ISSN 2347 - 3983

Volume 10. No.2, February 2022

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter111022022.pdf

https://doi.org/10.30534/ijeter/2022/111022022

Evaluation of Support Vector Machines SVM for Classification with Imbalanced Datasets

Cesar A. Perdomo Ch.[1], **Oscar D. Flórez C.**[2], **Julián R. Camargo L.**[3]

[1]Universidad Distrital “Francisco José de Caldas”, Bogotá, Colombia, cperdomo@correo.udistrital.edu.co

[2] Universidad Distrital “Francisco José de Caldas”, Bogotá, Colombia, odflorez@udistrital.edu.co

[3] Universidad Distrital “Francisco José de Caldas”, Bogotá, Colombia, jcamargo@udistrital.edu.co

Received Date : December 31, 2021 Accepted Date : January 27, 2022 Published Date : February 07, 2022

ABSTRACT

This paper presents the effect of unbalanced data sets on the training of classification models. For this purpose, sensor readings from a wall-following robot dataset available in Kaggle are used. The data was collected as the SCITOS G5 navigated the room using 24 ultrasonic sensors and following the wall clockwise, with 5,456 records distributed unbalanced into four classes. Training is performed by making a 70% to 30% split of the data for training and testing, initially using all the records. The data set is then balanced by sampling and equalizing the records by class. The models are trained with the same percentages for training and testing.

Key words: Classification, Imbalanced Data, Robot, Support vector machine

1. INTRODUCTION

Machine learning models focused on classification present an interesting data-driven alternative to be used in robots since their training facilitates the development of control that allows decision-making based on the records used in the training of the models.

Therefore, and considering the massive use of robots in different industrial, military, healthcare and educational applications and sectors, it is important to develop models that allow autonomous driving and data-driven decision making. In this particular case, data obtained by the Department of Teleinformatics Engineering of the Federal University of Ceará (Fortaleza, Ceará, Brazil) with the SCITOS-G5 mobile robot will be used.

Figure 1: SCITOS-G5 developed by MetraLabs

The SCITOS-G5 robot performs measurements from 24 ultrasonic sensors and records the decision to be made by the robot in its navigation. Each of the sensors is located fifteen degrees from the next and thus a sweep of the entire robot environment is achieved. This way, a dataset containing 5,456 records with four output classes was produced [1].

With this dataset, some classification options have been established, the main problem is that the classes are not linearly separable, as shown in Figure 2.

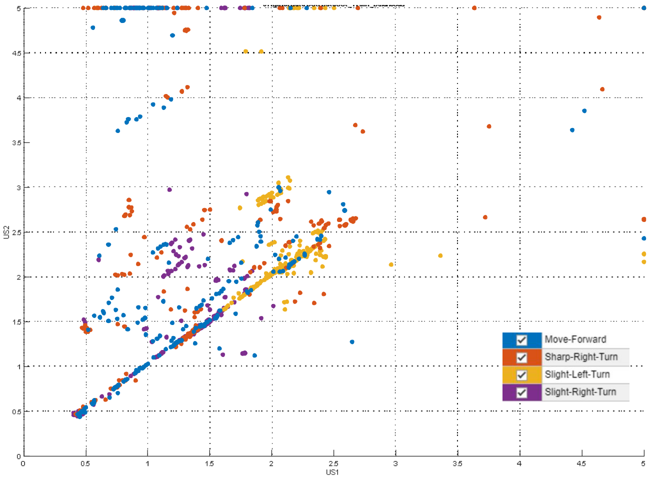

Figure 2: Scatter plot of readings from two adjacent sensors

Due to the above, machine learning techniques have been used as an interesting tool to explore as an alternative for developing models to enable decision-making [2].

2. METHODOLOGY

While the data set has twenty-four sensor data entries identified as US1 to US24 of record, where US1 is an ultrasonic sensor at the front of the robot (reference angle: 180°) [3].

All sensors provide numerical reference values and their maximum value is approximately 5.0 and their minimum value in some situations is 0.3. The classes presented are

Move-Forward, Slight-Right-Turn, Sharp-Right-Turn and Slight-Left-Turn (see Table 1).

