Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 62 (2022) 221–228

Transportation Research Procedia

www.elsevier.com/locate/procedia

24th EURO Working Group on Transportation Meeting, EWGT 2021, 8-10 September 2021, Aveiro, Portugal

Using Supervised Machine Learning to Predict the Status of Road Signs

Roxan Saleh[ab*], Hasan Fleyeh[b]

[a]Swedish Transport Adminstration, Röda vägen 1, SE-781 89 Borlänge, Sweden
[b]School of Information and Technique, Dalarna University, Borlänge, Sweden

Abstract

There is no data collected and saved about road signs in Sweden and the status for these signs is unknown. Furthermore, the status of the sign colors, the quality of the sign, the type of the retroreflection material, and age of the road signs are unknown. Therefore, it is difficult to know the status (approved or not) of any road sign without performing a costly inspection. The aim of this study is to predict the status of the road signs mounted on the Swedish roads by using supervised machine learning. This study investigates the effect of using principal component analysis (PCA) and data scaling on the accuracy of the prediction. The data were prepared before using then scaled using two methods which are the normalization and the standardization. The three algorithms that tested in this study are Random Forest, Artificial Neural Network (ANN), and Support Vector Machines (SVM). They are invoked to predict the status of the road signs. The algorithms exhibited overall high predicting accuracy (98%), high precision (98%), high recall (98%), and high F1 scores (98%). Random forest showed the best performance with 4 PC components on the normalized data with a highest accuracy of 98%. Using PCA showed different impacts on the performance of different techniques. In the case of ANN, invoking PCA improves the accuracy, while for SVM the accuracy decreases when PCA is used. On other hand, PCA has no effect on the accuracy of the random forest model when scaling is invoked. The effect of the data scaling using normalization and standardization is also investigated in this study, and it is noticed that scaling of the data increases the accuracy of the prediction for all the three models (ANN, SVM and Random Forest). Furthermore, better accuracy is achieved when the standardization is invoked compared with normalization.

Keywords: Road signs; supervised machine learning, principal component analysis, prediction

1. Introduction

Road safety plays an important role in the transportation systems (Madleňák et al. 2018) and road signs play a crucial role in the safety and efficiency of any road (Berces and Robertson 2012). The low visibility of the road signs

10.1016/j.trpro.2022.02.028

causes a high risk for accidents and cause material damage, human injuries and even deaths (Ž. Šarić et al. 2018) . Therefore, road signs must be recognizable, and the information given on the signs should be read from an adequate distance.

The colors of each road sign help the drivers to recognize the signs and distinguish them from each other in the daytime. It is even more important to recognize and read the signs at nighttime when no or only little light is available. In the nighttime, the retroreflective sheeting material on the road sign helps to increase visibility (Khrapova 2019). For this reason, the retroreflective material used for the road signs should meet an accepted level of retroreflectivity. The type of retroreflective material and its color has a significant influence on the degree of degradation under different conditions (Khrapova 2019).

Predicting the status of road signs is an important issue and can be used for different purposes. Judging if the road sign is approved or not, from the visibility point of view, helps in deciding whether to replace the road sign or not.

There is limited information on the status of the road signs, their placement, and conditions in Sweden. It is even unknown if the existing retroreflective road signs on the Swedish roads meet the requirements for the acceptable visibility performance or not. No inventory of road signs mounted on state roads is done and no registry for road signs type, date of mounting or material type are available (Kjellman, Fors, and Lundkvist 2018).

The classical approach is to evaluate the condition of road signs through inventory the retroreflectivity and color coordinates and then compare the invented values with the accepted levels in the regulations. Problems arise then within the manually compilations that take time and can cause fail evaluations. This paper presents an alternative solution in terms of machine learning algorithms for accurate and fast evaluation of road signs. Therefore, this paper aims to predict the status of the road signs by using three different machine learning algorithms which are Random Forest, Artificial Neural Network (ANN), and Support Vector Machines (SVM). The performance of these algorithms is compared, and their predictions are evaluated. The data invoked to train the learning algorithms comprises measurements of different features such as sign color values, retroreflection, age, placing of the sign, and its direction, and the learning algorithm should predict whether the road sign is approved or not.

