Journal of Marine Science and Technology

Volume 23 | Issue 3 Article 5

APPLICATION OF A DECISION TREE METHOD WITH A SPATIOTEMPORAL OBJECT DATABASE FOR PAVEMENT MAINTENANCE AND MANAGEMENT

Chien-Ta Chen
Department of Civil Engineering, National Central University, Taoyuan County, Taiwan, R.O.C., ewap1114@gmail.com

Chia-Tse Hung
Department of Civil Engineering, National Central University, Taoyuan County, Taiwan, R.O.C.

Jyh-Dong Lin
Department of Civil Engineering, National Central University, Taoyuan County, Taiwan, R.O.C.

Po-Hsun Sung
Department of Civil Engineering, National Central University, Taoyuan County, Taiwan, R.O.C.

Recommended Citation

Chen, Chien-Ta; Hung, Chia-Tse; Lin, Jyh-Dong; and Sung, Po-Hsun (2015) "APPLICATION OF A DECISION TREE METHOD WITH A SPATIOTEMPORAL OBJECT DATABASE FOR PAVEMENT MAINTENANCE AND MANAGEMENT," *Journal of Marine Science and Technology*: Vol. 23: Iss. 3, Article 5.
DOI: 10.6119/JMST-014-0327-5
Available at: https://jmstt.ntou.edu.tw/journal/vol23/iss3/5

Journal of Marine Science and Technology, Vol. 23, No. 3, pp. 302-307 (2015)
DOI: 10.6119/JMST-014-0327-5

APPLICATION OF A DECISION TREE METHOD WITH A SPATIOTEMPORAL OBJECT DATABASE FOR PAVEMENT MAINTENANCE AND MANAGEMENT

Chien-Ta Chen, Chia-Tse Hung, Jyh-Dong Lin, and Po-Hsun Sung

Key words: pavement management system, data mining, decision tree, SVM, pavement maintenance, management.

ABSTRACT

In recent years, pavement engineering has gradually shifted from new construction work to pavement maintenance and management. Since pavement engineers of the Taipei City Government change frequently, objective data is used to make decisions pertaining to road maintenance in Taipei City instead of relying on engineers' experience. In this study, three methods (ID3, C5.0 and SVM) have been chosen to test for use in the decision-making process related to road maintenance of Taipei City. The results show the correct classification rates of the decision trees are 76.67% (C5.0), 64.52% (ID3), and 66.67% (SVM). The decision tree of C5.0 was compared with engineer's experience, with 70% conformity between these two methods. Although the accuracy of the classification could be further improved, the decision tree of C5.0 could be used for pavement maintenance instead of human judgment.

I. INTRODUCTION

Generally, people need the smooth road without any distress (e.g., cracks, distress, sinking). Road maintenance became a public concern and road quality can be affected by the government's administrative efficiency. In order to promote road maintenance projects, this study sought to integrate road management and maintenance information on the roads in order to help the authorities improve traffic efficiency and road service quality.

A pavement management system was first developed 15 years ago in Taiwan, and since then the execution of budget priority and decision-making has remained unchanged. However, the authority's management policies are restricted by human resources and budget. In this study a database was built, including a road roughness index and raw data on pavement distress. The first step was the classification of all data, and was carried out through software programs via data mining and knowledge analysis rules. These results include the accuracy of the data in the pavement management system and the implementation of system management operations.

The collection of pavement information in an automatic or artificial way is indeed for pavement maintenance and rehabilitation. However, the amount of data in the pavement system database is too large for useful information to be located quickly, so data mining technology is particularly useful for maintenance and rehabilitation (Quinlan, 1993; Clair et al., 1998).

Amado (2000) used data mining technology to analyze pavement data, which had been collected from 1995 to 1999 to the Missouri State Department of Transportation (MoDOT). Amado's purpose was to predict the Pavement Serviceability Rating (PSR) in the future with former database, which contains 28,231 pen data and 49 data fields in the study.