Table 1: Class distribution data set

Class	Number of samples
Move-Forward	2205 samples (40.41%)
Slight-Right-Turn	826 samples (15.13%)
Sharp-Right-Turn	2097 samples (38.43%)
Slight-Left-Turn	328 samples (6.01%)

The data set is unbalanced and contains two classes associated with making a right turn (Slight and Sharp).

The histogram in Figure 3 classes contains very few records of left turns with 328 samples.

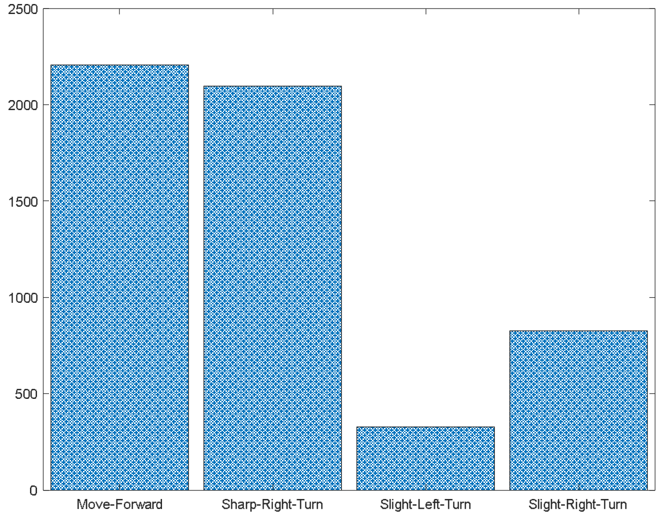

Figure 3: Histogram of classes in the data set

2.1 Imbalanced data split

Initially, the training and test data sets are created with all the records from the original data set. It is established that 70% of the data from each class will be used for training and the remaining 30% for each class's evaluation of the trained model.

For this purpose, a random selection of the records is made, taking into account the class and thus allowing the data sets obtained to have the same proportion between the classes [4].

Table 2 shows the class distribution of the unbalanced data set; Figure 4 shows the corresponding histogram.

Table 2: Class distribution in the unbalanced train data set

Class	Number of samples
Move-Forward	1544 samples (40.41%)
Slight-Right-Turn	578 samples (15.13%)
Sharp-Right-Turn	1468 samples (38.42%)
Slight-Left-Turn	230 samples (6.02%)

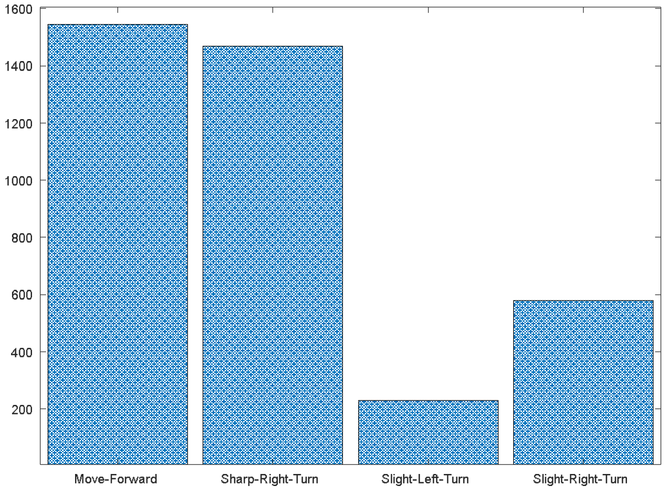

Figure 4: Histogram of classes in the unbalanced training data set

The same procedure is performed for the test data set, as shown in Table 3 and the histogram in Figure 5.

Table 3: Class distribution in the Imbalanced test data set

Class	Number of samples
Move-Forward	661 samples (40.40%)
Slight-Right-Turn	248 samples (15.15%)
Sharp-Right-Turn	629 samples (38.44%)
Slight-Left-Turn	98 samples (5.99%)

