Machine Learning in the Climatic Database: Support Vector Machine's Algorithm to Map Environmental Events in Months of the Year.

Gilvandro C. de Medeiros[1], Orivaldo V. de S. Junior[1], John V. A. Luiz[1]

[1]Escola de Ciências e Tecnologia – Universidade Federal do Rio Grande do Norte (UFRN)
Caixa Postal 1.524 – 59.078-970 – Natal – RN – Brasil

{gilvandrocesar,johnvictor}@ufrn.edu.br, orivaldo.santana@ect.ufrn.br

Abstract. *With the dissemination of Artificial Intelligence (AI), it becomes common the application of machine learning algorithms (ML) to model and solve problems. In this context, we intend to validate the performance of the ML Vector Support Machine (SVM) algorithm using a public climatic database for the city of Natal. The methodology for this consisted of using the data of said base to train and test the algorithm, placing the information referring to the month of the year in function of the other variables of a given climatic event. Once validated, it is considered promising to deepen the study and application of computational intelligence for meteorological and environmental purposes.*

1. Introduction

Artificial Intelligence (AI) quickly gains space in a variety of applications. Major companies invest in this technology, especially in the area of Machine Learning (ML), among them Apple, Google, Microsoft and IBM. ML's applications present themselves as potential engines for business development, and this is not a trend just for big business. The Startups market also invests in this paradigm, bringing intelligent solutions that stimulate economic, technological and scientific development.

Before proceeding, some definitions may be appropriate. Intelligence, according to Negnevitsky, can be understood as the ability to understand and learn ways to solve problems and make decisions. In computing intelligence, according to Alan Turing, machines can not "think", but are able to pass on a "behavior test for intelligence" (NEGNEVITSKY, 2011). Thus AI can be understood as the ability to make machines to make intelligent decisions, to have intelligent "behavior".

Learning is also defined as the ability to understand and classify new data from previous experiences. ML can be seen as "the science of programming computers so they can learn from data" (GÉRON, 2017). Therefore, ML is defined as the area of the AI that applies algorithms to analyze and model data, thus making possible predictions of unknown variables from a set of known variables, even if related in a complex way.

With this in mind, the following problem is identified: can climate variables, which are commonly considered "unpredictable", actually follow a complex mathematical or statistical modeling involving parameters that conventional meteorology has not been able to equate? With this research, we hope to raise this question, as well as promote in the long term, from a differentiated and complementary

approach, a more intelligent environmental planning, both in terms of scientific and business meteorology and in terms of identifying possible environmental trends.

This work is therefore available to study and analyze results of SVM applied in a database with information on environmental conditions. Therefore, we intend to use ML to identify relationships among these variables, stimulating computational models of meteorological forecasting. In order to do so, we used the climatic variables of Natal, a city with a tropical climate, assigning the vector months of the year as a function of the matrix formed by all other characteristics vectors extracted from the database. Thus, it was sought, based on the knowledge of the other climatic variables, to predict what month of the year this event is associated with, validating if said algorithm can perceive any relation between the variables that will stimulate future work.

2. Literature Review

Support Vector Machine (SVM) is an ML algorithm that is characterized as a computational model capable of performing linear or nonlinear classification, regression and detection of eventual outliers (GÉRON, 2017). For this, this algorithm uses transformations in the variables applied by a function called Kernel, allowing the separation of the data to consolidate the learning of the network.

The Kernel function of the SVM model must be defined as a parameter for the network. In general, such functions fit well or poorly depending on the database used or the expected modeling, but the Kernel Gaussian Radial Basis Function (RBF) presents a satisfactory database behavior in which the variables are related in an unknown way, that works with polynomial mathematical modeling with the function of minimizing the error (DUDA; HART; STORK, 2000).

In open source programming, Python gains prominence due to its libraries and strong virtual network of contributors, making troubleshooting easier. Because of its simple, high-level syntax, Python allows the developer to focus on data for maximizing results, making efficient preprocessing where desirable. Among the most relevant libraries to work with SVM, we highlight Pandas and Numpy, aimed at data analysis and preprocessing, Scikit-Learn, a library that offers Machine Learning methods, models and algorithms, and the Matplotlib library, useful in visualizing possible correlations of variables, as well as graphical plots.

3. Metodology

In the present work, the SVM algorithm was used for the modeling of meteorological data of Natal from a database of Sistema de Monitoramento Agrometeorológico (AGRITEMPO). The variables extracted from this database are related to the climatic events from 01/01/2015 to 07/31/2017, with reference to: month in which an event occurs, minimum, average and maximum temperature, rainfall, minimum and maximum relative humidity, potential evapotranspiration, solar radiation, average wind speed, minimum and maximum dew point, minimum and maximum atmospheric pressure, real evapotranspiration and soil water availability.

In pre-processing, all cases of events where there were missing variables were removed from the database, so that values read as NaN (Not a Number) do not affect modeling. With regard to the training and prediction process of the modeling using SVM, we made use of the Scikit-Learn library, reducing possible overfitting errors, as well as the use of the matrix of confusion matrix, responsible for validating the model

hit rate based on the test and training sets used. The code and database used can be found in the author's GitHub repository[1].

For such modeling, the parameters for the SVM algorithm were obtained empirically, from test and error using values that varied in logarithmic scale, keeping the parameters that made such algorithm to obtain better performance. As for the Kernel, the RBF (radial basis function) kernel was used, due to its popularity and to be satisfactory for several applications.

4. Discussions

Proper preprocessing with the database was achieved, with a hit rate of approximately 80% over the exact month in which an event occurred. Therefore, the algorithm can predict with 80% certainty about exactly of which month it treats a certain configuration of climatic events.

