Impact of Data Normalization on Global Terrorism Database

Ranjit Kathiriya

Abstract: Machine learning is an rising field in computer innovation. In this research Assignment first, the data analysis is done and extracting the important features or truncating the unwanted features. After that, the data has fed into the different models for obtaining the accuracy for model selections. Then, the evaluation for the research part is gained with different data normalization techniques like min-max, standard, Tanh estimators and many more. At last, the hyperparameters tuning processes to acquire more accuracy for evaluating our model.

1 Introduction

In this project, I have selected the Global Terrorism Database (GTD) from Kaggle[1]. The dataset is from 1970 to 2017, and it is also open source. The National Consortium maintained this dataset for the Study of Terrorism and Responses to Terrorism (START). This dataset contains 180,000 attack held in the world in a particular period. This dataset is too huge for my Assignment, so i have decided to pick a chunk of it based on top 20 terror groups, and i also have selected important features from 135 columns.

My Assignment is based on a classification model means based on features we have to predict target value. In our case based on features, like country, target type, weapon type, city etc. the model will predict the group responsible for the terrorist attack. I have picked this project because i have listened much time how brutally the terrorist has made an attack on 26th November in different parts of Mumbai and the shocking thing was 166 were found dead by only nine terrorists. Then after have decided to analyze the dataset and find some unique thinks based on that group who has organized an attack. My other motivations for picking this project was the attacked on 11th September almost 3,000 thousand people lost their lives in the world trade center.

There are many research paper based on this like Predict Terrorism and Threat, and Prevention [2] and Prediction for Combating Terrorism [3], but up till now no one has found about the group responsible for upcoming terrorist attacks based on a specific technique, date, weapons or place. By the end of 2017, there were more then 300 terrorist groups were active, and the attack structure was almost similar to a particular group.

2 Research [Section 2.6 of code]

This part, the various form of data normalization techniques will be discussed. Then after the analysis will be done on each technique of normalization for achieving good accuracy score. There are many techniques for normalization of data apart from all only fews are to be discussed.

2.1 Min-Max Normalization

The data normalization using min-max normalization linearty. The data are normalized into a particular range like in between 0 and 1 or -1 to 1. This normalization scale the dataset in the form of min and max. All data are scaled and not a single dataset will be considered as an outlier. In sklearn, there is a predefined class named MinMaxScaler[4].

$$Yi = [Xi - \min(X)]/[\max(X) - \min(X)]$$

Where, Xi is i th data point and min represents the minimum and Maximum represents maximum. So Xi converts to Yi

2.2 Z-Score Normalization

Z- score is a machine learning technique for resolving the outlier issue. It performs normalization based on mean and standard deviation. In this the value is above the mean means that is a positive number or else it is a negative number.

$$Zi = (Xi- mu)/\sigma$$

Here σ represents standard deviation and mu represents mean and Zi is the value of Xi after standardization.

2.3 Max-ABS Normalization

MaxAbs Normalization values are mapped in the range of 0 and 1. It behaves the same as MinMaxScalar. It scales every feature by maximum absolute value.

$$Yi = Xi / X.max()$$

Here the value of Xi is divided by X max value and store it into Yi.

2.4 Median Normalization

This normalizes of each data is done by taking the mean of a feature and then dividing it with particular data. This technique is most useful while performing distribution. It overcomes the problem of outliers.

$$Yi = Xi / mean(X)$$

Here the value of Xi is divided by mean of X and store it into Yi.

2.5 Sigmoid Normalization

This scaling method is very simple most of the data is normalized with this method. In this method, the sigmoid curve data scaling is done.

$$Yi = 1/(1+ exp(-Xi))$$

Here, expansional e to the base Xi + 1 and this divide by 1 and output is store in Yi respectively.

2.6 Tanh estimators

Tanh estimator is a highly efficient method. It was introduced by Hampel. It scales the series data and element the outliers.

$$Yi = 0.5[\tanh(0.01(Xi-\mu)/\sigma) + 1]$$

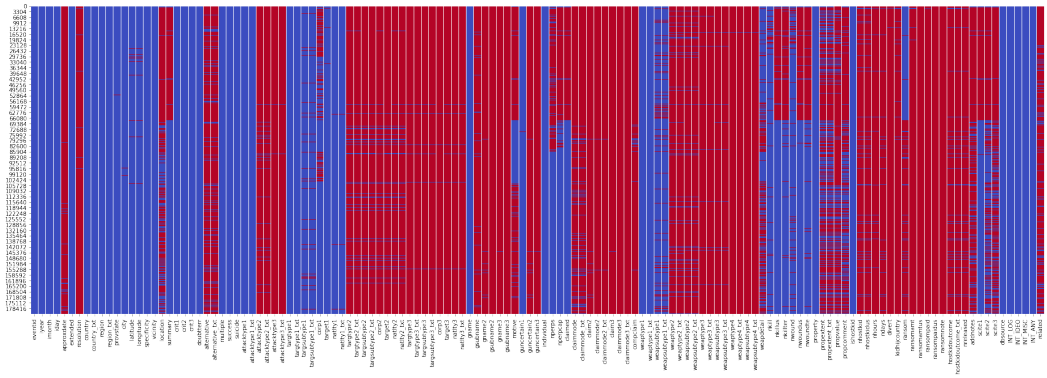
Where μ is the mean value and σ is the standard deviation and Xi is the data and store into Yi

3 Methodology

This parts focuses on the methods and tools used to obtain accuracy. Further, how data is observed and fixed using different parts of pre-processing and observation of an accurate model is also described in this section. This section is divided into four parts 1) analysis, 2) Pre-processing, 3) Model Selection and 4) Hyper Parameter Tuning.