Nassar (2007) used data mining to analyze the pavement data for a pavement-project management system in the Illinois Department of Transportation. The data contained 21 types of information pertaining to normal information, projects, and traffic control. The study was a success and established nine rules about pavement maintenance.

Khattak and Airashidi (2013) used Long-Term Pavement Performance (LTPP) distress data to evaluate actual pavement performances of various rehabilitation strategies for flexible pavements, and their study indicated the significance of the effective use of the LTPP distress data and provided a robust technique to evaluate the performance of various rehabilitation actions, thus allowed the state highway agencies to choose the best rehabilitation alternatives based on the actual pavement performance.

Ker et al. (2008) tried to find a mechanistic-empirical model to include several variables such as pavement age, yearly

Paper submitted 12/23/13; revised 01/30/14; accepted 03/27/14. *Author for correspondence: Chien-Ta Chen (e-mail: ewap1114@gmail.com).*
Department of Civil Engineering, National Central University, Taoyuan County, Taiwan, R.O.C.

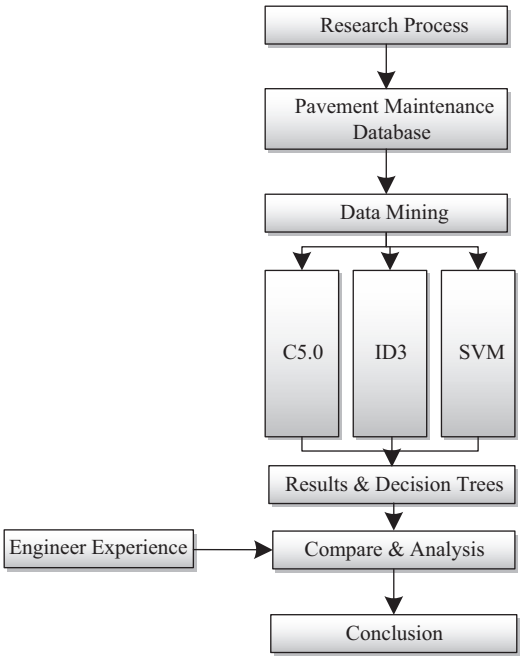

Fig. 1. The process flow of this research.

ESALs, bearing stress, annual precipitation, base type, subgrade type, annual temperature range, joint spacing, modulus of subgrade reaction, and freeze-thaw cycle for the prediction of joint faulting. The goodness of fit was further examined through the significant testing and various sensitivity analyses of pertinent explanatory parameters. The tentatively proposed predictive models appeared to reasonably agree with the pavement performance data and their further enhancements were possible and recommended.

The aforementioned studies showed that data mining could be used to quickly find useful information from a pavement system database. Therefore, to get a decision for pavement maintenance by International Roughness Index (IRI), crack, and distress was an attempt in this study. The process flow of this research project is shown in Fig. 1.

II. EXPERIMENTAL PROGRAM

Several studies proved that soft computing techniques may be used as tools in solving problems where conventional approaches fail or poorly perform (Mirzahosseini et al., 2011).

Soft computing was used in this study including evolutionary algorithms and combined all of their different methods with decision trees ID3, C5.0, and SVM. Soft computing techniques have widespread applications and many important tools used for approximating a nonlinear relationship between model inputs and corresponding outputs.

1. ID3 Decision Tree

The ID3 algorithm begins with the original set S as the root node, which consist of road data. It uses statistical property call information gain to select which attribute to test at each node in the tree. Each iteration of the algorithm works through every unused attribute of the set S and calculates the entropy H(s) (or information gain IG(A)) of that attribute. The algorithm selects the attribute that has the smallest entropy (or largest information gain) value. The set S is then split by the selected attribute (e.g. age < 50, $50 \leq$ age < 100, age $\geqq 100$) to produce subsets of the data.

The algorithm continues to recur on each subset, considering unselected attributes. Recursion on a subset may stop in one of these cases.

Every element in the subset belongs to the same class (+ or -), then the node is turned into a leaf and labelled with the class of the examples.

