CS631 - Home work 4 : SWAT Data Analysis

Abhishek Kumar (18111002), Aneet Kumar Datta (18111401),

Dixit Kumar (18111405), Komal Kalra (18111032),

Nitin Vivek Bharti (18111048), Riya James (18111054)

October 2018

1 Introduction

We have been provided with two data sets, one with all normal data and second with data with 6 type of attacks. Since, normal data has a lot of data points we are going to select randomly 15000 tuples from the it, which we have stored in file group2_data.csv. After reading the data from both the files we combine them as one.
To start with, we will first visualize the data in hand. The combined data has 19237 data points and 53 features. The correlation between the features can be visualized as:

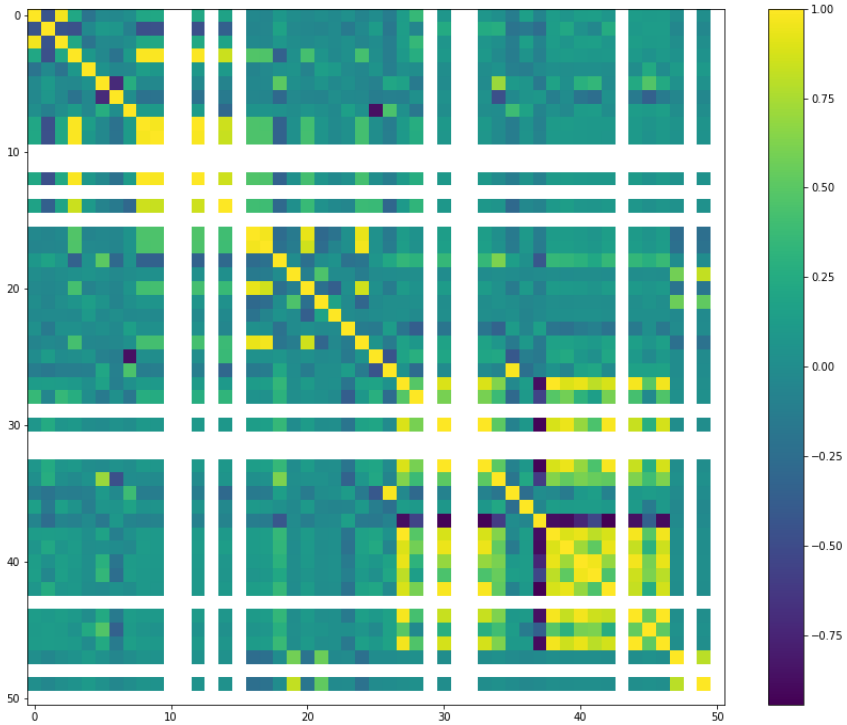

Figure 1: Correlation matrix visualization of original data

2 Feature Selection and Reduction

From visualizing the correlation between the features we can see that many features are very correlated, hence we can reduce the size of feature space.
We are selecting top 10 features from the given data using chi square method as we are trying to maximize the accuracy and we observed that with 10 features our prediction accuracy is good. So using chi-square

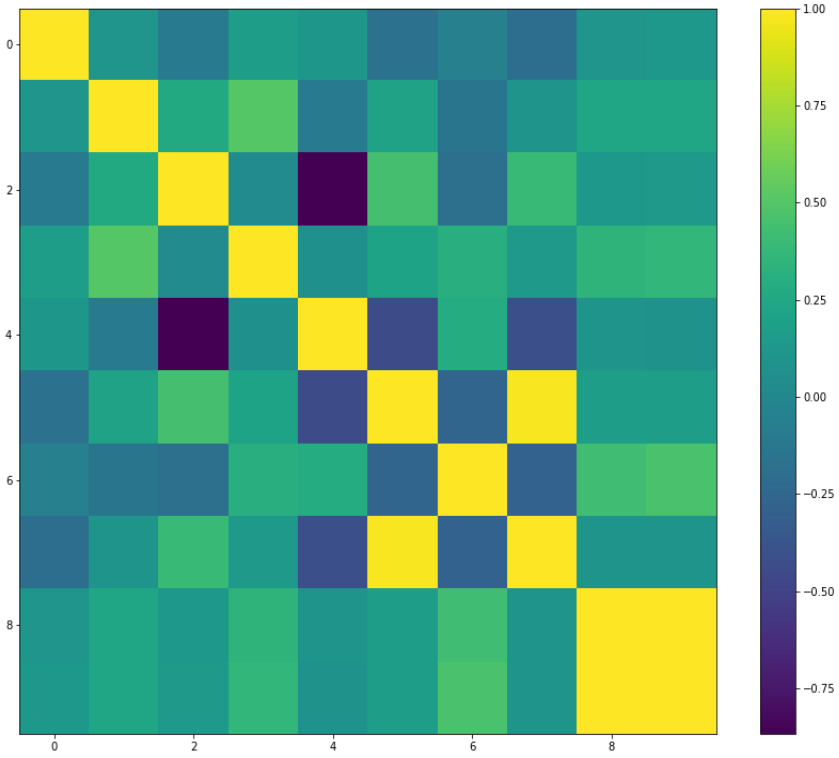

Figure 2: Correlation matrix visualization of selected data

method we are going to select top 10 features of the data.
The selected features are :

