Application of K-Nearest Neighbor Technique to Predict Severe Thunderstorms

Himadri Chakrabarty
Associate Professor and Head, Dept. of Computer Sc.,
Surendranath College, Calcutta University, Visiting Professor, Inst. of Radiophysics and Electronics
Calcutta University
Kolkata, India

Sonia Bhattacharya
Contractual Whole Time Teacher
Dept. of Computer Science
Panihati Mahavidyalaya
Barasat State University
Kolkata, India

ABSTRACT

Machine learning techniques are used in different types of pattern recognition works. Nowadays, these techniques are applied in meteorological fields for prediction purpose. In this paper, the pattern to be recognized is the severe weather event of squall-thunderstorms. Prediction of severe thunderstorms are done here by applying K-Nearest Neighbor (K-NN) technique. K-NN is a very good classifier which can classify two classes of events ‘storm days’ and ‘no storm days’. It is a non-parametric method. Three types of weather parameters such as moisture difference, dry adiabatic lapse rate and vertical wind shear are considered here as predictors. Both surface as well as upper air data which are measured by radiosonde/ rawindsonde in the early morning are used in this case. Weather forecasting is a challenging job because of the dynamic behavior of the atmosphere. ‘Storm days’ are predicted correctly more than 91% and both ‘storm and no storm days’ are classified more than 82% accuracy, having a lead time around 12 hours.

General Terms

Classifier, RSRW, Squall

Key Words

Squall-thunderstorm, Machine Learning, K-Nearest Neighbor and Similarity Measure

1. INTRODUCTION

Natural calamities cause heavy destruction to both life and property. Prediction of such calamities well in advance is inevitable. A squall-thunderstorm is severe weather phenomenon characterized by the presence of strong gusty wind, lightning, thunder and sometimes hails. The strong wind is known as squall, which has speed of at least 45 kilometers or more per hour with minimum duration of 1 second or above. Severe thunderstorm is a highly destructive force of nature and the timely tracking of the thundercloud direction is of paramount importance to reduce the property damages and human casualties. Substantial research work was carried out in the last two decades about the understanding of the life cycle of thunderstorm. Dynamic behavior of weather makes the forecasting a formidable challenge, [1]; specially, the

prediction of thunderstorm is still a very challenging problem [2]. Annually, it is estimated that thunderstorm related devastation causes billion dollars of damages worldwide through forest fires, shutdown of electrical plants and industries, property damages etc.,[3]. Generally, various surface as well as upper air weather data are required to predict squall-thunderstorms, [4], as these weather parameters have effect on the genesis of such type of severe weather phenomenon.

Several climatic parameters play roles to generate squall-storms. In this paper, three types of weather variables such as moisture difference, adiabatic lapse rate, and wind-shear are considered as input parameters or predictors. Here, the predictand is the squall-storm. Moisture difference indicates the measurement of atmospheric humidity. Air mass advection can influence air temperature and humidity [5]. The temperature change in unsaturated air can be predicted by the dry adiabatic lapse rate [5]. Atmospheric instability can be indicated by moisture difference and adiabatic lapse rate [5]. Chakrabarty1 et al predicted the ‘occurrence’ and ‘no occurrence’ of squall-storms in their previous work in 2013 using the weather data of moisture difference and dry adiabatic lapse rate. These weather data were sensed by radiosonde and measured at different geopotential heights of the upper atmosphere. In that previous work [4] there were ten input variables, where five variables are moisture differences at five different heights from the surface level up to the height of 4.5 kilometers of the upper atmosphere and the other five variables are the dry adiabatic lapse rates from the surface up to the height of 9.6 kilometers of the upper air. Here, in this present paper, the additional three input variables of vertical wind shear, sensed by rawindsonde are considered which are measured at three different altitudes up to 35 kilometers of the upper atmosphere. A clockwise turning of the wind shear vector with height favors the development of a cyclonic, right-moving storm; while conversely, a counterclockwise turning favors the anticyclonic, left-moving storm [6]. So, this present paper is based on the forecasting of severe thunderstorms using total thirteen RSRW input data recorded at 6.00 a.m. local time (00.00UTC) in Kolkata (22.3^{o}N/88.3^{o}E), India. Chakrabarty2 et al. cited in the previous work in 2013 that cumulonimbus clouds (i.e., thunderclouds) are developed through the levels of strong vertical wind shear [7]. All these meteorological data on squall-storms at the location of Kolkata were procured during the period of 18 years from 1990 to 2008 for the months of March-April-May (MAM). These three months are known as pre-monsoon season over North-East India.

