Experiment on Waveform data using K-NN

Md Nur Amin[1],Berisha Erblin[2]

Abstract

We are performing an experiment on a data set of waveform where it has 40 attributes describing the wave form. It generates the three classes of waves and each class is generated from a combination of 2 or 3 base waves. A non-parametric method K-Nearest Neighbor is used to classify the waveform.

1. Core

A series of experiments is conducted on the data set starting with tuning of k to find the best accuracy,analyzing bias-variance trade off to depict the Bayes error over the change of training size,model evaluation and comparison in an imbalanced data set by reducing the complexity, lastly testing on imbalance data set and comparing with the original model.

1.1. Tuning the best k of a kNN classifier

We are considering the 4000 samples as training set and rest 1000 as test set.Cross validation method is performed to tune the best value of k. For k=1 the model outputs the best accuracy of 86% and then decreases slowly.

1.2. Analysis of the bias-variance trade-off.

For the analysis of bias-variance trade-off different configuration of training and test set was formed and change in the accuracy was observed. Starting with a vary small size of train set 100, the accuracy was dropped to 70 resulting in a Bayes error of 30%. As the training size increases Bayes error starts to decrease. Train size ranges from 2000 to 4000 Bayes error stays around 14%. When the train size(4950) is close the actual size, the Bayes error is 8%. Initially when K=1, the bias is 0.With the increase of K, the model becomes more complex.

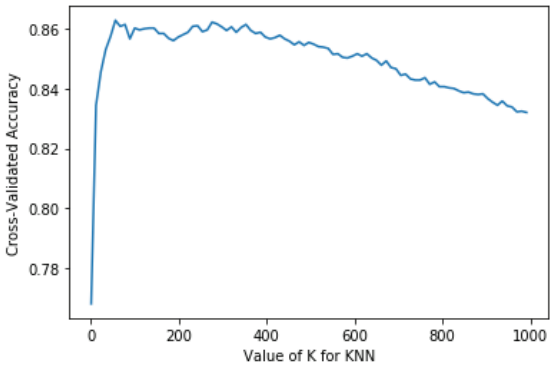

Figure 1. Accuracy over value of K

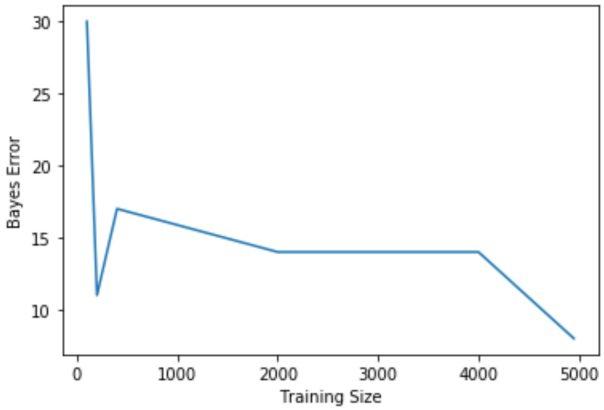

Figure 2. Bayes error for training size

1.3. Reducing the complexity

Keeping the pattern intact of the training data, CNN has the same classification with respect to original data set. It constructs a reduced subset of the training data and still capable of correctly classifying the original data set using a 1-NN algorithm. While performing CNN with the data set, number of examples was reduced down to 2413 by the algorithm and 77% of accuracy was obtained without cross validation and 79%. The reduced accuracy shows the shortcoming of the approach because it fails to find an absolute minimal classification set and outliers are retained as of original data set. With the RNN the data set was

even reduced to 1707 by the algorithm and an accuracy of 96% was achieved with cross-validation and 97% without cross-validation.

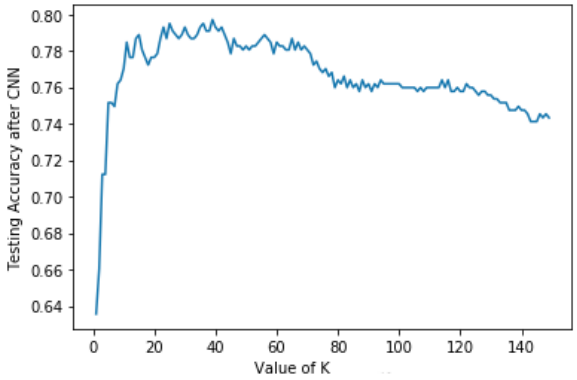

Figure 3. Accuracy over value of K in CNN

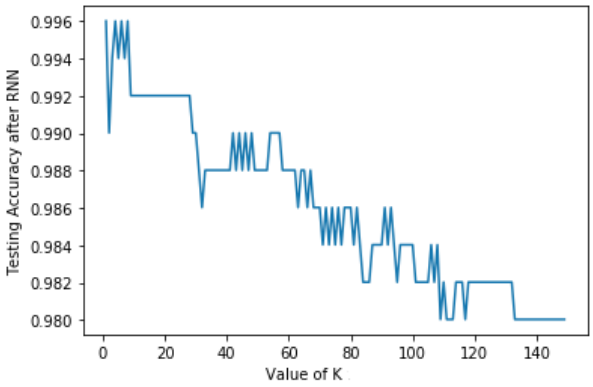

Figure 4. Accuracy over value of K in RNN

1.4. Performance on imbalance data

Imbalance was generated artificially t o o bserve t he accuracy compared to the original data set. `Make_imbalance` function from `imblearn` library was used to generate the imbalance and the data set was shrunk down to 60 by the algorithm. After performing the K-NN on this reduced data set, we achieved the best accuracy of 84% with K value of 4. The F- measure is 1 when the value of K is 1. It remains almost unchanged between 10 to 45 for k and then drops dramatically afterwards flattens out.

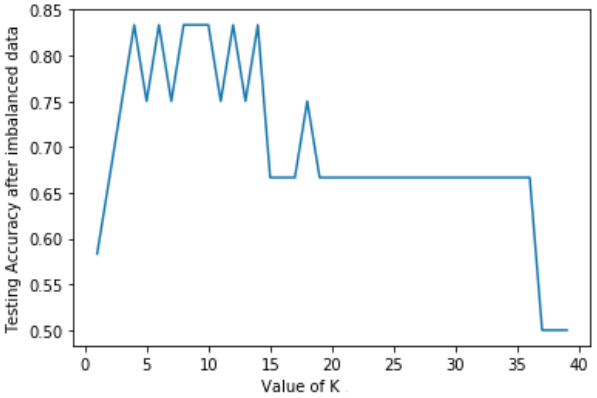

Figure 5. Accuracy over value of K on imbalanced data

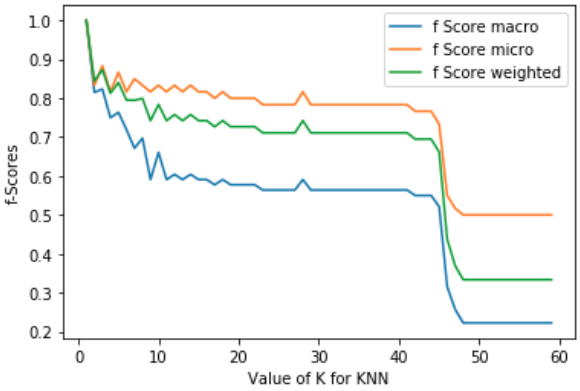

Figure 6. F-measure on imbalanced data

2. Conclusion

From the experiments we can see that K-Nearest neighbor performs quite well to classify the waveform. Over the course of these experiments, though the model suffers in an imbalanced data, RNN achieves a very good accuracy for the reduced data size.