Application of Machine Learning Technique to Predict Severe Thunderstorms using upper air data

Dyuti Chatterjee, Dr. Himadri Chakrabarty

Abstract— Squall-thunderstorm is a mesoscale severe weather event. The weather parameters like pressure, temperature from surface level to different levels of the upper atmosphere, cloud coverage, sunshine hour and lapse rate play an important role to generate the Thunder cloud. In this article, our focus is to establish a possible connection between the occurrence of the thunderstorms in Kolkata, North-East India and different types of atmospheric predictors. Here data have been considered both for 'squall-storm' and 'no-squall-storm' days. The total data set has 485 data from the year 1969 to 2002. K-Nearest Neighbor (K-NN) techniques have been applied in this research work. The application of K-NN method has been found to be useful in classifying the 'squall-storm' and 'no-squall-storm' days.

Index Terms— K-NN method, lapse-rate, nearest-neighbor, radiosonde, similarity measure, squall, Thunderstorm

1 INTRODUCTION

Squall-thunderstorm is an extreme weather phenomenon where a sudden and high speed wind (not less than 45 km/hr) is present usually along other severe activities like smart shower, lightning, thunder and sometimes hail [1]. Severe thunderstorm causes various types of damages of mankind creating socio-economic loss. The timely tracking of different weather parameters and thundercloud direction is very much necessary 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 very challenging [2] specially the prediction of thunderstorms. Superposition of upper-level divergence over lower-level convergence is the favorable process for convective storm generation [3].

It has been observed that a cluster of cumulonimbus (Cb) cells form thundercloud. The Cb cloud cells get arranged along a line namely the "squall line". Squall lines typically bow out due to the formation of a mesoscale high pressure system which forms within the convective area. This is formed due to strong descending motion behind the squall line, and could come in the form of a downburst [4].

Here the weather data considered are measured by radiosonde over Kolkata (22.3°N/88.3°E). Generally, Radiosonde sends data twice a day, one is at 5.30 am and the other is at 5.30 pm. Here we consider the morning (5:30 am) data on which K-Nearest Neighbor Technique has been applied. Here the lead time for nowcasting is 12 hours to 14 hours which is the sufficient time to make the people alert from the catastrophic effect of the severe weather event.

In this paper, eight types of upper air weather parameters such as Cloud coverage (Nh), Sunshine hour(SSH), Pressure at freezing point (FRZ), adiabatic lapse rate at different geopotential heights of the atmosphere (X4, X5, X6, X7 and X8) are considered as input parameters or predictors. Here, the predictand is the 'squall-storm' and 'no storm'.

Cloud coverage, sunshine hour and lapse rate play important role to generate the Thunder cloud. The cloud coverage (Nh) has an essential role in determining the local climate. Sunshine hour (SSH) indicates the duration of solar radiation on the earth. Surface heating, as well as the local atmospheric heating can be determined by this. The pressure at the freezing level (FRZ) of the upper atmosphere is one of the important features for thunder cloud formation.

The atmospheric lapse rate is the rate at which the temperature of the atmosphere changes with changing altitude. Adiabatic lapse rate of the atmosphere is the measure of the instability of the atmosphere [5]. The more unstable is the atmosphere; much moisture will enter from the sea to the land generating the thundercloud. In this study, dry adiabatic lapse rate is considered. Chakrabarty 1 et al. predicted the 'occurrence' and 'no occurrence' of severe thunderstorms in their previous work in 2013 using the weather data of moisture difference and dry adiabatic lapse rate. The vertical stability of the atmosphere largely depends upon the lapse rates. Thus forecasting of thunderstorms depends upon the lapse rates.

As warm air rises to upper atmosphere through the freezing level below 0˚C it gets condensed and goes through ice nucleation and multiplication processes. Clouds at higher altitude, usually called "High clouds" (usually above 7 Km

- *Dyuti Chatterjee is a PhD student under Institute of Radio Physics and Electronics, University of Calcutta, India, PH-+91 9748028103. E-mail: dyuti3@gmail.com*
- *Himadri Chakrabarty is currently an Associate Professor in Dept. of Computer Science, Surendranath College, Calcutta University, Kolkata, India, PH-91 9433355720, E-mail:hima.c@rediffmail.com*

height from ground level of the Earth) sometimes contain ice crystals or supercooled water. Clouds of vertical extent like cumulus also contain ice particles. At a temperature which is not far below 0˚C, supercooled liquid droplets and ice can exist simultaneously [6]. In fact, the transition of liquid water to solid state by the effects of mechanical pressure and the solutes in water particles is not fully understood. In previous experimental works of Koop et al. [7], it is found that the homogeneous ice formation from water solutions are independent of the solute natures, rather dependent upon the water vapor pressure of the solution and of pure water at that condition.

The homogeneous freezing of cloud evolution happens at a very low temperature (<-35˚ C), thus when the air parcel reaches at that temperature, the major fraction of the parcel becomes already frozen by heterogeneous freezing, [8]. It can be said that the ice formation rate is a function of water activity and pressure. The process involved is Bergeron process [8], which is a vapor diffusion process of growing ice crystals at the expense of droplets which evaporate. As some liquid water drops are nucleated, the drops grow by a process called 'riming'. It happens when the supercooled liquid drops of water vapor gets deposited on it by collision, [9].

