Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence

An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Exploiting Machine Learning for the Identification of Locomotives' Position in Large Freight Trains

Helder Arruda, Orlando Ohashi, Jair Ferreira, Cleidson De Souza & Gustavo Pessin

To cite this article: Helder Arruda, Orlando Ohashi, Jair Ferreira, Cleidson De Souza & Gustavo Pessin (2019) Exploiting Machine Learning for the Identification of Locomotives' Position in Large Freight Trains, Applied Artificial Intelligence, 33:10, 902-912, DOI: 10.1080/08839514.2019.1646019

To link to this article: https://doi.org/10.1080/08839514.2019.1646019

Published online: 26 Jul 2019.

APPLIED ARTIFICIAL INTELLIGENCE
2019, VOL. 33, NO. 10, 902–912
https://doi.org/10.1080/08839514.2019.1646019

Exploiting Machine Learning for the Identification of Locomotives' Position in Large Freight Trains

Helder Arruda[a], Orlando Ohashi[b], Jair Ferreira[a], Cleidson De Souza[c], and Gustavo Pessin[d]

[a]Applied Computing Lab, Instituto Tecnológico Vale, Belém, Brazil; [b]Cyberspace Institute, Amazon Federal Rural University (UFRA), Belém, Brazil; [c]Institute of Exact and Natural Sciences, Federal University of Pará, Belém, Brazil; [d]Robotics Lab., Instituto Tecnológico Vale, Ouro Preto, Brazil

ABSTRACT

Accurate identification of locomotives' position in large freight trains is important due to maintenance and management aspects. Current solutions employ RFIDs, image cameras or GPS, while the first two are expensive, the third is not an off-the-shelf hardware for all locomotives. In this paper we investigate a data driven solution to automatically identify locomotives' position in large freight trains. We take into account off-the-shelf hardware alone (that gather instant fuel consumption) seeking for a less expensive solution. We evaluate different machine learning approaches and algorithms and different inputs attributes, achieving significant results.

Introduction

Long freight trains can use several locomotives. These locomotives might be distributed in different positions along the composition. Accurate identification of a locomotive position in large freight trains is important due to maintenance and management aspects. For instance, the erosion will be different on the wheels of the locomotives depending on their position. Solutions employing RFIDs or image cameras (Martin 2008; Zhijun, Zhongpan, and Shaozi 2008) are expensive and are not off-the-shelf hardware in the locomotives. GPS is another potential solution, although, it is also not off-the-shelf hardware for all locomotives. In this paper we investigate a data driven solution to automatically identify locomotives' position in large freight trains. We take into account off-the-shelf hardware alone, that gather instant fuel consumption, seeking for a less expensive solution.

The gathered instant fuel consumption is a time series of fuel consumption in a given part of the railroad. We evaluate different machine learning approaches (multi and binary-class classifiers) and five machine learning algorithms: random forest, multilayer perceptron neural network,

CONTACT Gustavo Pessin gustavo.pessin@itv.org Robotics Lab., Instituto Tecnológico Vale, Ouro Preto, Brazil

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uaai.

convolutional neural network, support vector machine, and gaussian process. Furthermore, we consider that the time series can be transformed in attributes using different spatial resolutions; hence, we evaluate attributes taking into account 3 different spatial resolutions for the machine learning algorithms.

In this work, we present an evaluation of machine learning methods to perform locomotives' identification taking into account instant fuel consumption alone. This work extends Arruda et al.(Arruda et al. 2016) presenting a more comprehensive evaluation of machine learning approaches for the task of locomotives' identification. We aim to answer three research questions in the paper: (RQ1) which is the best machine learning approach to identify locomotives' position. (RQ2) is it better to employ a single (binary) class approach or a multi-class approach? (RQ3) which is the best way to treat the instant consumption data (as a 3 km, 6 km or 9 km spatial resolution)?

This paper is organized into 5 sections. Section 2 brings the related works to our approach. Section 3 treats of our standpoint for locomotive identification, including the data gathering, the system modeling, and the machine learning techniques. Section 4 describes the results, and Section 5 addresses the discussion and future works.

