ATLANTIS PRESS

Advances in Engineering Research, volume 118

2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017)

Application of Decision Tree C4.5 Algorithm in Air Quality Evaluation

Nan Hu[1, a], Qing Li[2, b]

[1,2]Key Laboratory of Fiber Optic Sensing Technology and Information Processing(Wuhan University of Technology), Ministry of Education, School of Information Engineering, Wuhan, Hubei Province, China

[a]695809317@qq.com, [b]654058182@qq.com

Keywords: C4.5 decision tree; air quality; classification prediction

Abstract. Air pollution has a significant impact on human production and life, the prediction of air quality's level is conducive to the relevant departments to take appropriate measures. This paper selects six basic pollutants PM10,PM2.5,SO_2,NO_2,CO and O_3, which are stipulated in the "Ambient Air Quality Standard" (GB3905-2012) issued by China in 2012. Based on the data of air pollutant concentration in Wuhan in October, 2016, a prediction model of air quality grade based on C4.5 decision tree algorithm was established. The experimental results show that the proposed algorithm has an ideal classification prediction effect.

Introduction

With the rapid development of economy, industrialized production of the many pollutants released into the atmosphere make the air quality significantly decreased, resulting in people's health by a certain threat. The establishment of ambient air quality level can provide reference for people's travel. Therefore, the uniform and precise classification of air quality will play a very important role for the rational planning of production and life, as well as the introduction of relevant policies and regulations for the management of air pollution by urban decision makers.

The decision tree is a representation of a decision process that determines the association of a given sample with a property. A particularly effective way to generate a classifier from data is to generate a decision tree that is widely used in data mining and machine learning to solve classification-related problem [1]. At present, the main algorithms of generating decision tree are CART algorithm [2], ID3 algorithm [3], C4.5 algorithm [4,5] and so on. Among them, the C4.5 algorithm has the characteristics of fast classification and high precision, and it is a kind of decision tree algorithm which is developed well.

Based on the decision tree C4.5 algorithm, this paper classifies and predicts air quality grade by the air quality pollutant concentration, and explores how to provide the reference for people to rationally plan the production and life under the current urban air quality condition, and further study of the effects of pollutants on air quality, in order to better improve the air quality.

C4.5 Decision Tree Algorithm

Decision tree algorithm was first proposed by Hunter in the CLS (Concept Learning System) in the 1950s and later developed by J.R. Quinlan in 1979, he put forward the famous ID3 algorithm. ID3 algorithm is based on information theory. To information entropy and information gain as a measure, so as to achieve the induction and classification of data. Its feature is mainly for discrete attribute data. C4.5 algorithm is developed on the basis of ID3 algorithm, which inherits all the advantages of ID3 algorithm, increases the discretization of continuous attributes, and uses the information gain rate as the standard of classification, thus overcoming the lack of tending to select an attribute with more values when using information gain to select an attribute [6]. The core idea of C4.5 algorithm is to use the principle of information entropy, and select the largest attribute of information gain ratio as a classification attribute, and construct a decision tree using recursive method.

Implementation Process of C4.5 Algorithm. (1) Create the root node N. (2) If the set D belongs to class C, then returned with N as the leaf node, marked with class C. (3) If the attribute list is empty, then N is the leaf node to return, and mark N as the most appearing class in D. (4) For each attribute in the attribute list, calculate the information gain rate, the test attribute of N is the attribute with the highest information gain ratio. (5) For each new leaf node generated by node N, if the leaf node corresponding to a subset of the sample is empty, then split the leaf node to generate a new leaf node. Otherwise, perform (1) again on the leaf node and continue splitting. (6) Calculate the classification error for each node to prune.

Calculation of C4.5 Algorithm's Information Gain Rate. Suppose D is the training set, |D| denotes the number of training example, and training set has m attributes. The class attribute C divides the training set into n subsets, each of which is $\{C_1, C_2, ..., C_n\}$, and the corresponding values of the class attribute are $\{c_1, c_2, ..., c_n\}$.Among them, $|C_i|$ represents the number of instances of subset c_i. Set A is one of the attributes (non-class attribute), A divides the training set into v subsets $\{A_1, A_2, ..., A_v\}$, and the corresponding values of attribute A are $\{a_1, a_2, ..., a_v\}$. Among them, $|A_j|$ denotes the number of data sample while the attribute A is a_j in training set D, $|A_{ij}|$ denotes the number of samples of class C_i in subset A_j(Among them, i=1,2,…,n; j=1,2,…,v).

