Research and Application of Summer High Temperature Prediction Model Based on CART Algorithm

Yujie Guan[1], Wei Wang[2], Fengchang Xue[3(✉)], and Shoudong Liu[1]

[1] College of Applied Meteorology,
Nanjing University of Information Science and Technology,
Jiangsu 210044, China
[2] College of Atmospheric Sciences,
Nanjing University of Information Science and Technology,
Jiangsu 210044, China
[3] College of Geographic Information and Remote Sensing,
Nanjing University of Information Science and Technology,
Jiangsu 210044, China
xfc9800@126.com

Abstract. In this paper, the average summer high temperature effective accumulated for many years is used as a judge of the extent of the hot summer temperatures of standards. Based on data mining, the CART algorithm is applied to analyze the relationship between high temperature and some climatic factors such as the East Asian summer monsoon index, summer India Burma trough, the summer North Atlantic Oscillation (NAO), Equatorial Pacific sea surface temperature and so on. The high-temperature forecasting model is established with the setup of the high temperature prediction rules. The data of summer maximum temperature in summer in Zhangzhou, Fujian Province from 1955 to 2012 are selected to calculate the summer hot temperature of 58a. Then, multiple climatic factor data of the same period is given to the input variable, and 46 years of data is randomly selected to get 10 classifications of rule sets, resulting in the achievement of the accuracy rate to 91.49%. With the remaining data of 12a test, the accuracy rate reaches 91.67%. In general, the results of this paper validate the feasibility and validity of the high temperature prediction model, and provide a new idea for the study of the catastrophic weather model.

Keywords: CART · High temperature effective accumulated temperature
Summer high temperature forecast

1 Introduction

In recent years, China is facing more frequent and severs meteorological disasters under global climate warming and urbanization expansion. As one of the common meteorological disasters in summer, high temperature will not only affect the social and economic development, but also affect people's daily life and safety of life [1–5].

K. Long et al. (Eds.): 5GWN 2017, LNICST 211, pp. 317–326, 2018.
https://doi.org/10.1007/978-3-319-72823-0_30

China Meteorological Administration provides the daily maximum temperature reached or exceeded 35 °C, known as high temperature. The occurrence of high-temperature events is the result of multi-factor and multi-system. At present, many factors have been studied on the influence factors of summer high temperature. Brabson et al. [5] studied the evolution of extreme temperatures in England and found that both cold winters and hot summer extreme temperatures were associated with changes in the lower atmospheric circulation; Lin et al. [6] and Yin et al. [7] found that summer high temperatures were associated with the West Pacific High anomalies; According to Sun [8], the extreme high temperature events in China and East Asia over the atmospheric circulation variation is consistent, and the role of low-level warm advection is rather important. Facing the high temperature caused by a variety of reasons, its influence cannot be underestimated. As a result, this calls for the establishment of high-temperature forecasting model.

In addition, the current high temperature of the degree of heat is no clear grade division. Some scholars use the 95th or 90th percentile of daily maximum temperature as a boundary [9, 10]. However, due to the climatic differences in regional level and people's long-term adaptation to local climate [11], the definition of the degree of hot does not only rely on the standard temperature, and the impact of high temperature on the human body or the extent of harm to develop should be taken into consideration. Therefore, according to the principle of high temperature warning and the impact of high temperature weather on the human body and crop hazards, high temperature effective accumulated temperature is adopted to distinguish the hot degree. The high temperature effective accumulated temperature (EAHT) is the sum of the difference between the daily maximum temperature and 35 °C. The larger the value is, the severe the hot event is, and the more harmful to human health, and vice versa. On this basis, the accumulated effective temperature can be calculated by month or year of high temperature to characterize the degree of hotness of the corresponding statistical period [12].

For the establishment of high-temperature forecasting model, an effective method of forecasting model is needed apart from the ideal high temperature index. Nowadays, there are mainly three modeling ideas [13]: The first is the statistical type. Through the statistical study of a large number of related data, we find the relationship between the climatic factors and the high temperature to establish the relationship model. And then use the change of factors to predict high temperature; The second is the theoretical analysis. Mainly through the basic principles of the weather science and other disciplines, use the numerical model to get the temperature of the forecast value and then forecast the high temperature [14, 15]. Even though the former is considered to be more comprehensive, the traditional methods of statistical analysis require numerous data and the process is too cumbersome and difficult to achieve the desired accuracy [16]. The latter can reflect the phenomenon of high temperature weather from a more essential level, but the parameters are not easy to obtain. Therefore, the third idea, which is, based on data mining methods, is used in the model.