Assuming that the distribution of months is only a reference, it is considered reasonable to think that initial days of a given month may have very similar climatic behavior to the previous month, and that final days may have similar behavior to the subsequent month. With this perspective, considering a margin of error of one month as acceptable for the network output, the model achieves a hit rate of approximately 96%.

Assuming a foundation and giving a view on the behavior of the variables, the plot of the points used is shown in Figure 01, placing variables separately according to the month of the year, as can be seen below:

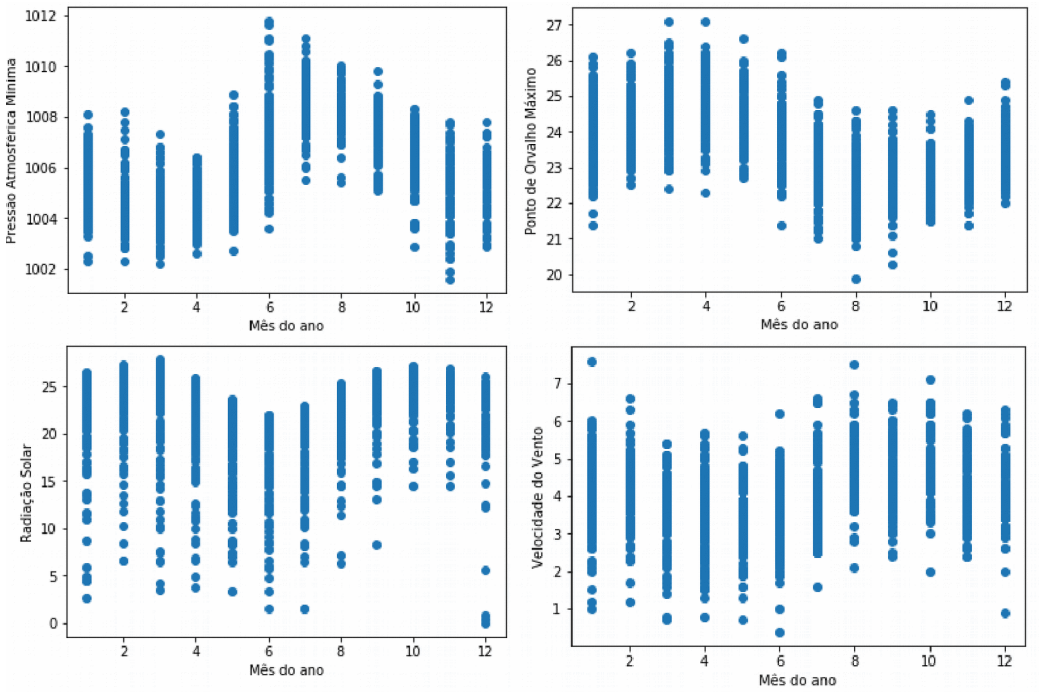

Figure 1. Charts of distributions of climatic variables in function of the month of the year

Observing the graphs, it is possible to notice, for the worked variables, a distribution similar to mathematical modeling of the variables throughout the months of the year. Roughly speaking, it seems reasonable to consider fair a cyclic modeling with a

1 https://github.com/gilvandrocesardemedeiros

corresponding correlated standard deviation. Thus, one can imagine that if there are possible mathematical relationships for variables taken two by two, there may be a larger modeling that enables a more comprehensive and reliable forecasting.

5. Conclusion

The accuracy rates of the algorithm for the exact match of the month is 80% and 96% when placed a tolerance of one month before or after. Relatively high values, which lead to the conclusion that the modeling approach can be treated as promising and that there may be complex mathematical modeling that adequately follows the behavior of variables. In addition, the objective character achieved with this work can be validated graphically from the analysis of Figure 01, which shows a remarkable expression of possible mathematical models for the climatic events discussed here.

By looking at even more efficient models, one can expect better results by improving the methodology and preprocessing, obtaining better quality databases, using optimization algorithms to choose the parameters for the network or even changing algorithms for computational modeling. It can also be commented that a study with scientific foundations in Meteorology was not done to analyze the variables with more property, assigning possible greater weights to certain data and smaller ones for others, which certainly improve the quality of the network and the respective forecast.

Previous studies in the researched literature have also demonstrated that climate variables can be modeled. Parman, for example, studied applications of different computational methods used to predict rainfall, highlighting the predictive approach by Artificial Neural Networks as preferable for rainfall forecasting (PARMA, 2017).

As motivation for future works, it is possible to see applications of this computational model for Precision Agriculture purposes, stimulating the economic and sustainable development of the country. It is also intended to improve this approach to be able to identify possible environmental projections and thus obtain a better quality environmental planning, combating the floods and drought that affects several locations in Brazil and in the world. Analyzing from the business point of view, one can also consider promising the development of a possible startup to provide services using ML and working, in a multidisciplinary way, with issues relating environment and technological development.

References

Duda, R. O., Hart, P. E. and Stork, D. G. (2000) “Pattern Classification”, ed. 2, New York: Wiley-Interscience.

GOVERNO FEDERAL; Sistema de Monitoramento Agrometeorologico. (2018) “Estatísticas”, www.agritempo.gov.br, August.

Negnevitsky, M. (2011) “Artificial Intelligence: A Guide to Intelligent Systems”, ed. 3, Canada: Pearson Education.

Géron, A. (2017) “Hands-On Machine Learning with Scikit-Learn and TensorFlow”, ed. 1, USA.

Parmar, Aakash & Mistree, Kinjal & Sompura, Mithila. (2017). Machine Learning Techniques For Rainfall Prediction: A Review.