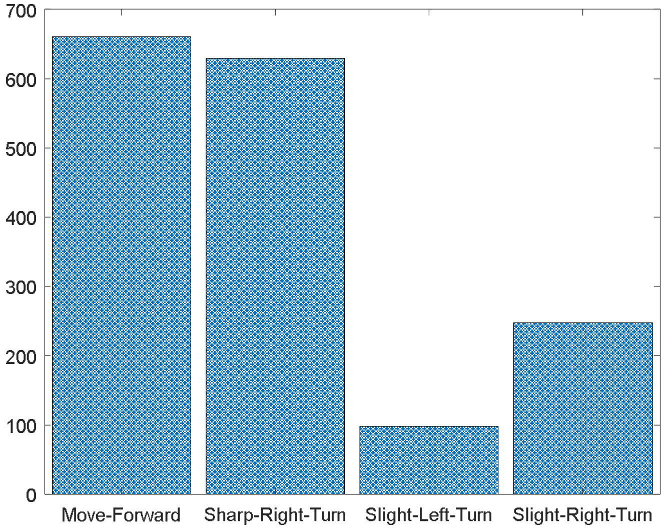

Figure 5: Histogram of classes in the Imbalanced test data set

2.2 Balanced data split

For data balancing, the Random Undersampling (RUS) technique is used, which is initially performed by fixing the class with the smallest number of records [5] [6]. In this case, the class Slight left turn has 328 samples. Therefore, 328 samples are randomly selected from the other classes and data splitting is performed for training and testing following the same 70%-30% proportions (See Table 4 and Figure 6).

Table 4: Class distribution in the balanced train data set

Class	Number of samples
Move-Forward	230 samples (25%)
Slight-Right-Turn	230 samples (25%)
Sharp-Right-Turn	230 samples (25%)
Slight-Left-Turn	230 samples (25%)

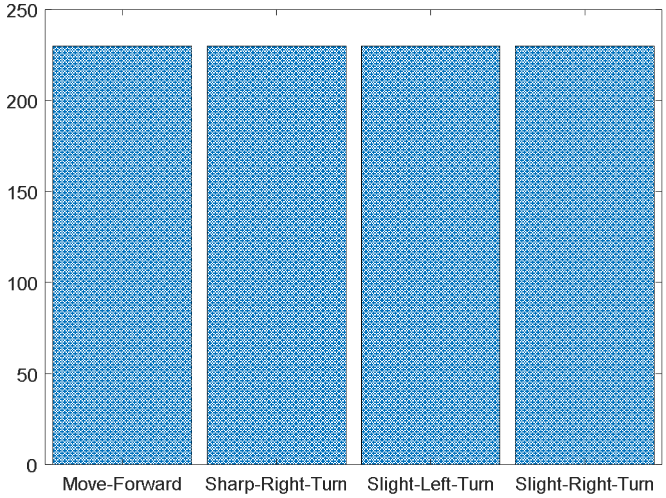

Figure 6: Histogram of classes in the balanced train data set

Table 5 shows the class distribution of the balanced data set; Figure 7 shows the corresponding histogram.

Table 5: Class distribution in the balanced train data set

Class	Number of samples
Move-Forward	98 samples (25%)
Slight-Right-Turn	98 samples (25%)
Sharp-Right-Turn	98 samples (25%)
Slight-Left-Turn	98 samples (25%)

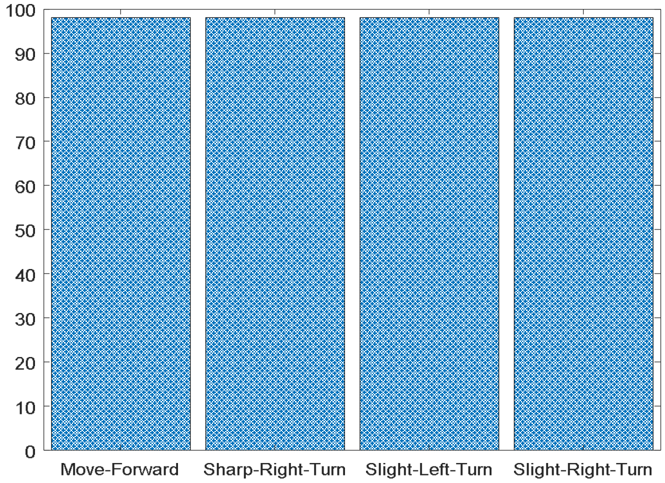

Figure 7: Histogram of classes in the balanced test data set

2.3 SVM Classification

SVM is a data mining technique well suited for statistical analysis of small samples. This technique allows establishing linearity in the classification problem by increasing the dimensionality of the samples through the spatial transformation of the samples [7]. Figure 8 shows a block diagram of the SVM model for classification.

