ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 9(38), DOI: 10.17485/ijst/2016/v9i38/101962, October 2016*

Analysis of Data Mining Techniques for Weather Prediction

Fahad Sheikh[1], S. Karthick[1]*, D. Malathi[2], J. S. Sudarsan[3] and C. Arun[1]

[1]Department of Software Engineering, SRM University, Kattankulathur, Chennai - 603203, Tamil Nadu, India; fahadsheikh1709@gmail.com , karthik.sa@srmuniv.ac.in, arun.c@ktr.srmuniv.ac.in
[2]Department of Computer Science and Engineering, SRM University, Kattankulathur, Chennai - 603203, Tamil Nadu, India; malathi.d@ktr.srmuniv.ac.in
[3]Department of Civil Engineering, SRM University, Kattankulathur, Chennai - 603203, Tamil Nadu, India; sudarsanjss@yahoo.com

Abstract

Background/Objectives: To forecast weather, which is one of the greatest challenges in meteorological department. Weather prediction is necessary so as to inform people and prepare them in advance about the current and upcoming weather condition. This helps in reduction in loss of human life and loss of resources and minimizing the mitigation steps that are expected to be taken after a natural disaster occurs. Methods/Statistical analysis: This study makes a mention of various techniques and algorithms that are likely to be chosen for weather prediction and highlights the performance analysis of these algorithms. Various other ensemble techniques are also discussed that are used to boost the performance of the application. Findings: After a comparison between the data mining algorithms and corresponding ensemble technique used to boost the performance, a classifier is obtained that will be further used to predict weather. Applications: Used to Predict and forecast the weather condition of specific region based on the available pre historical data which helps to save resources and prepare for the changes forth coming.

Keywords: Data Mining, Decision Tree, Ensemble Technique, Pre-Processing, Weather Prediction

1. Introduction

Weather prediction has been a challenging problem in meteorological department since years. Even after the technological and scientific advancement, the accuracy in prediction of weather has never been sufficient. Even in current date this domain remains as a research topic in which scientists and mathematicians are working to produce a model or an algorithm that will accurately predict weather. There have been immense improvements in the sensors that are responsible for recording the data from the environment and cancel the noise present in them; along with this new models have been proposed which include different attributes related to weather to make accurate prediction.

Currently one of the most widely used techniques for weather prediction is data mining. Data mining offers a way of analysing data statistically and extract or derive such rules that can be used for predictions. Presently it is being used in many domains such as stock market, sports, banking section, etc. Scientists have now realized that data mining can be used as a tool for weather prediction as well. The basic entity of data mining is data itself. It is defined as raw set of information which can be used

*Author for correspondence

to extract meaningful information depending upon the requirements of the application. Data can be stored in an organized manner which is known as database.

The term data mining refers to the techniques that are used to extract the required information from the given set of data that might be useful for statistical purpose or making predictions by learning patterns in data and correlation between different parameters. Data mining has now been adopted by many domains such as sports, banking, meteorological department, etc., and because of this, scientists, mathematicians and researchers have come up with a wide range of algorithms for finding solution.

2. Materials and Methods

Weather is one of the most influential factors in our daily life, to an extent that it may affect the economy of a country that depends on occupation like agriculture. Therefore as a counter measure to reduce the damage caused by the uncertainty in weather behaviour, there should be efficient ways to predict weather. Usually two main techniques are used for weather forecasting, one involves usage of large amount of data to gain knowledge about future weather and the other involves construction of equations that will help predict weather by identifying different parameters and substituting the values to obtain desired result. The decades of research work has been done in the field of meteorology. Recently researchers have started highlighting the effectiveness of data algorithms in predicting weather. One of the latest research works includes a paper[1]. In this paper makes a mention of Artificial Neural Networks[2] and Decision Tree algorithms and their performance in prediction of weather. ANN finds the relation between the weather attributes and builds a complex model, whereas C5 decision tree learns the trend of data and accordingly builds a classifier tree that can be used for prediction. Another well-known data mining technique, CART was used in her paper[3]. A decision tree was produced as an output and its performance was calculated using evaluation metrics which included parameters like precision, accuracy, FP rate, TP rate, F-measure, and ROC Area.

Since numerous data mining algorithms are available for use, it is necessary to find the appropriate technique that will be suitable for the domain it is being applied to. In certain cases regression technique proves to be more effective whereas in other cases, rule based technique and decision tree algorithms give accurate result with a low computational cost. In[4] have reviewed various data mining techniques and gave a performance comparison between algorithms like C4.5, CART, k-means clustering[5], ANN, and MLR when they were used for weather prediction. They made a conclusion that k-means and decision tree algorithms perform better than other algorithms in case of weather prediction. In[6–8] in-depth performance comparison[9] has been between C4.5 and Naïve Bayes algorithm, includes discussion over the suitability of algorithm when applied to different dataset.

