10th Brazilian Congress on Computational Intelligence (CBIC'2011), November 8 to 11, 2011, Fortaleza, Ceará Brazil

LIGHTNING FORECAST USING DATA MINING TECHNIQUES ON HOURLY EVOLUTION OF THE CONVECTIVE AVAILABLE POTENTIAL ENERGY

J. A. S. Sá[1], A. C. Almeida[2], B. R. P. Rocha[1], M. A. S. Mota[3], J. R. S. Souza[3] and L. M. Dentel[1]

1. Electrical Engineering Post Graduation Program - Institute of Technology - Federal University of Pará
e-mails: jalbertosa@ufpa.br; brigida@ufpa.br; lauredentel@ufpa.br
2. Mathematics Faculty - Federal University of Pará
e-mail: arthur@ufpa.br
3. Geosciences Institute – Federal University of Pará, Belém, Pará, 66075-110, Brazil
e-mails: aurora@ufpa.br; jricardo@ufpa.br

Abstract – This study presents a method developed for lightning forecasting in eastern Amazonia, based on the estimates of the hourly evolution of the convective available potential energy (CAPE). The CAPE is a computed index of the air stability situation over a given area of the Earth. This parameter is determined from vertical profiles of temperature and humidity of the atmosphere, obtained through radiosondes. The CAPE values may also be estimated during the period between soundings, by using the meteorological variables observed continuously at surface weather stations. Two data mining techniques were used for the forecasts: k-Nearest Neighbor and Decision Tree. For the calculation of the CAPE and its estimated hourly evolution, we used radio soundings data made available by a site of the University of Wyoming, in addition to surface temperature data provided by the METAR code, both collected at the Belém- Brazil airport, during 2009. The CAPE index levels, indicative of strong convection in the area were compared to data of actual lightning activity, provided by the STARNET detection system, in a circular area of 100 km radius, centered at that airport. The angular coefficient of the adjusted line equation to the hourly evolution values of the CAPE and the average value of the CAPE were the predicting attributes, while the number of lightning flashes detected by the STARNET was the classification attribute. The results indicated that it is possible to predict the lightning class of occurrences with an accuracy of the 70%, in this research area.

Keywords – Lightning Forecast, K-Nearest Neighbor, Decision Tree, Convective Available Potential Energy (CAPE).

1 Introduction

The lightning forecast allows prior warning of the risks associated to that atmospheric phenomenon, at a given location and time, in order to reduce its potential damage. [1] described in detail, with an engineering approach, various types of lightning damage (direct and indirect) on humans, which emphasized its risks and confirmed the importance of trying to predict them. Recent research have suggested the use of the Convective Available Potential Energy (CAPE) as one of its key elements, indicative of impending electrical storms [2]-[5]. This research presents a method developed for lightning forecasting in eastern Amazonia, based on the estimates of the hourly evolution of the Convective Available Potential Energy [6]. The study analyzed the degree of applicability of lightning prediction software that uses the angular coefficient of the adjusted line equation to the hourly evolution values of the CAPE and the average value of the CAPE as predictor attributes. The classifiers were created through of two data mining techniques: K-Nearest Neighbor (the most used instance-based machine learning method) and Decision Tree.

2 Instance-Based Machine Learning

An instance-based machine learning method, essentially uses data stored for the classification of a new element, without an induced set of rules, i.e., classifies a new example based on similar examples that exist in the database, using a particular metric [7]. The basic idea of these methods is shown in Figure 1.