The high temperature forecasting model based on climatic factors is established by using a large amount of relevant data. The data mining method has many categories, which can effectively and quickly extract effective information from a large number of related data to establish the model. Moreover, the operation is relatively simple, and the accuracy can reach a higher benchmark [17, 18] in the reasonable set of parameters.

The methods of data mining commonly used in meteorological forecasting model compose neural network, support vector machine and decision tree algorithm [13]. The decision tree algorithm is widely used because of its simple calculation, fast processing speed and easy explanation [13], which has a good application prospect in dealing with meteorological problems.

Therefore, in this paper, the classification regression tree algorithm (CART) is used to study the relationship between high temperature in summer and some climatic factors such as the East Asian summer monsoon index, summer India Burma trough, the summer North Atlantic Oscillation (NAO), Equatorial Pacific sea surface temperature, the landing typhoon, Nino3, Nion4, and Nino3.4 and so on. Then according to the obtained rule set, a high temperature forecast model based on climatic factors is established.

The following section will elaborate more on the data and analysis approaches and the results will be present in Sects. 3 and 4. In the last section, a brief discussion as well as the concluding remarks will be given.

2 Data and Methods

2.1 Data

The daily maximum temperature and daily mean temperature from the summer of 1955 to 2012 in Zhangzhou, Fujian Province were selected from the meteorological station data, and the missing data in the original data were excluded. (If no special instructions, the following research in this article are 6, 7 and 8 months for the study period)

The index data of the western Pacific subtropical high index, intensity index, west ridge point, ridge line position, north boundary position and Indo-Burmese trough index were derived from the National Climate Center of China from 1955 to 2012. And other climatic factors such as monthly mean sea surface temperature data for Nino1 + 2, Nino3, Nino4 and Nino 3.4 in 1955–2012 are obtained from NOAA website.

2.2 Research Methods

In this paper, high temperature effective accumulated temperature (EAHT) in summer is used to judge the hot degree of hot weather. The summer high temperature effective accumulated temperature is the sum of the daily maximum temperature and the difference of 35 °C in summer. The mean temperature of effective accumulated temperature in summer is −71.04 which is taken as the boundary of whether or not it is hot. This accumulated temperature considers the climate and physiology of the human body, which can better show that the high temperature hot degree.

Decision tree algorithm is an important classification method in data mining. It aims at obtaining decision-making steps, discovering rules, patterns and knowledge from archived databases [19]. The root node, the branch, and the leaf node are the necessary components of the decision tree. Where each interior node represents a detection on an attribute, each branch representing a detected output, and a leaf node of each tree

represents a class or class distribution. In this paper, the classification and regression tree algorithm (CART) which is a classic decision tree algorithm [20] proposed by Breman et al. in 1984 is selected. The algorithm is a nonparametric statistical method for classifying discrete or continuous dependent variables.

CART is a supervised learning algorithm. Before CART is used to predict, the user must first provide a set of learning samples to build and evaluate CART. CART uses a learning sample set in the following structure:

$$\begin{aligned} &L := \{X_l, X_{2\ldots}X_m, Y\} \\ &X_1 := (x_{11}, x_{12\ldots}x_{1t1)}, \ldots, X_m := (x_{m1}, x_{m2\ldots}x_{mtn}) \\ &Y := (Y_1, Y_{2\ldots}Y_k) \end{aligned}$$

Where $X_1 \ldots X_m$ is called the attribute vectors, and its attributes can be ordered or discrete. Y is called the label vectors, and its attributes can be ordered or discrete. When Y is an ordered quantity, it is called a regression tree; when Y is a discrete value, it is called a classification tree [22]. Since the target variables of the high temperature prediction model are discrete, a classification decision tree is generated [21].

