Classification and Forecasting of Weather using ANN, k-NN and Naïve Bayes Algorithms

Nishchala C. Barde[1], Mrunalinee Patole[2]

[1]RMD Sinhgad School of Engineering, Savitribai Phule Pune University, Pune, India

[2]Professor, RMD Sinhgad School of Engineering, Savitribai Phule Pune University, Pune, India

Abstract: *Weather forecasting is a way to predict future weather. It is widely researched area due to the fact that human life on earth is affected by the global climate. In this paper, we have proposed a comparative study between various techniques for prediction. This work focuses on developing an optimized system model which predicts future weather. The frequency of natuaral hazards occurring due to unpredictable weather conditions have been seen to be increasing causing damage to human life. There are some models that predict weather during real time, or monthly or annual period. This paper presents a system that carries out weather prediction using previous or historical weather data having attributes (Date, Temperature, Humidity, Wind Chill (WC) and Stn Pressure(SP). Various data mining techniques are used for this purpose of weather forecasting such as Multi-layer Perceptron, K-Nearest Neighbor and naive Bayes. Comparison based on evaluation parameters identify which model has more accurately performed the predictions.*

Keywords: Machine Learning, Artificial Neural Network, Naïve Bayes, K Nearest Neighbors, Weather Prediction

1. Introduction

Since long it has been observed and experienced that the weather prediction is becoming very crucial. The success of prediction solely depends upon the exactness of the model. There are many experiments carried out to develop appropriate model using various techniques. The India Meteorological Department (IMD) has developed different models like Parametric, Power Regression, Multiple Regression, Dynamic Stochastic Transfer, Power Transfer, Principal Component Regression and Neural Network models for issuing these long-range seasonal forecasts. A number of different Neural Network architectures and machine learning methods have been applied to the time series prediction problem with varying degrees of success. However, there still exists a need for satisfactory guidelines when choosing a model size for a particular prediction problem. In this paper we propose the methodology for weather forecasting by using the historical data of the weather.

The objective of model development is to minimize the cost function and error which is the difference between desired and calculated values. In recent years, long range monsoon rainfall forecast based on the power regression model has been found to be comparatively more accurate. Therefore the result of the power regression model has been given weightage in this forecast.

Some areas in India have high humidity while some are dry. In this paper, we are using the techniques of machine learning algorithms namely k-Nearest Neighbors, Naive Bayes and Multi layer perceptron.The multi layer perceptron is a technique of Artificual Neural Network. The fields of neural network approach are pattern recognition, system modeling and prediction. System-modeling capability of neural network approach is important. There are some advantages to modeling, design, problem solving and approximation of the mapping. By using Neural network, it is possible (i) to work without the knowledge of the physical process, assuming the internal parameters or boundary conditions, (ii) to train neural network itself by using massive amount of the data, and (iii) to use a complex non-linear mapping, such as precipitation phenomena.

The process of data organization into categories such that data objects of same category are more similar and data objects from different groups are not same is called as classication. Classification algorithm works to assign each instance to a particular class such that classification error will be lowest. Hence it is used to extract models that accurately define important data classes within the given dataset. The techniques of classification divided into two parts namely - training and testing.

2. Literature Survey

The section gives literature survey on ANN prediction methologies and other machine learning techniques. Neural network theory started with the first discoveries about brain cellular organization and the challenge is still with us to discover the principles that would make a complex interconnection of relatively simple elements produce information processing at an intelligent level. At the core of neural computation are the concepts of distributed, adaptive and nonlinear computing.

2.1 Artificial Neural Network

Artificial neural network is inspired by biological neuron model. A PE or the processing element simply multiplies an input by a set of weights, and nonlinearly transforms the result into an output value (table lookup).
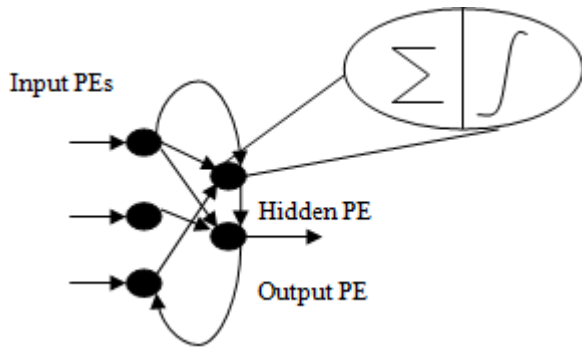
Figure 1: Building blocks of ANN

The principles of computation at the PE level are simple. The power of neural computation comes from the massive interconnection among the PEs, which share the load of the overall processing task, and from the adaptive nature of the parameters (weights) that interconnect the PEs. Normally, a neural network will have several layers of PEs. There are a number of neural network topologies formed by different combinations of arrangements of these layers. The diagram below illustrates a simple MLP. The circles are the PEs arranged in layers. The left column is the input layer, the middle column is the hidden layer, and the right column is the output layer. The lines represent weighted connections (i.e., a scaling factor) between PEs.

