Machine Learning

Traffic accident data multiclass classification with Support Vector Machines (SVMs)

Guillermo Villanueva Benito

Thrusday, 16 June

Contents

1 Abstract 2

2 Introduction 3

2.1 SVMs for multiclass classification problems . . . 3

2.1.1 One-versus-all (OVA) . . . 3

2.1.2 All-versus-all (AVA) . . . 3

2.2 Traffic accident data set . . . 3

2.2.1 Process data . . . 4

3 Software 4

4 Results 5

5 Conclusion 7

1 Abstract

Multiclass classification is the problem of classigying instances into one of three or more classes. The multiclass classification problem can be solved by naturally extending the binary classification technique for some algorithms, [1].

Traffic accidents are classified in 4 classes according to its impact on traffic, [2]. In this project, we apply two classification algorithms (kernel SVMs and random forest) to the traffic accident data set from [2]. The goal is to predict the degree of impact of an accident given the value of a set of explanatory variables and study the performance of the predictors.

2 Introduction

In this section, we introduce how multiclass classification problems are managed in SVMs. Moreover, the briefly describe the traffic accident data set.

2.1 SVMs for multiclass classification problems

Multiclass classification problems can be naturally tackled by extending the binary classification technique for some algorithm, such as support vector machines (SVMs), [1]. Strategies for reducing the problem of multiclass classification to multiple binary classification problems are categorized into *one vs all* and *all vs all*.

2.1.1 One-versus-all (OVA)

For this approach, each problem discriminates a given class from the other $K - 1$ classes. It is required K binary classifiers to complete the $K-$class classification problem.

2.1.2 All-versus-all (AVA)

In this approach, each class is compared to each other class. A binary classifier is built to discriminate between each pair of classes, while discarding the rest of the classes. This requires building $K(K-1)/2$ binary classifiers.

Python library scikit-learn implement SVM for muticlass classification following the *all-vs-all* approach. The built-in function *SVC* solves the problem.

2.2 Traffic accident data set

The data set has been downloaded from `https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents` and was presented in [2].

It is a countrywide car accident data set, which covers 49 states of the USA. The accident data are collected from February 2016 to Dec 2021. There are about 2.8 million accident records in this data set.

For each car accident, there are 47 data attributes, ranging from categorical/integer to real attributes. The *Severity* of each accident (a number between 1 and 4) defines the class to which it belongs. Class 1 corresponds to car accidents with low impact on traffic (i.e. short delays) while class 4 indicates car accident with significant impact on traffic (i.e. long delays).

We consider only a subset of the explanatory variables. Moreover, all explanatory variables considered are real. The following table illustrates them.

Explanatory variables	
Start_time	Shows start time of the accident in local time zone.
Start_Lat	Shows latitude in GPS coordinate of the start point.
Start_Lng	Shows longitude in GPS coordinate of the start point.
Temperature(F)	Shows the temperature (in Fahrenheit).
Humidity(%)	Shows the humidity (in percentage).
Visibility(mi)	Shows visibility (in miles).
Wind_Speed(mph)	Shows wind speed (in miles per hour).
Precipitation(in)	Shows precipitation amount in inches, if there is any.

Therefore, an overall of 8 real explanatory variables have been considered.

Applying two classification algorithms, we explore one of numerous applications of this data set: real-time accident prediction. We have not found any previous work tackling this prediction task.

2.2.1 Process data

Before solving the multiclass classification problem, the data set has been processed and mainly two tasks have been done: drop rows with missing values and convert the *Start_time* variable (string) into a real variable.

3 Software

We have decided to use python to carry out the multiclass classification problem. Firslty, because of its popularity among data scientists and secondly, because python has not been used throughout the course. Thus, we find this final project an ideal opportunity to implement a kernelized algorithm in python.

4 Results

Due to the large size of the data set and in order to avoid computational issues we have only considered 20000 random samples. Moreover, a quick inspection on the data set shows that classes are not proportionally represented. Fig. 1, shows the number of samples from each class in the whole data set.

