Compare support vector machines to a 3 layer neural networks with Titanic dataset

Gaurav Shrivastava
Dept. of AI and Machine Learning

AI-Tech Systems

ai-techsystems.com

Morena,India

Gaurav.shrivastava.15csc@bml.edu.in

***Abstract*— this paper shows the comparison of SVM to 3 layer Neural Network with titanic dataset. As we all know the sinking of titanic ship is one of big tragedy in the history which led to killing about 1500 passenger. Even now researcher are trying to what factor could lead more passenger to survival. In this paper. I have implemented the comparison b/w SVM and 3 layer Neural Network to predict survival of passenger.**

Keywords—Support Vector Machine (SVM), Neural Network (NN), style, Python, Classification, Titanic dataset.

I. INTRODUCTION

This paper shows the comparison of SVM to 3 layer Neural Network with titanic dataset. As we all know the sinking of titanic ship is one of big tragedy in history which led to killing about 1500 passenger. Even now researcher are trying to what factor could lead more passenger to survival. In order to find out, They are trying different algorithms such as Logistic Regression, Random Forest etc. But we are implementing Support Vector Machine and 3 layer Neural Network with titanic dataset. We can find this dataset on Kaggle where it is available for all people. Then we will compare which algorithms work best on the basis of accuracy

II. DATASET

We can find our dataset on kaggle website.in this website we can also find a lot of dataset besides this Titanic dataset. Titanic dataset contain all the necessary detail about passenger that were board on that day. Such as their Age whether they were adult or child, their Sex whether they were male or female, their Survival whether they were survive or not, Passenger Id, Name, of the passengers Number of parents and children, Price of the tickets, Cabin Number, Port of Embarkation. Dataset contains two files named as training data, testing data respectively. Training data contain 891 rows and 12 columns while testing contain 418 rows and 12 columns. Both file are in the format of CSV (Comma Separated value).

	PassengerId	Survived	Pclass	Name	Sex
0	1	0	3	Braund, Mr. Owen Harris	male

Figure 1 Titanic dataset

Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
22.0	1	0	A/5 21171	7.25	NaN	S

Figure 2 Titanic dataset continuation.

III. METHODOLOGY

Our Approach to build the project is divided into three points. First is to preprocess the data. This is most important step in order to build any type of good model, we can say this Foundation of our project the better we process the data the better we can get our accuracy then we will apply SVM and 3 layer Neural Network and at the end, we will compare with each other in order to find which algorithm work best.

.

A. *Data Preprocessing*

Data Preprocessing is one of the most important part in the machine learning. The developer has to spend more time on data preprocessing in order to gain more benefits. There are conditions where we need to do preprocess data. Following are some of them.

- Missing value in the dataset.
- Categorical variable in the dataset.
- Noisy data
- Data Transformation(Normalization)

Above given are some of those conditions where we need to apply preprocessing in order to build a good machine learning model. So we performed data preprocessing in our titanic dataset. Following are place where we applied data preprocessing.

- In the Age columns we found missing values. There were 177 and 86 missing values in the training and testing dataset respectively and Embarked column has 2 missing value. After finding them we solve the problem. For example

```
x_train["Age"]= x_train.fillna(x_train["Age"].median())
x_test["Age"]= x_test.fillna(x_test["Age"].median())
```

```
x_train['Embarked'] = x_train['Embarked'].fillna('S')
```

- In the Sex and Embarked Column, There were categorical variable so we changed them integer 0 and 1.For example

```
d = {'male':0,'female':1}
x_train["Sex"]= x_train["Sex"].apply(lambda x:d[x])
x_test["Sex"]= x_test["Sex"].apply(lambda x:d[x])
```

```
d = {'S':0,'C':1,'Q':2}
x_train["Embarked"]= x_train["Embarked"].apply(lambda x:d[x])
```

- In the Age and Fare column, there are lots of fluctuation in the values so we binning the numerical values into ranges.

```
# Binning numerical columns
x_train['Age'] = pd.qcut(x_train.Age, q=4, labels=False )
x_train['Fare']= pd.qcut(x_train.Fare, q=4, labels=False)
```

After performing all the data preprocessing. On the dataset, we got the data in the below given form.

	Pclass	Sex	Age	Fare	Embarked
0	3	0	0	0	0
1	1	1	3	3	1
2	3	1	1	1	0
3	1	1	2	3	0
4	3	0	2	1	0

Figure 3 processed data

B. *SVM(Support Vector Machine)*

SVM stands for Support Vector Machine.it is supervised machine learning algorithm. It is very powerful algorithm for classification that not only aims to classify data but it also aims to find best possible boundary, namely that maintain the max dist. b/w points .Since we know it is good for classification. Then it's time build up SVM model. We can build it with help of sklearn library.

```
from sklearn import svm
classifier = svm.LinearSVC()
```

```
classifier.fit(X_train,y_train)
```

As Above mentioned, we just need to implement library then just follows the predefined steps for library, After that we can get our accuracy will help of .predict() function. Our classification accuracy is 78.18%.

C. *3 layer Neural Network*

Neural network are designed based on human brain .it consists of densely interconnected neurons which are connected to each other. So neural network are designed based on this parallel computing..3 layer perceptron means multi-layer perceptron where one is input layer, other is hidden layer and last one is output layer. We can use 3NN with the Keras Library.

```
import keras
from keras.models import Sequential
from keras.layers import Dense
```

```
classifier = Sequential()
```

- Input Layer:

  classifier.add (Dense (output_dim = 3,

  init = 'uniform', activation = 'relu', input_dim = 5))

- Hidden Layer

  classifier.add (Dense (output_dim = 2,

  init = 'uniform', activation = 'relu'))

- Output layer

  classifier.add(Dense(output_dim = 1,

  init = 'uniform', activation = 'sigmoid'))

- compile Layer

  classifier.compile (optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])

- Fit Layer

  classifier.fit (X_train, Y, batch_size = 10, nb_epoch = 100)

  After completing the all predefined steps, we will just need to find accuracy which is 86.12%

CONCLUSION

Our SVM's accuracy is 78.18 while 3NN's accuracy is 86.12.So 3NN is better model than SVM.

REFERENCES

[1] 1. Kaggle, Titanic: Machine Learning Form Disaster [Online]. Available: http://www.kaggle.com

[2] 2. Vyas, K., Zheng, Z. and Li, L, "Titanic-Machine Learning From Disaster," pp. 6, 2015. K. Elissa, "Title of paper if known," unpublished

.

[3] 3. Cicoria, S., Sherlock, J., Muniswamaiah, M. and Clarke, L, "Classification of Titanic Passenger Data and Chances of Surviving the Disaster," pp. 4-6, May 2014..

[4] 4.Corinna Cortes, Vlasdimir Vapnik, "Support-vector networks", Machine Learning, Volume 20, Issue 3,pp 273- 297.