Related Works

The identification of rail cars is tried since the 60s. However, the usage of RFID tags was only implemented from the 80s. The wagons could be equipped with the tags both underneath as on the sides, though the usage on the sides is the most recommended in order to ease the maintenance (Levine 2015). For the tracking of rail cars in rail transportation systems, Hendrickson et al.(Hendrickson et al. 2005) proposed a system using low on-car power and with location-reporting accuracy by means of GPS and RFID.

The obstacle detection by digital image processing was used to help locomotive drivers choose the best speed at certain stretches of the railway, or to use the brakes. A problem to deal with is the image noise caused by light, fog, and vibration since they are real-time (Xue et al. 2008). Another approach related to images is the use a camera on board of a locomotive to determine its location by means of video frames (Yu et al. 2009). The ALVeRT framework uses static images and videos to recognize and track on-road vehicles by means of active learning. Factors such as illumination can be a problem for this kind of approach, once the quality of images and videos are primordial to avoid false positives (Sivaraman and Trivedi 2010).

To avoid collisions between locomotives, a method using GPS data was proposed. Railroad vehicles exchange data frames with its geographic positions, speed, and direction of travel, among other data. The data are processed onboard and if there is a collision risk, the system triggers alarms as a first warning. At a second stage, the vehicle speed is reduced (Hsu 1996). In another work, Doner et al.(Doner, Diana, and Clyne 2002) used GPS coordinates to identify locomotives and determine his operation state.

On-board data is used for locomotive's diagnosis and monitoring. One example is the ICARUS case-based reasoning tool. The software uses the fault codes generated by the locomotives (Varma 1999). Using historical data, a predictive system was proposed to detect the malfunction of machines (locomotives), by means of repair and fault log data (Varma and Roddy 2002).

In the machine learning field, a work with vertical acceleration of ballast wagons, processed by regression algorithms, was used to forecast exceeded limits and to send messages warning the driver (Shafiullah et al. 2010). Prediction of train wheel's failures was used with data from WILD systems that measure the impact of each wheel on the rail. Each axle of a car is treated as a time series. Decision Trees and Naive Bayes algorithms were used. Multiple Classifier Systems were implemented (Yang and Létourneau 2005).

Being that one of the great expenditure for rail organisations is the cost of fuel, previous works have been trying to minimize the fuel consumption of long-haul trains that use diesel-electric locomotives, by means of Lagrangian analysis (Howlett 1996). Speed profiles of long-haul locomotives were used to minimize local energy usage (Howlett, Pudney, and Vu 2009). To minimize the costs of fuel purchasing, of the shipment delay, and of the fuel station contracts, a mathematical model was used. The work proposes a Lagrangian relaxation framework (Nourbakhsh and Ouyang 2010).

For our purpose, we take into account off-the-shelf hardware alone, that gather instant fuel consumption, seeking for a less expensive solution. The usage of RFID, GPS, images, and videos are good options, as mentioned previously. However, they increase the implementation costs. As our aim is a low cost solution to identify the locomotive position, we used log data with machine learning techniques.

Our Approach for Locomotive Identification

Railway Environment and Data Gathering

The train which we refer to is composed by 4 locomotives and 330 train wagons (Figure 1). This type of formation is a default adopted by Vale S.A. (Matos et al. 2015). The positions of the locomotives, in order, are: Leader, Commanded, Remote B and Remote C. We employ a dataset from large freight trains from Vale S.A., which has the largest trains in operation in the world with 330 wagons, 3.3 kilometers long, and usually 4 to 6 locomotives. The Carajás Ridge presents the largest iron ore deposit in the world (Piló,

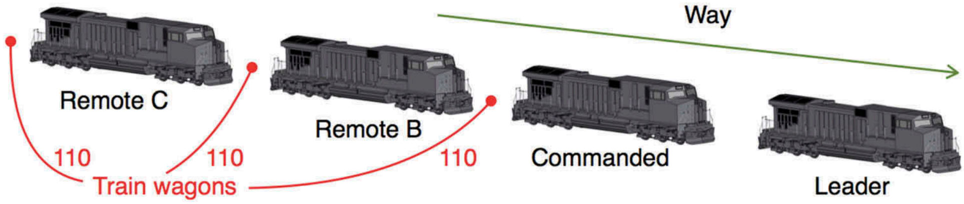

Figure 1. The train formation composed by 4 locomotives and 330 train wagons. The locomotives' positions from the beginning to the end of the train are: Leader, Commanded, Remote B and Remote C. There are 110 train wagons between Commanded and Remote B, other 110 train wagons between Remote B and Remote C, and 110 more train wagons afer Remote C.

