Respected Sir,

A Happy Republic day to you! It think I have got some interesting results from my revision of the RBM/DBN part. I have got some significant improvements.

Here's what I did:

1. I had read up on some tricks and tips to train a DBN. Extracted 100 features for the training set.

2. I decided to quickly run these extracted features through my MATLAB classification toolbox. It allows us to quickly experiment

with a number of classifiers including neural networks, SVM's and tree ensembles.

So, the training data was 770 samples each reduced to a 100 dimensional feature vector.

a. First tried using a shallow neural network. But the training accuracy itself was not going above 65%. Realized that the network was probably underfitting

because of lack of data. I thought , "Well then lets try the other standard shallow methods instead". I knew that when data is less shallow methods become more

competitive with the neural networks.

b. I took the samples and started running then through an SVM(quadratic, cubic, Gaussian) and some Decision Tree Ensembles. The training accuracy suddenly just

to 95% in case of the simpler models to 100%. This meant that it was the lack of data which was holding back the nueral net. It also meant that the features extracted

using DBN's could be used by the shallow classifiers for discrimination.

c. But there was still a catch. When I then tested the trained models, the overall accuracy was still 65%. The training data had 4 classes, 0(negative),1(ADHD),2(ADHD2),3(ADHD 3). When I looked at the confusion

confusion matrix of the test, I noticed that the classifiers were classifying most of the test samples as healthy. There were very few instances of the ADHD positive categories mixing up. This was because the training data was skewed. Out of the 770 samples, over 80% of the samples were of healthy individuals. From the remaining 20%, most very of label 1.

d. To see if this was the problem, I decided to level the playing field. I took 100 healthy samples and 100 label 1 ADHD samples. I knew that training would be more difficult as I was decreasing the size of the training set even further. The problem was now of binary classification.

e. I knew, I had to make sure that the accuracy is above 50%. Because here, even a useless classifier could get 50% accuracy by chance. I again ran the shallow tests and again the training accuracy was in the range 95-100 %. I then decided to hold around 5% of the data that 5% of 200 samples. The network was no longer biased against one category.

I got the following results-

for 5% hold out –

1. Cubic svm

training accuracy: 100%

testing set size- 20 (10 healthy and 10 AHDD\_1)

testing accuracy: 75%- overall

ADHD (1 incorrect)- 90 %

2. Ensemble tree classifier - 80% overall

80% overall accuracy (all adhd samples correct, misclassifications in healthy samples).

f. I then decide to increase the hold out percent. However, this meant taking out training data from the 202 samples available. Thus for lager testing sets, the overall accuracy was falling towards 60%, but still the ADHD samples were being classified correctly to a great extent.

Conclusion-

1. The extracted RBM features clearly have discriminatory qualities.

2. The less accuracy was primarily a result of skewness in the available data. The classifiers were seen a lot of one type of data.

3. The relatively small training set available imply that the shallow classifiers performed better than the generally more powerful neural networks.

4. Currently the shallow classifiers are learning the data perfectly and for smaller test sets are doing quite well.

5. The accuracy for true positives was higher than true negatives. The classifers find it easier to classify sick samples. Maybe there is something which sticks out in these samples making classification easier.

In the ADHD dataset .mat file, there are two more matrices UY\_1 and UY\_2. Maybe clubbing them together will give me a larger dataset to work on.