In this present work, K- nearest neighbor (K-NN) model is used for the prediction purpose. K-NN is an important pattern recognition technique in soft computing. Sharma et al. [8] forecasted storms using soft computing method. Chakrabarty 1 et al. [4] nowcasted severe storms using K-NN models having different values of K. K-nearest neighbor (K-NN) is one of the best data mining algorithms for classification, which is used in different applications. K-NN algorithm was originally suggested by Cover in 1968. This algorithm operation is based on comparing a given testing data point with training data points and finding the training data points (neighbors) that are similar to it, and then predict the class label of these neighbors [10]. K-nearest neighbor algorithm is a non-parametric method for classifying objects based on closest training examples in the feature space. It is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. K-NN technique is applied by Li et al. [11], to forecast solar flare. Brath et al. [12] and Jayawardena et al. [13] applied K-NN method for flood forecasting. Jan et al. [14] used data mining technique for the seasonal to Inter-Annual Climate prediction.

Here in this paper, K-NN technique has been applied on the weather data to forecast the 'occurrence'/'no occurrence' of squall-storm having around 12 hours lead time.

2. DATA

2.1. Data Collection

All the weather data were collected from India Meteorological Department, Govt. of India during the period of 18 years from 1990 to 2008 for the months of March-April-May. The data were recorded at 06.00 a.m. local time (00:00 UTC) by radiosonde and rawindsonde over Kolkata, North-East India. Here data have been considered both for 'squall' and 'no-squall' days. The numbers of 'squall-storm' days are 69 and 'no squall-storm' days are 315.

2.2. Data Description

There are thirteen weather variables which are used in this work. Three types of weather parameters such as moisture difference, adiabatic lapse rate and wind shear are considered for this analysis. These variables were observed from the surface level to different geopotential heights of the upper atmosphere. First five weather parameters x_1, x_2, x_3, x_4, and x_5 indicate vertical moisture difference profile at surface level, and at 1000hpa (approximately 75 meters), 850hpa (approximately 1500 meters), 700hpa (approximately 3100 meters), and 600hpa (approximately 4500 meters) respectively. Moisture difference is the difference between dry bulb temperature and the dew point temperature. This moisture, when carried out to the upper atmosphere by vertical wind shear forms the thundercloud [5]. Other five weather predictors x_6, x_7, x_8, x_9, and x_{10} indicate dry adiabatic lapse rates, i.e., the dry bulb temperature difference between surface and 850hpa (approximately surface to 1500 meters), between 850hpa and 700hpa (approximately 1500 to 3100 meters), between 700hpa and 600hpa (approximately 3100 to 4500 meters), between 600hpa and 400hpa (approximately 4500 to 7500 meters), and between 400hpa and 300hpa (approximately 7500 to 9600 meters) respectively. Dry adiabatic lapse rate is the measure of the conditional instability of the atmosphere [16]. The more the atmosphere is unstable, more moisture would be carried out to the upper atmosphere from the surface level to form thunderclouds [16]. The last three atmospheric input parameters x_{11}, x_{12}, and x_{13} indicate vertical wind shear at 900hpa to 700hpa (approximately 980 meters to 2500 meters), at 700hpa to 500hpa (approximately 2500 meters to 12340 meters), and at 500hpa to 200hpa (approximately 12340 meters to 35000 meters) respectively. Vertical wind shear plays an important role to carry the moisture to the upper atmosphere. This is very significant for the formation of thundercloud.

3. METHODOLOGY

K-nearest neighbor technique has been applied here for correct prediction of 'squall-days' and 'no squall days'. Detail description of this method is given below.

3.1. K-Nearest Neighbor (K-NN)

Yakowitz extended the K-nearest neighbor method constructing a robust theoretical base for it and introduced it into the successful forecast in the hydrological research. K-nearest neighbor method is applied to recognize the 'squall' class pattern and as well as the 'no squall' class pattern in this paper. The total data set is divided into two classes and these are training dataset and test dataset. The total number of squall data is 69 and total number of no squall data is 315. The number of squall data in the training data set is 35 and in the test data set is 34. The number of no squall data in the training dataset is 35 and in the test data set is 280. In the training set total number of squall and no squall data is 70. In the test set the total number of squall data is 34. The training data set is arranged consecutively by squall and no squall data vector. The similarity measure has been taken between each data vector of test set with each data vector of training set. Similarity between two observation vectors say,