Given an instance, the probability that the instance belongs to the class C_i is

$$P(C_i) = \frac{|C_{ij}|}{|D|} \quad (i = 1,2,\dots,n) \tag{1}$$

The probability of its attribute A is a_j

$$P(A_j) = \frac{|A_j|}{|D|} \quad (j = 1,2,\dots,v) \tag{2}$$

In the sample set, when the value is a_j of attribute A, The probability of the class attribute C_i is

$$P(C_i|A_j) = \frac{|A_{ij}|}{|A_j|} \quad (i = 1,2,\dots,n; j = 1,2,\dots,v) \tag{3}$$

The amount of information required for the decision tree to divide the category C, namely the information entropy is

$$Info(D) = -\sum_{i=1}^{n} P(C_i) * \log_2 P(C_i) \tag{4}$$

The information entropy required for the attribute A to classify is

$$E_A(D) = \sum_{j=1}^{v} P(A_j) * Info(A_j) \tag{5}$$

Among them,

$$Info(A_j) = -\sum_{i=1}^{n} P(C_i|A_j) * \log_2 P(C_i|A_j) \quad (j = 1, 2, ..., v). \tag{6}$$

Then, the information gain of the attribute A with respect to the category set C is

$$Gain(A) = Info(D) - E_A(D). \tag{7}$$

The information gain rate of the attribute A relative to the category set C is

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}. \tag{8}$$

Among them,

$$SplitInfo_A(D) = -\sum_{j=1}^{v} P(A_j) * \log_2 P(A_j) \tag{9}$$

The Environment of Experiment

Analysis Tools and Platforms. The tool of experiment is weka3.8.1, the platform is windows7 System, the address of downloading the weka software is http://www.cs.waikato.ac.nz/ml/weka/downloading.html.

The Data of Experiment. This paper selects the data provided by Qingyue Open Environment Data Center for mining analysis [7]. The data set is the one-hour average concentration of air pollutants in Wuhan in October 2016. The data set is named airData_201610, with 7 attributes and 638 data. According to China's urban air quality classification standards [8], the air quality grade is divided into six levels, including excellent (I), good (II), mild (III), moderate (IV), severe (V), serious(VI). Thus the data that was pretreated by classifying included one category item(air rating - last column in TABLE 1) and six attribute items affecting the classification(air pollutants). TABLE 1 shows the partial data after the pretreatment by classifying(The unit of pollutant's concentration: ug/m^3).

TABLE 1
The relationship table between air pollutants and grade

SO_2	NO_2	CO	O_3	PM10	PM2.5	Air level
7	32	0.556	28	25	19	I
7	29	0.611	30	27	21	I
7	28	0.456	30	23	22	I
17	29	0.563	26	26	22	I
8	28	0.625	26	36	28	I
7	25	0.513	29	27	25	I
7	26	0.538	26	30	26	I
7	31	0.588	19	35	28	I
7	37	0.563	18	27	26	I
7	36	0.567	20	36	28	I
7	37	0.656	22	39	28	I
8	37	0.722	24	44	30	I
9	44	0.967	21	59	30	II
9	45	1.178	23	77	38	II
10	55	1.4	27	75	45	II

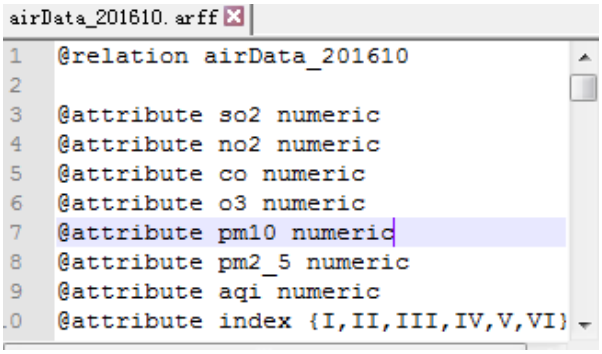

Figure 1 Header Information of airData_201610 Data Set

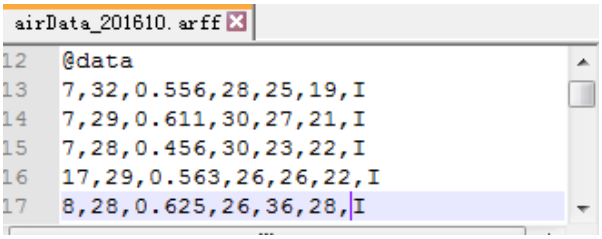

Figure 2 Data Information of airData_201610 Data Set

The original way of weka to store data is ARFF format file [9]. The original data in this paper is a .csv format file, which can be converted into .arff format file by weka software. Saving the converted file, and then imported weka software for experiment. The data in ARFF format file is composed of several instances, which are independent of each other and are not intrinsically linked, and the file does not need to give a specific ordering for these instances. An ARFF file is usually composed of three parts: header information, instance data information and comment lines.