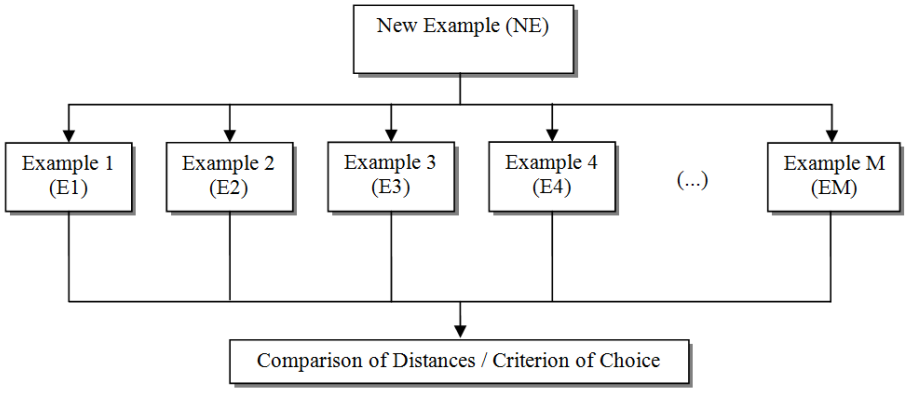

Figure 1 – Flow chart for instance-based machine learning methods [7]

It is perceived that a new example is compared to all existing examples, by calculating a distance measure between each pair. Among the new examples were selected those with the shortest distance, to predict a class of the new instance (the default is the class of most examples selected). Examples of these techniques are for instance: the K-NN (K-Nearest Neighbors), and Case-Based Reasoning, among others.

The most used technique is the K-NN. Since 1960, with the increased processing power of computers, this method has become widely used in the field of pattern recognition [7]. The K-NN is based on learning by example.

Basically, the database consists of examples described by *n* attributes. The proximity of the attribute values is usually measured in terms of Euclidean distances, being selected as k closest examples those which have the smallest Euclidean distances to the new example. Consider the points $\boldsymbol{X}_1 = (x_{11}, x_{12}, ..., x_{1n})$ and $\boldsymbol{X}_2 = (x_{21}, x_{22}, ..., x_{2n})$. Both have n attributes and the Euclidean distance between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is obtained by equation 1:

$$d(\boldsymbol{X}_1, \boldsymbol{X}_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \quad (1)$$

The use of the K-NN technique for building a classifier requires some details. The first, concerns the value of k. One way to determine this is through incremental recursive tests. Starting with k = 1 it is checked the error rate of the classifier. The process stops (at a certain value of k) when the error is acceptable. The second refers to the validation of the predictor. In the technique are used predictive attributes and known classes that represent the features of the examples contained in the database. It is necessary to divide the data into two sets (training and validation) to obtain greater reliability on the predictor generated, i.e., the predictor will only be considered effective if it properly classifies a large amount of unused data, in the training process.

3 Decision Tree Induction

Another technique frequently used in data mining is the decision tree induction. A decision tree is a structure for data classification, by recursive composition of elements to reach a logical decision. Basically, it has the function to stratify continuously a data set in order to generate subsets formed by elements belonging to a single class, allowing at the end to create a set of rules for future classifications [8].

Figure 2 (a) describes the components of a decision tree: the Nodes represent the possible attributes associated with an event. The first node is called root and represents the attribute with largest information gain; Branches represent the attributes values; and Leaves represent the classes. When building a decision tree, it is possible to create a set of rules, as shown in Figure 2 (b). These are written considering the trajectory of the root node to the leaf nodes. Figure 2 (c) shows the input data partitioning.