Auler, and Martins 2015). From the mine to the harbor, the Carajás railway has 897 km and crosses two of the largest Brazilian states, Pará and Maranhão, connecting Carajás ridge to Ponta da Madeira harbor (Figure 2).

We obtained diesel consumption from 3 Vale' locomotives from May 2014 to December 2015. This data was integrated with information from the Vale operational system which describes the locomotive position in the train. We have 3 locomotives with a total of 63 trips. A locomotive could be a Leader, Commanded, RemoteB or RemoteC, from these trips we have, respectively 5, 11, 16 and 31 trips.

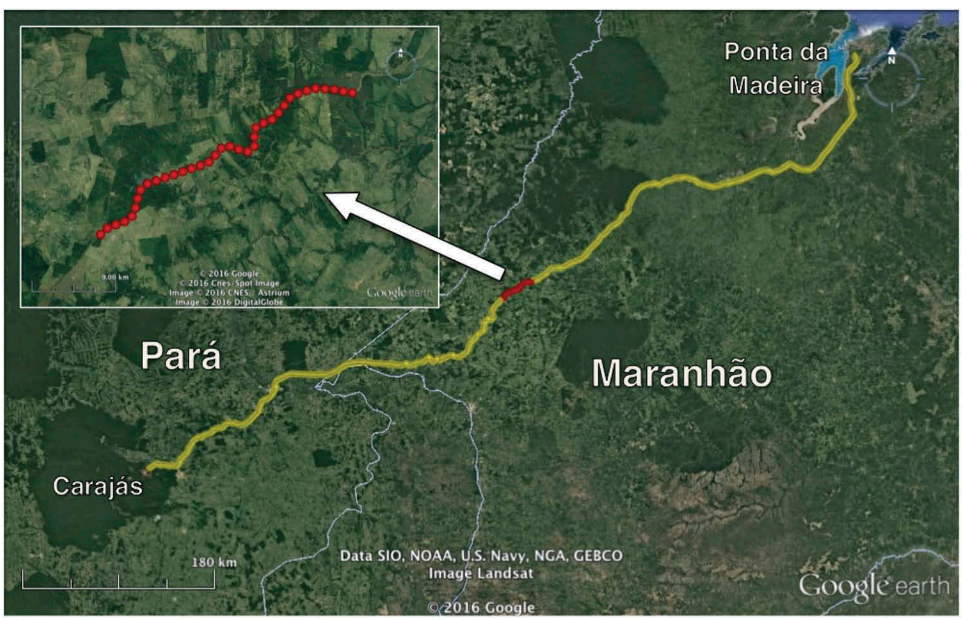

Figure 2. The Carajás railway is 897 km long and connects Carajás ridge, in Pará state, to Ponta da Madeira harbor, in Maranhão. The yellow line shows the railway, highlighting a small stretch in red (36 km), which is used in this work (Images from Google Earth, adapted by the authors).

System Modeling

We employed a subset of 36 km of each trip to perform the classification because we had access only to a subset of the trips. Noteworthy to point out that, for all trips, the same 36 km were employed. The data coming from the monitored locomotives are in plain text format and require preprocessing before being imported into a database. The R programming language (Lantz 2013) was used to deal with raw data and to populate a PostgreSQL database (Vohra 2016). Once the data was stored in the database, SQL queries embedded in Python scripts helped the extraction of feature vectors to be used in machine learning algorithms (Garreta and Moncecchi 2013; Richert and Coelho 2013). A feature vector (Figure 3) with the mean, standard deviation, trend, sum, and median of fuel total consumption, was used instead of raw data according to the best results obtained using this kind of approach in previous work (Furquim et al. 2014). The algorithms used are part of Weka software (Hall et al. 2009). As the chosen stretch is 36 km long, three spatial resolutions were used: 4 stretches of 9 km, 6 stretches of 6 km and 12 stretches of 3 km. The first investigation involved all the classes (Figure 5). The second investigation used each of the classes against the rest (Figure 6), which is a binary classification (Lorena and de Carvalho 2010).