2.1 Approach

The methodology used in this paper consists of certain steps that are usually used in data mining applications the steps are as follows[10,11]:

- Data Collection and Retrieval:
  Data used for the research work was obtained from meteorological tower of SRM University Chennai, India. The format of data was in CSV format and included parameters like humidity, temperature, cloud cover, wind speed, etc[12].
- Data Transformation:
  The CSV file was first converted to .arff format to feed it into the Data Mining tool – WEKA. The conversion to .arff format was implemented through code written in java. Two separate files were maintained as weather.arff and predictweather.arff in which weather.arff consists of the actual data collected over a period of 2 years and predictweather.arff file contained sample data used for prediction.
- Data Pre-processing:
  The weather.arff file was used as a source file and then Resample technique was applied to the data present in it. Resample technique involves choosing instances from dataset on a random basis with or without replacement.
- Feature Extraction:
  Among all the parameters considered which consisted of max temperature, min temperature, mean temperature, max humidity, min humidity,

mean humidity, wind speed, cloud cover, and rainfall. The maximum humidity was overlooked since the data was filled with noise and was not accurate. Rest of the parameters were used for further processing in application as they were mutually exclusive and no redundancy was present between them.

- Data mining:
  This stage consisted of analysing the given dataset with different algorithms like Naïve Bayes and C4.5 (J48) algorithm and then choosing the better one for further predictions. Then the dataset was split into training set for making the machine learn and the testing dataset along with cross validation. Then the patterns were recorded to make further predictions. Additionally few ensemble algorithms like boosting and bagging were applied to improve the results which was shown in Table 1.

Table 1. Attributes of weather.arff

Min Temperature	Numeric
Min Temperature	Numeric
Min Temperature	Numeric
Min Humidity	Numeric
Min Humidity	Numeric
Min Humidity	Numeric
Wind Speed	Numeric
Cloud Cover	Numeric
Rainfall	Boolean

2.2 Comparison between Naïve Bayes and C4.5 Decision Tree Algorithm

2.2.1 VNaïve Bayes

Naïve Bayes algorithm belongs to the family of probability based classifiers and revolves round the concept of Bayes theorem. The probabilistic model consists of vector containing features with a probability assigned to it. The estimation of class condition probability is done by the classifier with the assumption that attributes are conditionally not dependent on each other. Construction of classifier model is done by combining the probability based model with decision rule.

$$p(C_k|x_1, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \qquad (1)$$

$$\hat{y} = \underset{k\in\{1,\dots,K\}}{\mathbf{argmax}}\, p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \qquad (2)$$

Equation (1) represents the conditional distribution, where ‘Z’ is the scaling factor and ‘C’ being the class variable. Equation (2) represents a Bayes classifier built using the probability model.

2.2.2 C4.5 Decision Tree

Unlike Naïve Bayes, the C4.5 is a classification algorithm used to generate decision tree for the given dataset. It is based on the information entropy concept. Construction of the decision tree is done by selecting the best possible attribute that will be able to split set of samples in most effective manner. The attribute having the highest entropy difference or normalized information gain is selected as the splitting criteria for that particular node. Similar fashion is followed and nodes are added to the decision tree. Each penultimate node carries the last attribute or multiple attributes for making the final decision of the problem.

```
Algorithm J48:
INPUT
D // Training data
OUTPUT
T // Decision tree
DTBUILD (*D)
{
T = Null;
T = Create root node and label with splitting attribute;
T = Add arc to root node for each split predicate and label;
For each arc do
```

```
    D = Database created by applying splitting predicate to D;
    If stopping point reached for this path, then
    T'= Create leaf node and label with appropriate class;
    Else
    T' = DTBUILD (D);
    T = Add T' to arc;
```

The Evaluation of training set for C4.5 (J48) algorithm with bagging was shown in the Table 2[13,14]. The Evaluation of training set for Naïve Bayes algorithm was shown in the Table 3.