There are no more attributes to be selected, but the examples still do not belong to the same class (some are + and some are -), then the node is turned into a leaf and labelled with the most common class of the examples in the subset.

There are no examples in the subset. This happens when no example in the parent set is found to be matching a specific value of the selected attribute; for example, if there were no examples with age $\geqq 100$. In this case, a leaf is created and labelled with the most common class of the examples in the parent set.

The following is the list of steps followed in the ID3 Algorithm:

Step 1. Add the training samples of raw data into the roots of the decision tree.
Step 2. Divide the raw data into two parts: one for training data set, and the other for test of data set.
Step 3. Use data to build decision trees in each internal node based on the information theory to evaluate the choice of which properties continue to be based on branches.
Step 4. Use test data to carry out decision tree pruning. Each classification tree is trimmed (pruned) to only one node in order to enhance the predictive power and speed.

The above steps 1~4 are continuously repeated until all the new internal nodes are leaf nodes.

Throughout the algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute on which the data was split and terminal nodes representing the class label of the final subset of this branch.

H(S) measures the amount of uncertainty in the (data) set S (i.e. entropy characterizes the (data) set S).

$$H(S) = -\Sigma_{x \in X} p(x) log_2 p(x) \quad (1)$$

where,

S - The current (data) set for which entropy is being calculated (changes with every iteration of the ID3 algorithm)
X - Set of classes in S
$p(x)$ - The proportion of the number of elements in class x to the number of elements in set S

When H(S) = 0, the set S is perfectly classified (i.e. all elements in S are of the same class).

In ID3, the entropy was calculated for each remaining attribute. The attribute with the smallest entropy is used to split the set S in this iteration. Information gain, IG(A), is the measure of the difference in entropy before and after the set S is split by an attribute A. In other words, it is the measure of how much uncertainty in S reduced after splitting set S by attribute A.

$$\mathrm{IG(A)} = \mathrm{H(S)} - \Sigma_{t\in T} p(t)H(t) \tag{2}$$

where,

H(S) - Entropy of set S

T - The subsets created from splitting set S by attribute A such that $\mathrm{S} = \bigcup_{t\in T} t$

p(t) - The proportion of the number of elements in t to the number of elements in set S

H(t) - Entropy of subset t

In ID3, information gain can be calculated (instead of entropy) for each remaining attribute. The attribute with the largest information gain is used to split the set S in this iteration.

2. C5.0 Decision Tree

C5.0 builds decision trees from the training data set in the same way with ID3 by using the concept of information entropy. The training data was a set of $\mathrm{S} = \mathrm{S}_1, \mathrm{S}_2, \ldots$ and already classified samples. Each sample S_i consists of a p-dimensional vector ($x_{1,i}$, $x_{2,i}$, ..., $x_{p,i}$), where the x_j represents attributes or features of the sample, as well as the class in which S_i falls.

The decision tree model is used to produce the rules, usually by internal node mapping of some test attributes. Every branch has a value, and every leaf node maps a Boolean function. This methodology creates a decision tree model including the following steps: 1. problem characteristics and collection; 2. dealing with the data; 3. finding association rules; 4. creating the decision tree. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. In this study, C5.0 methods were used to create a model based on the ID3 (Interactive Dichotomize) algorithm first proposed by Quinlan (1993). This method is based on information theory. The attribute entropy is found by using the following equation:

$$E_a = -\sum_{i=1}^{m}\sum_{j=1}^{n} P_{ij} \log P_{ij} \tag{3}$$

The ID3 finds the minimum entropy attribute as the next node of the decision tree; it is a so-called 'Greedy Algorithm'. After the decision tree C5.0 was proposed by Quinlan, the Boosting algorithm was used to improve model accuracy. This method has several advantages:

- It is accurate when using C5.0 to deal with missing values in the input field.
- It does not take a lot of time to carry out estimations.
- It is easier to use compared to other methods.
- It uses accurate technology.

