The Preventive Maintenance of Highway Based on Data Mining

Shunzhou Xiao[1,2],and *Mingxin* Nie[1,2]

[1]School of Information Engineering, Wuhan University of Technology, Wuhan, 430070, China

[2]Key Laboratory of Fiber Optic Sensing Technology and Information Processing, Ministry of Education, Wuhan University of Techn ology, Wuhan, 430070, China

Abstract: Judging from the current situation of Chinese highway maintenance,only after there is highway distress would the staff have a repair, that results in the poor efficiency of highway maintenance. In order to improve the efficiency of highway maintenance,this paper will use data mining technology to predict the pavement performance of highway and analyze the main factors of pavement performance attenuation,so that the preventive maintenance can be carried out.We will provide data support to the preventive maintenance of highway by using the isolation Forest anomaly detection algorithm to have a data pretreatment, the regression model and time series GM (1,1) model to predict the pavement performance and the association rule analysis and isolation Forest to analyze the main factors of pavement performance attenuation.

1.Introduction

In recent years, with the rapid development of information technology, the highway maintenance management has realized informatization.Technology status assessment of highway, traffic flow and other information have accumulated a large amount of data with the highway operation.On the other hand, our road maintenance management is still using traditional maintenance methods--passive maintenance, only if the pavement performance dropped to a level of fairly low will a series of repair measures be carried out. This kind of maintenance dose not make full use of our existing data information and it can just improve the pavement performance a little, what's more, the rate of performance decaying again is also very fast. That results in the extremely low efficiency of highway maintenance and the waste of maintenance funds. According to these problems above, this paper will apply data mining technology to highway maintenance work to fully explore the potential value of technology data that already existed for the purpose of improving the efficiency of highway maintenance[1].

In this paper, the iForest anomaly detection algorithm is used to preprocess the existed data of highway pavement performance. Then, we predict the pavement performance like as pavement damage index (PCI), riding quality index (RQI), road rutting depth index (RDI) and pavement comprehensive quality index (PQI) by using the regression model and the time series model GM (1,1)[2]. Finally, it uses the association rule algorithm and iForest algorithm to analyze the main factors of pavement performance decay and have practical application with existed pavement performance data of

Hubei Province.The pavement performance prediction results and the mining of main factors of highway performance decay can allow us to locate the target that needing maintenance and carry out preventive maintenance.Further,preventive maintenance can prevent or delay the rapid deterioration of highway diseases, effectively extend the road life and save a lot of road maintenance funds.

2.Key technology of preventive maintenance

2.1IForest anomaly detection model

The iForest algorithm consists of a large number of binary trees, called isolation tree and iTree for short, which are the basis of the iForest algorithm and are constructed as follows.

Assuming that the data set has N records,randomly selecting ψ records from the data set as the training samples to build an iTree,usually sampling without replacement.

In the samples, an attribute q and a split value p is randomly selected within the range of all values of the attribute(between the minimum and maximum). Then the samples is divided into two parts by attribute q and split value p. The instance whose q is smaller than p will be divided into left subtree and the instance whose q is greater than or equal to p will be divided into right subtree.That results in a splitting condition and a data set on the left and right subtree. We recursively divide subtree according to the method above,until either:(i)the tree reaches a height limit,(ii)the subtree has only one node,(iii)all data in subtree have the same values.

After the amount of iTree reach to t, iForest is ready to work. We can evaluate the degree of anomaly of test instance x by using the generated isolation forest above, the process of testing as follows: (i) instance x traverses an iTree from the root node until the traversal is terminated at an external node,(ii)make the test data x in each iTree along the corresponding branch conditions to go down to the leaf node and calculate the height $h(x)$ of the end node in the iTree,(iii)calculate the anomaly score of instance x using the following Equation(1).

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (1)$$

where $E(h(x))$ is the average of $h(x)$ from a collection of isolation trees.$c(n)$ is the average height of the unsuccessful search in binary search tree, and can be calculate by the equation as follows:

$$c(n) = 2H(n-1) - (2(n-1)/n) \quad (2)$$

where $H(i)$ is the harmonic number and it can be estimated by $ln(i) + 0.5772156649$.We can calculate the anomaly score by Equation(1).Then we are able to make the following assessment:

(a) if instances return s very close to 1, then they are definitely anomalies[3],

(b) if instances have s much smaller than 0.5, then they are quite safe to be regarded as normal instances[3],

(c) if all the instances return $s \approx 0.5$, then the entire sample does not really have any distinct anomaly[3].

