RBRH

Revista Brasileira de Recursos Hídricos
Brazilian Journal of Water Resources
Versão On-line ISSN 2318-0331
RBRH, Porto Alegre, v. 22, e58, 2017
Scientific/Technical Article

http://dx.doi.org/10.1590/2318-0331.0217170029

Comparison of data mining models applied to a surface meteorological station

Comparação de modelos de mineração de dados aplicados a uma estação meteorológica de superfície

Anderson Cordeiro Charles[1], Anderson Amendoeira Namen[1,2] and Pedro Paulo Gomes Watts Rodrigues[1]

[1]Universidade do Estado do Rio de Janeiro, Nova Friburgo, RJ, Brazil
[2]Universidade Veiga de Almeida, Rio de Janeiro, RJ, Brazil
E-mails: andersoncordeiro@iprj.uerj.br (ACC), aanamen@iprj.uerj.br, anamen@uva.br (AAN), pwatts@iprj.uerj.br (PPGWR)

Received: February 21, 2017 - Revised: July 17, 2017 - Accepted: September 08, 2017

ABSTRACT

This paper presents the application of data mining techniques for pattern identification obtained from the analysis of meteorological variables and their correlation with the occurrence of intense rainfall. The used data were collected between 2008 and 2012 by the surface meteorological station of the Polytechnic Institute of Rio de Janeiro State University, located in Nova Friburgo - RJ, Brazil. The main objective is the automatic prediction related to extreme precipitation events surrounding the meteorological station location one hour prior its occurrence. Classification models were developed based on decision trees and artificial neural networks. The steps of consistency analysis, treatment and data conversion, as well as the computational models used are described, and some metrics are compared in order to identify their effectiveness. The results obtained for the most accurate model presented a rate of 82. 9% of hits related to the prediction of rainfall equal to or greater than 10 mm h^{-1} one hour prior its occurrence. The results indicate the possibility of using this work to predict risk events in the study region.

Keywords: Data mining; Climate prediction; Surface meteorological station.

RESUMO

Este artigo apresenta a aplicação de técnicas de mineração de dados visando à identificação de padrões, a partir da análise de variáveis meteorológicas e sua correlação com a ocorrência de precipitações intensas. Os dados observados foram coletados pela estação meteorológica de superfície do Instituto Politécnico da Universidade do Estado do Rio de Janeiro, localizada na cidade de Nova Friburgo - RJ, Brasil, entre os anos 2008 e 2012. O principal objetivo é a previsão automática, com antecedência de uma hora, relacionada a eventos extremos de precipitação no entorno do local onde se encontra instalada a estação meteorológica. Foram desenvolvidos modelos de classificação baseados em árvores de decisão e redes neurais artificiais. As etapas de análise de consistência, tratamento e conversão de dados, bem como os modelos computacionais utilizados são descritos, sendo comparadas algumas métricas relacionadas à sua eficácia. Os resultados obtidos para o modelo mais preciso apresentaram uma taxa de 82.9% de acertos relacionados à previsão, com antecedência de uma hora, de precipitações iguais ou maiores que 10 mm h^{-1}, indicando a possibilidade de utilização desse trabalho para previsão de eventos de risco na região de estudo.

Palavras-chave: Mineração de dados; Previsão climática; Estação meteorológica de superfície.

INTRODUCTION

Flooding and landslides are some of the consequences related to high precipitation rates. Predicting these volumes automatically and in advance has been the focus of several studies (AMENT et al., 2011; ALFIERI et al., 2012; KNEIS et al., 2012; VERKADE et al., 2013).

According to NOAA (2010), a warning system for heavy rains must be differentiated when compared with other systems related to hydrometeorological events. Most current alert systems are focused on predicting events that involve a broader area; however, more effective warning systems are needed to cover smaller geographical areas (HAYDEN et al., 2007). Borga (2008) shares this view by suggesting that warning systems for heavy rains should be developed on a local scale, since the meteorological phenomena that cause very heavy rains usually have scales of less than 100 km^2. Kobiyama et al. (2006) argue that the main factors that cause disasters must be continuously monitored. The data must, in parallel, feed a model capable of simulating phenomena in real time, so that, when the system identifies the possibility of occurrence of a critical condition, alerts are issued and the withdrawal of the population and assets from the risk location begins.

According to Fayyad et al. (1996), Knowledge Discovery in Databases (KDD) is a non-trivial process of identifying valid, new, potentially useful and understandable patterns, embedded in the data. This process involves three phases: data collection and preparation; data mining, with the application of techniques that allow the extraction of patterns; and, finally, the post-processing, where the validation of the results is done.