In this study, the following process and equations are used:

Step 1. Calculate the entropy and measurement degree

$$E(S) = -\sum_{c=1}^{m} P(Sc) \times \log_2 P(Sc)\,; \tag{4}$$

Step 2. Calculate the information gain

$$Gain(S,V) = E(S) - \sum_{V\in Values(v)} \frac{|Sv|}{|S|} \times E(Sv)\,; \tag{5}$$

Step 3. Use SplitInfor to calculate every variable

$$SplitInfor(S,V) = \sum_{i=1}^{m} -\frac{|Si|}{|S|} \times \log_2 \frac{|Si|}{|S|}\,; \tag{6}$$

Step 4. Use the gain division SplitInfor to obtain the gain value

$$GainRatio(S,V) = \frac{Gain(S,V)}{SplitInfor(S,V)} \tag{7}$$

The C5.0 algorithm was utilized to create a model in hope that the database could be used to obtain a good gain value; but, when the database must be input into a larger dataset, useful information can be obtained through this model.

At each node of the tree, C5.0 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C5.0 algorithm then recuses on the smaller sub lists.

3. Support Vector Machines (SVM)

SVM is a machine-learning algorithm based on statistical learning theory (Cortes and Vapnik, 1995). The main idea of SVM is to transform the input space into a high-dimensional space by a nonlinear transformation defined by an inner product function. SVM calculation takes the form of a convex quadratic optimization problem, ensuring that the solution is optimal. The SVM has a good ability to generalize and resolve some practical problems such as small samples, nonlinearity, and high-dimensional input spaces (Smola and Scholkopf, 2004; Maalouf et al., 2008).

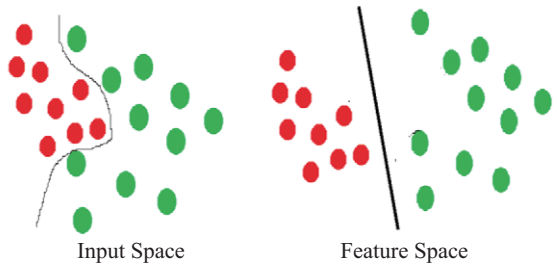

Fig. 2. The basic theory behind Support Vector Machines

For this type of SVM, training involves the minimization of the error function:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \tag{8}$$

subject to the constraints:

$$y_i(w^T \varnothing(x_i) + b) \geq 1 - \xi_i, \text{ and } \xi_i \geq 0, 1, \ldots, N$$

Here, C is the capacity constant, w is the vector of coefficients, b is a constant, and ξ_i represents the parameters for handling non-separable data (inputs). The index i labels the N training cases. Note that $y \in \pm 1$ represents the class labels and x_i represents the independent variables. The kernel $\varnothing$ is used to transform data from the input (independent) to the feature space. It should be noted that the larger the value of C, the more the error is penalized. Thus, C should be chosen with care to avoid over-fitting.

Fig. 2 shows an illustration of the basic theory behind Support Vector Machines. The original objects (left side of the schematic) are mapped (i.e. rearranged) using a set of mathematical functions known as kernels. The process of rearranging the objects is known as mapping (transformation). Note that in this new setting, the mapped objects (right side of the schematic) are linearly separable; thus, instead of constructing the complex curve (left schematic), all that needs to be done is to find an optimal line that can separate the GREEN and the RED objects.

III. DATA AND DATABASE

Every field of database was collected to understand the association of every field. Fig. 3 shows the flowchart of the database association of the Taipei pavement management system. The databases include information with regards to pavement such as: basic information, maintenance data, and statistics (IRI, distress, crack, pavement contract, etc.). Different people have different competencies to create, delete, and revise the data. This step can be useful to realize what is in this database and find more information from it.

