CS631 : Homework 3

Abhishek Kumar (18111002),
Aneet Kumar Datta (18111401), Dixit Kumar (18111405),
Komal Kalra (18111032), Nitin Vivek Bharti (18111048),
Riya James (18111054)

October 2018

Understating Data

To work efficiently with any data we must understand its properties and visualization becomes a key part of any machine learning problem. The data consists of 4150 system operating points with 132 features each. A visualization of the correlation between features of the data is as follows:

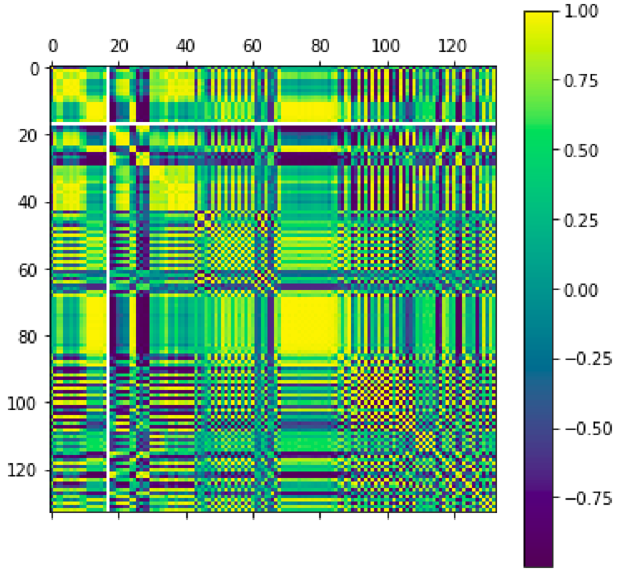

Figure 1: Correlation matrix plot

Feature Selection and Reduction

Feature selection

Looking at the high correlation between features one can conclude a feature selection can be very help full in reducing the number of features we need to observe.
We have used **chi square** testing over the data to choose 5 best features which are sufficient enough to classify data. chi square test tells how much class distribution depends on each feature and we can discard the features with lower dependence.

Selected features were:
RINALDI, TOLUCA, MIGUEL, MIGUELMP, VALLEYSC.
So we only need to use 5 sensors for classification purpose.

Feature Reduction

Further more we can reduce the model size by reducing number of features to be used by feature reduction technique such as principal component analysis (PCA). We plot this - the accuracy of data that can be reconstructed with the first few eigen values: The fraction of data retained by projection on the first 4 out of 5 eigen vectors (ordered) is 0.999999 and that for the first 5 is also the same (up to 20 digits of precision). Therefore we only consider the first 4. The variance between each pair of these five features:

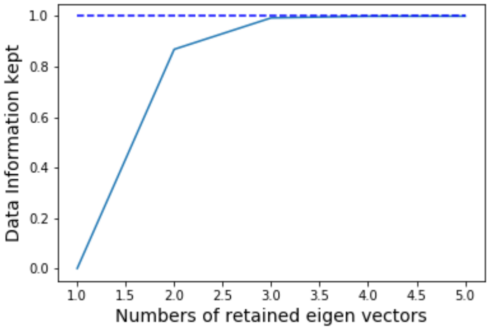

Figure 2: Number of transformed features vs the retained variation in data using PCA

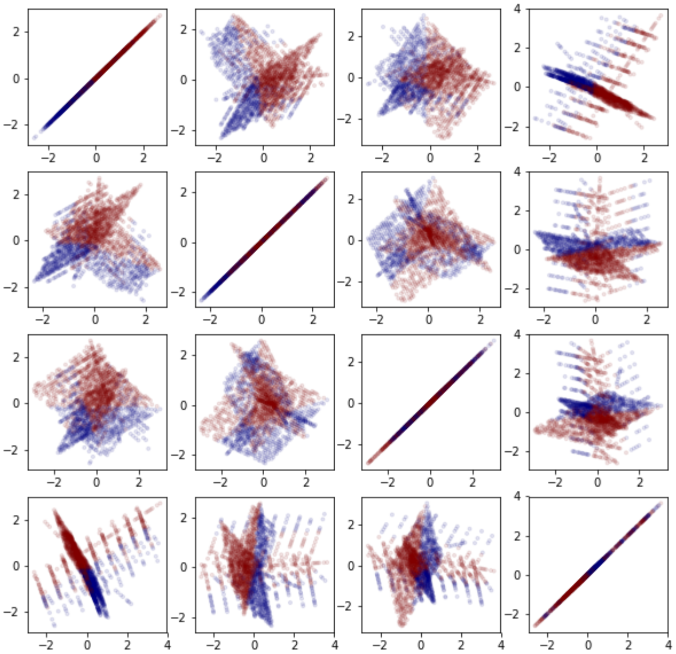

Figure 3: Scatter plot of new features

Modelling

We divide the data into three parts training, test and validation set.

Split of data

The splitting of data was done in a stratified way based on class labels.