Works like those developed by Onwubolu et al. (2007), Petre (2009), Salvador et al. (2009), Olaiya and Adeyemo (2012), Pessoa (2004), Pessoa et al. (2012), Joshi et al. (2015), Taksande and Mohod (2015) and Yadav and Khatri (2016), can be cited as examples of the application of KDD in climate-related studies. In a recent study, Krishna (2015) reviews meteorological forecasting surveys that apply different data mining models in KDD processes. This work, developed in the city of Nova Friburgo, located in the mountain region of the State of Rio de Janeiro in Brazil, aims to apply the KDD process to identify patterns of correlation between different meteorological variables that can indicate, with one hour in advance, high precipitation rates. In January 2011, the city was affected by an event considered one of the biggest climatic tragedies in Brazil (MEDEIROS; BARROS, 2011).

MATERIAL AND METHODS

The work used the historical records of the automatic surface meteorological station maintained by the Center for Technology in the Environment (Cetema), located at the Polytechnic Institute of the Rio de Janeiro State University (IPRJ) building, in the central region of Nova Friburgo. These records were collected from November-2008 to April-2012.

It was decided to create models that would allow the prediction of rainfalls equal to or greater than 10 mm in the hour following the collection of meteorological data. In situations where there is a large accumulation of rainfall in a short span of time, this level of precipitation may be a risk factor for the population. Therefore, these models should be a relevant contribution to the authorities responsible for the prevention of natural accidents, and can be applied for forecasting and issuing alerts.

This work included the tasks of data collection, data preparation, data mining and the validation of the results obtained, aiming at the identification of patterns related to the occurrence of high precipitation rates.

Data collection and preparation

The IPRJ weather station is an automatic surface station that collects the meteorological data (temperature, humidity, wind speed and direction, atmospheric pressure, solar radiation and precipitation) every 10 minutes using its sensors. Data are sent automatically to the servers of IPRJ Information Technology Laboratory (LTI). This data is stored in a database and a website (CETEMA, 2009) has been built since 2009. This website provides many facilities, including glossary, statistics and graphs. This station is located at latitude -22º 17' 10,59", longitude -42º 32' 31,86" at an altitude of 846 meters above sea level.

Because it is an automatic station, the IPRJ weather station has a data logger that stores all the information collected by the sensors. This ensures that the data will be saved, even if the connection to the server is lost for some reason. This redundancy made it possible, for example, that records referring to the heavy rains that hit the region in January 2011 were stored, despite the loss of connection to the database in that period, allowing subsequent access to those records. Table 1 presents the variables provided by the IPRJ meteorological station sensors.

One of the tasks implemented in the preparation of data was the grouping of the records in larger time intervals. As the focus of analysis was on hourly variations, the records, originally divided into 10-minute time intervals, were grouped in one-hour intervals. In order to avoid the insertion of noise into the values, as a consequence of the grouping process, it was established that the accumulated precipitation value would be stored (precipitation attribute) while all other attributes would store their averages at the respective time. The original number of 165539 records, corresponding to the observations made by the meteorological station between the period of 10/13/2008 and 04/27/2012, was reduced by 83.5%, to 27232 records.

Works such as that of Redman (2001) and Wang et al. (2001) present different aspects related to data treatment and the guarantee of data quality. In this sense, inconsistent information in the databases was corrected so that the quality of the models was not compromised. Records that indicated a possible failure in

Table 1. IPRJ meteorological station variables.

Variable	Unit
Temperature	ºC
Relative air humidity	%
Precipitation	mm
Atmospheric pressure	mb
Solar radiation	W m^{-2}
Wind Speed	m/s
Wind direction (angle)	º

the collected data by the station, such as negative rainfall values, or off-range precipitation values (e.g. records with rainfall indices of 425 mm h^{-1}) were deleted. The same procedure was adopted in all variables of the meteorological station. Registers having any of the variables not filled (null value) were considered inconsistent and were also excluded. More details on the process can be found in Charles (2015).

It is known that precipitation caused by frontal displacements is usually preceded by changes in atmospheric pressure (RODRIGUES et al., 2004). Thus, understanding the relationship between pressure and the occurrence of precipitation is very important. Within this perspective, it became necessary to create new attributes that indicated the changes suffered by the meteorological variables in the analysed period of time. In addition to the creation of attributes which represented the meteorological variations at the same measuring hour, new attributes were created expressing the behaviour of such variables one, two and three hours prior to measuring. The updating of the created attributes was done through the execution of SQL queries (Structured Query Language), as detailed in Charles (2015).

In the present work, it was defined that, in circumstances such as a rainy sequence that causes soil flooding, an accumulation of precipitation equal to or greater than 10 mm in the interval of one hour should be considered a threat. Therefore, a last variable was created, here denominated intense rain, indicating if the accumulation of precipitation would exceed this threshold in the next hour. This binary variable indicated whether there would be precipitation values equal to or greater than 10 mm h^{-1} or values less than 10 mm h^{-1}- It was used as a target variable in the classification algorithms used, since the research's main objective is to predict the occurrence of precipitations with accumulation equal to or greater than 10 mm in the hour after the measurement of the variables. The prediction would allow the generation of alerts for the population, related to the occurrence of risk events.

