifier outputs and optimal hyperplane categorizing examples in the data. Various types of kernels can be used in SVM like, linear, polynomial, rbf.

The Regularization parameter (C) tells the SVM optimization upto what extent misclassification of each data sample is avoided. For larger value of C, optimization will choose smaller margin hyperplane if it does a better job classifying. Conversely, for smaller values of C, optimizer will look for larger margin separating hyperplane, even if it misclassifies few data samples.

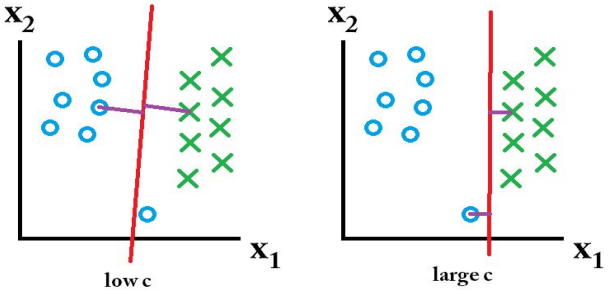

Fig 5. Comparison between two values of C.

v) Neural Network:

A Neural Network consists of neurons (units) arranged in layers. Each unit takes an input, applies a function and then passes the output on to the next layer.

It uses back propagation during training. Back propagation is a method used to tune the weights of the neural network depending on the error of the previous epoch. The weights are tuned in such a way that the error decreases and the model is more generalised.

C. Model Building

i) Decision Tree:

We used DecisionTreeClassifier provided by sklearn. We obtained the accuracy of 77%.

ii) Boosted Tree:

We used all boosted algorithms. With AdaBoostClassifier, XGBoostClassifier provided by sklearn and xgboost respectively, we obtained the accuracy of 78%. With GradientBoostingClassifier provided by sklearn, we obtained accuracy of 79%.

iii) Random Forest:

We used the RandomForestClassifier provided by sklearn. We set number of trees in the forest equal to 100. Instead of training it with training and testing set, we used k-fold cross validation. With that we got an accuracy of 81%.

iv) Support Vector Machines:

We used the Support Vector Classifier (SVC) provided by sklearn. Instead of training it with training and testing set, we used k-fold cross validation. With that we got an accuracy of 82%.

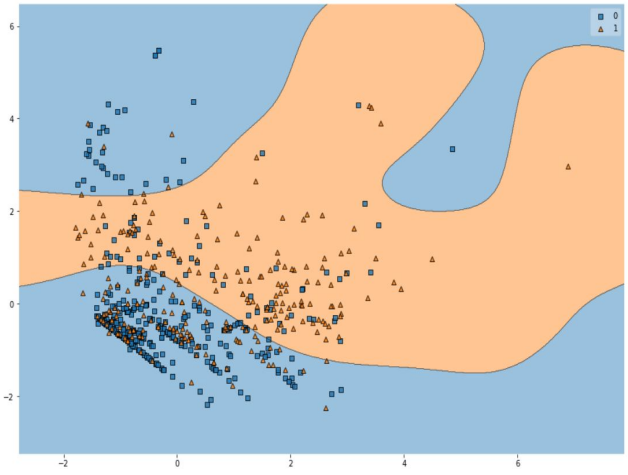

Fig 6. Hyperplane generated by SVC.

v) Neural Network:

We used Sequential model provided by keras. We used 2 hidden layers, and a dropout layer of dropout 0.01. We obtained the accuracy of 83%.

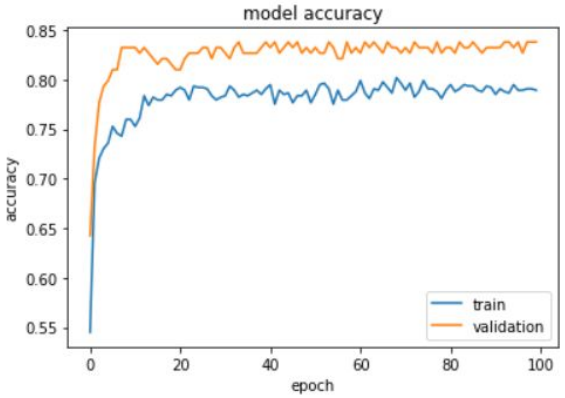

Fig 7. Accuracy vs Epoch for training and validation set

Algorithm	Time	Accuracy	Classification
Decision Tree	0.004	77%	Tree
Boosted Tree	0.07	78%	Tree
Random Forest	3.35	81%	Tree
SVM	0.012	82%	Function
Neural Network	3.37	83%	Function

We can observe that it takes a long time to train a neural network as compared to other algorithms.

III. Conclusion

We learnt that if the number of parameters affecting the output are more then, decision tree and boosted tree are not suitable algorithms to use as calculations get very complex if the parameters are interdependent. Random Forest on the other hand can be used for dataset with multiple features. In SVM finding optimum values of the hyperparameters can be difficult.

After studying the algorithms and comparing their performances we observe that neural network provided the highest accuracy.