1. LIT101
2. AIT201
3. AIT203
4. LIT301
5. AIT401
6. AIT402
7. LIT401
8. AIT502
9. PIT501
10. PIT503

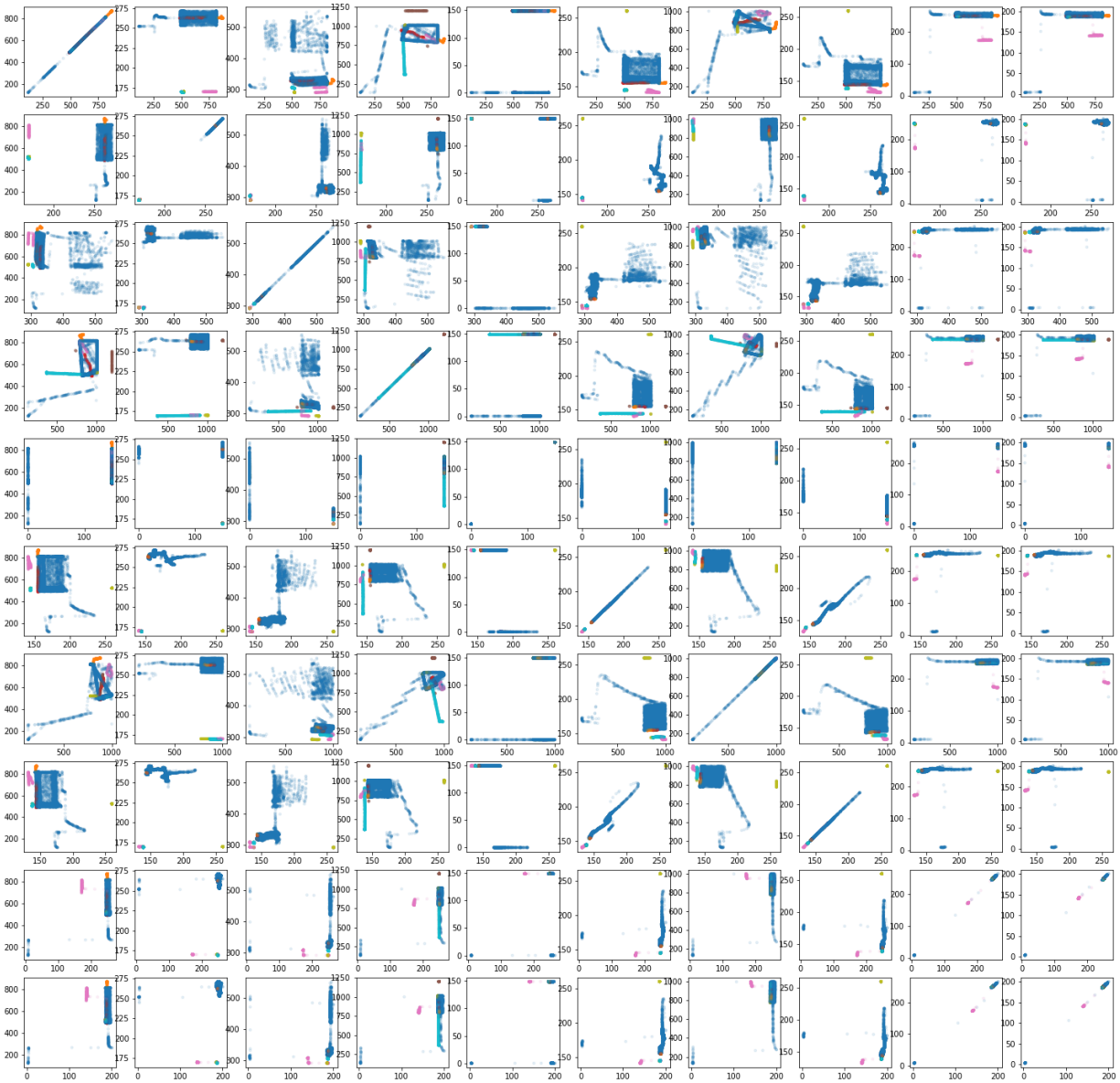

Figure 3: Scatter plot between each feature with respect to other

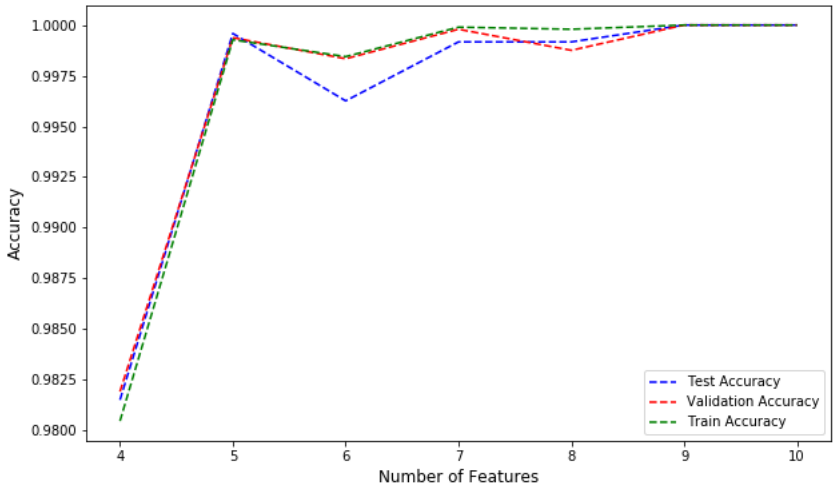

Figure 4: Accuracies with respect to number of features

3 Modelling

We divide the data into :

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

with top 10 features from original features. We are dividing the data such that ratio of data points with normal data and each type of attack data remains same in all three partitions.