2.2The model of regression prediction

Through research and practical observations, we find that the attenuation of pavement performance is not linear, and it is very slow in the beginning period of highway operation, but it will begin to drop sharply when the performance drops to a certain value. According to this characteristic, we use the nonlinear regression equation shown in Equation(3) as the model[4].

$$PPI = PPI_0\{1 - exp[-(\frac{\alpha}{t})^{\beta}]\} \quad (3)$$

where PPI_0 represents the initial value of the pavement performance (usually 100), α is the parameter which controls the time that the pavement performance decays to 63.2% of the initial value, and β is the parameter which controls the decay nature of the curve, t is the age of highway.Then Equation(3) can be transformed to the format as Equation (4):

$$\beta \ln t + \beta \ln\alpha^{-1} = \ln[\ln(1 - \frac{PPI}{PPI_0})^{-1}]^{-1} \quad (4)$$

Then we can do variable substitution on Equation(4), we replace lnt with variable x and replace β with variable

a,replace $\beta ln\alpha^{-1}$ with *b*, replace $\ln[\ln(1-\frac{PPI}{PPI_0})^{-1}]^{-1}$ with y, then the Equation(4) can express as $y = ax + b$. According to our historical pavement performance data *PPI* and *t*, we can get the corresponding(*x, y*), so we can estimate the parameters *a* and *b* through the least squares method as follows Equations[5] :

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad (5)$$

$$b = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \quad (6)$$

After we get the values of *a* and *b*, we can estimate the values of α and β according to the equation as follow.

$$\beta = a \quad (7)$$

$$\beta \ln \alpha^{-1} = b \quad (8)$$

then we can use the Equation(3) to estimate the future trend of highway pavement performance.

2.3The model of time series prediction

The factors that influence pavement performance are too many and uncertain[6], so grey theory can be used to predict and analyze the pavement performance. And GM (1,1) is a more commonly used prediction model, with the modeling process as follows[7].

Given an original discrete sequence X_0={$x_0(1),x_0(2),x_0(3),...,x_0(n)$},making a cumulative calculation to generate a new series X_1={$x_1(1),x_1(2),x_1(3),...,x_1(n)$} through the Equation(9),

$$x_1(k)=\sum_{i=1}^{k} x_0(i) \quad (9)$$

then calculating generation sequence of consecutive neighbors of X_1,as Z_1={$z_1(1),z_1(2),...,z_1(n)$},by the Equation(10) as follow.

$$z_1=0.5x_1(k)+0.5x_1(k-1) \quad (10)$$

Then basic equation of the GM (1,1) model and the albinism differential equation are shown as follows:

$$x_0(k) + az_1(k) = b \quad (11)$$

$$\frac{dX_1}{dt} + aX_1 = b \quad (12)$$

where *a, b* are the parameters need to be estimated.We can use the least squares method to estimate the value of *a* and *b* by Equation(13)(14) as follows.

$$\bar{a}=\frac{\sum_{i=2}^{n} z_1(k)\sum_{i=2}^{n} x_0(k)-(n-1)\sum_{k=2}^{n} z_1(k)x_0(k)}{(n-1)\sum_{i=2}^{n} z_1^2(k)-(\sum_{i=2}^{n} z_1(k))^2} \quad (13)$$

$$\bar{b}=\frac{\sum_{i=2}^{n} z_1^2(k)\sum_{i=2}^{n} x_0(k)-\sum_{k=2}^{n} z_1(k)\sum_{k=2}^{n} z_1(k)x_0(k)}{(n-1)\sum_{i=2}^{n} z_1(k)^2-(\sum_{i=2}^{n} z_1(k))^2} \quad (14)$$

Then solve the differential Equation(12), and take $x_1(0) = x_0(1)$, we can obtain the prediction model of X_1:

$$\bar{x}_1(k+1) = (x_0(1) - \frac{b}{a})e^{-ak} + \frac{b}{a} \quad (15)$$

According to the relationship between X_1 and X_0,we can estimate the value of $\overline{x_0}(k+1)$ through $\bar{x}_1(k+1)$ using the formula as follow:

$$\overline{x_0}(k+1) = \overline{x_1}(k+1) - \overline{x_1}(k) \quad (16)$$

3.Practical application of data mining in preventive maintenance

Taking the data set from a common trunk highway section of Hubei Province as an example, the regression model and the GM (1,1) model are used to forecast the performance of the pavement. The pavement performance index from 2011 to 2016 as follows.