The criteria for variable classification in CART are the Gini Impurity Criterion and the Goal Dichotomy Criterion. Whether or not high temperature is a binary classification problem is a special case of multivariate classification. Given a node t, the estimated class probability p (j|t) represents the probability that the node belongs to class j ($j = 1, 2, 3 \ldots J$). The formula for determining the impurity of a given node is

$$i(t) = \phi[p(1/t), \ldots, p(J/t)]$$

Where Φ is the impurity function, the optimal partition is obtained when constructing the decision tree nodes, so that the impurity degree of each child node is the lowest. Impurity function is generally expressed as:

$$i(t) = \phi(p_1, p_1, \ldots, p_1) = -\sum_{j=1}^{J} p_j \log p_j$$

With the Gini diversity index, the form of the function is as follows

$$i(t) = \left[\sum_{j=1}^{J} p(j \mid t)\right]^2 - \sum_{j=1}^{J} p^2(j \mid t)$$

This paper discusses the binary classification problem; the index can be simplified as

$$i(t) = 2p(1 \mid t)p(2 \mid t)$$

At node t, randomly selected objects are assigned to class i according to probability p (i|t), and the estimated probability of the object actually belongs to class j is p (i|t). The estimated probability of misclassification under this rule is the Gini index,

$$Gini\ Index = \sum_{i \neq j} p(i \mid t\ ) p(j \mid t) = 1 - \sum_{j=1}^{J} pj^2$$

In the process of classifying the decision tree with high temperature, the optimal partitioning threshold and the best test variable are selected according to the Gini coefficient of the computing node. And the optimal decision tree is generated by recursive call until the end rule is satisfied.

3 Temporal Characteristics and Influencing Factors of Summer High Temperature

In this paper, data preprocessing in Zhangzhou area of Fujian Province was used to analyze the change trend of summer high temperature effective accumulated temperature. And the impact factors of summer high temperature were studied to determine their influence on high temperature trend.

3.1 Time Distribution of Summer High Temperature

As shown in Fig. 1, the effective accumulated temperature of high temperature in summer showed an upward trend during the period of 1955–2012. It is worth

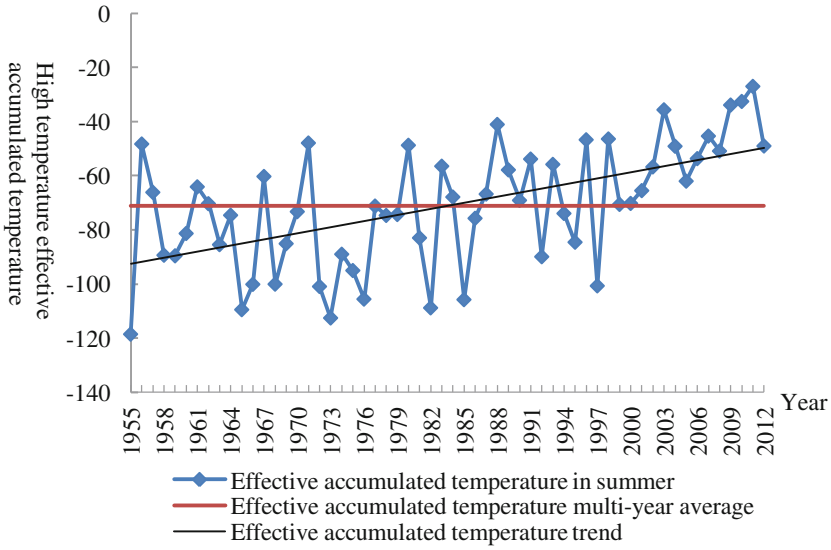

Fig. 1. The change trend of effective accumulated temperature in summer in Zhangzhou, Fujian

mentioning that the effective accumulated temperature in summer was higher than the average for many years after 2000, where the hot weather appeared.

When constructing a decision tree using CART algorithm, it is necessary to perform a preprocess analysis of the data set, which is the target variable and the input variable needed to be determined. In order to study the influence of climatic factors on this trend, we use the multi-year mean summer effective accumulated temperature (−71.04) as the target variable.