3.1 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.

3.2 Models Used

We chose the models which were providing good accuracy for the test data. The models used are:

1. Logistic Regression
2. Random Forest
3. SVM with polynomial kernel of degree 5
4. Multi Layer Perceptron (Deep Learning) with limited memory bfgs solver and 3 layers with 5,5 and 2 nodes in respective layers.

Each algorithms used has its own advantages like logistic regression can learn linear separation very efficiently and similarly random forests 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.

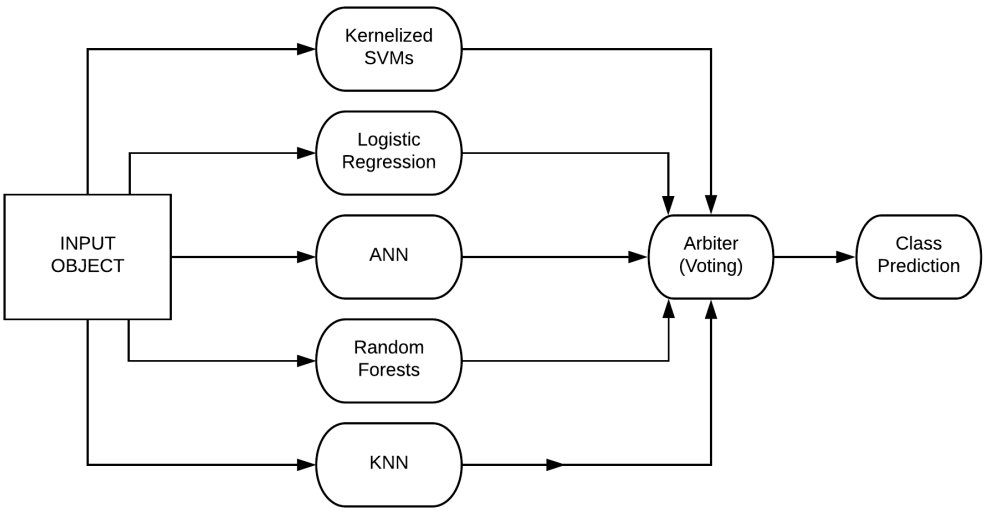

Figure 5: Ensemble model representation

4 Results

Number of tuples for partitions:

Data Type	Number
Training data	9618
Validation data	4809
Test data	4810

Table 1: Number of tuples for for each partitions

By applying our model on the above data and iterating it multiple times, in maximum instances of splits we observed accuracy for the predicted vs actual labels on all training, test and validation set as :

Case	**Number of Tuples**	**Accuracy**
Training	9618	1.0
Validation	4809	1.0
Testing	4810	1.0

Table 2: Accuracy Table

The confusion matrices are :

Attack1	Attack2	Attack3	Attack4	Attack5	Attack6	Normal
470	0	0	0	0	0	0
0	221	0	0	0	0	0
0	0	215	0	0	0	0
0	0	0	234	0	0	0
0	0	0	0	141	0	0
0	0	0	0	0	837	0
0	0	0	0	0	0	7500

Table 3: Confusion matrix for Training data

Attack1	Attack2	Attack3	Attack4	Attack5	Attack6	Normal
235	0	0	0	0	0	0
0	111	0	0	0	0	0
0	0	107	0	0	0	0
0	0	0	117	0	0	0
0	0	0	0	70	0	0
0	0	0	0	0	419	0
0	0	0	0	0	0	3750

Table 4: Confusion matrix for Validation data

Attack1	Attack2	Attack3	Attack4	Attack5	Attack6	Normal
235	0	0	0	0	0	0
0	111	0	0	0	0	0
0	0	107	0	0	0	0
0	0	0	118	0	0	0
0	0	0	0	70	0	0
0	0	0	0	0	419	0
0	0	0	0	0	0	3750

Table 5: Confusion matrix for Test data

4.1 Observation

By adding more models with different parameters we saw that the accuracy starts reducing. Random forests gives us accuracy of 100% in most cases when used alone. To avoid over fitting by Random forest we are using some more models along with that which does not over fit (accuracy $< 100\%$) in ensemble model. The combination reduces the probability of over fitting the given data.

5 Conclusion

We observed that given data had high correlation among features and could be represented in less feature space,i.e, from 52 to 10 and last one being state of machine(Attack / Normal). We used the reduced data for training and testing purpose.
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.