Table 2 presents the attributes generated from the originally existing ones (Table 1) and which were used as input to the data mining process. In the table are presented several attributes related to the hour before the measurement of the variables (e.g. rain_1hourbefore); however, attributes referring also to the period of 2 and 3 hours prior to the observation were created (e.g. rain_2hoursbefore, rain_3hoursbefore), although they are not presented in the table.

Data mining

The second phase of the KDD process consists of data mining. In this stage we used the *Weka (Waikato Environment for Knowledge Analysis)*, open source software developed in Java, consisting of a set of implementations of algorithms, corresponding to different techniques of data mining (BOUCKAERT et al., 2010).

Identifying interesting patterns of correlation between data is the primary goal of data mining. Two classification models were used in this work: decision trees and artificial neural networks.

According to Han et al. (2012), classification is the process of finding models that describe and distinguish data classes. Classes represent the values of a given attribute named as the target attribute. For the case in question, the target attribute was related to the accumulated rainfall in the hour after the measurement of the meteorological variables, having a value indicating rainfall equal to or greater than 10 mm h^{-1} or rainfall less than 10 mm h^{-1}. Thus, the goal was to identify the combination of attributes that led to the classification of precipitation as intense (> = 10 mm) in the hour after its registration.

Table 2. Set of attributes created from existing attributes.

Attribute	Meaning
rain_1hourbefore	Accumulated precipitation in the previous hour
humidity_1hourbefore	Average humidity in the previous hour
variation_humidity1hour	Variation of humidity in the previous hour
temperature_1hourbefore	Average temperature in the previous hour
variation_temperature1hour	Variation of temperature in the previous hour
solarradiation_1hourbefore	Average solar radiation in the previous hour
variation_solarradiation1hour	Variation of solar radiation in the previous hour
windspeed_1hourbefore	Average wind speed in the previous hour
variation_wind_speed1hour	Variation of average wind speed in the previous hour
winddirection_1hourbefore	Average wind direction in the previous hour
variation_winddirection1hour	Variation of wind direction in the previous hour
averagepressure_1hourbefore	Average pressure in the previous hour
variation_averagepressure1hour	Variation of average pressure in the previous hour
occurrence_rainfall_1hourbefore	Binary attribute, indicating the occurrence of rainfall in the previous hour
average_humidity_variation	Average humidity variation in the last three hours
average_temperature_variation	Average temperature variation in the last three hours
average_solarradiation_variation	Average solar radiation variation in the last three hours
average_pressure_variation	Average pressure variation in the last three hours
average_winddirection_variation	Average wind direction variation in the last three hours
average_windspeed_variation	Average wind speed variation in the last three hours
intense_rain	Binary attribute, indicating the occurrence of precipitation greater than/equal 10 mm h^{-1} or less than 10 mm h^{-1} in the hour after the measurement. Target attribute for the classification algorithms.

Decision trees

A decision tree is a tree-shaped structure, where each inner node denotes a test, related to a variable (attribute).

Each branch refers to the test result; and each leaf node stores a class label. The top node in a tree is its root (HAN et al., 2012). In the present work the target class refers to the attribute intense_rain.

The decision tree construction procedure is called induction. There are many decision tree induction algorithms; most of them adopt a recursive and top-down approach, that is, from the root node towards the leaf nodes. Among these, we find the Hunt Algorithm, which serves as the basis for many other induction algorithms. According to Tan et al. (2009), this algorithm consists basically of two steps:

1) If all records belong to the same class, then associate this class with a leaf node;
2) If the set contains records that belong to more than one class, then:

- Select an attribute to partition records into smaller subsets;
- Create a child node for each value of this attribute and distribute the records to these nodes according to the value they satisfy;
- Apply the algorithm recursively to each child node.

Weka provides several decision tree algorithms in its environment. We used the Decision Tree algorithm J48, similar to the known induction algorithm called C4.5 (details in WITTEN et al., 2011).

Artificial neural networks

An artificial neural network simulates the biological neural system in a simplified way, constructing mathematical models capable of learning, generalizing and making associations. Artificial neural networks, like the human brain, have a parallel structure, composed of processing units, or neurons, connected by communication channels associated with weights. The function of these weights is the same as the dendrites, responsible for the synapses in the human brain. The weights, when have their values alternated during the stimuli, influence the result of the output signal (TAFNER, 1998). Neural networks can be composed of several layers, being formed by:

- Input layer: receives the input data to the neural network;
- Intermediate layers: are responsible for processing;
- Output layer: presents the final result processed by the neural network.

The multilayer perceptron (MLP) was used in our work. It is a neural network that has several layers, being an input layer, one or more hidden layers and an output layer, which are connected through their nodes, to which the weights are assigned (HAYKIN, 1999). Further details regarding the MLP elaborated in the present application can be found in Charles (2015).

Class imbalance

It should be noted that there was a large difference between the number of records classified as intense precipitation ($> = 10$ mm h^{-1}) compared with the number of the other class (<10 mm h^{-1}). The former contained a much lower number of occurrences (99 records out of a total of 27232), due to their own characteristics related to the history of precipitation. According to Han et al. (2012) this situation, known as class imbalance can compromise the accuracy of the classificatory model.