3.1 Analysis of Data

Firstly, the missing values have found based on the heat map and then have extracted features based on this. Secondly, observation of every column and checking the percentages of lost data of each feature. if the data is more then 40% missed then elements that columns. The below graph shows the feature here red line means missing and blue means the data is present on that feature. As it is to be seen that more number of data are missing, if the data is less messing further the missing value is filled using pre-processing section. In code the analysis of this is observed in 1 part. The dataset was to big so based on top 20 groups who were responsible for terror attack that data is only picked.



3.2 Pre-Processing

For obtaining a good accuracy and f-score the pre-processing is an important step. In this section there are much technique is used like feeling the missing values, handling the categorical data, and scaling the data that is to be discussed in 2 sections of this document.

3.2.1 Dealing with Missing Values [Section 2.1 of code]

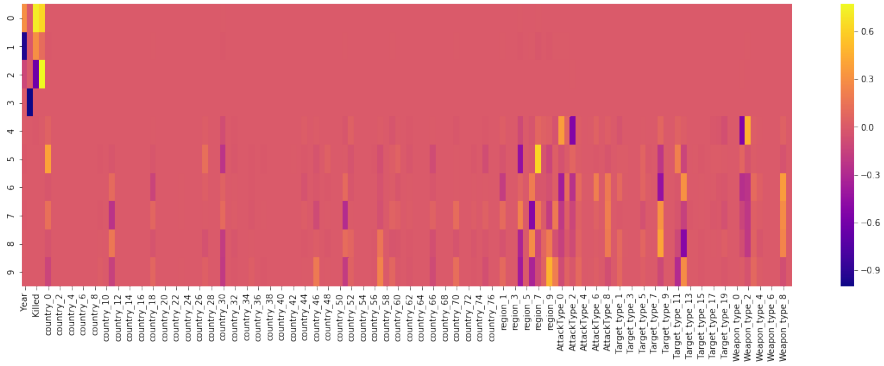
Firstly, from the given dataset there were so many missing value some were adjusted based on the mean or medium but the column like city were filled manually from GPS location. The column name Killed and Wounded were filled by taking the mean of of all killed and wounded based on attacked type. At last the columns which has more categorical data like type which was removed because while applying categorical encoding it would be much difficult for it.

3.2.2 Handling Categorical Data [Section 2.2 of code]

In this the mostly data are in the form of categorical, how to deal with categorical data is to be discussed. Mainly for dealing with this data the technique named one-hot encoding is used so in this all data will become in the form of 1 or 0 concerning unique date into the column. For group column Ordinal Encoder is used so we can get the group of terrorist organization who is responsible for the attack and this will be in the form of numbers.

3.2.3 Dimensionality Reduction [Section 2.5 of code]

After one-hot encoding the dimensional were huge and it was also time taking while training the model. So, the most important step was to dimensionality reduction like for example ten feature to two feature. In is to be observed with many PCA components but by 10 the accuracy score is good. The components less then 7 gives very bed accuracy score in all model.



3.2.4 Scaling Data [Section 2.6 of code]

In scaling the data are scaled and this was my research topic more explanation is given in section 2 of the documents. In this the research paper named Impact of Data Normalization on Stock Index Forecasting [5] the various technique for normalizing the data is there this technique is to be used in code each formula is maintained into the code and above in section 2.

3.3 Model Selection [Section 3 of code]

In this I have combine my research work with the model selection. I have taken 9 different model like SVM, KNN and so on with including each normalization technique like Min-man, Z-score etc. From given table below the accuracy is observed for each model with different normalization technique. I have also user cross value score for running my model into 3 different iteration and taking the mean of all three accuracy.

Model / Normalization	Accuracy
Logistic Regression	0.636
KNN	0.255
Naive Bayes	0.493
SVM	0.311
Random Forest Classifier	0.701
Decision Tree Classifier	0.628
Ridge Complexity	0.519
Stochastic Gradient Descent	0.351
Multi-layer Perceptron	0.662

3.4 Hyper Parameter Tuning [Section 4 of code]

Based on the above table the accuracy is observed for all normalization and model. From this I am going to select best 3 model including normalization method and Hyper parameter is optimized concerning this model. Above table we observed the most accuracy and the best hyper parameter tuning is given below of each model. I have not done full tuning because it was time taking.