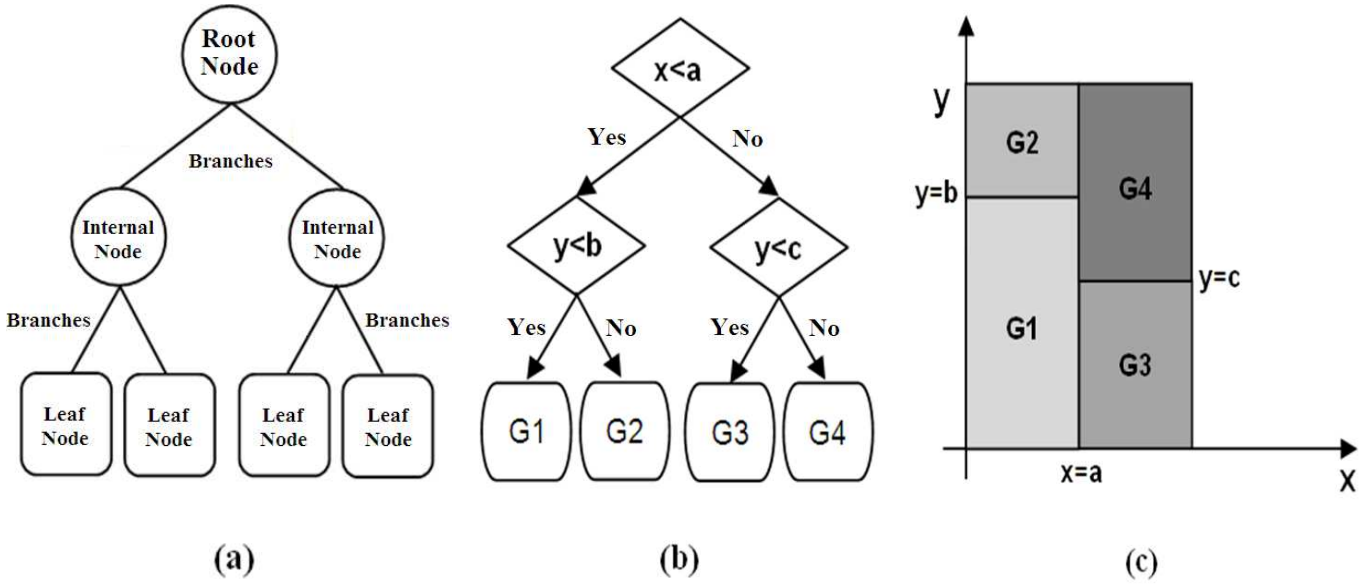

Figure 2 – (a) Components of a decision tree; (b) Flow chart of the rules of a binary decision tree; and (c) Input data partitioning (groups).

4 Convective Available Potential Energy (CAPE)

The Convective Available Potential Energy (CAPE) is the positive area of a Thermodynamic Diagram, obtained by equation 2:

$$CAPE = g \int_{LFC}^{EL} \left(\frac{\theta_{e(LFC)} - \theta_{es}}{\theta_{es}} \right) dZ \tag{2}$$

where:

- LFC: Level of Free Convection;
- EL: Equilibrium Level;
- θe: Equivalent Potential Temperature of the Air Parcel;
- θes: Saturated Equivalent Potential Temperature of the Atmospheric Environment.

The CAPE exists when the difference between the equivalent potential temperature of the air parcel (θe) and the saturated equivalent potential temperature of the environment (θes) is positive. This means that the pseudo-adiabatic of the displaced air parcel is warmer than the environment (unstable situation). Then the area between the pseudo-adiabatic and the sounding profile is proportional to the amount of kinetic energy that the parcel receives from the environment.

It is possible to estimate the CAPE hourly values between the soundings carried out at 00:00 UTC and 12:00 UTC, by observing the variations of pressure, air temperature and dew point temperature, measured at the surface, i.e., through estimates of the new potential temperature of the parcel [8]. These estimated values of CAPE allowed the construction of the lightning predictor, and they were used as predictive attributes.

5 Methodology

Information used to create the lightning predictor software:

- Radio soundings data collected at the airport in Belém, Brazil, for the year 2009 (available at the website of the Wyoming University);
- Hourly variations of the pressure, air temperature and dew point temperature, all measured at the surface (provided by the METAR code);
- Data of the occurrence of lightning (available from the STARNET network database), observed within a circular area of 100 km, centered at the Belém airport.

Only the radio soundings made at 12:00 UTC (9:00 LT) were used, due to the interest in making predictions of lightning strikes to the area surveyed, in the afternoon. The lightning incidence data were limited to the interval from 15:00 UTC to 22:00 UTC (12:00 LT to 19:00 LT), which is the hourly interval when most local lightning events occur.

Initially, the software performed the collection and stratification of the necessary information from the radiosonde data and surface variables (METAR code), available at a site of the University of Wyoming; and lightning data, available at the STARNET data base, for the period and local considered (circular area of 100 km radius, centered at the Belém airport, Brazil, during 2009).

Then, the value of CAPE (calculated for 12:00 UTC) and the estimated values of CAPE (calculated for 13:00 UTC, 14:00 UTC and 15:00 UTC) were used to calculate the angular coefficient of the adjusted line equation to the hourly evolution values of the CAPE and the average value of the CAPE which were used as predicting attributes, while the number of lightning flashes detected by the STARNET system was the classification attribute.