- Training data: 50%
- Validation data: 25%
- Test data: 25%

Ensemble modelling

Many Classification models can over fit and gives complex decision boundaries over data but assuming that every model overfits the data in a different way we have applied ensemble method over all the models by taking the majority voting of models (arbiter over all models). This approach guarantees to reduce the over fitting of individual models.

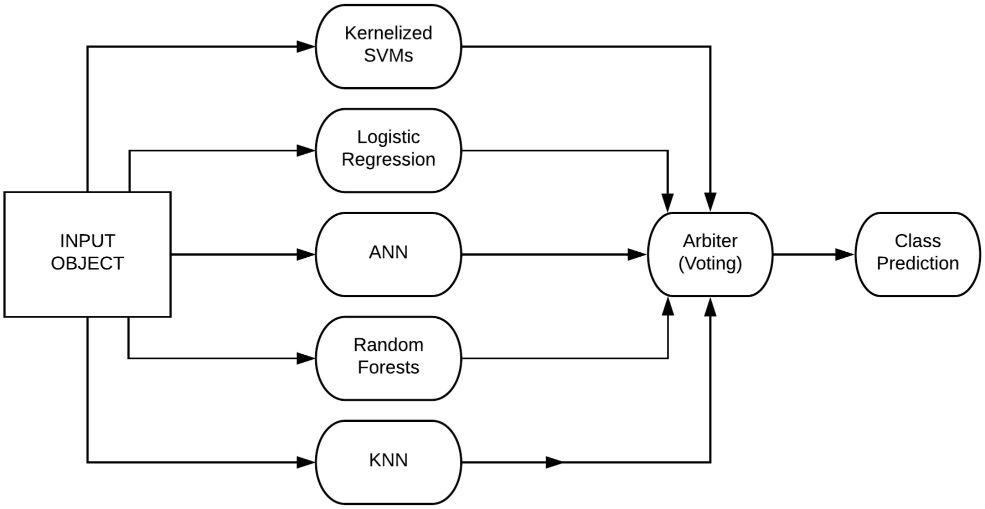

Figure 4: Ensemble model representation

Individual Algorithms used for Ensemble Model

We have tried with different combinations of individual models. The individual models used were:

- Logistic Regression
- Random Forrest Classifier
- K Nearest Neighbours
- Kernelized SVMs
- Multi Layered Perceptrons (ANNs)

Each algorithms used has its own advantages like logistic regression can learn linear separation very efficiently and similarly random forests and KNNs can learn non linear combination of linear classifiers, kernelized SVMs and ANNs can learn highly non linear curves as they transform input into Hilbert Space with infinite features.

The final model configuration that we came up with included Logistic regression,Random forest classifier, 2 polynomial kernel SVMs with degrees 5, 8 and 3 Artificial Neural Networks.

Results

Test data size was 1038, training size was 2075 and validation size was 1037. So we got our false positives, false negatives and accuracy's for the predicted vs actual labels on all training, test and validation set as follows:

Case	**Size**	**Accuracy**	**FP**	**FN**	**TP**	**TN**
Training	2075	1.0	0	0	818	1257
Validation	1037	0.9971	1	2	407	628
Testing	1038	0.9971	1	2	407	627

Here FP, FN, TP, TN are false positives, false negatives, true positives and true negatives respectively.
Although the models are different we try to justify why it is useful by showing how number of models in ensemble learning effects test accuracy and train accuracy.

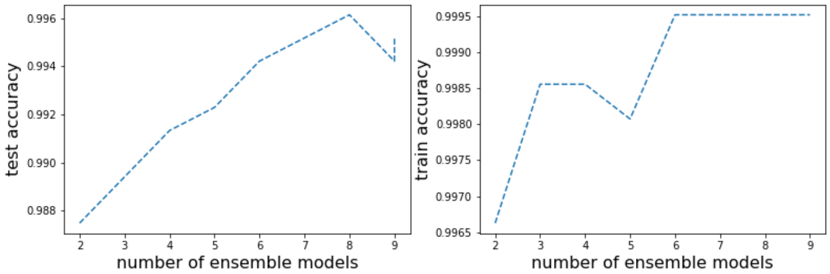

Figure 5: Number of ensemble models vs Accuracy

Observation: Test accuracy increases with increase in number of models used up to certain extent and then saturates whereas it does not effect training accuracy much.

Conclusion

We observed that the data had high correlations among 132 features only 5 features were enough to classify based on chi-square test and after transforming only 4 transformed features using principal component analysis were sufficient. So we only need to use 5 sensors.

Also we observed that using models that can model linear, nonlinear and highly nonlinear decision boundaries in ensemble method reduces over fitting and increases test accuracy and does not effect the training accuracy significantly. The accuracy obtained is better in comparison with Random Forest(CART) which has been used in [1]. So our models predicts a bit more accurate than CART with similar number of features to be tested or obtained.

References

[1] Methodology for a Security/Dependability Adaptive Protection Scheme Based on Data Mining Emanuel E. Bernabeu, Member, IEEE, James S. Thorp, Life Fellow, IEEE, and Virgilio Centeno, Senior Member, IEEE