To reduce the problem of class imbalance, we used the SMOTE algorithm, available in the *weka* tool. This algorithm adds to the set a number of fictitious records with behavior similar to the records of the class with lower number of occurrences (*oversampling*). More details about the algorithm can be found in Blagus and Lusa (2013).

Classifier performance evaluation

A classificatory model is composed of the learning phases, where a classifier algorithm is applied on a set of training data; and test, which aims to evaluate the effectiveness of the results through a set of test data. The training set is used to construct the classification model (decision tree or MLP). Subsequently, this model is applied to the test set, containing records with unknown class labels. Here the Holdout Method, known as the Test- Training method, was adopted. In this method 2/3 of the data are destined for training; the remaining base, corresponding to 1/3 of the data, is used for model validation.

The effectiveness of the classifier represents the percentage of observations from the test set that are correctly classified by the model. If this efficacy is high, the classification model is considered efficient and can be used to classify new cases.

For a binary classification problem, where the target attribute contains only two values, it is usual to denote the classes as positive and negative. As previously mentioned, the target attribute used in the model corresponded to the precipitation in the hour after the measurement of the variables, which could be classified as intense precipitation, with values equal to or greater than 10 mm h^{-1}, or not (precipitation less than 10 mm h^{-1}).

An important measure is that which indicates the number of true positives (TP), and rresents the number of true occurrences correctly classified. In the present case, the true positives would be the cases extracted from the test database correctly classified as from intense precipitations. Another important indicator is the number of true negatives (TN). These would be the cases correctly classified as not occurrence of intense precipitations, that is, the forecast obtained by the application of the model would be similar to the information of the instance collected in the test base, which would indicate no rainfall equal to or greater than 10 mm in the next hour.

Measures concerning false positives (FP) and false negatives (FN) would indicate inaccuracies in the predictions made by the classification model, representing, respectively, cases of badly anticipated intense precipitations and the failure to predict intense rainfall occurrences, when in fact they should be foreseen.

The following measures of performance evaluation were collected and presented as results in the software *Weka*:

$$\text{TP Rate}\left(\text{Recall}\right) = \frac{\text{TP}}{\text{TP} + \text{TN}} \quad (1)$$

$$\text{FP Rate} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (2)$$

$$\text{F} - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

where

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

As can be seen in Equations 1 to 4, the F-measure is given by the harmonic average between precision and recall. Given that the harmonic mean between two numbers tends to have a value closer to the smallest of the numbers, then larger values of the F-measure indicate higher values of both recall and precision, characterizing better model performance.

Here it was also applied the ROC (Receiver Operating Characteristics) analysis, a graphical method for evaluating, organizing and selecting classifiers based on their performances. The space where the ROC is plotted is two-dimensional for a two-class problem. In the axis of the abscissa (x) is plotted the value of the true positive rate (TP Rate) and in the axis of ordinates the value of the false positive rate (FP Rate), defined previously in Equations 1 and 2. The performance of a classifier is evaluated by means of the ROC curve, analyzing how distant it is from the diagonal $y = x$, where the classifier presents a stochastic behavior (Prati et al., 2008). In addition, through the ROC curve you can define an evaluation measure, which is the area below the curve (ROC Area). In practical terms, the closer this number is to the unit, the better the classifier performance. More details on ROC analysis can be found in Majnik and Bosnić (2013), Fawcett (2006) and Tan et al. (2009).

RESULTS AND DISCUSSION

Decision tree models (J48 algorithm) and neural networks (MLP) were developed and tested, initially using records containing a large imbalance between classes. Subsequently, the SMOTE method was applied to the training base, which allowed the use of records containing classes in a balanced way, for the construction of the models. Table 3 presents the values of the main evaluation metrics, obtained from the test base, being related to the class that indicates the occurrence of intense precipitations (> = 10 mm) in the hour after the measurement. These data were extracted from the generated models, allowing the comparative analysis of the results related to the different methods.

According to the data of Table 3, it is verified that the algorithms J48 and MLP behaved very poorly, when applied to the original database. It can be noticed a great imprecision in the forecast of intense precipitations, from the rate of true positive (TP Rate), with only 2.9% of hits for the J48 algorithm and no adjustment for the MLP. It should be noted that the true positive and false positive rates related to the MLP algorithm applied to the original (unbalanced) database had zero value, since there was not even any prediction of intense precipitation made by the model when applied to the test set.

It can also be observed that the application of the SMOTE algorithm, which added to the data set effective records with similar behavior to the records of the class that presented lower number of occurrences, gave a significant improvement in the results of both models. The MLP obtained a rate of 82.9% of hits related to the forecast of intense rainfall (> = 10 mm), higher than the rate of 71.4% obtained by the J48 algorithm.

