Analysis of Performance of KNN on Waveforms Dataset

Mohammad Sadil Khan , Hanok Mesa

December 2020

1. Abstract

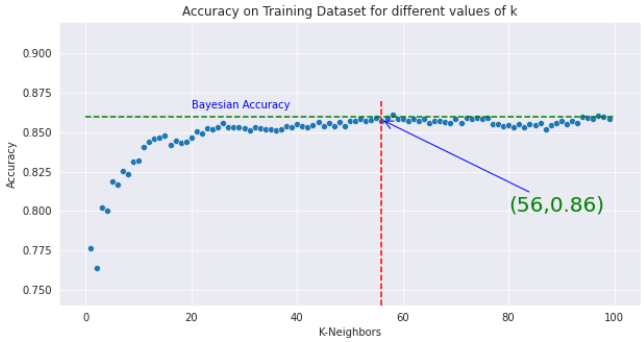
Despite its simplicity, the KNearest Neighbours classifier has surprised machine learning researchers by exhibiting good performance on a variety of learning problems. The article analyses the performance of KNN on Waveforms dataset[1]. We compare the empirical risks of the model by experimenting with various values of k (neighbors) and estimate of the true risk using Cross-Validation. We also analyse the effect of training sample size in bias-variance trade-off and present the reduction of training dataset using CNN and RNN. Finally we generate artificial imbalancy in dataset and compare the performance.

2. Core

Waveform is a multivariate dataset of 5000 instances. It contains 21 attributes and 3 class labels corresponding to three waves. Each class is generated from a combination of 2 of 3 "base" waves. The dataset has no missing values and is balanced. For training dataset, we took 4000 random samples(Train_{NN}), and for the test set 1000 samples(Test_{NN}).

2.1. Tuning optimal K

To tune the value of k, we have used K-fold cross-validation approach with K=5 and analysed the accuracy for different neighbor values from 1 to 100.



From the graph ,for optimal neighbor, we have chosen 56 and the estimate of the true accuracy is 0.86.

2.2. Bias-Variance Trade Off

In this section, we present the survey of the gap between the Bayes error and KNN model with various size training samples and different neighbor values. We observe that for smaller training size and smaller k, the model overfits and larger the k the model underfits and risk is much more than bayes error and as the training size increases the training error curve smooths down to bayes error and the test error curve also merges with the training error. So it removes the overfitting issue. The obervation gives an intuitive idea about KNN's capability in converging $P_n(Y=y_j|X=x)$ to $P(Y=y_j|X=x)$ as $n \to \infty$.

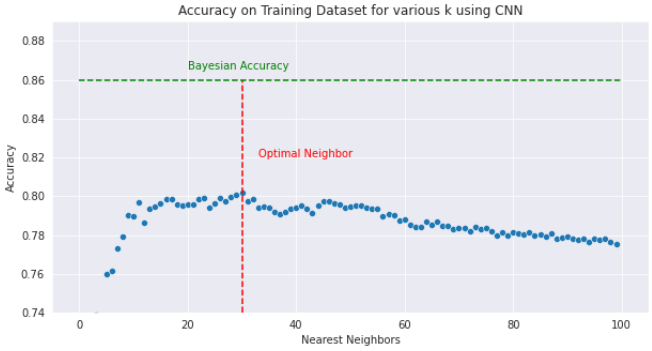
2.3. Reduction of Complexity

KNN's simplicity and effectiveness are its advantage but its disadvantages can't be ignored even. The memory requirement and computation complexity also matter. Many techniques have been devolped. We present the survey of two such algorithms - CNN and RNN.