2. Literature review

One of the recent studies from 2020 pointed out the importance of the retroreflectivity and luminance on improving the detectability and legibility of the road traffic signs (Salih and Fleyeh 2021). Furthermore, this study found out that the retroreflectivity depends on vehicle factors (headlights color and angle of illumination), environmental factors (weather and ambient conditions) and sign factors (type and color of retroreflective sheeting material).

The retroreflective material was developed since 1930 (Berces and Robertson 2012) and used for road signs since 1970. Almost all the road signs that are in use today contain retroreflective material. The retroreflective material used for the road signs have been developed recently and a new technique is used to manufacture this material.

Previous research has shown poor correlations and predicted unrealistic service life on the basis on the minimum retroreflectivity levels (Ré, Miles, and Carlson 2011) (Babić, Ščukanec, and Fiolić 2016). One of the problems was the significant variability in the environmental conditions when the sign data are collected (Swargam 2004).

A literature review, presented in a paper from 2011, showed that developing a robust prediction of traffic sign retroreflectivity based on in-service measurements can be difficult (Ré, Miles, and Carlson 2011).

Babić et al conducted a study to develop a model to predict the status of traffic signs retroreflectivity based on their age. The linear models for predicting status of signs developed using binary logistic regression. Statistical analysis showed that the age of traffic signs contributes significantly to the predictive ability of the model. At the same time, the study concluded that functional service life of the sign could not be sufficiently accurately determined based only on sign age (Babić, Ščukanec, and Fiolić 2016).

3. Data pre-processing

The data used in this study was collected by Road and Transport Research Institute (VTI) in Sweden. The purpose of the collected data was to analyze the life cycle costs for road signs (Kjellman, Fors, and Lundkvist 2018). The dataset consists of different parameters including retroreflectivity values, the color values measured on different locations on the sign board, the geographical coordinates of the road sign, the year of manufacturing, the geographic direction of the road sign, the surrounding environment, the manufacturer name, and the type of retroreflective material. The color coordinates and retroreflection were measured in three different points in the background and the boarder of the road signs.

The requirements for the lowest retroreflection and colors for road signs judged according to the Swedish Transport Administration regulations. The approved road signs were coded as 1 and the disapproved ones as 0.

Five different types of road signs mounted on the Swedish roads were invented as shown in Table 1.

Table 1. Type and number of measured road signs.

Road sign	Loftbacken	NYKÖPING 23	NYKÖPING 23			Total
Number of studied signs	74	81	10	79	58	302

3.1. Data cleaning

To be able to use the collected data, it was cleaned from incomplete and unnecessary data. Manufacture name, class of retroreflection, the surrounding environment and other parameters were removed from the data because they are not suitable for this study. Only 10 relevant features remained after cleaning which are listed in Table 2 in addition to the ground truth (approved - not approved) as annotated by the field expert.

Table 2. The Input features.

Feature	Description
The color value (x, y) for background and border in the CIE	Measured according to CIE standard
Retroreflection values of the background and border	cd/lx.m2
Year of manufacturing	1983-2018
Direction	0-359 degrees
GPS coordinates where the sign is mounted	Latitude and longitude

After cleaning, the dataset comprised 906 records (three readings for each of the 302 road signs). By removing the records with missing data, the final number of records invoked in this study are 753.

3.2 Data scaling

Data scaling is an important step in in data preprocessing and has an impact on the performance of machine learning algorithm (Ambarwari, Adrian, and Herdiyeni 2020). Many machine learning methods expected to be more effective if the data features have the same scale.

This study invokes and compares two methods in data scaling, namely normalization and standardization.

Normalization is a process in which features within a model are categorized to increase the bond of input variables(Raju et al. 2020). The Min-Max Scaler is used to fit the data within a scale of 0-1 as given in Eq.1.

$$x_{in} = \frac{x_i - min\ (x)}{max(x) - min\ (x)} \tag{1}$$

Where x_i is the input variable, x_{in} is the normalized variable, min(x) is the minimum value of the variable x, and max(x) is the maximum value of this variable.