Figures 1 and 2 show the graphs indicating the application of the MLP and J48 models, integrated to the SMOTE algorithm to the test base, which considers a period of 42 months, comprising the months of November 2008 to April 2012. Looking at the graphs is possible to verify the actual occurrences of intense rains and the correct forecasts made during the evaluated 42 months.

Despite to present a higher precision in the forecast of precipitation equal or higher than 10 mm the MLP algorithm applied to the balanced base presented higher number of false alerts (FP Rate equal to 12.6%) when compared to the algorithm J48 (FP rate equal to 9.9%).

According to Andrews, Diederich and Tickle (1995), artificial neural networks can be considered black-box processing models, that is, despite solving highly complex problems, they do not provide adequate interpretation of the results due to the incomprehensibility of model generated. Decision trees, on the other hand, have an intuitive representation, facilitating the interpretation of the model. In this perspective, a rule extracted from the decision tree generated by the J48 algorithm will be presented with the application of the SMOTE method. The presentation of a single rule allows to illustrate the correlation between some microscale variables obtained in the surface meteorological station and the occurrence of heavy rains. Further details on the various rules extracted from the decision tree can be found in Charles (2015). The rule presented here is in the form X => Y, where X is composed of the different meteorological variables that imply Y, the latter representing a situation of intense precipitation in the hour after the measurement of the variables.

Table 3. Metrics related to the different classification methods.

Method/data base	TP Rate	FP Rate	F measure	ROC Area
J48 with original data	0.029	0.000	0.050	0.541
MLP with original data	0.000	0.000	0.000	0.873
J48 after SMOTE application for class balance	0.714	0.099	0.052	0.842
MLP after SMOTE application for class balance	0.829	0.126	0.048	0.927

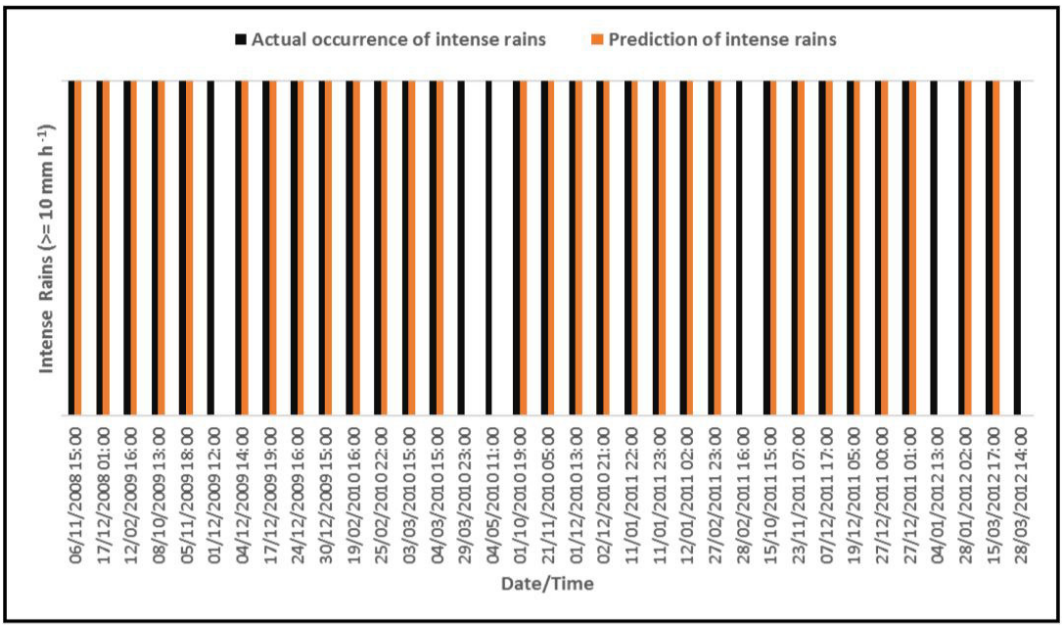

Figure 1. Predictions of intense rains compared with actual occurrences (MLP after SMOTE application).