2.3.1. Condensed Nearest Neighbours(CNN)

CNN allows to choose a subset of the training data which will work as well as the training set in classifying unknown patterns. In this case, a possible drop in performance is being traded for greater efficiency, both in the amount of memory required to store the training set and in the compu-

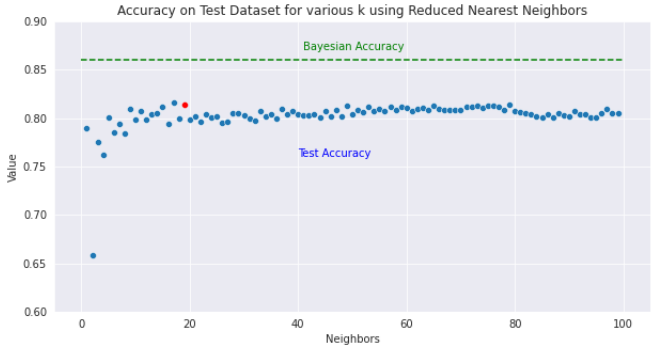
tation time required to reach a decision. After analysing the accuracies for various k, we choose k=30. We achieve 0.82 as training accuracy and 0.79 as testing accuracy. In case of data reduction, we achieve 53% reduction in samples.



Despite the efficiency, CNN has introduced imbalancy in the dataset. The f1-score is 0.79 compared to 0.86 for model trained on Train_{NN}.

2.3.2. Reduced Nearest Neighbours(RNN)

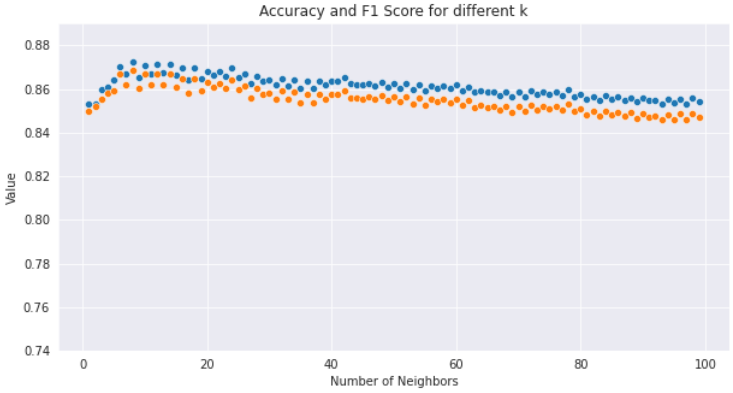
RNN is an extension of CNN and it futher reduces the data. Since it depends upon CNN, we use 1NN to reduce the training samples. We use the RNN reduced dataset as training sample and use Test_{NN} for testing. We plot the accuracies on the test set for every k value from 1 to 100. For optimal k, we choose 21 (red point in the graph). We achieve almost 82% reduction in sample points. Unlike CNN, RNN reduced dataset isn't much imbalanced.



Observation:
Since RNN induced Dataset(T_{RNN}) $\subseteq$ CNN induced Dataset (T_{CNN}), if T_{CNN} doesn't contain minimal consistent subset, neither does T_{RNN}. The reduction in the accuracy is most likely due to the inconsistent training samples[2].

2.4. Generation of Artificial Imbalancy

We present the analysis of how KNN works in imbalanced dataset. We create artificial imbalancy using SMOTE approach. We double the number of a class label and plot the accuracy and F-score for choosing the best k. We observe that as we increase the value of k, after k=20, the F-Score keeps reducing. So we choose k=20 and find that the model achieves 81% accuracy in test and 80% as F- score.



3. Conclusion

The main objective of the article is to analyse the performance of KNearest Neighbors on Waveform Dataset. We applied various reductions on dataset to reduce bayesian errors and time complexity and based on the analysis, CNN and RNN works well and it will be a useful tool if we can manage to make T_{CNN} and T_{RNN} minimal consistent dataset. Also in case of imbalancy , lower k's work better in KNN.

4. References

[1]. http://archive.ics.uci.edu/ml/datasets/waveform+database+generator+(version+2)

[2]. The Reduced Nearest Neighbor Rule ,GEOFFREY W. GATES https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.294.4537&rep=rep1&type=pdf