Evaluating Different Classifiers

Anshul Jain
CSE Department, The LNM Institute of Information Technology
Rupa Ki Nangal, Post-Sumel,Via-jamdoli, Jaipur-302031, (Rajasthan) INDIA
anshul.lnmiit@gmail.com

November 24, 2011

Abstract

This survey report focuses on the results obtained by different classifying and pre-processing techniques for a given dataset .

1 Keywords

PCA, Naive Bayes, K- Nearest Neighbour (KNN), ID3, Supervised Binning, Un-supervised Binning.

2 Introduction

There are different pre-processing techniques that can be applied on the dataset to reduce features that might not contribute to train the model. This can have an adverse effect on the process of training data also. Different pre-processing techniques works differently for different classifier. Some may increase the accuracy of classifier while another may decrease.

2.1 The Dataset

Given Dataset contains data of 20 different fonts with 20,000 unique stimuli each containing 16 feature values which are scaled in the range of 0 to 15. Each of these feature value depends on the black-and-white rectangular box that contains a letter of a particular font. Each class is uniformly distributed.

3 Description of Methods

We have considered 70% training data and 30% testing data. The classifier is trained on the training data and the accuracy is found for the testing data. We have used WEKA to perform our experiment.
We started our work on the dataset by reducing the attributes using either of the 3 different methods namely Principal Component Analysis, Supervised Binning, Un-Supervised Binning and then applying 3 classifying algorithms namely K-Nearest Neighbours, ID3 and Nave Bayes.
We started the classification by applying KNN algorithm (which belong to lazy classifier class) on discretised and Non-discretised value of dataset with different values of k (neighbours). We used 11 different values of k.
Next we applied ID3 (Decision Tree) which required the dataset to be discretised. To get a discretised dataset we applied un-supervised discretisation and supervised discretisation methods. For un-supervised discretisation we used 11 different values of bin size.
The KNN method was also applied on the dataset pre-processed by using PCA for different values of k. The third classifying algorithm we used was Nave Bayes (which belong to Bayes class) on the raw data and the data pre-processed by using PCA and un-supervised discretisation.

4 Results

The results obtained showed a large variation in the accuracy of different classifiers for different type of pre-processed data. Accuracy varied from 59% for 5 bin Nave Bayes to 95% for KNN with k value 1 and unprocessed data. For the data pre-processed by PCA, the variation in accuracy of KNN was not much.
On applyinf PCA, the features reduced from 16 to 13.
Accuracy of KNN was best for raw data with value of k=1

k	Recall	Precision	Accuracy
1	0.935	0.935	93.8833%
3	0.938	0.939	93.7667%
5	0.937	0.938	93.7%
10	0.935	0.935	93.35%
15	0.923	0.925	92.2667%

Table 1: PCA & KNN for different values of k

k	Recall	Precision	Accuracy
1	0.906	0.909	90.6337%
3	0.894	0.898	89.367%
5	0.883	0.888	88.2667%
10	0.86	0.867	86.0333%
15	0.852	0.844	84.35%

Table 2: KNN & supervised discretisation for different values of k

# Bins	Recall	Precision	Accuracy
5	0.773	0.783	77.3167%
8	0.836	0.845	83.5667%
10	0.8445	0.847	84.567%

Table 3: KNN & un-supervised discretisation for different number of bins

# Bins	Recall	Precision	Accuracy
5	0.808	0.81	78.5833%
8	0.846	0.846	80.55%
10	0.839	0.84	73.1167%
supervised	0.835	0.835	77.2167%

Table 4: ID3 for different number of bins

Pre-processing	Recall	Precision	Accuracy
Nothing	0.64	0.655	64.01%
PCA	0.645	0.65	64.4833%
supervised	0.731	0.749	73.1167%
un-supervised 5 bins	0.595	0.618	59.4667%
un-supervised 8 bins	0.696	0.713	69.233%
un-supervised 10 bins	0.696	0.716	69.5667%

Table 5: Naive Baye's for different pre-processing methods

5 Inferences

It is observed that for KNN, we get best accuracy for k = 1. For k 1, the accuracy keeps on decreasing. This might be because for a given test stimuli, there are some training stimuli of the same class that exactly resembles the test stimuli, but there are many other training stimuli that belong to different class and are similar to our test stimuli which affect the result for large value of k.
Applying KNN after Binning is not a good option as the accuracy further decreases. The eucledian distance is generalised after binning and therefore the accuracy went down. As the bin size increases, the eucledian distance becomes more and more generalised.
Apart from comparing a single algorithm for different values, we now compare different algorithms. It was observed that, accuracy for ID3 was less than that of KNN. This might be because, the dataset we are considering has feature values on the basis of pixels in black-and-white rectangular box. There might be many features that are similar for different classes and some that are different. If these features are removed, the accuracy goes down.

In KNN, we consider all the features whereas in ID3, we start our classification from the root. In this case, there might be certain features that are ignored and therefore the accuracy of our classifier goes down.

6 Conclusion

It's is quite clear from the above inferences that the performance of KNN is best among all the classifiers and even better if none of the features are ignored i.e. if KNN is applied on the raw data. Effect of bin size is observed for ID3, i.e. Decision tree and for this dataset, bin size=4 gave the best accuracy.Considering Naive Bayes, which gave the worst result among all the three classifiers, the best accuracy value was obtained for supervised discritisation. A general trend in recall and precision is also observed. Generally for all the algorithm we used, recall value is less than or equal to precision value. For the similar dataset like ours, the best way to train a classifier is to consider every attribute of the dataset.