Table 1. The pavement performance index of one section

index	2011	2012	2013	2014	2015	2016
PQI	95.11	94.93	93.62	89.73	85.54	82.27
PCI	94.30	91.22	88.97	86.63	84.12	80.93
RQI	96.38	95.29	93.55	90.61	86.86	84.17
RDI	95.03	94.11	93.01	90.45	85.86	82.98

3.1 Pavement performance prediction based on regression model

According to the existed data of pavement performance index of highway,we apply the regression model described in subsection 2.2. Because the application of the four index are similar, we only select the pavement condition index (PCI) as an example to predicting.

Using the PCI value of 2011 as the initial value PPI_0,according to the principle of variable substitution about Equation(4),we can get value pairs(x_i,y_i) as follows:{(0,-1.509),(0.693,-1.207),(1.098,-1.301),(1.386, -0.889),(1.609,-0.742)};then calculate the estimates of a and b according to Equation(5) and Equation(6) as $a\approx$ 0.4668, $b\approx$-1.5227;next,we can estimate the values of α and β according to the relation between α, β and a,b.as a result that $\alpha\approx$26.0985,$\beta\approx$0.4668;finally,the predictive values of PCI index can be calculated according to Equation(3).The prediction results of PCI from 2012 to 2016 are shown in Table 2.

Table 2. Predictive values of PCI based on regression model

year	actual values of PCI	predictive values of PCI	differrences
2012	91.22	91.28	-0.06
2013	88.97	88.87	0.10
2014	86.63	86.30	0.33
2015	84.12	83.86	0.26
2016	80.93	81.62	-0.67

The comparison between the predicted value and the actual value of PCI and variation trend based on regression model are shown in Fig 1.

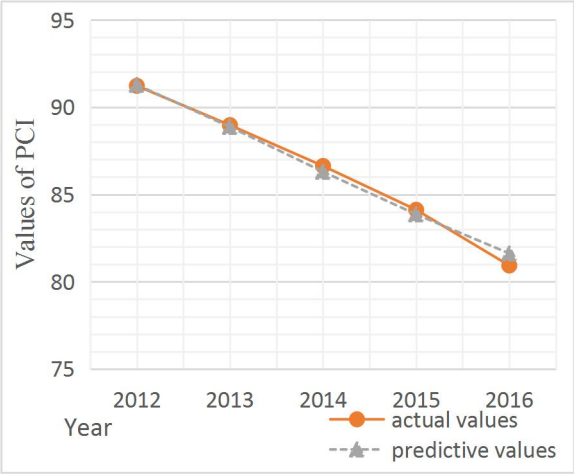

Fig 1. Comparison of regression prediction and actual values

Table 2 and Figure 1 show that the predictive values from regression model is basically consistent with the actual values of pavement condition index, and the max differences do not exceed 1%,futher, the PCI index is declining year by year.According to the current trend，it will be reduced to under 80 in the next year, so we have to take measures to have a preventive maintenance in advance, for the purpose of preventing further deterioration of road performance[8].

3.2Pavement performance prediction based on time series model

In order to ensure the credibility of the prediction results, the time series GM (1,1) model is also used to predict the pavement performance indicators, so that the two predictions can be confirmed mutually. We still used the PCI values in Table 1 as an example to forecast PCI's future tend[9].

According to Table 1, there are original discrete sequences X_0 = {94.30,91.22,88.97,86.63,84.12,80.93}; next,calculating the accumulated sequences of X_0 as a result X_1={94.30,185.52,274.49,361.12,445.24,526.17}; then we can obtain generation sequence of consecutive neighbors of $X1$ as $Z1$={47.15,139.91,230.005,317.805, 403.18,485.705};futher,we can estimate the parameteres of model through Equation(12)(14) as a result a=0.029,b=95.633;finallly,we estimate the predictive values of PCI according to Equation(15)(16) and shown in Table 3.