3.4.1 KNN [Section 4.1 of code]

Best hyper parameter tuning for this is:

```
{'n_neighbors': 1,
'p': 2,
'weights': 'uniform'}
```

3.4.2 Random Forest [Section 4.1 of code]

Best hyper parameter tuning for this is:

```
{'n_estimators': [200, 205, 250],
 'max_features': ['auto', 'sqrt'],
 'bootstrap': [True, False]}
```

3.4.3 SVM [Section 4.3 of code]

Best hyper parameter tuning for this is:

```
{'kernel': ['linear', 'poly', 'rbf', 'sigmoid'],
'C': [1.0, 2.0, 1.5, 2.5]}
```

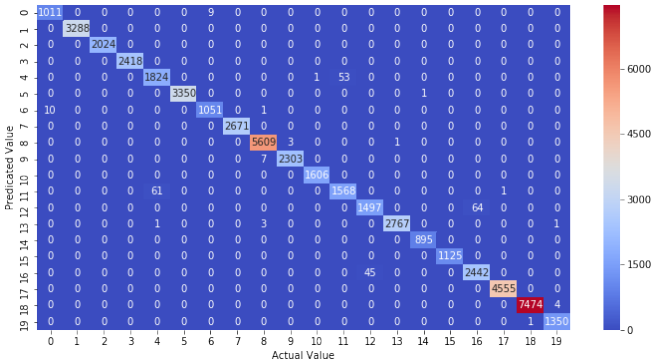
4 Evolution

In a different model, the model perfection is not only based on the accuracy but precision, f1 score, and classification matrix, the confusion matrix plays an important role.

Model	Before Hyper parameter tuning	After Hyper Parameter Tuning
KNN	0.73	0.833
Random Forest	0.701	0.721
SVM	0.71	0.742

From this table, it is to be observed that accuracy is increased if the hyperparameter tuning is applied. From this above table in KNN, almost a 10% increase is found while on the other model 2-3% of the increase is there.

Confusion matrix of KNN model:



4.1 Impact of research :

Model / Normalization	**Without Normalization**	Min-Max	Z - Score	Max-ABS	Median	Sigmoid	Tanh	**Accuracy After Hyper tuning (3 only)**
Logistic Regression	0.636	0.600	0.640	0.612	0.505	0.601	0.254	-
KNN	0.255	0.739	0.695	0.731	0.601	0.562	0.695	0.833
Naive Bayes	0.493	0.493	0.493	0.493	0.493	0.545	0.496	
SVM	0.311	0.595	0.717	0.647	0.149	0.604	0.146	0.742
Random Forest Classifier	0.701	0.689	0.698	0.699	0.695	0.696	0.696	0.721
Decision Tree Classifier	0.628	0.633	0.628	0.625	0.625	0.644	0.627	-
Ridge Complexity	0.519	0.518	0.518	0.519	-7.380	0.525	0.500	-
Stochastic Gradient Descent	0.351	0.551	0.589	0.547	0.200	0.593	0.339	-
Multi-layer Perceptron	0.662	0.698	0.774	0.774	0.669	0.679	0.532	-

I have watched from the taking after table that the normalization of information is critical since the precision influences this. For distinctive show there's diverse normalization strategy is effective like for KNN demonstrate Min-Max or Tanh method is nice. For great exactness with regard to normalization procedure at that point tuning is done and the more exact precision is picked up.

5 Conclusion

In this paper describe about different classification algorithm performed with Machine Learning technique. The dataset was the challenging task because of every record was missed or in the categorical form. In this the data cleaning is most important because if the data is not cleaned properly then it won't be optimized further. Then after pre-processing of data like cleaning, categorical feature extraction, and then I have trained with different model. But after normalizing the dataset I have seen a good jump in an accuracy so have check with different types of normalization of data method. The dimensionality reduction technique is applied for extracting the data. To extend the model beyond the wall, the Hyperparameter Tuning was done in this the model check every different parameter and gives the best score.

References [1]

1. Kaggle.com. (2019). Global Terrorism Database. [online] Available at: https://www.kaggle.com/START-UMD/gtd.
2. Ijeat.org. (2019). Using Global Terrorism Database (GTD) and Machine Learning Algorithms to Predict Terrorism and Threat. [online] Available at: https://www.ijeat.org/wp-content/uploads/papers/v9i1/A1768109119.pdf.
3. Arxiv.org. (2019). A Conjoint Application of Data Mining Techniques for Analysis of Global Terrorist Attacks Prevention and Prediction for Combating Terrorism. [online] Available at: https://arxiv.org/pdf/1901.06483.pdf.
4. Scikit-learn.org. (2019). sklearn.preprocessing.MinMaxScaler — scikit-learn 0.22 documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html?highlight=minmax#sklearn.preprocessing.MinMaxScaler.

5. Mirlabs.org. (2019). Impact of Data Normalization on Stock Index Forecasting. [online] Available at: http://www.mirlabs.org/ijcisim/regular_papers_2014/IJCISIM_24.pdf [Accessed 4 Dec. 2019].
6. Research.ijcaonline.org. (2019). [online] Available at: https://research.ijcaonline.org/volume32/number10/pxc3875530.pdf [Accessed 6 Dec. 2019].