On the other hand, standardization refers to shifting the distribution of each feature so that the mean of the variable is zero and the standard deviation is one (Jayalakshmi and Santhakumaran 2011). It is useful to standardize mixed data into numerical data when applying machine learning models (Modarresi and Munir 2018). The mean and standard deviation are calculated for each feature and then the standardized feature is calculated as given in Eq. 2.

$$x_{is} = \frac{x_i - mean\ (x)}{\sigma\ (x)} \tag{2}$$

Where x_i is the input variable, x_{is} is the standardized variable, and σ is the standard deviation.

4. Data analysis

Before using the data to train and test the machine learning algorithms, there will be a number of preliminary steps to be applied to prepare this data. This includes data splitting and correlation analysis.

4.1. Data Splitting using K-fold Cross Validation (CV)

Cross-validation is an important concept in machine learning, and it helps to ensure that the data used in training will not be used in the testing of the model (Arlot and Celisse 2010). The dataset invoked in this study were split using K-fold cross validation where K= 3 so that 2 subsets are invoked for training and 1 subset is used for testing. This process was repeated 3 times and the performance parameters are calculated each time. The final performance parameters are the average of the 3 cases.

4.2. Correlation Analysis

A correlation analysis is performed to find the dependent features and extract the possible relationships between them. The correlation between the features in the dataset is represented in a heatmap, as depicted in Fig. 1. The heatmap is used to give a primary idea about correlation between the input variables listed in Table 2. Only two features were found to be correlated which are the color values in the CIE color system and the GPS coordinates of the road sign. These correlations are natural and did not affect the results. The features in the dataset were uncorrelated, therefore, the PCA can be used to find the representation of the data.

4.3. Principal Component Analysis (PCA)

A common way of speeding up a machine learning algorithm and help classifiers to make accurate decisions is by using Principal Component Analysis (PCA) (Salo, Nassif, and Essex 2019). PCA uses orthogonal transformation of the correlated features into principal components that are linearly uncorrelated.

The way PCA works is it finds principal components (PC) in the descending order by the amount of variation each component explains. The maximum number of PC restricted by the number of features (10 in this study) and the choice of the number of PCs affects the accuracy of the results. An important factor in choosing the number of PC is the cumulative variance explained by the PC. In this study, PCA was implemented, and three respective four principal components were used to train the different classification algorithms. According to results in Table3, the three PC that were invoked explain 94.55% of the variance of the unscaled data and 99.99% of the variance of the unscaled data. To improve the accuracy of the results, four PC were invoked in which 99.99% of the variance explained for both the scaled and unscaled data as shown in Fig. 2.

4.4. Machine learning algorithms

Three algorithms are used in this study to predict the status of the road sign:

Artificial Neural Network (ANN)

A neural network with an input layer, an output layer, and a hidden layer, is used in this study.

The architecture of the NN is 10-30-2 is employed where the two classes represent (approved/disapproved) . The NN parameters used in this study are: Learning rate = 0.01, Max iteration = 1000 Activation function = logistic, and the optimization method is Stochastic gradient descent.

Support Vector Machines (SVM)

SVM is a supervised machine learning model that used in classification problems. It defines a hyperplane between the two classes and extends the margin in order to maximize the distinction between these classes, which results fewer close miscalculations. The performance of SVM largely depends on the kernel and choosing the right kernel can improve the performance of the classifier. In this study, Radial Basis Function (RBF) kernel is used which was found to give the best accuracy. The two parameters to be considered with RBF kernel function are C which is selected to be 2 and gamma which was auto selected.

Random forest

Random forests are a widely used decision-tree-based ensemble method for supervised learning. The method operates by constructing many decision trees from the input data and aggregating over the output to generate predictions. The number of trees in the forest used in this study were 100.