Header information usually includes a relationship name and attribute list and so on. The header information of ARFF file includes two parts, @relation and @attribute. Among them, @relation declares the relationship name of a data in the form of @relation <relation name>, where relation name is a string that contains quotation marks when it contains spaces; @attribute declares a property in the text, and the declared attributes are arranged in order of precedence. Similarly, attribute name is also a string, when it contains spaces, you must use quotation marks. ARFF file, starting the line with @data, followed by examples of data, one line is an instance, each instance has a number of properties, attributes for the column. Among them, in order to facilitate the distinction between

attributes, with "," separated. If a line starts with "%", the line is the comment line for the data set. Figure 1 and Figure 2 shows the head data and data information of airData_201610.

Application of C4.5 Algorithm and Experimental Results

In this paper, the C4.5 decision tree algorithm in data mining is used to establish the model of air quality classification and prediction. After the data preprocessing, by the aforementioned C4.5 algorithm, the six air pollutant attributes are regarded as the object attributes of the algorithm, and the attribute of air quality level is regarded as the target attribute. The attribute is ranked by the definition of the information gain rate, and the attribute with the highest information gain rate is selected as the test attribute of the given set. Use this attribute tag to create a branch for each value of the attribute, and then recursively build tree, which a decision tree can be constructed.

After importing the file of airData_201610.arff into weka, selecting J48 algorithm, namely C4.5 decision tree algorithm, and then setting the algorithm parameters and experimenting. Select Cross-validation, and set the parameter to 10, that is, 10-fold cross-validation. The method divides the data set into ten parts, nine of them as training data and one part as test data by turn. Each test will get a corresponding correct rate, the accuracy of 10 times results' average as an estimate of the accuracy of the algorithm. The method uses pessimistic estimation method, which can be closer to the real prediction rate. The parameters of C4.5 algorithm are set as follows [10].

Testing parameters of ConfidenceFactor forms TABLE 2, the TABLE 2 shows that when the confidence factor is 0.25, the accuracy of classification prediction model reaches the maximum. So take ConfidenceFactor equals 0.25.

Set the confidence factor to 0.25, testing parameters of minNumObj forms TABLE 3, the TABLE 3 shows that when the smallest number of leaf nodes, that is, the branch number is 4, the accuracy of classification prediction model reaches its maximum. So take minNumObj equals 4.

TABLE 2 The accuracy about ConfidenceFactor

confidence factor	0.05	0.1	0.2	0.25	0.3	0.4
accuracy	96.0815 %	96.0815 %	95.7650 %	96.395 %	96.395 %	96.395 %

TABLE 3 The accuracy about minNumObj

number of branches	2	3	4	5	6
accuracy	96.3950%	96.8652%	97.0219%	96.0815%	95.7660%

After the parameters are set up, generating a new file airData_201610Result.arff by classification and prediction in weka. The changes are shown in Figure 3.

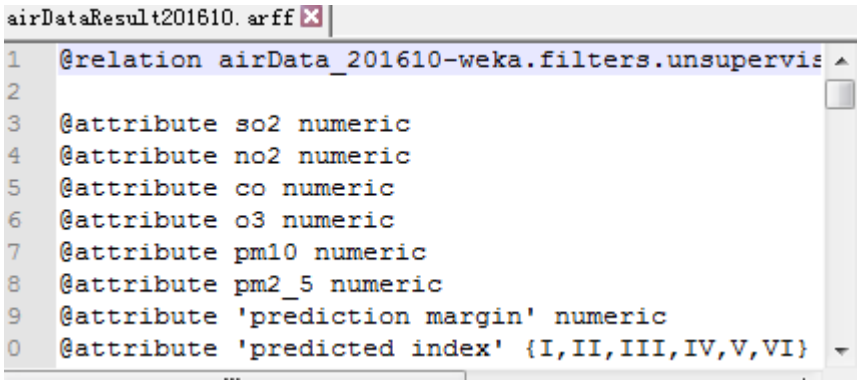

Figure 3 New file generated after the classification prediction

There are two new properties added to the file, where the value of the predicted index is the prediction value of the original index attribute. We can draw the air quality decision tree shown in Figure 4.