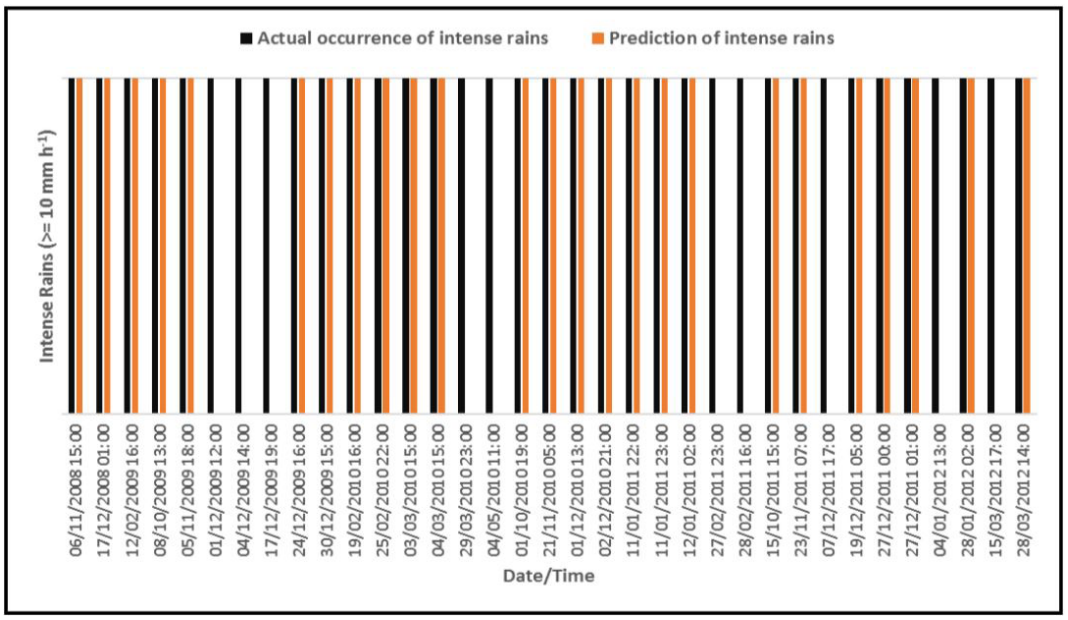

Figure 2. Predictions of intense rains compared with actual occurrences (J48 after SMOTE application).

- Average temperature 3 hours before the measurement >16.4; solar radiation variation 2 hours before the measurement < 22.5; it was recorded any precipitation in the current hour = not; average pressure 1 hour before the measurement < 901.3; average temperature 1 hour before the measurement > 18.9; it was recorded any precipitation 1 hour before = yes => intense precipitation.

As can be observed in the above rule, if no precipitation is occurring at the moment of the measurement, but 1 hour before rainfall is registered and, concurrently, the average temperature and pressure recorded in the previous hour is, respectively, higher than 18.9 °C and lower than 901.3 mb. If, in addition, the average temperature in the preceding three hours and the solar radiation variation in the preceding two hours are respectively above 16.4 °C and below 22.5 W m^{-2}, the model will forecast precipitation equal or higher than 10 mm. It should be noted that this rule referred to 96 instances of the database, 93 being correctly classified and 3 incorrectly classified by the model.

Analyzing Figures 1 and 2, it can be observed that the events studied here are concentrated between the months of October and March. There is only one occurrence of rains larger than 10 mm outside these months, precisely on 05/04/2010. It can be seen that neither of two models was able to correctly predict the event that occurred on that date. It is believed that this fact is due to the reduced number of records in the database containing variables related to the occurrence of intense rains outside the period of october-march. In this context, the constructed models were unable to identify patterns related to events occurring outside

this period. When analyzing the above-mentioned rule and, more specifically, the variable indicating the temperatures at the time of data collection and three hours before, it can be seen that temperatures (higher than 18.9 ºC and 16.4 °C, respectively), are not characteristics in the month of May in the city of Nova Friburgo. In short, it is understood that the models were adjusted based on the seasonality of the occurrences, not being able to predict situations outside the period characterized by the occurrence of intense rains.

It should be noted that the models developed in the present work differ in relation to other models related to the forecast of extreme rains due to the use of micro scale data collected in a surface meteorological station. More recent works, such as those by Ruivo et al. (2015), who study extreme precipitation events in Santa Catarina and Amazonia, or Yucel and Onen (2014), focusing on events of the same category in different regions of Turkey, or even from Pessoa et al. (2012), who tried to predict occurrences of severe convective events in three mini-regions of the Brazilian territory, use as base mesoscale data, involving a broader area. It is believed that the contribution of the present proposal is to enable the more precise elaboration of forecasts of extreme phenomena, considering regions with smaller area of coverage.

CONCLUSIONS

The present work had the objective of analyzing the data recorded by a surface meteorological station in the period between 2008 and 2012 to extract patterns of behavior of the meteorological variables that indicated the formation of precipitations equal or greater than 10 mm for the hour after the collection of the data.

Among the obtained conclusions, it can be mentioned that the class balancing carried out by the SMOTE algorithm made the efficiency in the identification of the cases of intense precipitations increase considerably; on the other hand, the number of false positives, i.e. false alerts, also increased considerably. It is believed, however, that from the perspective of population safety that may potentially be affected by the precipitations being predicted, the emission of false alarms has fewer consequences than the omission of correct alarms.

Another finding was that both models were not able to make correct forecasts outside the rainy season. It is believed that this fact would not have so much relevance, since between the months of April and September does not occur a great accumulation of rains. Failure to predict precipitation in excess of 10 mm h^{-1} during these months would not have as much impact as in the period from October to March, characterized by a higher accumulation of precipitation and, consequently, a greater probability of flooding or landslides as consequences of heavy rains.