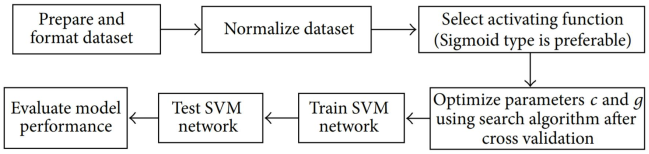
Prepare and format dataset
Normalize dataset
Select activating function (Sigmoid type is preferable)
Optimize parameters *c* and *g* using search algorithm after cross validation
Train SVM network
Test SVM network
Evaluate model performance

Figure 8: Flow chart of SVM classification model [8]

According to the data sets presented above, the SVM multiclass classification model is used [9], [10]. For this purpose, the MATLAB tool is used, which allows the training and testing of the models from its Classification Learner app.

3. RESULTS

The results obtained in training with the imbalanced dataset are presented below. The confusion matrix is used and the True Positive Rates (TPR) and False Negative Rates (FNR) are shown (Figure 9).

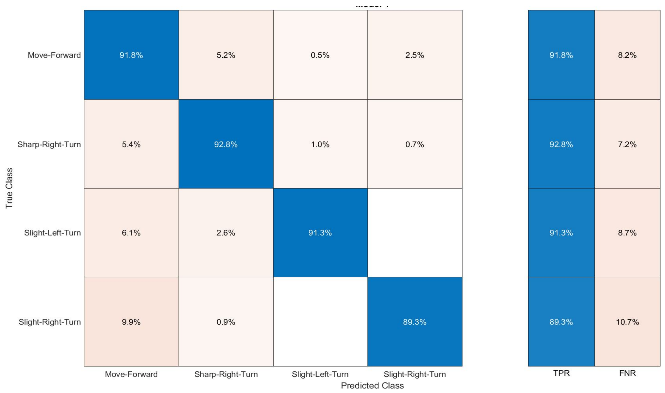

Figure 9: Validation Confusion Matrix Imbalanced Data

In the same way, Figure 10 shows the confusion matrix for the test data and the True Positive Rates (TPR) and False Negative Rates (FNR) are presented for the four classes.

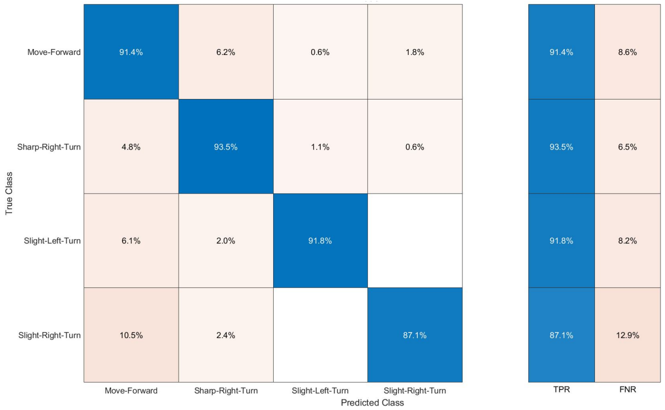

Figure 10: Test Confusion Matrix Imbalanced Data

After balancing and partitioning the data, the model has trained again from the same initial values of the model, with the results shown in Figure 11.