Table 3.Predictive values and differences based on GM(1,1)

year	actual values of PCI	predictive values of PCI	differrences
2012	91.22	91.51	-0.28
2013	88.97	88.87	0.10
2014	86.63	86.29	0.34
2015	84.12	83.8	0.32
2016	80.93	81.37	-0.44

The comparison between the predicted value and the actual value of PCI and variation trend based on GM(1,1) model are shown in Fig 2.

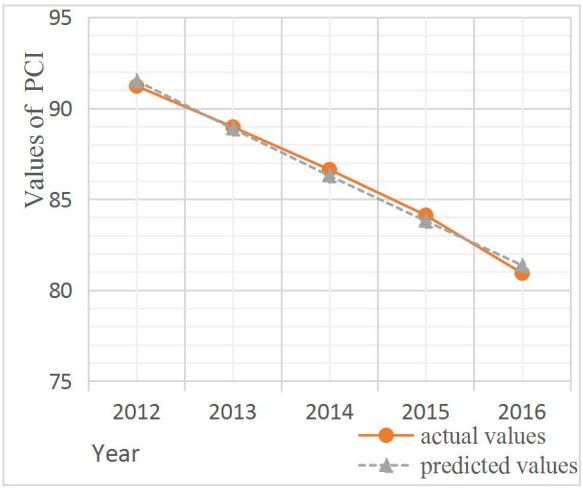

Fig 2. Comparison of GM(1,1) prediction and actual values

It can be seen from Table 2 and Figure 2 that the predictive values from GM(1,1) model is basically consistent with the predictive values from regression model and it also quite similar with the actual value, the max differences do not exceed 1%,futher, the PCI index is declining year by year.According to the current trend，it will be reduced to under 80 in the next year, so we have to take measures to have a preventive maintenance in advance, for the purpose of preventing further deterioration of road performance.we get the same conclusion with regression model.

4. Conclusions

The combination of data mining technology and highway maintenance will bring great surprises to China's highway maintenance and management system, which will change our traditional passive maintenance mode into the period of active preventive maintenance. As a result,we predict the future trend of the highway and the main factors of influencing pavement performance to find out the potential or possible highway distress in advance,so that we can locate the particular disease of highway and be well prepared before the highway pavement perormance reach to a bad status.

The research in this paper shows that the pavement performance predictive values based on the regression model and the time series model is consistent with the actual value and the error is very small. To a certain degree,it reflects the fact that the data mining technology can accurately predict the variation tendency of pavement performance and provide reliable data support for the highway preventive maintenance[10]. In short, Preventive maintenance of highway can greatly improve the efficiency of highway maintenance and save a lot of funds used in highway maintenance.

Acknowledgements

This work has been supported by Project of Hubei Provincial Highway Bureau under Grant Numbers 20141h0288.

References

1. S.L. Hong,Y.H. Zhuang,K. Li,*Data mining technology and engineering practice*. Beijing: China Machine Press,105-136,(2014).
2. *Ministry of Transport of the People's Republic of China*,technical specifications of Highway maintenance (JTG H10-2009).
3. F.T. Liu,K.M. Ting,Z.H. Zhou.*Isolation Forest*.Eighth IEEE. ICDM.(2008).
4. L.J. Sun,*Asphalt pavement structure behavior theory*,People's Publishing Press,380-472,(2005).
5. X.J. Liu,Y.J. Zheng.*Study on multi-index forecasting of asphalt pavement performance based on gray theory*.HighWay,**04**: 233-237, (2012).
6. G. Wang.*The applications of data mining in urban highway traffic*. Chongqing: Chongqing University, (2016).
7. T. Cheng,*The application of data mining in traffic accident*. Harbin Institute of Technology ,(2009).
8. M.G. Tan,*research and application of data mining in Urban highway traffic*. Fudan University, (2010).
9. Q. Ma,*the research and application on Association rules algorithm*, Taiyuan University of Technology,(2007).
10. B.B. Meng,*The application of data mining technology in intelligent traffic detection system*,Wuhan University of Technology ,(2008).