Anaysis of WaveForm data for KNN

Sri Kalidindi/Guillaume Sacchetti/Shamprikta Mehreen

***Abstract*— This report illustrates the KNN behavoiur on WaveForm Data from UCI Repository, and presents the results on Cross Validation, Bias-Variance Tradeoff, CNN, RNN and Imbalancy.**

I. INTRODUCTION

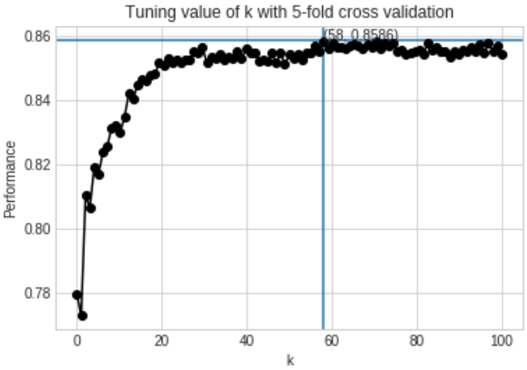
In the waveform data set, we have data 5000 waveforms data for 3 classes, and 21 attributes which include noise.The interesting aspect of this dataset is we know the Optimal Bayes classification rate which 86

This report includes the following:

- Tuning the best k of a KNN Classifier
- Analysing of bias-variance trade-off
- Reducing the complexity(Using RNN and CNN)
- Analysing artificial imbalancy

II. TUNING THE BEST K OF KNN CLASSIFIER

After performing KNN for values of k from 1 to 100 with validating the performance using 5 Fold cross validation(4000 samples for training and 1000 samples for testing). Below is the graphical representation of performance with respect to k.

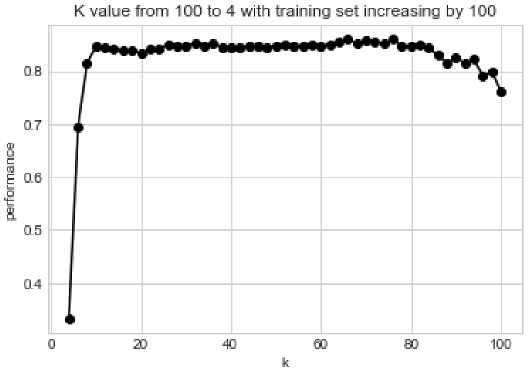


The best value of K is found to be 58 with a performance of 0.8586.

III. ANALYSIS OF BIAS-VARIANCE TRADEOFF

A. Experimental Setting

For demonstrating Bias-Variance trade off we have performed KNN with K ranging starting from 100 to 4 and training set starting from 100 with increasing training examples with multiples of 100. Over 50 iterations



B. Results

As expected Bias-Variance nature, the performance is very poor with K=100 and training examples of 100, while the performance increased with decreased k value and increased training examples.

Below are the intermediate results of the experiment

K value	Training Examples	Performance
100	100	0.331
38	3100	0.8557
4	4900	0.76

- The setting of k=100 and training examples=100 has huge bias and low variance due to which performance suffered.
- The setting of k=38 and training examples=3100 has the best balance of low bias and variance which resulted in a good performing KNN.
- The setting of k=4 and training examples=4900 has small bias and huge variance which also resulted in poor performance.

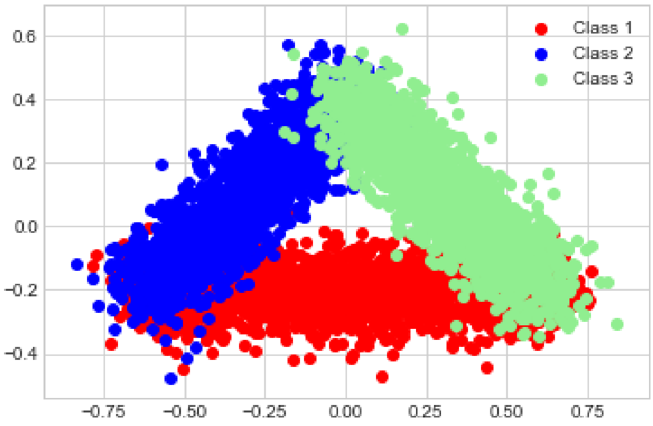
Note: As observed above it is important to choose the setting with low bias and low variance as possible.

IV. REDUCING THE COMPLEXITY USING RNN AND CNN

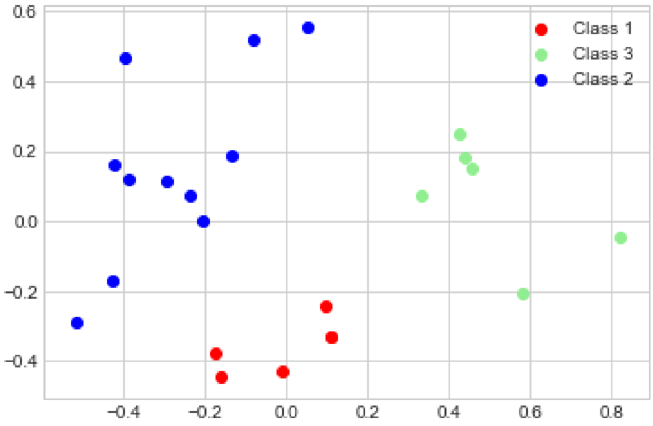
In this section we will present the results for performing RNN and CNN on the WaveForm Dataset.

RNN: After performing RNN we have reduced the data size to 2015 samples from 5000, eliminated 2985 points representing the region of bayesian overlap and outliers

CNN: We have taken 2015 examples from RNN and reduced the data set to 23 data points after performing CNN, which represent most relevant examples for learning each class.



Original Data



Data after RNN and CNN

Above figure represents the distribution of Original data and After RNN and CNN

	Precision	recall	f1-score	support
0	1.00	0.11	0.20	1657
1	0.65	0.74	0.69	1647
2	0.58	1.00	0.73	1696
accuracy			0.62	5000
macro avg	0.74	0.62	0.54	5000
weighted avg	0.74	0.62	0.54	5000

Note: Above table represents the confusion matrix for the setting Training Data: RNN and CNN data set and Testing Data: 5000 examples of full dataset

V. Artificial Imbalance

We have created imbalance among the classes by removing 95 percent of class 0, 85 percent of class 1. Leaving class 2 untouched, and after tuning the best value of k with respect to f1-measure which is 12, the performance is as below

	Precision	recall	f1-score	support
0	0.99	0.35	0.52	1657
1	0.73	0.78	0.76	1647
2	0.64	1.00	0.78	1696
accuracy			0.71	5000
macro avg	0.79	0.71	0.69	5000
weighted avg	0.79	0.71	0.69	5000

VI. Conclusions

As proved above the best KNN for this data set is with k value 58 as the performance 0.8586 is very close to the Opimal bayes classification rate of 86 percent.