$\underline{p} = (p_1, p_2, \ldots\ldots, p_\gamma)$, $\underline{q} = (q_1, q_2, \ldots\ldots, q_\gamma)$ is defined as

$$\frac{\sum_{i=1}^{\gamma} p_i q_i}{\sqrt{[\sum_{i=1}^{\gamma} p_i^2 \sum_{i=1}^{\gamma} q_i^2]}} \qquad (1)$$

The similarity measures between two vectors reflect the cosine of the angles between them. The similarity is more if the angle is smaller. The similarity measure indicates vicinity between the two vectors (one test vector and one training vector) with each other. The cosine angle for each of the test data vector with each of the training data vector is determined. These cosine angles are arranged in the decreasing order. As the number of training data vectors is 70, the numbers of cosine angles are also 70. Half numbers of cosine angles are considered for analysis at first. For each of the squall data vector, if maximum number of 'squall' data appears within half of the set of cosine angles then it is to be considered as properly classified as 'squall' class. Similar thing happens for 'no squall' class. Here the value of k is 35.

4. RESULT

Table 1:

Squall Class	**Total number of squall days in the test data set = 35**	**Number of accurately classified days with % in test data set**	**31, 91.42%**
		Misclassification rate	**0.08**
No Squall Class	**Total number of no squall days in the test data set =**	**Number of accurately classified days with % in test data set**	**246, 87.63%**
		Misclassification	**0.12**

	280	**rate**	
Total Dataset	**Total number of squall days and total no squall days in the test data set = 315**	**Total Number of accurately classified days with % in test data set**	**277, 88.21%**
		Misclassification rate	**0.11**

It is observed from this analysis that applying K-nn technique more than 91% of 'squall storm' days, and more than 87% of the 'no squall storm' days are properly classified for the prediction purpose. When both the 'squall storm' and 'no squall storm' data are combined more than 88% are properly classified. The misclassification rate is also very low for 'squall storm' days (0.08), for 'no squall storm' days (0.12). Chakrabarty1 et. al., in 2013 applied K-nn technique for the prediction of severe thunderstorm having around 12 hours lead time. They used two types of weather parameters such as moisture difference and dry adiabatic lapse rate. These parameters are considered from the surface level up to the five different geopotential layers of the upper air. So there are 10 weather parameters. They got only 55.55% of the 'squall storm' days which are properly classified. When they applied modified K-nn technique (where k = 3) they obtained more than 87% accurate classification of the 'squall storm' days and more than 71% accurate classification for 'no storm' days. But here in this paper K-nn method (where k = 35) are used on the 13 weather variables where five variables represent the moisture difference, other five variables represent the dry adiabatic lapse rate and the rest 3 variables represent the vertical wind shear measured at different layers of the upper air. K-nn technique yields a better result in this case using 13 weather variables than that obtained by the same technique on the 10 weather variables in the paper of Chakrabarty1 et. al [4]. Misclassification rate for squall class is 0.08 and for no squall class is 0.12, and for both squall and no squall data the misclassification rate is 0.11, which is a very significant result.

5. DISCUSSION AND CONCLUSION

The challenge that has been undertaken for this forecasting work is the proper selection of the machine learning technique to get accurate prediction using only the three types of input weather variables: moisture difference, dry adiabatic lapse rate and vertical wind shear at different heights recorded in the early morning (00:00UTC). Instability decreases as low-level moisture decreases. Instability occurs when a parcel of air is warmer than the environmental air and rises on its own due to positive buoyancy. Instability is often expressed using positive CAPE. Instability allows air in the low levels of the atmosphere to rise into the upper air. Without instability, the atmosphere will not support deep convection which generates thunderstorms. Instability can be increased through daytime heating. The wind shear carries the moisture from the surface level to the upper atmosphere to form thunderclouds. Regardless of the direction of shear, this process induces flow perpendicular and to the left of the shear vector, and thus by itself introduces a leftward deflection of the motion from the shear [17]. The result of the model shows that K-nn technique can classify 91.42% squall data, 87.63% no squall data and 88.21% of both 'squall-no squall' data accurately with a lead time of around 12 hours. Generally atmospheric surface parameters, upper air parameters measured by radiosonde, Doppler radar and satellite imageries are required to predict severe storm in a conventional way. Selection of the input parameters is an important factor in the prediction purpose. The lead time of 12 hours is sufficient to make the people alert from the destructive event of severe thunderstorms.