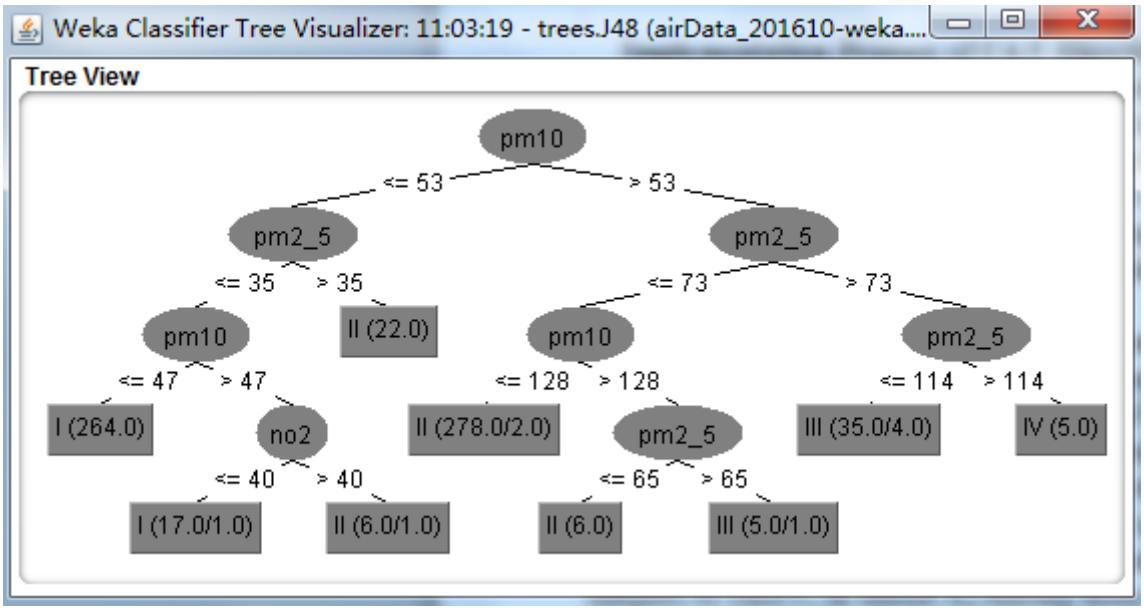

Figure 4 Model of Decision Tree

Conclusion

In this paper, the C4.5 decision tree is applied to the classification and prediction of air quality grade, and a large amount of data is transformed into classification rules. From the experimental results, we can know that choosing appropriate confidence factor and the number of branches can improve the classification accuracy of the algorithm. The accuracy of classification prediction reached to 97.0219% in this paper, which indicates that the air quality assessment model based on C4.5 decision tree algorithm can achieve good classification and prediction effect. At the same time, from the decision tree can be drawn that PM10, PM2.5 and NO_2 have a great impact to the air quality in Wuhan, which will guide the urban decision-making management to pay special attention to pollution sources of PM10, PM2.5, NO_2 in air quality's improvement.

References

[1] S.B.Kotsiantis. Decision trees: a recent overview[J]. Artificial Intelligence Review, 2013, 39(4)261-283.

[2] Leszek Rutkowski, Maciej Jaworski, Lena Pietruczuk, Piotr Duda. The CART decision tree for mining data streams[J]. Information Sciences, 2014,266.

[3] Wang Xiaohu, Wang Lele, Li Nianfeng. An Application of Decision Tree Based on ID3[J]. Physics Procedia, 2012,25,1017-1021.

[4] Hong Yan Zhao. The Application and Research of C4.5 Algorithm[J]. Applied Mechanics and Materials, 2014, 1285-1288.

[5] Li Dongming, Li Yan, Yuan Chao, Li Chaoran, Liu Huan, Zhang Lijuan. The application of decision tree C4.5 algorithm to soil quality grade forecasting model[C]. Computer Communication and the Internet, 2016.

[6] Miao Hongxing, Yu Jiankun. The Comparisons Between ID3 and C4.5 Algorithm Based on Decision Tree[J]. Modern Computer, 2014, 7-10.

[7] https://wat.epmap.org/.

[8] Weather network. http://www.tianqi.com/news/14127.html..

[9] Ian H.Witten. DATA MINING: Practical Machine Learning Tools and Techniques[M]. Beijing: Machinery Industry Press, 2006.

[10] Shih-Wei Lin, Shih-Chieh Chen. Parameter determination and feature selection for C4.5 algorithm using scatter search approach[J]. Soft Computing, 2012, 16(1), 63-75.