By making use of the data collected from the years 2010 to 2012, an attempt was made to classify ten roads in Taipei City

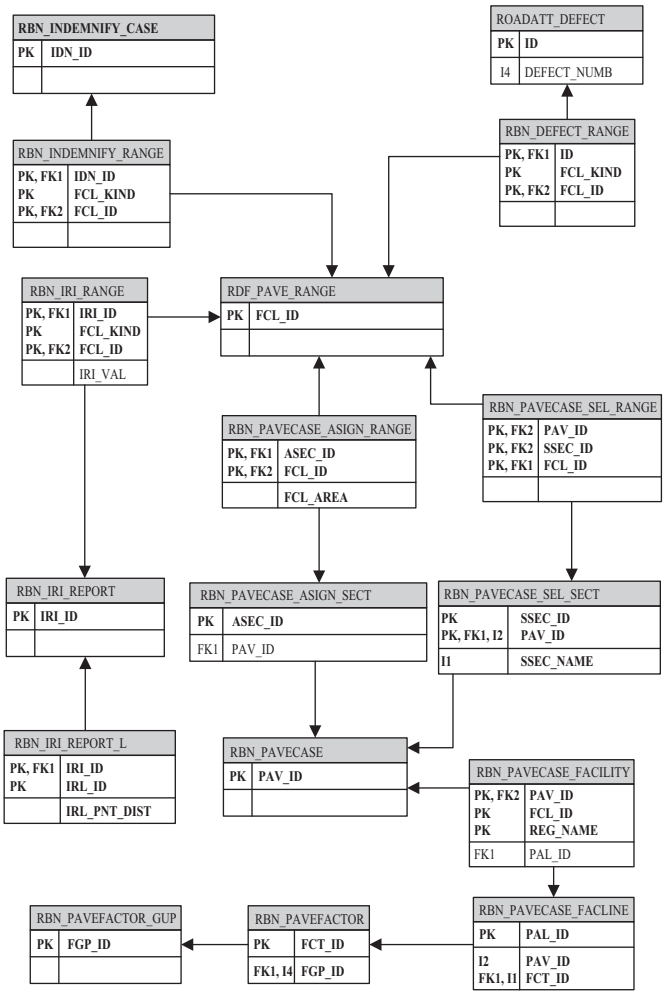

Fig. 3. Database association.

using the C5.0 decision tree, and compared the results with the decision of engineer's experience. The C5.0 algorithm has been used to generate decision trees and association rules. It is known that pavement database A (provided by the ITRE at NCDOT) covers four counties in the state of North Carolina in the US.

This dataset was used to test the proposed method. The C5.0 algorithm was also used for a Taiwan pavement database. Through this method, it was sought to create useful decision rules that can be used to identify the relationship between pavement rehabilitation and the occurrence of various road distress conditions.

It was hoped that this could help units concerned with maintenance and management in the future. The collection of raw data includes information from 2010 to 2012 which is compared as shown in Table 1, where the average data falls can be found. In actuality, road maintenance was carried out by people and suppliers. For the reasons outlined above, the raw data in the database was checked. Certainly, one must deal with complete information so as to avoid the creation of incomplete situations.

Table 1. 2010-2012 raw data.

Road's NO	Pavement Situation								
	Distress			Cracks			IRI		
	2010	2011	2012	2010	2011	2012	2010	2011	2012
Road 1	12	8	13	25	20	23	4.666	4.524	4.655
Road 2	10	7	5	22	12	15	4.64	4.241	4.929
Road 3	14	10	10	17	9	6	4.197	4.192	4.486
Road 4	19	17	7	2	0	1	5.544	5.355	5.833
Road 5	64	40	55	50	32	40	6.674	5.235	6.963
Road 6	10	4	6	10	10	12	5.535	4.201	5.824
Road 7	16	8	9	13	8	10	5.732	4.198	6.021
Road 8	15	7	12	10	6	5	4.695	5.121	5.41
Road 9	15	7	8	10	6	6	8.445	5.201	5.49
Road 10	4	2	4	3	0	1	9.379	5.102	5.379
Average	17.90	11.00	12.90	16.20	10.30	11.90	5.95	4.74	5.50

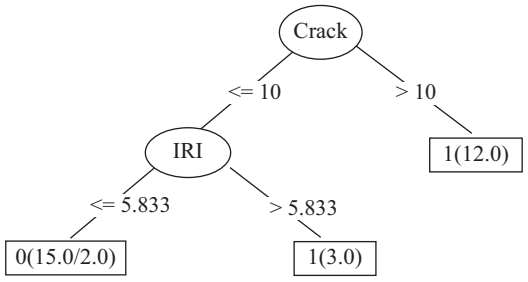

Fig. 4. Decision tree model (C5.0).