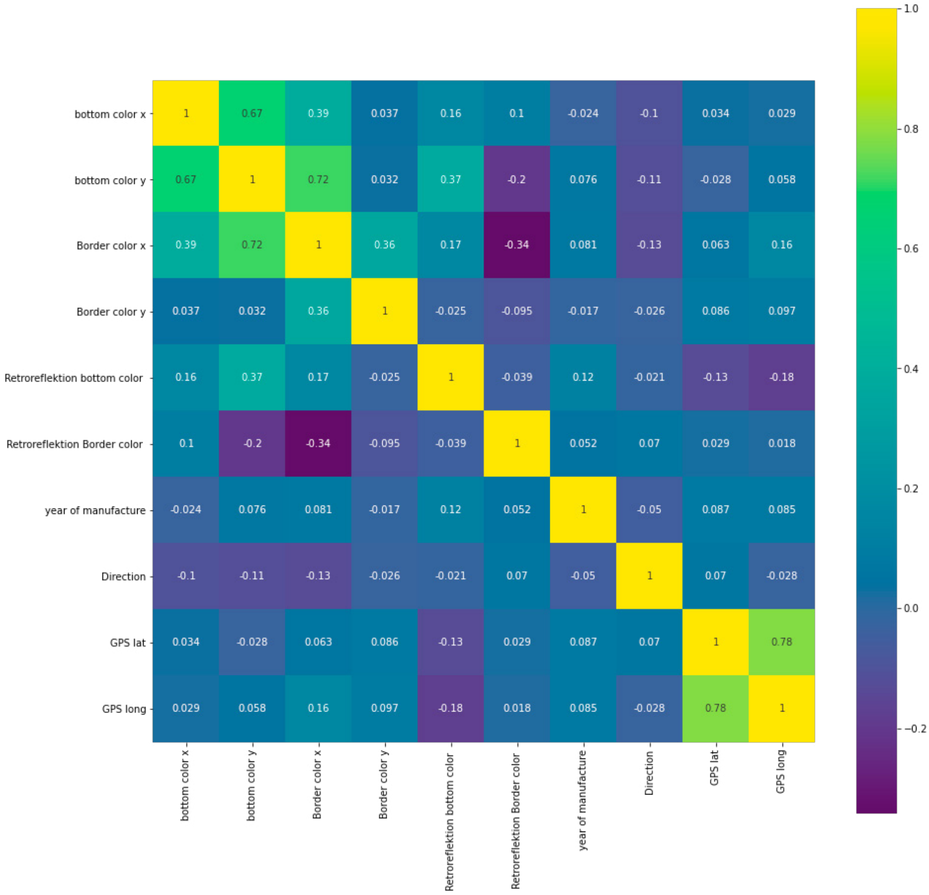


Fig. 1. Correlation analysis heatmap of the dataset.

Table 3. Variance explained.

PC	Variance explained %	
	Unscaled data	Scaled data
1	42.26	99.99
2	77.73	99.99
3	94.55	99.99
4	99.99	99.99

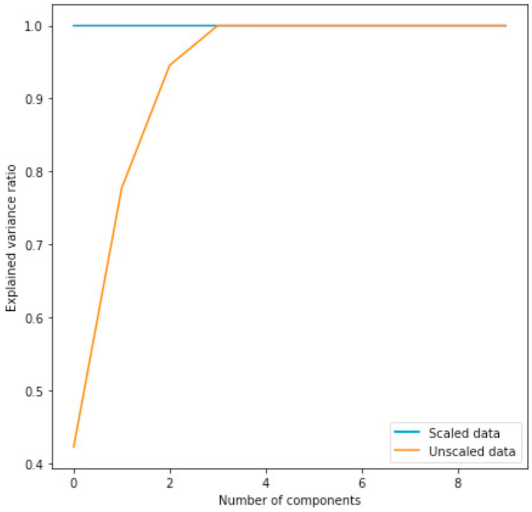


Fig. 2. Cumulative variance explained by principal components on unscaled dataset.