There is a certain height or pressure in the upper air, above which the air becomes saturated, also called the dew point temperature. This is the temperature where the relative humidity becomes 100%. Convective condensation level (CCL) is the imaginary line, which is usually the base of clouds, i.e., the lowest level of the clouds at which the water vapor in air starts to condense and form cloud. Bhattacharya et al [10] predicted the squall-storms applying neural network on the upper air data. The parameters of cloud used in that work were cloud coverage and pressure at freezing level (FRZ). In the previous works of Chakrabarty et al. [11], there were thirteen input variables such as moisture differences, adiabatic lapse rates, and wind-shears at different geopotential heights of the upper atmosphere, which were used to predict severe thunderstorms.

Neural network (MLP) was applied in that work for the prediction purpose. All the parameters are sensed by radiosonde and rawindsonde. This paper considers four parameters, Cloud coverage(Nh), Sunshine hour(SSH), Pressure at freezing point(FRZ), and dry adiabatic lapse rates at different geopotential heights of the atmosphere.

Present work is based on the application of Machine Learning Technique on the weather data to predict thunderstorms. Here K-Nearest Neighbor (K-NN) method has been adopted for this purpose. Chakrabarty 2 et al. applied this method in their previous work. K-NN algorithm was first suggested by Cover in 1968 [12], and it is one of the best data mining procedure for pattern classification. K-NN 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, where the function is only approximated locally and all computation is deferred until classification, [2].

Jayawardena et al. [13] applied K-NN method for flood forecasting. Li et al [14] applied it to forecast solar flare.

2. DATA

2.1. Data Collection

All the weather data were collected from India Meteorological Department, Govt. of India during the period of 33 years from 1969 to 2002 for the months of March-April-May. The data were recorded at 05.30 a.m. local time (00:00 UTC) by radiosonde over Kolkata, North-East India. Here data have been considered both for 'squall' and 'no-squall' days. The total data set has 485 data from the year 1969 to 2002. The numbers of 'squall-storm' days are 160 and 'no squall-storm' days are 325. Total 485 data have been taken for analysis.

2.2. Data Description

Here the data of dry adiabatic lapse rate at different heights of the upper air are procured in this work. The variables X4 to X8 represent the data of adiabatic lapse rate at five different altitudes of the atmosphere.

The different input variables are:

- X1 (NH): It stands for the cloud coverage over the sky, represented in octas.
- X2 (SSH): It represents sunshine hours measured in hours.
- X3 (FRZ): It represents pressure at the freezing point level of the atmosphere, measured in hPa. Dry adiabatic lapse rate is the difference in dry bulb temperatures at respective heights of the atmosphere.
- X4: Dry adiabatic rate from surface to 850 hectapascal (hPa) (from mean sea level to 1500 meters height)
- X5: Dry adiabatic lapse rate from 850 hPa to 700 hPa (approx. 1500 m to 3100 m height)
- X6: Dry adiabatic lapse rate from 700 hPa to 600 hPa (approx. 3100m to 4500 m height)
- X7: Dry adiabatic lapse rate from 600 hPa to 400 hPa (approx. 4500 m to 7500 m height)
- X8: Dry adiabatic lapse rate from 400 hPa to 300 hPa (approx. 7500 m to 9600 m height)

Here, we have used total 485 data. Out of total 485 data, 160 data of them are squalls (high wind thunderstorms) and 325 data of them are no squalls (normal days)

- The squalls and the no squalls data have been classified into 5 different sets.
- Each dataset consists of total 97 data, of which 32 data are for squall-storm days, and 65 data are for normal days.
- Analysis is done taking each dataset as training set and the remaining four datasets as testing set. Thus, as there are five different datasets, analysis is done five times.

3 METHODOLOGY

The concept of machine learning originates from artificial intelligence. Recognition of data and patterning becomes easier with the use of a certain type of machine learning. Machine learning is also known as predictive analytics or data mining. It includes the study of pattern recognition as well as computational learning theory in artificial intelligence. Machine learning has become a universal tool to analyze many real life data. These algorithms are used to make data-driven predictions or decisions by building a model from training data inputs.

This process is performed when the algorithms figure out the way to perform tasks by generalizing from examples. It is so vastly used in solving many practical problems that many researchers think it is the best way to make progress towards human-level artificial intelligence. Machine learning tasks can be classified into three broad categories, namely the Supervised learning, Unsupervised learning, and Reinforcement learning [15], K-NN method belongs to the Unsupervised learning category [16].

Pattern recognition finds its application largely in computer vision, mostly in formalizing and visualizing patterns. In large class of data analysis, the pattern recognition provides a generalized approach to solutions. There are properties of objects which are not directly measurable and must be found with experimental measurements. The pattern recognition technique helps to recognize and classify such parameters.