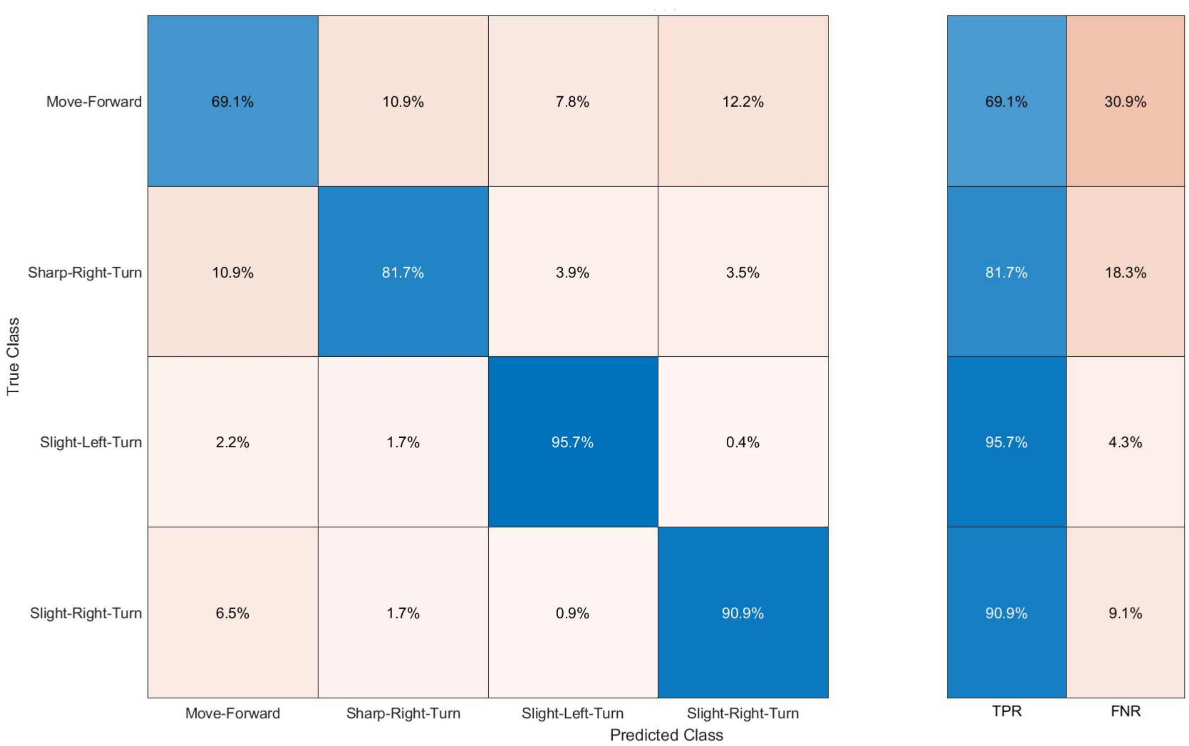

Figure 11: Validation Confusion Matrix Balanced Data

Figure 12 shows the results of the balanced test data set and the true positive rates (TPR) and false-negative rates (FNR) for each class are presented.