5. Results and discussion

As mentioned before, three classification algorithms were tested to predict the status of the road signs and different measurements are used to evaluate the performance of them:

- Accuracy: The correct number of predictions made by the model over all the observed values. The accuracy is calculated by Equation (3), where TP refers to true positive, TN refers to true negative, FP refers to false positive and FN refers to false negative:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

- Precision: Precision gives how many of the correctly predicted cases actually turned out to be positive. The proportion is calculated with the formula shown in Equation (4):

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

- Recall: Recall gives how many of the actual positive cases the model is able to predict correctly. It is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

- F1-score: F1-score is a metric which takes into account both precision and recall and is defined as follows:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

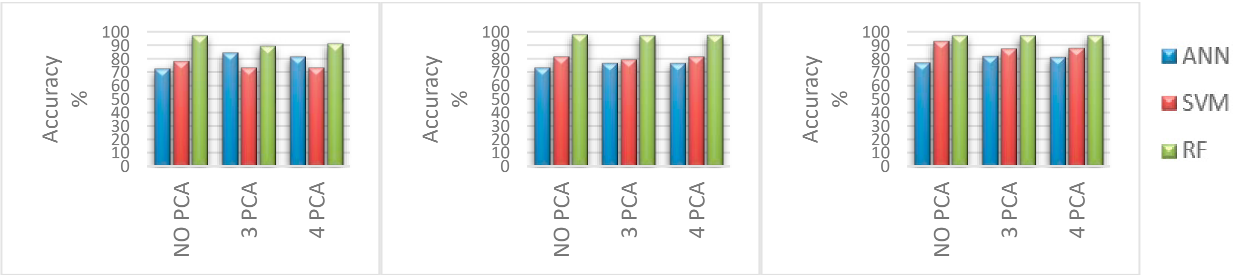
To evaluate the performance of the three classification algorithms, the predicted values are compared with the ground truth values in the test data. The performance results from the ANN, SVM, and random forest with/without scaling and with/without PCA are presented in Table 4.

Table 4. Classification results.

Algorithm	Scaling	PCA	Accuracy	Precision	Recall	F1-Score
ANN	No scaling	No	72.5	73.8	73	72.2
ANN	Normalization	No	73.3	75	73.4	73.1
ANN	Standardization	No	76.9	78.4	76.7	76.6
ANN	No scaling	3	**84.3**	84.9	84.4	84.4
ANN	Normalization	3	76.4	77.3	76.9	76.5
ANN	Standardization	3	81.9	82.4	82.3	82.2
ANN	No scaling	4	81.5	82.4	81.5	81.8
ANN	Normalization	4	76.5	76.6	77.3	76.9
ANN	Standardization	4	81.1	81.9	81.7	81.4
SVM	No scaling	No	78.1	85.6	76	75.8
SVM	Normalization	No	81.4	81.4	81.6	81.3
SVM	Standardization	No	**93**	93	93.3	92.9
SVM	No scaling	3	73.3	82.7	70.8	69.6
SVM	Normalization	3	79.4	79.3	79.4	79.3
SVM	Standardization	3	87.3	87.4	87.3	87.2
SVM	No scaling	4	73.2	83.5	70.6	69.3
SVM	Normalization	4	81.4	81.3	81.4	81.3
SVM	Standardization	4	88	88.1	88.2	88
Random Forest	No scaling	No	97.2	97.2	97.3	97.2
Random Forest	Normalization	No	97.9	97.7	97.8	97.7
Random Forest	Standardization	No	97.3	97.3	97.5	97.3
Random Forest	No scaling	3	89.2	89.3	89.3	89.2
Random Forest	Normalization	3	97.3	97.4	97.4	97.3
Random Forest	Standardization	3	97.1	97.1	97.2	97.1
Random Forest	No scaling	4	91.2	91.2	91.3	91.2
Random Forest	Normalization	4	**98**	98	98.1	98
Random Forest	Standardization	4	97.2	97.2	97.3	97.2

From Table 4, it is clear that random forest performed better than the other algorithms, with accuracy of 98% followed by the SVM which performed 93% and the ANN that performed 84%. The highest predicting accuracy was established by using Random Forest on four PC and on normalized data. This algorithm achieved accuracy of 98% and 98% (F1 score) of the positives were successfully predicted by this model. SVM comes in the second place by establishing an accuracy of 93% with no PC and using normalization while ANN gives the lowest accuracy (84%).

