TEAM 51

Foundation of Data Mining

Task 2

-By
Sheetal Tukaram Budbadkar

# Experimental Setup

The task was solved using two classification methods:

1. **Support Vector Machines:** SVM is used because it's easy to find complex relationships between data
2. **Naïve Bayes Classifier:** Naïve bayes classifier is a simple classifier. I think as specified by Occam's Razor there are high chances this might give the best results

Feature Selection

First, over the entire dataset of 80 features, we select 40 best features using the statistical property of Information Gain.

Over this 40 features we trained both the classifiers for the following situations:

1. Training the classifier with each class containing exactly set of 3,5,10,15 training data

2. Over one split of train and test set over the complete data.

3. Training the classifier over entire data without any train test split.

Also just to check we have also trained both the models over entire 80 features just to check the difference in how well classifier performs over 40 selected features and over all 80 features

Feature Engineering

Since the features represent an Edge histogram Descriptor, we have created 16 bins(features) over 80 features considering and merging set of 5 features for the five orientations of the edge.

Also for each set we selected the maximum value which would represent the dominant edge.

Over these 16 features we train the classifiers for following situations:

1. Over one split of train and test set over the complete data.

2. Training the classifier over entire data without any train test split.

#Results

Feature Selection

The result for classifiers are as follows:

1. SVM Classifier

Situation	**Accuracy**
Exactly set of 3,5,10,15 images in training set	Train accuracy: 0.963 Test accuracy: 0.441
One split of train and test set over the complete data	Train accuracy: 0.949 Test accuracy: 0.506
Over entire data without any train test split	Accuracy: 0.9424
Over entire data with all 80 features	Accuracy: 0.578

2. Naive Bayes Classifier

Situation	**Accuracy**
Exactly set of 3,5,10,15 images in training set	Train accuracy: 0.570 Test accuracy: 0.411
One split of train and test set over the complete data	Train accuracy: 0.473 Test accuracy: 0.426
Over entire data without any train test split	Accuracy: 0.465
Over entire data with all 80 features	Accuracy: 0.484

Feature Engineering

The result for classifiers are as follows:

1. SVM Classifier

Situation	**Accuracy**
One split of train and test set over the complete data	Train accuracy: 0.608 Test accuracy: 0.322
Over entire data without any train test split	Accuracy: 0.578

2. Naive Bayes Classifier

Situation	**Accuracy**
One split of train and test set over the complete data	Train accuracy: 0.245 Test accuracy: 0.231
Over entire data without any train test split	Accuracy: 0.244

#Conclusion:

Overall, even though the accuracy of SVM shows bigger value for SVM it is overfitting in most of the situations. The best performance is given for all 80 features for the Naive Bayes classifier.

Also the assumption that the 16 features extracted using feature engineering should have a better performance could not be validated. Maybe having more data could give better highlights over this.