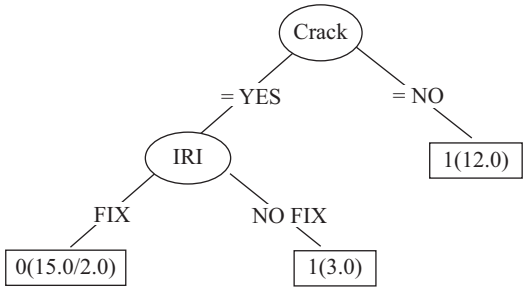

Fig. 5. Decision tree model (ID3).

IV. ANALYSIS AND COMPARISON

When this model was created, not all of the data were used. Therefore, it was classified into two types of samples. The real data compared with the database shows that the 2012 results were better than the results from 2010 and 2011. This situation makes it possible, for the area was the same, which results in double input data. Another thing to keep in mind is that data taken as averages, which results in overvaluing for the pavement situation. Roads of different widths must be classified separately as the extent of the damage will not be the same. The classification tree models of C5.0 and ID3 in this study are shown in Fig. 4 and Fig. 5.

Decision tree model analysis was used to analyze the original data. The results are shown in Table 3, which shows correct classification rates of about 76.67% with C5.0, 66.67% with SVM, and 64.52% with ID3. The results show that C5.0 become the best way to classify the data, and ID3 become the worst.

Table 2. Kappa statistic classification.

Kappa statistic	Accuracy
<0	Less than chance agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement

Note. From "Understanding interobserver agreement," by Viera and Garrett, 2005, The kappa statistic Family Medicine Journal, 37(5), 360-363.

Table 3. Decision trees comparison.

Project	Algorithms Results		
Results	C5.0	ID3	SVM
Correctly Classified Instances	23	19	20
Incorrectly Classified Instances	7	11	10
Kappa Statistic	0.5291	0.1263	0.2534
Correctly Classified Rate	76.67	64.52	66.67
Incorrectly Classified Rate	23.33	35.48	33.33

Cohen's kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories. The first mention of a kappa-like statistic is attributed to Galton and Smeeton (Galton, 1892; Smeeton, 1985).

The equation for k is:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{9}$$

Where, $Pr(a)$ is the relative observed agreement among raters and $Pr(e)$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer by randomly saying each category. If the raters are in complete agreement, then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (as defined by $Pr(e)$), $\kappa = 0$.

Viera and Garrett found that the correct degree of κ values is classified as follows in Table 2. From Table 3, it can be seen that the kappa statistic with C5.0 is 0.5291, which falls in the range of moderate agreement. However, the kappa statistics of the results shown in ID3 and SVM are 0.1263 and 0.2534, respectively, which fall in the range of slight agreement and fair agreement, respectively (Viera and Garrett, 2005).

An attempt was made to classify ten roads in Taipei city with the data from 2010 to 2012 using the C5.0 decision tree, and compare the results with the decision of engineer's

Table 4. Engineer's experience & C5.0 decision comparison.

Maintenance recommendations	Engineer's experience decision (Road Number)	C5.0 decision (Road Number)
2010	No. 1, No. 4, No. 5, No. 9, No. 10	No. 1, No. 2, No. 3, No. 5, No. 7, No. 9, No. 10
2011	No. 5	No. 1, No. 2, No. 5
2012	No. 5, No. 7	No. 1, No. 2, No. 5, No. 6, No. 7

Table 5. Confusion matrix.