Table 2. Evaluation of training set for C4.5 (J48) algorithm with bagging

Correctly Classified Instances			459			73.322%	
Incorrectly Classified Instances			167			26.677%	
Kappa statistic				0.4654			
Mean absolute error				0.3634			
Root mean squared error				0.4236			
Relative absolute error				72.7753%			
Root relative squared error				84.7653%			
Total Number of Instances				626			
Detailed Accuracy By Class:							
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.753	0.288	0.737	0.753	0.745	0.807	yes
	0.712	0.247	0.729	0.712	0.720	0.807	no
Weighted Avg.	0.733	0.268	0.733	0.733	0.733	0.807	
Confusion Matrix							
a	b	← Classified as a=yes b=no					
244	80						
87	215						

Table 3. Evaluation of training set for Naïve Bayes algorithm

Correctly Classified Instances	343			54.7923%			
Incorrectly Classified Instances	283			45.2077%			
Kappa statistic	0.1088						
Mean absolute error	0.4781						
Root mean squared error	0.5061						
Relative absolute error	95.8786%						
Root relative squared error	101.3492%						
Total Number of Instances	626						
Detailed Accuracy By Class:							
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.41	0.3	0.603	0.41	0.488	0.569	yes
	0.7	0.59	0.517	0.7	0.595	0.569	no
Weighted Avg.	0.548	0.437	0.562	0.548	0.539	0.569	
Confusion Matrix							
a	b	← Classified as a=yes b=no					
135	194						
89	208						

After the performance comparison, the J48 algorithm was chosen for further implementation which involved study of the legacy data about the weather. The resample filter was used for data pre-processing, furthermore for the data selection step, from a total of 8 parameters the maximum humidity was neglected due to noise present in it which was affecting the accuracy of the system. This resulted in consideration of linear parameters like max temperature, min temperature, mean temperature, min humidity, mean humidity, wind speed, cloud cover, and rainfall.

The next step in the implementation was to apply the decision tree algorithm to the dataset. The filtered data was given as an input to the algorithm and a decision tree was expected as output. This decision tree will contain collection of nodes and each node consists of attributes or set of attributes as criteria to split the node for further classification. To improve the accuracy of the algorithm ensemble technique, bagging was chosen in the final application of algorithm along with C4.5 decision tree to get the final decision tree[15]. The performance comparison of both algorithm was shown in Table 4.

Table 4. Performance comparison

Parameters	Naïve Bayes	C4.5 Decision Tree
Correctly Classified Instances	343	549
Incorrectly Classified Instances	283	167
Kappa Statistic	0.1088	0.0.4654
Mean Absolute Error	0.4781	0.3634
Root Mean Squared Error	0.5061	0.4236
Relative Absolute Error	95.89%	72.7753%
Root Relative Squared Error	101.34%	84.7653%
F-Measure	0.539	0.733
Precision	0.562	0.733
Time taken to Build Model	0.01	0.12

3. Resultant C4.5 Decision Tree

```
Mean Humidity <= 85
|  Min Temp <= 20
|  |  Max Humidity <= 78
|  |  |  Cloud Cover <= 1: no (2.8/0.91)
|  |  |  Cloud Cover > 1: yes (14.64)
|  |  Max Humidity > 78
|  |  |  Mean Temp <= 24
| | | | Max Humidity <= 94
| | | | | Mean Humidity <= 74
| | | | | | Max Temp <= 29: no (8.78/0.99)
| | | | | | Max Temp > 29: yes (9.76/1.1)
| | | | | Mean Humidity > 74: yes (9.81/0.96)
| | | | Max Humidity > 94: yes (236.19/77.68)
| | | Mean Temp > 24
| | | | Max Temp <= 35
| | | | | Max Temp <= 32: no (7.48)
| | | | | Max Temp > 32
| | | | | | Max Humidity <= 89: no (9.25/0.95)
| | | | | | Max Humidity > 89: yes (5.43/0.28)
| | | | Max Temp > 35: yes (4.29)
| Min Temp > 20
| | Wind Speed <= 15: no (115.17/40.35)
| | Wind Speed > 15
| | | Wind Speed <= 23
| | | | Max Humidity <= 73: no (3.8)
| | | | Max Humidity > 73
| | | | | Mean Temp <= 31
| | | | | | Cloud Cover <= 4
| | | | | | | Mean Temp <= 27
| | | | | | | | Wind Speed <= 18: yes (6.38)
| | | | | | | | Wind Speed > 18: no (6.69)
| | | | | | | Mean Temp > 27
| | | | | | | | Wind Speed <= 22
| | | | | | | | | Min Humidity <= 33
| | | | | | | | | | Min Humidity <= 27: yes (6.64)
| | | | | | | | | | Min Humidity > 27: no (3.9)
| | | | | | | | | Min Humidity > 33: yes (12.71)
| | | | | | | | Wind Speed > 22
| | | | | | | | | Mean Humidity <= 70: yes (7.4/0.06)
| | | | | | | | | Mean Humidity > 70: no (3.63/0.4)
| | | | | | Cloud Cover > 4: yes (12.06)
| | | | | Mean Temp > 31: no (3.84)
| | | Wind Speed > 23: no (109.88/49.37)
Mean Humidity > 85
| Max Temp <= 24
| | Mean Temp <= 21: no (2.0)
| | Mean Temp > 21: yes (2.37)
| Max Temp > 24: no (21.1/1.79)

Number of Leaves:       25
Size of the tree:       49
Weight: 0.95.
```