For the Gaussian Process there was an extra step onto preprocessing the dataset. In the comparison among all the classes, each class was represented by a number. Leader was number 1, Commanded was number 2, RemoteB was number 3 and RemoteC was number 4. In the comparison of each class against the rest, Leader, Commanded, RemoteB and RemoteC were number 6 while the other classes were represented by number 5. After classification, if the value stood between 0.50 and 1.49, we assumed that the class was Leader. If the value stood between 1.50 and 2.49, we assumed that the class was Commanded. If the values stood between 2.50 and 3.49, we assumed that the class was RemoteB. If the values stood between 3.50 and 4.49, we assumed that the class was RemoteC. This is for all classes' comparison. For each class against the rest, if the values stood between 4.50 and 5.49, we assumed that the class was Other. If the values stood between 5.50 and 6.49, we assumed that the class could be Leader, Commanded, RemoteB or RemoteC, depending on the choice of the binary classification.

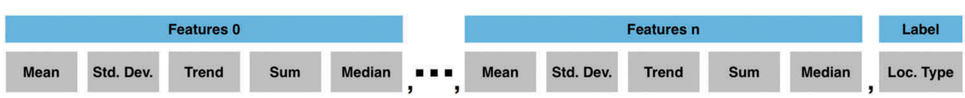

Figure 3. The feature vector is composed by repetitions of mean, standard deviation, trend, sum and median. The last element of the vector is the locomotive type as label.

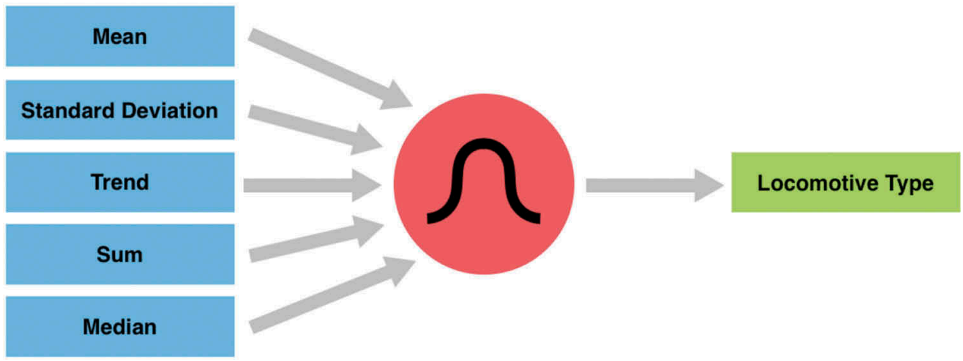

Figure 4. The model is represented by the red circle that receives as input the values from the feature vector, which are denoted by blue rectangles with mean, standard deviation, trend, sum and median. The output is the classification of the locomotive type, represented by the green rectangle.

Independently of the employed model, the workflow is the same. The feature vector is submitted to the model and the presented result is the locomotive type as a percentage (Figure 4). The only exception is the gaussian process, as explained above.

Employed Machine Learning Techniques

Five machine learning methods were used in the investigation: random forests (RF), a bagging of multilayer perceptron (MLP) neural networks, convolutional neural networks (CNN), support vector machines (SVM) and gaussian process (GP). All methods were executed using cross-validation with leave-one-out technique, due to the small amount of data (Cawley and Talbot 2003).

Random Forest (RF)

Random Forest is an algorithm that works as a collection of decision trees, which chooses the best option among the classes. Each decision tree is independently initialized and has a unique vote for the best class (Breiman 2001).