For the classifier attribute (three categories of classes values known) were adopted: Class 1 (no lightning), Class 2 (occurrences between 1 and 500 lightning flash) and Class 3 (number of occurrences greater than 500 lightning flashes per day).

In this study, 222 examples were obtained (valid days). The elimination of 143 days of 2009 was due to factors such as lack of radio soundings and/or surface data, and filtering of the lightning recorded by STARNET with less than five sensors. From the examples selected, the software chose, at random, 50 examples to compose the validation set, the remainder being for the training of the classifiers (K-Nearest Neighbor Classifier and Decision Tree Classifier).

6 Results

To analyze the results of the lightning predictors, the visualization tool called confusion matrix was used. In this technique, commonly used in supervised learning, each column of the matrix represents the instances of expected results, while each row corresponds to instances of actual results.

The results of the decision tree classifier are shown in Table 1. For this predictor 70% of the forecasts were correct.

Table 1 – Confusion Matrix for the Decision Tree Classifier

		Prediction outcome		
		Class 1	**Class 2**	**Class 3**
Actual value	**Class 1**	*1*	*1*	*0*
	Class 2	*3*	*31*	*2*
	Class 3	*2*	*7*	*3*

For K-Nearest Neighbor Classifier was used as the criterion of choice: the first k with level of accuracy equal or superior to the prior technique. Following this procedure it was obtained the value k = 3 to a level of correct predictions also of 70%. Table 2 shows the confusion matrix of this predictor.

Table 2 – Confusion Matrix for K-Nearest Neighbor Classifier

		Prediction outcome		
		Class 1	**Class 2**	**Class 3**
Actual value	**Class 1**	*0*	*1*	*1*
	Class 2	*1*	*30*	*5*
	Class 3	*0*	*7*	*5*

7 Conclusions

The results showed that the lightning classifiers, based on CAPE estimated values, represents a subsidy for the studies on lightning forecasts. The errors associated with the predictors, possibly are related to the fact that CAPE is a necessary but not a sufficient parameter index for the formation of deep convection, suggesting the need of future studies of these predictors, associated with other indicators of convection, as for example the CINE (Convective Inhibition).

8 References

[1] Cooray, V., C. Cooray, and C. J. Andrews, Lightning caused injuries in humans, Journal of Electrostatics, vol. 65, Issues 5-6, 27th International Conference on Lightning Protection, pp. 386-394, 2007.

[2] Frisbie, P. R., J. D. Colton, J. R. Pringle, J. A. Daniels, J. D. Ramey JR., and M. P. Meyers, Lightning prediction by WFO Grand Junction using Model Data and Graphical Forecast Editor Smart Tools. In: Conference on the Meteorological Applications of Lightning Data, 4., 2009, Phoenix. Elec. Proceedings Phoenix: MAS, 2009.

[3] Kaltenbock, R., G. Diendorfer, and N. Dotzek, Evaluation of thunderstorm indices from ECMWF analyses, lightning data and severe storm reports, Atmospheric Research, Volume 93, Issues 1-3, 4th European Conference on Severe Storms, pp. 381-396, 2009.

[4] Liou, Y.-A. and S. K. Kar, Study of cloud-to-ground lightning and precipitation and their seasonal and geographical characteristics over Taiwan, Atmospheric Research, vol. 95, Issues 2-3, pp. 115-122, 2010.
[5] Yamane, Y., T. Hayashi, A. M. Dewan, and F. Akter, Severe local convective storms in Bangladesh: Part II.: Environmental conditions, Atmospheric Research, vol. 95, Issue 4, pp. 407-418, 2010.
[6] Mota, M. A. S., I. M. O. Silva, and J. R. S. Souza, Atmospheric thermodynamic conditions leading to severe lightning storm cases in Eastern Amazônia, Internal Report, Geosciences Institute – Federal University of Pará, Belém, 2011.
[7] Han, J. and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann Publishers, 2006.
[8] Witten, Ian H. and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Elsevier Inc., 2005.