6. ACKNOWLEDGEMENT

The authors are thankful to India Meteorological Department, Govt. of India for providing atmospheric data to do this research work.

7. REFERENCES

[1] Maqsood, Imran, Khan M. R., Huang G.H., and Abdalla R "Application of soft computing models to hourly weather analysis in southern Saskatchewan, Canada", 2005, Engineering Applications of Artificial Intelligence,Vol. 18, Issue 1, pp. 115-125.

[2] Ludlam F.H "Severe Local Storms: A Review", Sept 1963, Meteorological Monographs, Vol. 5, pp. 1-30, American Meteorological Society.

[3] Chakrabarty Himadri 2, Murthy C. A., Bhattacharya Sonia and Das Gupta Ashish, "Application of Artificial Neural Network to Predict Squall-Thunderstorms Using RAWIND Data", May-2013 International Journal of Scientific & Engineering Research (IJSER),Volume 4, Issue 5, pp. 1313-1318, ISSN: 2229-5518.

[4] Chakrabarty1 Himadri, Murthy C. A., and Das Gupta Ashish, "Application of pattern recognition techniques to predict severe thunderstorms", 2013, International Journal of Computer Theory and Engineering (IJCTE), Vol. 5, No. 6, pp. 850-855, ISSN: 1793-8201.

[5] Moran, J.M., Moran M.D. and Pauley P.M., Meteorology: "The Atmosphere and the Science of Weather", 5th Edition, Chapter 6, Prentice Hall, 1997.

[6] Rotunno, Richard and Klemp Joseph B "The Influence of the Shear-Induced Pressure Gradient on Thunderstorm Motion", 1982, Monthly Weather Review, Vol. 110, pp. 136-151.

[7] Fujita Tetsuya, "Analytical Mesometeorology: A Review", 1963, Meteorological Monographs, Vol.5, No. 27, pp. 77-125, American Meteorological Society.

[8] Sharma Sanjay, Dutta Devajyoti, Das J, and Gariola R.M., "Nowcasting of severe storms at a station by using the Soft Computing Techniques to the Radar Imagery", 5thEuropean Conference on Severe Storms, Landshut-Germany, 2009.

[9] Cover, Thomas M., "Estimation by the Nearest Neighbor Rule, 1968 IEEE Transactions on Information Theory", Vol. IT-14, No. 1,pp,.50-55.

[10] Moradian, Mehdi and Baraani Ahmad, "KNNBA: K-Nearest-Neighbor Based Association Algorithm, 2009, Journal of Theoretical and Applied Information Technology", Vol. 6, No.1, pp. 123-129.

[11] Rong Li, Wang Hua-Ning, He Han, Cui Yan-Mei and Du Zhan-Le, "Support Vector Machine combined with K-Nearest Neighbors for Solar Flare Forecasting", Chinese Journal of Astronomy and Astrophysics,,2007 Vol. 7, pp. 441-447.

[12] Brath, A., Montanari A and Toth E, 2002, "Neural networks and non-parametric methods for improving real-time flood forecasting through conceptual hydrological models", Hydrology and Earth System Sciences, Vol. 6 (4), pp.-627-640.

[13] Jayawardena, A.W., Fernando D.A.K. and Zhou M.C., 1997, "Comparison of Multilayer Perceptron and Radial Basis Function networks as tools for flood forecasting", Proceedings of the Conference Water-Caused Natural Disasters, their Abatement and Control, held at Anaheim, California, Publ. no. 239.

[14] Zahoor Jan, Abrar Muhammad, Bashir Shariq and Mirza Anwar M., 2008, "Seasonal to Inter-Annual Climate Prediction Using Data Mining KNN Techniques", Communications and Computer and Information Science, Vol. 20, ISSN: 1865-0929, PP. 40-51.

[15] Yakowitz, S., 1987, "Near neighbor or method for time series analysis", J, Time-series Analysis, 8, 235-247.

[16] Volland, Hans, 1995, "Handbook of Atmospheric Electrodynamics", Vol. 1, ISBN: 0-8943-8647-0(V. 1)

[17] Kristen L. Corbosiero and Molinari John "The Relationship between Storm Motion, Vertical Wind Shear, and Convective Asymmetries in Tropical Cyclones", Journal of Atmospheric Sciences, 2002 Volume 60 Page No. 366.