3.2 The Correlation Analysis of Summer High Temperature Index and Different Factors

Previous studies have found that there are significant correlations between the high temperature and some climatic factors, such as the western Pacific subtropical high, the intensity index, the west extension point, the landfall typhoon etc. In this paper, the correlations of summer high temperature effective accumulated temperature and the data of each factor in the same period in Zhangzhou, Fujian from 1955 to 2012 were analyzed. The results are as follows:

Table 1. 1955–2012 summer climatic factors and the summer high temperature effective accumulated temperature correlation coefficient set

Climatic factor	Correlation coefficient
Nino3	−0.023
Summer NAO	−0.331*
Nino4	0.079
Nino3.4	−0.037
Summer Western Pacific Subtropical High Strength Index	0.417**
Summer Western Pacific Subtropical High Area Index	0.413**
Summer Western Pacific Subtropical High West Stretch Point	−0.372**
Nino1+2	−0.026
Landing Typhoon	0.102
Summer East Asian Monsoon Index	−0.383**
Summer India Burma Trough	0.399**

Note: ** for the correlation coefficient through the $\alpha = 0.01$ significant test.
* for the correlation coefficient through the $\alpha = 0.05$ significant test.

It can be seen from Table 1 that Nino3, Nino3.4, summer NAO, summer western Pacific subtropical high west ridge point, Nino1+2 and East Asia summer monsoon index were negatively correlated with summer high temperature effective accumulated temperature; Besides, what is positively correlated is Nino4, summer western Pacific subtropical high Intensity index, the summer western Pacific subtropical high area index, landing typhoon, summer India and Myanmar and summer high temperature

effective accumulated temperature. The correlation coefficients of summer western Pacific subtropical high intensity index, summer western Pacific subtropical high area index, summer western Pacific subtropical high west ridge point, East Asian summer monsoon index, summer Indian Burma trough and summer high temperature effective accumulated temperature were adopted $\alpha = 0.01$ significance test. The correlation coefficient between summer NAO and summer high temperature effective accumulated temperature was tested by $\alpha = 0.05$. The results showed that the summer high temperature is highly related to these, and other factors had some correlation with the summer high temperature. In this paper, these climatic factors are used as input variables when constructing decision tree using CART algorithm.

4 Construction and Application of High Temperature Forecast Model Based on CART Algorithm

Data mining methods (Fig. 2) are used in this research to identify and analyze the relation between summer temperatures effective accumulated temperature and climatic factors. The accessibility of the forecast data regarding climatic factors from air-sea coupled model forecasting together with the use of forecasted factors as input variables ensure the make prediction of the high temperature in summer.

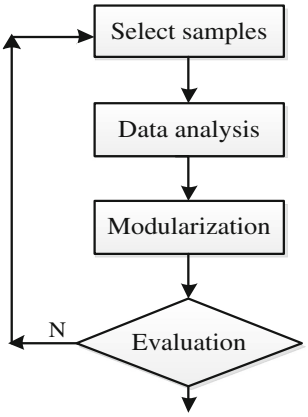

Fig. 2. Basic flow chart of data mining

4.1 Construction of Decision Tree Model Based on CART

In order to make the high temperature forecast more reasonable, this paper uses the multi-year average of summer high temperature effective accumulated temperature (−71.04) as the hot index, which is also the target variable when constructing the model.

The process of establishing a high temperature forecasting model based on the CART model is as follows: firstly, the input variable data is obtained (A number of

climatic factor data, and pretreatment); secondly, using the temperature data, calculating the index of summer hot degree and judging whether it is hot in summer; then selecting some data randomly as training set, taking the climate data as the input variables and whether the summer high temperature as a target variable. Using the data mining software (IBM SPSS Modeler) and the CART algorithm to build the model, then obtain the forecast rule set. Finally, select the remaining data as the test set to verify the accuracy of the model.

4.2 Application of High Temperature Prediction Model

Taking Zhangzhou of Fujian Province as an example, the CART algorithm was selected by IBM SPSS Modeler to build a high temperature forecasting model.