2.2 K- Nearest Neighbors

The base of Nearest-neighbor classifiers is learning by resemblance, which is by comparing a given test sample with the available training samples which are similar to it.
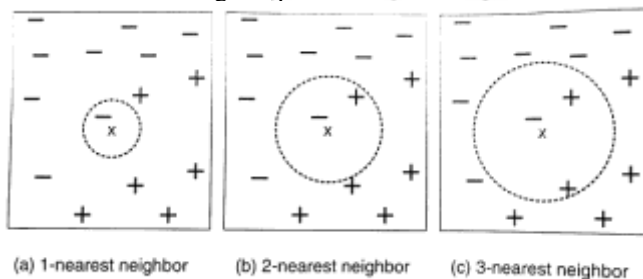

Figure 2: K-NN

To classify a data sample X, search is done for its K-nearest neighbors and then X is assigned to a class label to which majority of its neighbors belong. In this method, the choice of k also affects the performance of k-nearest neighbor algorithm. If the value of k is too small, then K-NN classifier may be vulnerable to over-fitting because of noise present in the training dataset. However on the other hand, if k is too large, then the nearest-neighbor classifier may not classify the test sample accurately. This is as the list of its nearest neighbors may contain some data objects that are located far away from its neighborhood.

2.3 Naïve Bayes

Naive Bayes Classifier is the simple Statistical Bayesian Classifier. As the name Naive suggests it assumes that all variables contribute towards classification and are mutually correlated or independent. This assumption is called class conditional independence. Hence it is also called Idiot's

Bayes or Simple Bayes, and Independence Bayes.They predict class membership probabilities. This classifier considers that the presence (or absence) of a particular feature(attribute) of a class is unrelated to the presence (or absence) of any other feature when the class variable is given thus identifying the correlation.

3. Related Work

In past there were number of models designed using Numerical Weather Prediction (NWP) as well as categorical weather prediction. Recently, with the development of artificial intelligence (AI), various new AI methods for wind speed and power prediction have been developed. However, these methods have more complexity of computations. Various other machine learning and statistical classifiers like J48, k-Nearest Neighbors etc. have been used for prediction purposes.

The new methods found include artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), fuzzy logic methods, support vector machine (SVM), neuro-fuzzy network, and evolutionary optimization algorithms. Artificial Neural Networks could effectively deal with non-linear and complex problems in terms of classification or forecasting. The ANN models can represent a complex nonlinear relationship and extract the dependence between variables through the training process Gowariker, V.,1991, has given a power regression model approach for predicting southwest monsoons in India. Nekoukar et al., 2010, used radial basis function neural network for financial time-series forecasting, the result of their experiment shows the feasibility and effectiveness. KOŠCAK et al., 2009, compared common meteorological forecasting method with ANN and he found the performance of ANN with high accuracy. Geetha and Selvaraj, 2011, have predicted Rainfall in Chennai using back propagation neural network model, and by their research the mean monthly rainfall is predicted using ANN model. This model can perform well both in training and independent periods. T. Auld, A. Moore, and Gill have found Bayesian Neural networks for internet traffic classication perform in a better way.

4. Implementation

4.1 Multilayer Perceptron

In this model of neural network, neurons are organized in layers and the ones which belong to the same layer have the same inputs. The outputs of the neurons of a layer are the inputs of the neurons of the next layer and there is no feedback in the network. The multilayer perceptron or MLP is trained with error correction learning, which means that the desired response for the system must be known.

Using the theory of gradient descent learning, every weight in the network can be adapted by correcting the present value of the weight with a term that is proportional to the present input and error at the weight, given as :

$$w_{k+1} = w_k + 2\mu\varepsilon_k x_k \qquad (1)$$

Where :

μ = the learning rate parameter),

ε $_k$ = error (the difference between the desired response and the actual system response) at iteration step k,

x$_k$ = input value to the weight i at iteration k, and

w$_k$ = the value of weight i at iteration k.