Multilayer Percepetron (MLP)

The Multilayer Perceptron is a system of interconnected neurons where each neuron is connected to the neurons of previous and next layer. The system is divided into input layer, hidden layers and output layer. For classification purposes, which is the aim of this work, the output layer is represented by each one of the classes as a node. The Multilayer Perceptron is a feed-forward artificial neural network and learns in supervised mode. The algorithm used

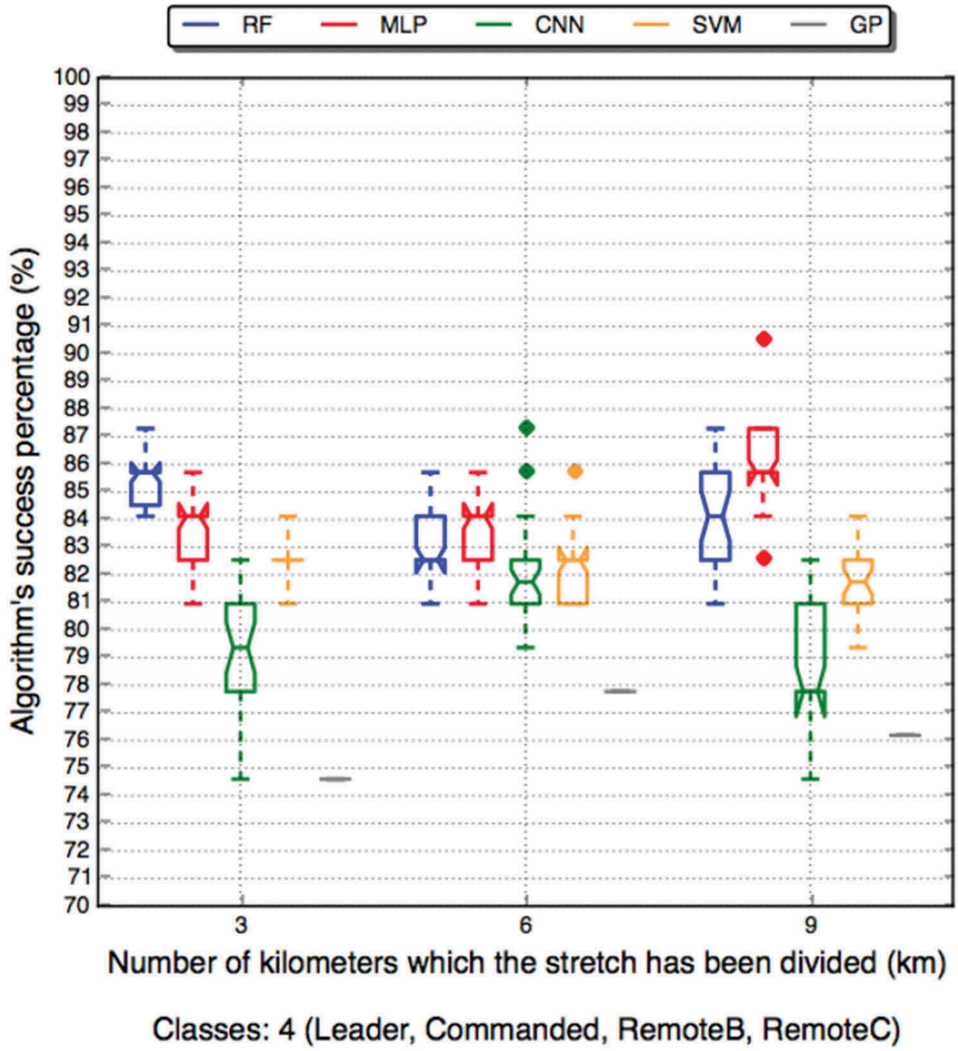

Figure 5. The boxplots show the comparison among 4 classes of locomotives (Leader, Commanded, RemoteB and RemoteC). There are 5 boxplots for each spatial resolution (3 km, 6 km and 9 km), as shown in the x axis. The blue ones indicate the Random Forest algorithm, the red ones show the Multilayer Perceptron Bagging, the green ones are the Convolutional Neural Networks, the yellow ones represents the Support Vector Machines, and the greys ones indicates the Gaussian Processes. In the y axis are the percentual of success.

for training was backpropagation, which is one of the most basic algorithms for this kind of task (Gardner and Dorling 1998).