Engineer's experience decision	Decision with C5.0	
	YES	NO
YES	7	1
NO	8	14

experience. The results are shown in Table 4, which also shows the maintenance recommendations for the three years. One can obtain a confusion matrix from the decision tree, which will show the correction of the maintenance recommendations with the C5.0 decision tree. The results of the confusion matrix show 70% correct classification rate between engineer's experience and artificial intelligence. The results of confusion matrix are shown in Table 5.

These results show that a decision of M&R could probably be made by decision tree of C5.0, which can be instead of human judgment.

V. CONCLUSIONS

The final goal in this PMS study was to integrate all data and search for useful information to combine spatial-temporal databases containing multinomial and complete data, and the use of an objective method to analyze the data to give maintenance recommendations for engineers.

The conclusions are summarized as follows:

Three methods of road maintenance classification were used, and the most suitable method is the C5.0 decision tree, which can classify the data with 76.67% accuracy. The classification rates of the ID3 algorithm and the SVM algorithm were not suitable for the road maintenance decision classification of the current road maintenance information. Their incorrect classification rates were higher than 30%, because the current description of the numerical data volume regarding road condition was insufficient; therefore, C5.0 was used as the most suitable prediction method.

The decision of C5.0 was compared to engineer's experience, and the results show that there is 70% accuracy with C5.0.

Temporal attribute data through spatial management were used to make the records. The multi-granularity method was normally used to discuss the following data base:

a. Whether the information in the database was correct.
b. Road pavement is never homogenized. Pavement was not uniformly consistent.
c. Milling process cannot avoid the human factor.

Finally, pavement maintenance management was usually carried out based on experience. A method to extract useful information from a database (including GIS information, pavement distress, PCI, IRI, etc.) was developed in this study. Certainly, the results will also improve the accuracy of the decision tree through various methods involving theoretical approaches, practical operations, and data collection.

REFERENCES

Amado, V. (2000). Expanding the Use of Pavement Management Data. Department of Civil and Environmental Engineering, University of Missouri-Columbia, MTC Transportation Scholars Conference, Ames, Iowa.

Clair, C., C. Liu and N. Pissinou (1998). Attribute weighting: a method of applying domain knowledge in the decision tree process. The Seventh International Conference on Information and Knowledge Management, 259-266.

Cortes, C. and V. Vapnik (1995). Support vector networks. Machine Learning 20, 1-25.

Galton, F. (1892). Finger Prints London: Macmillan and co.

Ker, H. W., Y. H. Lee and C. H. Lin (2008). Prediction models for transverse cracking of jointed concrete pavements: Development with long-term pavement performance database. Transportation Research Record 2068, Journal of the Transportation Research Board, 20-31.

Khattak, J. and M. Airashidi (2013). Performance of preventive maintenance treatments of flexible pavements. International Journal of Pavement Research & Technology 6(3), 184-196.

Maalouf, M., N. Khoury and T. B. Trafalis (2008). Support vector regression to predict asphalt mix performance. International Journal for Numerical and Analytical Methods in Geomechanics 2(16), 1989-1996.

Mirzahosseini, M. R., A. Aghaeifor, A. H. Havi and A. H. Gandomi (2011). Permanent deformation analysis of asphalt mixtures using soft computing techniques. Expert Systems with Applications 38(5), 6081-6100.

Nassar, K. (2007). Application of data-mining to state transportation agencies' projects databases. Journal of Information Technology in Construction ITcon 12, 139-149.

Quinlan, J. R. (1993). C4.5: Programs for machine learning, San Mateo: Morgan Kaufmann.

Smeeton, N. C. (1985). Early History of the Kappa Statistic. Biometrics 41, 795.

Smola, A. J. and B. Schölkopf (2004). A tutorial on support vector regression. Statistics and Computing 14, 199-222.

Viera, A. J. and Garrett J. M. (2005). Understanding interobserver agreement: The kappa statistic Family Medicine Journal 37(5), 360-363.