4.2 k-Nearest Neighbors

The fundamental belief on which k-Nearest Neighbors works is that the data is connected in a feature space. Therefore, to find the distance among data points, all the points are considered in order. The distance is calculated using Euclidian distance based on the data type of data classes used. A single value of K is given and it is used to find the total number of nearest neighbors that determine the class label for unknown sample.

If the value of K=1, then it is called as nearest neighbor classification. The working of K-NN classifier is as follows:
1) Value of K is initialised.
2) The distance between input sample and training samples is calculated:

$$dist((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

(2)

3) Sort the distances.
4) Take top K- nearest neighbors.
5) Apply simple majority.
6) Predict class label with more neighbors for input sample.

4.3 Naïve Bayes implementation

The Naive Bayes algorithm is based on Bayesian Theorem. It is used when the inputs have high dimensionality. It calculates the probability of a new event on the basis of earlier probability estimates derived from empirical data. To calculate the probability of hypothesis Y, where the observed training data is X is given by –

$$p\left(\frac{y}{x}\right) = \frac{p\left(\frac{x}{y}\right) p(y)}{p(x)}$$

(3)

The Naive Bayes works as follows –

1. Let D be the training dataset associated with class labels. Each tuple is represented by n-dimensional element vector, X=(x1, x2, x3,.....,xn).

2. Consider that there are m classes C1, C2, C3...., Cm. Suppose that we want to classify an unknown tuple X, then the classifier will predict that X belongs to the class with higher posterior probability, conditioned on X. i.e., the Naive Bayesian classifier assigns an unknown tuple X to the class Ci if and only if $P(\_{Ci}|X) > P(\_{Cj}|X)$ For $1 \leq j \leq m$, and i≠j, above posterior probabilities are computed using Bayes Theorem.

5. Results

The results of three models were compared based on the parameters of accuracy and time taken. Following results were observed using proposed methods for prediction

trained using time series historical weather data of 999 instances.

Table 1: Results of Accuracy, AME and Time

Parameter	KNN	Naïve Bayes	MLP
Accuracy	100%	42.54%	40.81%
Time(sec)	1.25	22.5	28.75
Absolute Mean error	9.5	7.5	9.0

6. Conclusions and Future Scope

In this study of comparison between MLP, k-NN and Naïve Bayes algorithms to predict and classify future weather, it shows that k-NN is more accurate in classification and also in terms of execution time required. However, training cost of k-NN increases with size of data set. The numeric prediction results shows Naïve Bayes gives better results compared to k-NN and MLP.

From the comparison results, in future we can develop a system which uses Naïve Bayes for numeric prediction and k-NN for classication and develop a hybrid system which would give more accurate results as single system can not satisfy all the constrains. Also it can be used for predicting thunderstorms and other similar applications of weather.

References

[1] S.Razavi et. al, emph. "A New Formulation for feed forward neural network, IEEE. Trans. Vol. 22, pp. 1588-1598, 10 October 2011.

[2] Tsang, S., et al., "Decision trees for uncertain data.", IEEE Transactions on Knowledge and Data Engineering, 2011. 23(1): p. 64-78.

[3] Haiyan Shu Rongshan Yu, Wenyu Jiang, Wenxian Yang, "Efficient Implementation of k -Nearest Neighbor Classifier Using Vote Count Circuit", IEEE Transactions on (Volume:61 , Issue: 6)24 April 2014.

[4] Ren, J., et al. "Naive bayes classification of uncertain data." Ninth IEEE International Conference on in Data Mining, 2009.

[5] Bin Yousuf A.H., Hossain M.N.,Asasduzzaman, M.M., Hossain, M.I.,"Artificial neural network based temperature prediction and its impact on solar cell.", IEEE conference on Neural Networks pp: 897 902,18-19, May 2012.

[6] Xinghuo Yu,A General Backpropagation Algorithm for Feedforward Neural Networks Learning , IEEE Trans. On Data Engineering. Vol. 13, No. 1,January 2002.

[7] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for internet traffic classification," IEEE Trans. Neural Netw., vol. 18, no. 1, pp. 223–239, Jan. 2007.

Author Profile

Nishchala Barde received the B.E. degree in Computer from Smt. Kashibai Navale College of Engineering in 2011. She has studied subjects on Data mining, distributed systems and communication systems. She has interest in studying Data mining domain.

Paper ID: NOV161672

1742