Convolutional Neural Network (CNN)

Convolutional Neural Networks belongs to Deep Learning techniques and are based on neuroscientific principles. They are artificial neural networks with at least one convolutional layer. This type of network usually deals

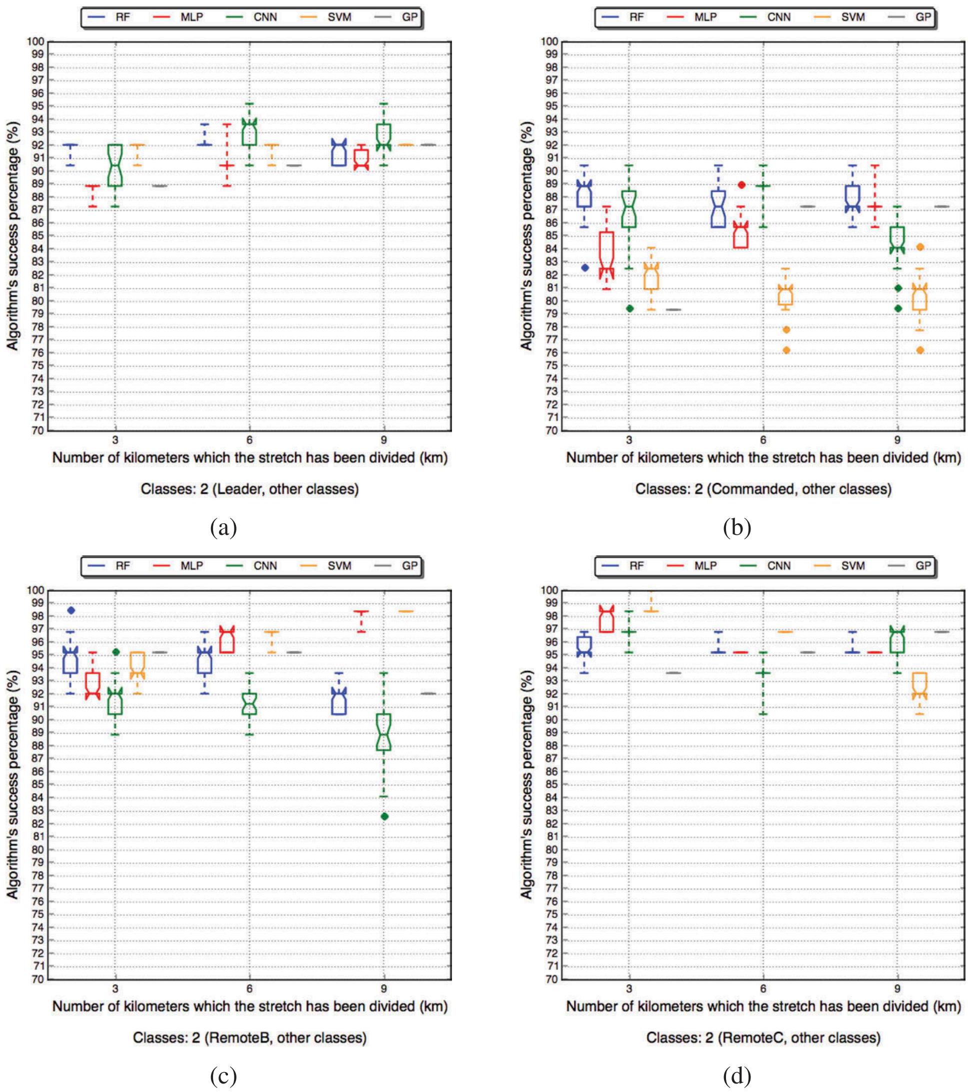

Figure 6. Each one of the locomotive's positions against the others, featuring comparisons among 2 classes. Figure 6a shows Leader (5 trips) against all (58 trips). Figure 6b shows Commanded (11 trips) against all (52 trips). Figure 6c shows RemoteB (16 trips) against all (47 trips). Figure 6d shows RemoteC (31 trips) against all (32 trips).

better with input data from images, that can be represented by a 2D grid, or regular time series intervals, represented as a 1D grid (Goodfellow, Bengio, and Courville 2016).

Support Vector Machine (SVM)

Support Vector Machines work with kernel substitution and comes from the statistical learning theory (Bennett and Campbell 2000). An SVM algorithm transforms the data to a higher dimension hyperplane and creates support vectors with the aim of maximizing the distance between the classes.