3.1 K-Nearest Neighbor (K-NN)

Yakowitz extended the K-nearest neighbor method [17]. 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 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,

a = (a1, a2,……,an), b = (b1,b2,…….., bn) is defined as (1)

$$\frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2 \sum_{i=1}^{N} b_i^2}} \quad (1)$$

The similarity measures between two vectors reflect the cosine of the angle 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. These cosine angles are arranged in the decreasing order.

As the number of training data vectors is 97 in each set, the numbers of cosine angles are also 97. For each of the squall data vector, if maximum number of 'squall' data appears within K number of cosine angles then it is to be considered as properly classified as 'squall' class. Similar thing happens for 'no squall' class.

4 RESULTS

The K-NN method has been applied on the eight types of atmospheric parameters taken into account for this research. They are cloud coverage (Nh), Sunshine hours(SSH), pressure at freezing level of the upper atmosphere(FRZ), and X4, X5, X6, X7, X8 where X4, X5, X6, X7, X8 are the adiabatic lapse rates measured at different geopotential heights of the atmosphere. The result obtained is shown in the table below. If we summarize the findings, depending upon the number of parameters involved, the performance of the K-NN method in respective calculation can be compared. The performance is compared for three categories. At first, a dataset of 32 days had been taken. The exact number of days which has been properly classified after applying the K-NN method is observed. The same process is repeated by taking larger data sets of 65 days and 97 days.

Methods applied-K-NN technique	No. of "squall-storm" days properly classified,(No. of storm data in test dataset=32)	No. of "No squall-storm" days properly classified,(No. of No storm data in test dataset=65)	No. of "squall-storm" as well as "No squall-storm" days properly classified,(No. of test dataset =97)
K-NN with 8 variables, (K=5)	11	51	62
K-NN with 8 variables, (K=17)	6	61	67
K-NN with 8 variables, (K=21)	6	64	70

We have got the 72.16% correct result.

5 DISCUSSION AND CONCLUSION

The proper selection of the machine learning technique is one of the very important matters in this type of work. The challenge here lies on another thing which is the lead time for prediction of thunderstorm. Here the lead time is twelve hours which is sufficient to take necessary measure from the natural hazard. A few atmospheric parameters (only four types) of upper air are considered for the prediction purpose. The result here shows that the K-NN technique can classify 72.16% of data correctly.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Chakrabarty 1, 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.

[2] Chakrabarty 2, Himadri, and Bhattacharya Sonia, "Application of K-Nearest Neighbor Technique to Predict Severe Thunderstorms", January, 2015, *International Journal of Computer Applications (IJCA)*, Vol.110, No. 10, pp. 1-4, ISSN: 0975-8887.

[3] Newton, C.W., Sept., 1963, "Dynamics of Severe Convective Storms", *Meteorological Monographs, American Meteorological Society*, Vol. 5, Number 27, pp. 33-58.

[4] Johnson, R. H., and Hamilton P.J., July, 1988. "The relationship of surface pressure features to the precipitation and airflow structure of an intense midlatitude squall line", *Monthly Weather Review*. 116 (7): 1444–1472.

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

[6] Hobbs, P.V., S. Chang, and J.D. Locatelli (1974), "The dimensions and aggregation of ice crystals in natural clouds, *J. Geophys*. Res., 79, 2199-2206.

[7] Koop, Thomas, Beiping Luo, Athanasios Tsias, and Thomas Peter, 2000, "Water activity as the determinant for homogeneous ice nucleation in aqueous solutions," *Nature*, 406, pp. - 611-614, 10 August.

[8] Bergeron, T.,"On the physics of clouds and precipitation," Procès-Verbaux Séances Assoc. Météor. l'U.G.G.I., Lisbon, Pt. II, 156-178., 1935.

[9] Kerkweg, A., S. Wurzler, T. Reisin, and A. Bott, 2013, "On the cloud processing of aerosol particles: An entraining air parcel model with two-dimensional spectral cloud microphysics and a new formulation of the collection cernel, *Q. J. R. Meteorol*. Soc.,129,1-18.

[10] Bhattacharya, Sonia, and Himadri Chakrabarty, July, 2015, "Forecasting of Severe Thunderstorms using Upper Air Data,", *International Journal of Scientific and Engineering Research*, ISSN: 2229-5518, Vol. 6, Issue. 7.

[11] Chakrabarty, Himadri and Sonia Bhattacharya, March, 2014, "Prediction of Severe Thunderstorms applying Neural Network using RSRW data," *International Journal of Computer Application*, ISSN: 0975-8887, Vol. 89, No. 16, pp. 1-5.

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

[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 Contro*l, held at Anaheim, California, Publ. no. 239.

[14] 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*, Vol. 7, pp-441-447,2007.

[15] Russell Stuart J., Norvig Peter, (2003) "*Artificial Intelligence: A Modern Approach*", 2nd Edition, Prentice Hall,1995.

[16] D. Coomans; D.L. Massart, "Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Nearest neighbour classification by using alternative voting rules," *Analytica Chimica Acta* 136: 15–27 (1982).

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