The obtained results showed the satisfactory predictive capability of the models used here, motivating their future incorporation to the servers that collect the data of the meteorological station used in this study. Thus, automatic predictions could be made, with an advance of one hour, related to extreme events of precipitation in the surroundings of the place where the meteorological station is installed. These forecasts could support decision-making by government agencies, such as Civil Defense, in addition to preventing the population in general, with the availability of information, in real time, on the site www.clima.iprj.uerj.br. It is also expected that the work may contribute to the development of new research related to meteorological forecasting based on microscale data analysis and to the development of more effective alert systems, considering smaller areas.

REFERENCES

ALFIERI, L.; THIELEN, J.; PAPPENBERGER, F. Ensemble hydro-meteorological simulation for flash flood early detection in southern Switzerland. *Journal of Hydrology*, v. 424-425, p. 143-153, 2012. http://dx.doi.org/10.1016/j.jhydrol.2011.12.038.

AMENT, F.; WEUSTHOFF, T.; ARPAGAUS, M. Evaluation of MAP D-PHASE heavy precipitation alerts in Switzerland during summer 2007. *Atmospheric Research*, v. 100, n. 2-3, p. 178-189, 2011. http://dx.doi.org/10.1016/j.atmosres.2010.06.007.

ANDREWS, R.; DIEDERICH, J.; TICKLE, A. B. A survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, v. 8, n. 6, p. 373-389, 1995. http://dx.doi.org/10.1016/0950-7051(96)81920-4.

BLAGUS, R.; LUSA, L. Smote for high-dimensional class-imbalanced data. *BMC Bioinformatics*, v. 14, n. 1, p. 106, 2013. PMid:23522326. http://dx.doi.org/10.1186/1471-2105-14-106.

BORGA, M. *Realtime guidance for flash flood risk management*. Padova: University of Padua, 2008. 84 p. v. 2.

BOUCKAERT, R. R.; FRANK, E.; HALL, M. A.; HOLMES, G.; PFAHRINGER, B.; WITTEN, I. H. WEKA: experiences with a Java open-source project. *Journal of Machine Learning Research*, v. 11, p. 2533-2541, 2010.

CETEMA – CENTRO DE TECNOLOGIA EM MEIO AMBIENTE. *Estação meteorológica*. Rio de Janeiro: CETEMA, 2009. Available from: <www.clima.iprj.uerj>. Access on: 10 oct. 2017.

CHARLES, A. C. *Mineração de dados para previsão de eventos extremos de precipitação*. 2015. 120 f. Dissertação (Mestrado em Modelagem Computacional) - Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2015.

FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861-874, 2006. http://dx.doi.org/10.1016/j.patrec.2005.10.010.

FAYYAD, U. M.; PIATETSKY–SHAPIRO, G.; SMYTH, P. Knowledge discovery and data mining: towards a unifying framework. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD-96), 2., 1996, Portland, Oregon. *Proceedings*… California: AAAI, 1996.

HAN, J.; KAMBER, M. E.; PEI, J. *Data mining*: concepts and techniques. 3. ed. Massachusetts: Morgan Kaufmann Publishers, 2012. http://dx.doi.org/10.1007/978-1-4419-1428-6_3752.

HAYDEN, M.; DROBOT, S.; RADIL, S.; BENIGHT, C. C.; GRUNTFEST, E.; BARNES, L. Information sources for flash flood warnings in denver, CO and Austin, TX. *Environmental Hazards*, v. 7, n. 3, p. 211-219, 2007. http://dx.doi.org/10.1016/j.envhaz.2007.07.001.

HAYKIN, S. *Neural networks*: a comprehensive foundation. New York: Prentice Hall, 1999.

JOSHI, A.; KAMBLE, B.; JOSHI, V.; KAJALE, K.; DHANGE, N. Weather forecasting and climate changing using data mining application. *International Journal of Advanced Research in Computer and Communication Engineering*, v. 4, n. 3, p. 19-21, 2015. http://dx.doi.org/10.17148/IJARCCE.2015.4305.

KNEIS, D.; BÜRGER, G.; BRONSTERT, A. Evaluation of medium-range runoff forecasts for a 50 km2 watershed. *Journal of Hydrology*, v. 414-415, p. 341-353, 2012. http://dx.doi.org/10.1016/j.jhydrol.2011.11.005.

KOBIYAMA, M.; MENDONÇA, M.; MORENO, D. A.; MARCELINO, I. P. V. O.; MARCELINO, E. V.; GONÇALVES, E. F.; BRAZETTI, L. L. P.; GOERL, R. F.; MOLLERI, G. S. F.; RUDORFF, F. M. *Prevenção de desastres naturais*: conceitos básicos. Curitiba: Organic Trading, 2006. 109 p.

KRISHNA, G. V. A review of weather forecasting models-based on data mining and artificial neural networks. *IJCSC*, v. 6, n. 2, p. 214-222, 2015.

MAJNIK, M.; BOSNIĆ, Z. ROC analysis of classifiers in machine learning: a survey. *Intelligent Data Analysis*, v. 17, n. 3, p. 531-558, 2013.