Gaussian Process (GP)

Gaussian Process is a supervised learning technique based on the Gaussian probability distribution. This technique presents computational tractability as a strong feature (Rasmussen and Williams 2006).

Results

The approach using binary classification showed better results for all the spatial resolutions. The median was taken into account for the comparisons. The best result of the multi-classifier was 85.71% for both RF, in 3 km, and MLP, in 9 km (Figure 5). The Leader locomotive against the other classes achieved 93.65% of success using CNN for the 6 km spatial resolution, which was the best result for this binary classification (Figure 6a). The Commanded locomotive presented a tie between RF, in 3 km resolution, and CNN, in 6 km resolution, with 88.89% of success (Figure 6b).

For RemoteB locomotives in 3 km spatial resolution, both RF and GP were better, achieving the same result of 95.24%. For the 6 km spatial resolution, MLP and SVM won with 96.83%. And in 9 km, which presented the best results, MLP and SVM won again with 98.41% (Figure 6c). The RemoteC locomotives also achieved the good mark of 98.41% in 3 km spatial resolution with MLP and SVM (Figure 6d).

Discussion and Future Works

All the machine learning methods showed good results, both working with all classes or dealing with binary classification. Although the use of binary classifiers showed better results than the multi-classifier, we did not find a pattern related to the spatial resolutions (3 km, 6 km or 9 km).

However, with a larger dataset the results should be better. In the next phases of this work, we intend to have more data to use with the proposed machine learning methods. Another point that could be taken into account is to analyze full trips of 897 km. Work with full trips brings to us the possibility of dealing with locomotive stops, which could be another variable to the model. In addition, we have interpolated data of temperature from 9 stations over the railway that also could be incorporated to the model

Acknowledgments

The authors would like to thank CNPq (processes 440880/2013-0 and 310468/2014-0).

References

Arruda, H., G. Pessin, O. Ohashi, J. Ferreira, and C. De Souza. 2016. Identifying locomotives' position in large freight trains: An investigation with machine learning and fuel consumption, Symposium on Knowledge Discovery, Mining and Learning (KDMILE), Recife – PE, p. 140–43.

Bennett, K. P., and C. Campbell. 2000. Support vector machines: Hype or hallelujah? *ACM SIGKDD Explorations Newsletter* 2 (2):1–13. http://portal.acm.org/citation.cfm?doid=380995.380999.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1):5–32. doi:10.1023/A:1010933404324.

Cawley, G. C., and N. L. C. Talbot. 2003. Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. *Pattern Recognition* 36 (11):2585–92. doi:10.1016/S0031-3203(03)00136-5.

Doner, J. R., D. L. Diana, and K. N. Clyne. 2002. Methods and apparatus for locomotive tracking. U.S. Patent n. 6,490,523, 3 dez. 2002.

Furquim, G., F. Neto, G. Pessin, J. Ueyama, J. P. D. Albuquerque, M. Clara, E. M. Mendiondo, V. C. B. D. Souza, P. D. Souza, D. Dimitrova, et al. 2014. Combining wireless sensor networks and machine learning for flash flood nowcasting. Proceedings - 2014 IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE WAINA 2014, IEEE, 67–72, Victoria, Canada.

Gardner, M., and S. Dorling. 1998. Artificial neural networks (the multilayer perceptron)âĂŤa review of applications in the atmospheric sciences. *Atmospheric Environment* 32 (14–15):2627–36. http://www.sciencedirect.com/science/article/pii/S1352231097004470.

Garreta, R., and G. Moncecchi. 2013. *Learning scikit-learn: Machine learning in python*. Birmingham, UK: Packt Publishing.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep learning*. Cambridge, MA: MIT Press.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations Newsletter* 11 (1):10–18. http://portal.acm.org/citation.cfm?doid=1656274.1656278.

Hendrickson, B. C., D. M. Ballesty, G. R. Shaffer, and J. J. Kisak. 2005. Rail car tracking system. U.S. Patent Application n. 11/065,423, 22 set. 2005.

Howlett, P. 1996. Optimal strategies for the control of a train. *Automatica* 32 (4):519–32. doi:10.1016/0005-1098(95)00184-0.