4. Conclusion and Future Enhancements

For the current application of data mining in weather prediction domain, the analysis of Naïve Bayes and C4.5 Decision Tree algorithm was done simultaneously with dataset containing weather data collected over a period of 2 years. It was found that the performance of C4.5 (J48) decision tree algorithm was far better than that of Naïve Bayes. The accuracy for C4.5 was 88.2% with respect to classifying the instances correctly. On the other hand, Naïve Bayes showed a poor performance of 54.8% while classifying the instances.

The confusion matrix also supported the above made statement of C4.5 being a better performer in case of weather dataset. The number of instances that were true positives, i.e., true instances and also were predicted true by C4.5 was higher than that of Naïve Bayes and in case of number instances that were true negatives, i.e., false and were predicted as false showed a similar result. Even the precision of C4.5 was considerable higher in this case. Only the time taken to build the model was less in case if Naïve Bayes as compared to C4.5 decision tree.

Previous research has highlighted that performance of C4.5 algorithm improves when the dataset used for application is quite large whereas the performance of Naïve Bayes is comparatively poor. Similarly when the number of attributes increase in the dataset the Naïve Bayes performance drastically affected but C4.5 handles this problem of more number of instances being present in a single dataset in a subtle manner. Therefore it can be said that the performance of C4.5 algorithm was better than that of Naïve Bayes in case of dataset dealing with weather. Further improvements can be made to improve the result of the algorithm by applying appropriate filter to the dataset in pre-processing stage as well as ensemble algorithms can be used along with the C4.5 to achieve a better result.

5. References

1. Jadhawar BAAM. Weather forecast prediction: A data mining application. International Journal Of Engineering Research And General Science. 2015 Mar-Apr; 4(3):19–21.
2. Gharehchopogh FS, Khaze SR, Maleki I. A new approach in bloggers classification with hybrid of k-nearest neighbor and artificial neural network algorithms. Indian Journal of Science and Technology. 2015 Feb; 8(3):237–46.
3. Petre EG. A decision tree for weather prediction. BULETINUL UniversităŃii Petrol – Gaze din Ploiesti; 2009 Apr. p. 77–82.
4. Chauhan D, Thakur J. Data mining techniques for weather prediction: A review. International Journal on Recent and

Innovation Trends in Computing and Communication. 2014 Aug; 2(4):2184–9.

5. Gokila S, Kumar KA, Bharathi A. Modified projected space clustering model on weather data to predict climate of next season. Indian Journal of Science and Technology. 2015 Jul; 8(14):1–5.
6. Joshi S, Pandey B, Joshi N. Comparative analysis of Naive Bayes and J48 classification. International Journal of Advanced Research in Computer Science and Software Engineering. 2015 Dec; 5(12):813–7.
7. Patil TR, Sherekar SS. Performance analysis of Naïve Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications. 2013 Apr; 6(2):256–61.
8. Goyal A, Mehta R. Performance comparison of Naïve Bayes and J48 classification algorithms. International Journal of Applied Engineering Research. 2012; 7(11):1–5.
9. Verma A, Kaur I, Arora N. Comparative analysis of information extraction techniques for data mining. Indian Journal of Science and Technology. 2016 Mar; 9(11):1–18.
10. Kalyankar MA, Alaspurkar SJ. Data mining technique to analyse the metrological data. International Journal of Advanced Research in Computer Science and Software Engineering. 2013 Feb; 3(2):114–8.
11. Taksande AA, Mohod PS. Applications of data mining in weather forecasting using frequent pattern growth algorithm. IJSR. 2015 Jun; 4(6):3048–51.
12. Olaiya F. Application of data mining techniques in weather prediction and climate change studies. I. J. Information Engineering and Electronic Business. 2012 Jul; 1:51–9.
13. Dutta PS, Tahbilder H. Prediction of rainfall using data mining technique over Assam. IJCSE. 2014 Apr/May; 5(2):85–90.
14. Weka Software. Available from: https://nl.wikipedia.org/wiki/Weka_(software)
15. 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. 2015 Mar; 4(3):19–21.