MEDEIROS, V. S.; BARROS, M. T. L. Análise de eventos críticos de precipitação ocorridos na região serrana do Estado do Rio de Janeiro nos dias 11 e 12 de janeiro de 2011. In: SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS, 19., 2011, Maceió, AL. *Anais*… Porto Alegre: ABRH, 2011.

NOAA – NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION. *Flash flood early warning system reference guide*. Washington: NOAA/COMET, 2010.

OLAIYA, F.; ADEYEMO, A. B. Application of data mining techniques in weather prediction and climate change studies. *International Journal of Information Engineering and Electronic Business*, v. 4, n. 1, p. 51-59, 2012. http://dx.doi.org/10.5815/ijieeb.2012.01.07.

ONWUBOLU, G. C.; BURYAN, P.; GARIMELLA, S.; RAMACHANDRAN, V.; BUADROMO, V.; ABRAHAM, A. Self-organizing data mining for weather forecasting. In: IADIS EUROPEAN CONFERENCE DATA MINING, 2007, Lisbon. *Proceedings*… Lisbon: IADIS, 2007. p. 81-88.

PESSOA, A. S. A. *Mineração de dados meteorológicos pela teoria dos conjuntos aproximativos na previsão de clima por redes neurais artificiais*. 2004. 132 f. Dissertação (Mestrado) - Instituto Nacional de Pesquisas Espaciais, São Paulo, 2004.

PESSOA, A. S. A.; LIMA, G. R. T.; SILVA, J. D. S.; STEPHANY, S.; STRAUSS, C.; CAETANO, M.; FERREIRA, N. J. Mineração de dados meteorológicos para previsão de eventos severos. *Revista Brasileira de Meteorologia*, v. 27, n. 1, p. 61-74, 2012. http://dx.doi.org/10.1590/S0102-77862012000100007.

PETRE, E. G. A decision tree for weather prediction. *Universitatea Petrol-Gaze din Ploiesti*, v. LXI, n. 1, p. 77-82, 2009.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Curvas ROC para avaliação de classificadores. *IEEE Latin America Transactions*, v. 6, n. 2, p. 215-222, 2008. http://dx.doi.org/10.1109/TLA.2008.4609920.

REDMAN, T. C. *Data quality*: the field guide. Boston: Digital Press, 2001.

RODRIGUES, M. L. G.; FRANCO, D.; SUGAHARA, S. Climatologia de frentes frias no litoral de Santa Catarina. *Revista Brasileira de Geofísica*, v. 22, n. 2, p. 135-151, 2004. http://dx.doi.org/10.1590/S0102-261X2004000200004.

RUIVO, H. M.; VELHO, H. F. C.; SAMPAIO, G.; RAMOS, F. M. Analysis of extreme precipitation events using a novel data mining approach. *American Journal of Environmental Engineering*, v. 5, n. 1A, p. 96-105, 2015.

SALVADOR, H. G.; CUNHA, A. M.; CORRÊA, C. S. VEDALOGIC: um método de verificação de dados climatológicos apoiado em modelos minerados (A method of checking climatological data based on mining models). *Revista Brasileira de Meteorologia*, v. 24, n. 4, p. 448-460, 2009. http://dx.doi.org/10.1590/S0102-77862009000400007.

TAFNER, M. A. Redes neurais artificiais: aprendizado e plasticidade. *Cérebro Mente*, v. 2, n. 2, p. 1, 1998.

TAKSANDE, A. A.; MOHOD, P. S. Applications of data mining in weather forecasting using frequent pattern growth algorithm. *IJSR*, v. 4, n. 6, p. 3048-3051, 2015.

TAN, P.; STEINBACH, M.; KUMAR, V. *Introdução ao Data Mining*: mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

VERKADE, J. S.; BROWN, J. D.; REGGIANI, P.; WEERTS, A. H. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, v. 501, p. 73-91, 2013. http://dx.doi.org/10.1016/j.jhydrol.2013.07.039.

WANG, R. Y.; ZIAD, M.; LEE, Y. W. *Data quality*. Hinghan: Kluwer Academic Publishers, 2001. (The Kluwer International Series on Advances in Database Systems, 23).

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data mining*: practical machine learning tools and techniques. 3. ed. Massachusetts: Morgan Kaufmann Publishers, 2011.

YADAV, R. K.; KHATRI, R. A. Weather forecasting model using the data mining technique. *International Journal of Computers and Applications*, v. 139, n. 14, p. 4-12, 2016.

YUCEL, I.; ONEN, A. Evaluating a mesoscale atmosphere model and a satellite-based algorithm in estimating extreme rainfall events in northwestern Turkey. *Natural Hazards and Earth System Sciences*, v. 14, n. 3, p. 611-624, 2014. http://dx.doi.org/10.5194/nhess-14-611-2014.

Authors contributions

Anderson Cordeiro Charles: Manuscript structure, literature review, methods and analysis of the results.

Anderson Amendoeira Namen: Study orientation, literature review, methods, analysis of the results and technical review of the manuscript.

Pedro Paulo Gomes Watts Rodrigues: Analysis of the results and technical review of the manuscript.