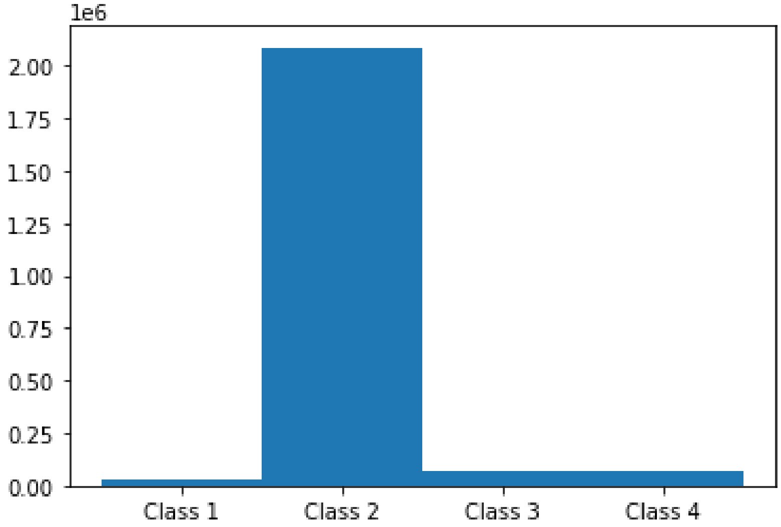

Figure 1: Number of samples from each class in the data set.

Train and test sets have been created taking into account the fact the class 2 is, by large, the most represented class in the data set. Thus, we have created the train and validation sets ensuring that each class is equally represented in these sets. Therefore, in each set around 1/4 of the samples correspond to each class.

Table 1 shows the obtained confusion matrix for the 4-class classification problem obtained by applying kernel SVM with the *RBF* kernel. Unfortunately, Table 1 shows that the solution obtained to the prediction task consist in classifying all samples to the same class. The accuracy of the prediction is around $1/4$. We notice that this accuracy is the expected one if we would assign classes to samples randomly.

		Predicted Classes			
		Class 1	Class 2	Class 3	Class 4
Actual Classes	Class 1	0	0	0	1000
	Class 2	0	0	0	1000
	Class 3	0	0	0	1015
	Class 4	0	0	0	985

Table 1: Confusion Matrix kernel SVM

Preliminary results show that the *RBF* kernel SVM does not seem a good approach to study this data set, since the random classification is as good as the prediction learned. We notice that it might be that the considered classification problem is not learnable. In order to verify learnability we also apply to the data set a simple random forest to see whether or not its performance is better.

Table 2 shows the obtained confusion matrix obtained when applying a simple random forest to the data set. The prediction is better and the accuracy has increases until around 0.5. Moreover, when we try to predict one class, the algorithm doubts between the correct class and the adjacent class. Therefore, the clearly the random forest outperforms the kernel SVM for the proposed prediction task.

		Predicted Classes			
		Class 1	Class 2	Class 3	Class 4
Actual Classes	Class 1	472	412	114	2
	Class 2	41	472	456	31
	Class 3	24	184	779	28
	Class 4	18	150	667	140

Table 2: Confusion Matrix random forest

5 Conclusion

Two classification algorithms have been tested to predict the degree of severity (4 classes) of a car accident given some explanatory variables: the kernel SVM and the random forest. It is worth mentioning that no previous work have been found doing a prediction of this type and that we did not know whether or not the prediction task as stated in this project was learnable,

One the one hand, the kernel (*RBF*) SVM has been tested and no satisfactory results have been found. The learned predictor in this case consists in assigning the same class to all samples. We have obtained the same performance as a random predictor.

On the other hand, the random forest has been tested to check the learnability of the prediction. In this case, a better performance has been obtained. The accuracy of the 4-class classification problem has increased until almost 0.5. We have also analyzed the prediction given by the random forest and have obtained a remarkable finding: the random forest classifier doubt whether or not choose the correct class or the adjacent class. This is an encouraging finding since it means that the random forest classifier is able to detect the severity, but not with the enough precision so as to classify correctly between adjacent degrees of severity.

In conclusion, we have firstly found that the kernel SVM seems not to be the proper classifier for predicting the severity of a car accident on traffic. And secondly, the proposed prediction task seems to be learnable as for the results when applying the random forest classifier. These results might be useful to predict the severity of a car accident in real time given some basic features of the car accidents, such as the location, time or weather conditions. Its application for authorities to take one action route or another, given a car accidents, are of big impact.

Finally, we notice that more work should be done to have a better predicting tool, such as assessing variable importance or tuning hyperparameter to get a better performance. Moreover, more kernel functions could be tested and more study should be done to know why the rbf kernel SVM does not work for this data set.

References

[1] Mohamed Aly. Survey on multiclass classification methods. *Technical Report, Caltech.*, pages 833–842, 01 2005.

[2] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. A countrywide traffic accident dataset. *CoRR*, abs/1906.05409, 2019.