Noise-Removal and Classification

Project Report

The aim of the project is to make the dataset better by removing the noise and outliers to improve the accuracy of a classification algorithm.

To achieve this, the given datasets were denoised based on two methods – knn and pca. And then the following algorithms were trained for the datasets.

1. Logistic Regression
2. KneighborsClassifier
3. SVM
4. RandomForestClassifier

The following experiments were conducted on the Synthetic and Iris Data –

1. Denoise by knn with k = 5 , xi = [4,3]
2. Denoise by knn with k = 6 , xi = [5,4]
3. Denoise by knn with k = 3 , xi = [2]
4. Denoise by pca with p = [0.75,0.9]

And for Wine data : -

1. Denoise by knn with k = 5 , xi = [4,3]
2. Denoise by pca with p = [0.75,0.9]

Along with the above experiments, a few selected models were hyper tuned for the parameters. It was found that hypertuning of the parameters does not make a big change in the accuracies of the models.

It was observed that after denoising the accuracies of the models increased slightly or remained same for most of the models.

For the Moon dataset:-

First the dataset was clustered in to 8 different clusters and the pca based denoising was performed.

The experiments run for the clustered moon data set were performed by taking different values of p [0.5,0.6,0.7,0.8,0.9,0.95,0.99] on the above mentioned classifiers.

After denoising, the accuracies improved for RandomForestClassifier.

Result:

The best chosen models for the datasets are :-

1. For Synthetic dataset - Logistic Regression and Knn by pca denoising.
2. For Iris dataset – Knn and Svm by both pca and knn denoising.
3. Wine dataset – RandomforestClassifier by pca and knn denoising.
4. Moon dataset – RandomForestClassifier by K-means clustering and denoising by pca.

Conclusion:

Thus, we can conclude that denoising the dataset makes it better. By denoising we get well separated clusters which is better for training a model. When the data is in better shape, we can get better accuracies on the trained models. We did not find any substantial increase in accuracy from the experiments because the chosen datasets did not have any extreme outliers in it. Denoising the data set which has extreme outliers will be beneficial as it will retain the distribution of the data.