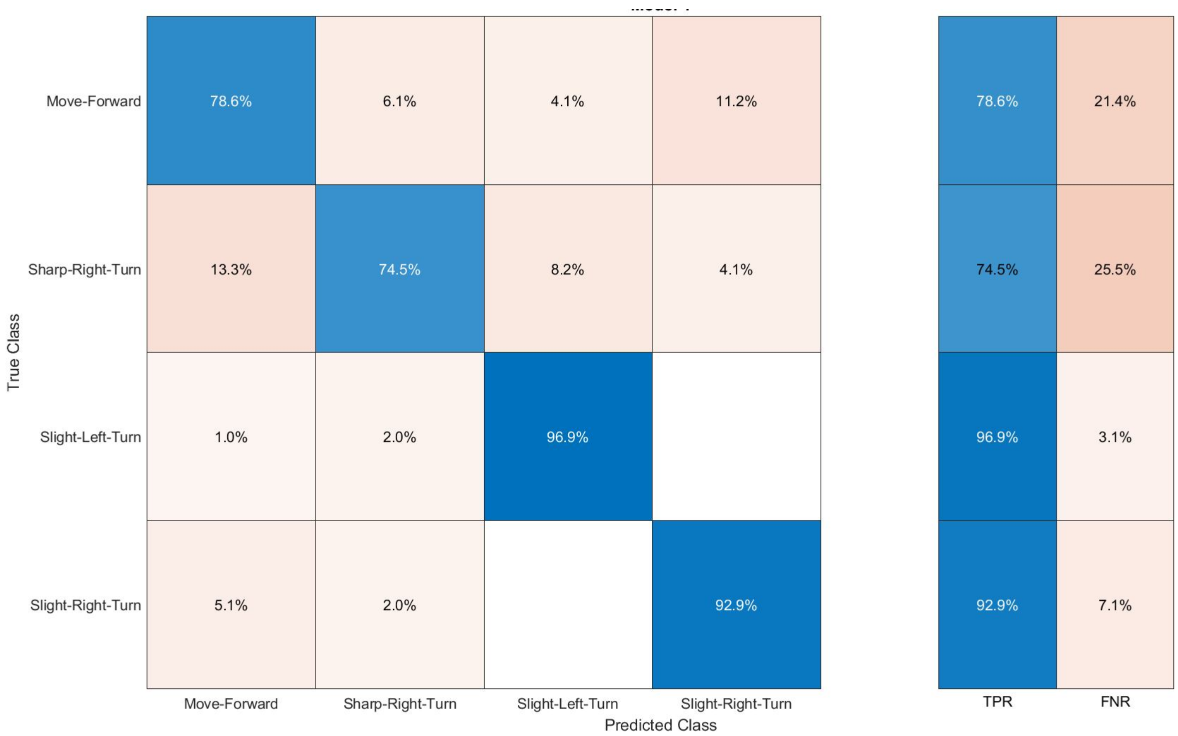

Figure 12: Test Confusion Matrix Balanced Data

A summary of the results obtained in the two training and testing models is shown in Table 6 below.

Table 6: Class distribution in the balanced train data set

	Imbalanced Data	**Balanced Data**
Training Results		
Accuracy	91.8%	84.3%
Total cost	315	144
Prediction speed	~14000 obs/sec	~12000 obs/sec
Training time	10.613 sec	2.0551 sec
Test Results		
Accuracy	91.6%	85.7%
Total cost	138	56

4. CONCLUSION

The handling of unbalanced data sets for the classification models should be taken into account in the data preparation phase, as this improves the model learning process and avoids overlearning or overfitting the model with the element containing the most significant number of records. At the same time, it disadvantages the characteristics of lower-class records, so the metrics used to establish model performance may contain biases.

In the results obtained as usual in these models, the accuracy is a little lower in the test data sets because they are new data for the model and may contain unknown information.

Due to the limited number of samples obtained when balancing the data with the Random UnderSampling technique (RUS), it is observed that the accuracy of the model that used the balanced data set decreases concerning the model that used the imbalanced data set. This is mainly because reducing the number of samples does not allow generalizing the complexity of the problem presented, increasing the number of errors per class.

REFERENCES

1. **Wall-Following Robot Navigation Data Data Set**, Machine Learing Repository, University of California, Irvine (UCI). https://archive.ics.uci.edu/ml/datasets/WallFollowing+Robot+Navigation+Data
2. Freire, Ananda L., et al. **Short-term memory mechanisms in neural network learning of robot navigation tasks: A case study.** *Robotics Symposium (LARS), 2009 6th Latin American. IEEE,* 2009.
3. HAMMAD, Issam; EL-SANKARY, Kamal; GU, Jason. **A comparative study on machine learning algorithms for the control of a wall following robot**. *IEEE International Conference on Robotics and Biomimetics (ROBIO).* IEEE, 2019. p. 2995-3000.
4. KOTSIANTIS, Sotiris, et al. **Handling imbalanced datasets: A review.** *GESTS International Transactions on Computer Science and Engineering,* 2006, vol. 30, no 1, p. 25-36.
5. MISHRA, Satwik. **Handling imbalanced data: SMOTE vs. random undersampling**. *International Research Journal of Engineering and Technology (IRJET),* 2017, vol. 4, no 8.
6. YEN, Show-Jane; LEE, Yue-Shi. **Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset**. *Intelligent Control and Automation. Springer, Berlin, Heidelberg,* 2006. p. 731-740.
7. V. Cherkassky and Y. Q. Ma, **Practical selection of SVM parameters and noise estimation for SVM regression”** Neural Networks, vol. 17, no. 1, pp. 113–126, 2004.
8. Zhang, Huan & He, Chunxia & Yu, Min & Fu, Jingjing. (2015). **Texture Feature Extraction and Classification of SEM Images of Wheat Straw/Polypropylene Composites in Accelerated Aging Test**. Advances in Materials Science and Engineering. 2015.
9. AKBANI, Rehan; KWEK, Stephen; JAPKOWICZ, Nathalie. **Applying support vector machines to imbalanced datasets**. En *European conference on machine learning*. Springer, Berlin, Heidelberg, 2004. p. 39-50.
10. TANG, Fa-ming; WANG, Zhong-dong; CHEN, Mian-yun. **On multiclass classification methods for support vector machines.** Control and Decision, 2005, vol. 20, no 7, p. 746.