Howlett, P. G., P. J. Pudney, and X. Vu. 2009. Local energy minimization in optimal train control. *Automatica* 45 (11):2692–98. doi:10.1016/j.automatica.2009.07.028.

Hsu, G. C. C., 1996. Locomotive collision avoidance method and system. U.S. Patent n. 5,574,469, 12 nov. 1996.

Lantz, B. 2013. *Machine learning with R*. Birmingham, UK: Packt Publishing.

Levine, T. 2015. Working on the Railroad. *IEEE Potentials* 34 (5):22–26. doi:10.1109/MPOT.2015.2410372.

Lorena, A. C., and A. C. P. L. F. de Carvalho. 2010. Building binary-tree-based multiclass classifiers using separability measures. *Neurocomputing* 73 (16–18):2837–45. doi:10.1016/j.neucom.2010.03.027.

Martin, S. E. 2008. Overview of the operation for the proposed locomotive worker identification with cameras. 2008 IEEE International Conference on Technologies for Homeland Security, HST'08, 351–52. IEEE, Waltham, MA.

Matos, J. C. L., V. H. L. Branco, A. N. Macêdo, and D. R. C. Oliveira. 2015. Structural assessment of a RC Bridge over Sororó river along the Carajás railway. *IBRACON Structures and Materials Journal* 8 (2):140–63.

Nourbakhsh, S. M., and Y. Ouyang. 2010. Optimal fueling strategies for locomotive fleets in railroad networks. *Transportation Research Part B: Methodological* 44 (8–9):1104–14. doi:10.1016/j.trb.2010.03.003.

Piló, L. B., A. S. Auler, and F. Martins. 2015. Carajás national forest: Iron ore plateaus and caves in Southeastern Amazon. In *Landscapes and Landforms of Brazil*, 273–83. Brasilia: Springer. doi:10.1007/978-3-319-11800-0.

Rasmussen, C. E., and C. K. I. Williams. 2006. Gaussian processes for machine learning. 14. http://www.gaussianprocess.org/gpml/chapters/RW.pdf.

Richert, W., and L. P. Coelho. 2013. *Building machine learning systems with Python*. Birmingham, UK: Packt Publishing.

Shafiullah, G. M., A. B. M. S. Ali, A. Thompson, and P. J. Wolfs. 2010. Predicting vertical acceleration of railway wagons using regression algorithms. *IEEE Transactions on Intelligent Transportation Systems* 11 (2):290–99. doi:10.1109/TITS.2010.2041057.

Sivaraman, S., and M. M. Trivedi. 2010. A general active-learning framework for on-road vehicle recognition and tracking. *Ieee Transactions on Intelligent Transportation Systems* 11 (2):267–76. doi:10.1109/TITS.2010.2040177.

Varma, A. 1999. Icarus: Design and deployment of a case-based reasoning system for locomotive diagnostics. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 1650, 581–96.

Varma, A., and N. E. Roddy. 2002. Method and system for processing repair data and fault log data to facilitate diagnostics. U.S. Patent n. 6,415,395, 2 jul. 2002.

Vohra, D. 2016. *Using PostgreSQL database. In: Kubernetes microservices with Docker*, 115–39. Berkeley: Apress.

Xue, J., J. Cheng, L. Wang, and X. Gao. 2008. Visual monitoring-based railway grade crossing surveillance system. Proceedings - 1st International Congress on Image and Signal Processing, 2, 427–31. CISP, Sanya, Hainan, China.

Yang, C., and S. Létourneau. 2005. Learning to predict train wheel failures. Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), 0–3. Chicago.

Yu, T., F. W. Wheeler, R. A. Kaucic, P. R. Mendonca, A. K. Kumar, G. R. Shaffer, and T. B. Sebastian. 2009. Method for locomotive navigation and track identification using video. U.S. Patent Application n. 11/904,761, 5 fev. 2009

Zhijun, S., Q. Zhongpan, and L. Shaozi. 2008. The application of UHF RFID technology in mine locomotive positioning system. Proceedings of 2008 IEEE international symposium on IT in medicine and education, 1079–82. IEEE, Xiamen, China.