The main contents of the decision tree are constructed by CART algorithm: the parent node recursively tests the random test variable and the segmentation threshold with the Gini coefficient until the best test variable and the segmentation threshold are generated. After entering the child node, the behavior of the parent node continued until the end condition is satisfied. The decision tree (Fig. 3) was obtained by using CART for a number of climatic factors during the summer of 1955–2000, and the accuracy of self-learning was 91.49% (Table 2). Each path from the root node to the child node represents a high-temperature prediction rule. The data in leaf nodes represent the high temperature, the total sample size and the number of misclassified samples respectively. Take "0 (1/0)" as an example: outside of the brackets represent non-high-temperature 0 (1 on behalf of high temperature), left side of the parentheses on behalf of a sample of the total number of brackets inside the right 0 represents the number of high-temperature samples. The difference, that is, 1 – 0 = 1, represents the

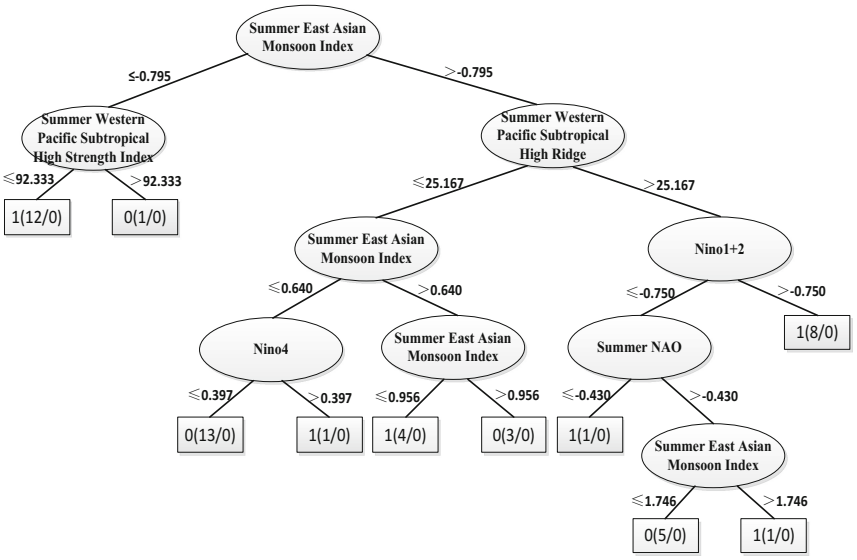

Fig. 3. The high temperature forecasting decision tree produced by CART algorithm

number of non-high temperature samples correctly classified. In order to further verify the reliability, the 2001–2010 climatic factor data was selected and the corresponding accuracy rate achieved 91.67% (Table 2).

Table 2. High temperature prediction model results table

	Training	Probability	Testing	Probability
Correct	43	91.49%	11	91.67%
Wrong	4	8.51%	1	8.33%
Total	47	100%	12	100%

With the high accuracy rate of 91.67%, the high temperature prediction model established by the data mining method can be considered valid and reliable. This also provides a new idea for the study of high temperature and other climate prediction.

5 Conclusion

This paper was based on the CART algorithm of data mining, and a high temperature climate prediction model was established. Taking Zhangzhou, Fujian as an example, and the following conclusions are obtained:

(1) The inter-annual variation of effective accumulated temperature of summer high temperature in Zhangzhou, Fujian is an upward trend, especially after 2000.
(2) There were significant correlations between summer NAO, summer western Pacific subtropical high intensity index, summer western Pacific subtropical high area index, summer western Pacific subtropical high west ridge point, East Asian summer monsoon index, summer Indian-Burmese trough and summer high temperature.
(3) Random selection of Zhangzhou from 1955 to 2012 of which 46 years of data, the establishment of classification decision tree to get the rule set, and classification accuracy rate of 91.49%. The remaining data for the inspection accuracy rate of 91.67%, indicating that the model has good reliability.

Although the prediction model established in this paper has reached certain accuracy, further studies are still needed. It is helpful to improve the accuracy of disaster prediction by discovering the potential information and laws concerning weather through the data mining method.

