CSCM45 Big Data and Machine Learning

Coursework 2: Object Recognition

**Introduction -** Provide an overview of the problem, your proposed solution, and your experimental results.

We have been provided with an image dataset set, which is a subset of CIFAR-10[1]. The aim of the experiment is to try and find out what machine learning technique or combination of techniques, creates the best prediction accuracies. Principle component analysis (PCA), linear discriminant analysis (LDA), support vector machine (SVM) and a neural network (NN) will be used with a bench mark of 44.68%. As well as just feeding the data directly into a NN for a direct comparison. Another metric that we used for comparison is the time it takes to compute, as we can assume that using a form of dimension reduction will lower the accuracy, but the potential faster time it takes to train the model might be a justifiable trade-off. (NEED TO ADD WHAT RESULTS THAT WERE FOUND!!!!!)

**Method-** Present your proposed method in detail. This should cover how the features are extracted, any feature processing you use (e.g. clustering and histogram generation, dimensionality reduction), which classifier(s) is/are used, and how they are trained and tested. This section may contain multiple sub-sections.

**Extracting the data from the Dataset**

We first need to extract the data from the two provided 4D arrays, one of size 32x32x3x10000, the training data, and the other 32,32,3,1000, the testing data. In order to be able to effectively use the data a provided function, from the SKLearn library, SKImage was used to extra the features. The features were then placed into numpy arrays shaped 324, 10000. The same was done to the testing set which created a numpy array shaped 324, 1000.

**Initial Benchmarks**

We first fed the extracted features into a NN. This was to provide an initial benchmark for comparison. The NN settings were changed using different number of layers, optermisers and activations to try and find the best outcome for the data. The same was done with SVM, again changing the kernel from linear, sigmoid and poly as well as changing the penalty of error.

**Dimensionality Reduction Techniques**

Both PCA and LDA will be used. First by themselves, fitting and then testing the data. Then by fitting the data with the output results being fed through a Neural Network.

**Supervised learning Techniques**

The data will also be fed through a NN without any reduction to see a comparison, along with the running time of the task processing. As well as SVM, to have a comparison of non-dimensionality techniques.

**Results**

Present your experimental results in this section. Explain the evaluation metric(s) you use and present the quantitative results (including the confusion matrix).

**Conclusion**

Provide a summary for your method and the results. Also, provide your critical analysis; that is the shortcomings of your method and how they may be improved.

**References**

Include references where appropriate. References are not included in the page limit.

[1] subset reference CIFAR-10

Submission

Submit your work electronically to Blackboard. Your report should be in PDF format only. Compress your Python source code and report into a Single Zip file. The deadline for this coursework is 11AM Monday 9th December.

Notes:

With the provided dataset being so big, with many features, it is potentially going to take a long time to process and train the Neural Networks. With using a dimensionality reduction techniques along side a NN, it will be compared to see if the reduction in potential accuracy is a valid trade off for processing time.

The aim of the project is to see what method creates the best prediction accuracy form the provided dataset. This will be using LDA, PCA, SVM and a Neural Network to see what provides the best results. The accuracy, as a percentage, and how long the algorithms take to run will be used as factors/benchmarks for comparison.

The first task will be using the provided function to extract the features from the dataset and then add them to a numpy array. The same will be done to the training data set provided. Training and testing data labels will be used for classifying the data and testing the accuracy.