The scaling of the data increased the accuracy of the prediction using ANN, SVM and Random Forest when no PCA is employed, see Fig.3. The effect of the PCA vary from increasing to decreasing the accuracy and sometime have no effect at all. For the ANN, the PCA increases the accuracy while for SVM the accuracy decreases when using PCA. The PCA has no effect on the accuracy when using random forest and scaled data. Using ANN and SVM with PC gives more accuracy with standardization than normalization.



(a) (b) (c)

Fig.3. Performance of the different classifier when using (a) No scaling; (b) Normalization; (c) Standardization.

6. Conclusion

This paper shows that the status of the road signs in Sweden can successfully be predicted using machine learning algorithms. Three prediction algorithms (ANN, SVM and random forest) exhibited high accuracy. Among them Random forest achieved the highest accuracy (98%), high precision (98%), high recall (98%), and high F1 scores (98%).

Scaling of the dataset improved the accuracy of the random forest prediction by giving the model a better chance to find the right patterns and adequate weights to features. Furthermore, the prediction results shows that standardization gives a better accuracy than normalization.

Using PCA helps in decrease the number of features and gives more prediction accuracy than when using all the features with the unscaled data. Decreasing the number of features helps in speeding up of the learning process. The data used in this study were not big and learning process was very fast therefore, the effect of PCA on the speed not evaluated.

The findings of this research can be used to predict the status of the road signs. In future work, a bigger dataset will be collected and tested. Furthermore, the status of the road signs will be predicted using transfer learning in which only small number of features will be invoked for classification.

References

Ambarwari, A, Q Adrian, and Y Herdiyeni. 2020. Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification, Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 4: 117-22.

Arlot, S, and A Celisse. 2010. A survey of cross-validation procedures for model selection, Statistics surveys, 4: 40-79.

Babić, D, A Ščukanec, and M Fiolić. 2016. Predicting state of traffic signs using logistic regression, International Journal for Traffic and Transport Engineering (IJTTE), 6: 280-88.

Berces, A, and S Robertson. 2012. Keeping people safer through better visibility: Advances in retroreflective technologies for road signage, pavement markings and vehicle visibility delivering safer roads. In Australasian Road Safety Research Policing Education Conference. Wellington, New Zealand.

Jayalakshmi, T, and A Santhakumaran. 2011. Statistical normalization and back propagation for classification, International Journal of Computer Theory and Engineering, 3: 1793-8201.

Khrapova, M. 2019. Determining the influence of factors on retroreflective properties of traffic signs, Agronomy Research, 17: 1041–52.

Kjellman, Erik, Carina Fors, and S Lundkvist. 2018. Analysis of life-cycle costs for road signs with focus on retroreflective sheeting materials. In.: Swedish National Road and Transport Research Institute, VTI.

Madleňák, R, D Hoštáková, L Madleňáková, P Drozdziel, and A Török. 2018. The analysis of the traffic signs visibility during night driving, Advances in Science and Technology. Research Journal, 12.

Modarresi, K, and A Munir. 2018. Standardization of featureless variables for machine learning models using natural language processing. In International Conference on Computational Science, 234-46. Springer, Cham.

Raju, V, K Lakshmi, V Jain, A Kalidindi, and V Padma. 2020. Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. In IEEE 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 729-35.

Ré, J, J Miles, and P Carlson. 2011. Analysis of in-service traffic sign retroreflectivity and deterioration rates in Texas, Transportation research record, 2258: 88-94.

Salih, R, and H Fleyeh. 2021. Factors Affecting Night-Time Visibility of Retroreflective Road Traffic Signs: A Review, International Journal for Traffic and Transport Engineering (IJTTE), 11: 115-28.

Salo, F, A Nassif, and A Essex. 2019. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection, Computer Networks, 148: 164-75.

Swargam, N. 2004. Development of a neural network approach for the assessment of the performance of traffic sign retroreflectivity, Louisiana State Univesity.

Ž. Šarić, X. Xu, L. Duan, and D. Babić. 2018. Identifying the safety factors over traffic signs in state roads using a panel quantile regression approach, Traffic injury prevention, 19: 607-14.