References

1. Eastcrling, D.R., Evans, J.L., Groisman, P.Y., et al.: Observed variability and trends in extreme climate events: brief review. Bull. Amerimete Soc. **81**(3), 417–425 (2000)
2. Changnon, S.A., Pielke, R.A., Changnon, D., et al.: Human factors explain the increased losses from weather and climate extremes. Bull. Amerimete Soc. **81**(3), 437–442 (2000)

3. Karl, T.R., Jones, P.D., Knighta, R.W., et al.: A new perspective on recent global warming: asymmetrie trends of daily maximum and minimum temperature. Bull. Am. Meteorol. Soc. **74**(6), 1007–1023 (1993)
4. Gruaz, G., et al.: Indictors of climate change for the Russian Federation. Clim. Chang. **42**, 219–242 (1999)
5. Brabson, B.B., Palutikof, J.P.: The evolution of extreme temperatures in the Central England temperature record. Geophys. Res. Lett. **29**(24), 2163–2166 (2002)
6. Lin, J., Bi, B., He, J.: Study on the variation of the western Pacific subtropical high and the formation mechanism of high temperature in Southern China in July 2003. Chin. J. Atmos. Sci. **29**(4), 594–599 (2005). (in Chinese)
7. Yin, J., Zhang, C., Zhang, C.: Analysis of rare high temperature climate in summer 2003 in Jiangxi Province. J. Nanjing Inst. Meteorol. **28**(6), 855–861 (2005). (in Chinese)
8. Sun, J., Wang, H., Yuan, W.: Decadal variability of the extreme hot event in China and its association with atmospheric circulations. Clim. Environ. Res. **16**(2), 199–208 (2011). (in Chinese)
9. Ding, T., Qian, W.H., Yan, Z.W.: Changes in hot days and heat waves in China during 1961–2007. Int. J. Climatol. **30**, 1452–1462 (2010)
10. He, S., Dai, E., Ge, Q., et al.: Spatiotemporal pattern prediction of high temperature disaster risk in China. J. Nat. Disasters **19**(2), 91–97 (2010). (in Chinese)
11. Ye, D., Yin, J., Chen, Z., et al.: Temporal and spatial characteristics of summer heat wave in China during 1961–2010. Chin. J. Clim. Chang. Res. **9**(1), 15–20 (2013). (in Chinese)
12. Chen, M., Geng, F., Ma, L., et al.: Analysis of high temperature heat wave in Shanghai area in recent 138 years. Plateau Meteorol. **32**(2), 597–607 (2013). https://doi.org/10.7522/j.issn.1000-0534.2012.00058. (in Chinese)
13. Shi, D., Geng, H., Chen, J.I., et al.: Construction and application of road icing forecast model based on C4.5 decision tree algorithm. J. Meteorol. Sci. **35**(2), 204–209 (2015). (in Chinese)
14. Gao, X., Zhao, Z.: The experiment of extra seasonal prediction in China by OSU/NCC GCM for flood season. J. Appl. Meteorol. **11**(2), 180–188 (2000). (in Chinese)
15. Wang, Q., Feng, G., Zheng, Z., et al.: Study on objective and quantitative prediction of multi-factor combination for precipitation optimization in flood season of middle and lower reaches of Yangtze River. Chin. J. Atmos. Sci. **35**(2), 287–297 (2011). (in Chinese)
16. Wang, W., Xue, F., Shi, D., et al.: Research on summer drought prediction model based on CART algorithm. J. Meteorol. Sci. **36**(5) (2016). https://doi.org/10.3969/2015jms.0067. (in Chinese)
17. Zhang, W., Gao, S., Chen, B., et al.: The application of decision tree to intensity change classification of tropical cyclones in western North Pacific. Geophys. Res. Lett. **40**(9), 1883–1887 (2013)
18. Liu, Z., Du, Z., Chen, H., et al.: Study on the land use and cover classification of Zhengzhou based on decision tree. Meteorolog. Environ. Sci. **31**(3), 48–53 (2008). (in Chinese)
19. Friedl, M.A., Brodley, C.E.: Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. **61**(3), 399–409 (1997)
20. Zhao, P., Fu, Y., Zheng, L., et al.: Land use/cover classification of remote sensing images based on classification regression tree analysis. J. Remote Sens. **9**(6), 708–716 (2005). (in Chinese)
21. Zhang, W., Leung, Y., Chan, J.C.L.: The analysis of tropical cyclone tracks in the western North Pacific through data mining. Part I: tropical cyclone recurvature. J. Appl. Meteor. Climatol. **52**(6), 1394–1416 (2013)
22. Ying, T.: Remote Sensing Image Classification Based on Texture Information and CART Decision Tree Technology. Nanjing Forestry University (2008)