Integration of GIS and Data Mining Technology to Enhance the Pavement Management Decision Making

Guoqing Zhou1; Linbing Wang2; Dong Wang3; and Scott Reichle4

Abstract: This paper presents a research effort undertaken to explore the applicability of data mining and knowledge discovery (DMKD) in combination with Geographic Information System (GIS) technology to pavement management to better decide maintenance strategies, set rehabilitation priorities, and make investment decisions. The main objective of the research is to utilize data mining techniques to find pertinent information hidden within the pavement database. Mining algorithm C5.0, including decision trees and association rules, has been used in this analysis. The selected rules have been used to predict the maintenance and rehabilitation strategy of road segments. A pavement database covering four counties within the state of North Carolina, which was provided by North Carolina DOT (NCDOT), has been used to test this method. A comparison was conducted in this paper for the decisions related to a rehabilitation strategy proposed by the NCDOT to the proposed methodology presented in this paper. From the experimental results, it was found that the rehabilitation strategy derived by this paper is different from that proposed by the NCDOT. After combining with the AIRA Data Mining method, seven final rules are defined. Using these final rules, the maps of several pavement rehabilitation strategies are created. When their numbers and locations are compared with ones made by engineers at the Institute for Transportation Research and Education (ITRE) at North Carolina State University, it has been found that error for the number and the location are various for the different rehabilitation strategies. With the pilot experiment in the project, it can be concluded: (1) use of the DMKD method for the decision of road maintenance and rehabilitation can greatly increase the speed of decision making, thus largely saving time and money, and shortening the project period; (2) the DMKD technology can make consistent decisions about road maintenance and rehabilitation if the road conditions are similar, i.e., interference from human factors is less significant; (3) integration of the DMKD and GIS technologies provides a pavement management system with the capabilities to graphically display treatment decisions against distresses; and (4) the decisions related to pavement rehabilitation made by the DMKD technology is not completely consistent with that made by ITRE, thereby, the postprocessing for verification and refinement is necessary.

DOI: 10.1061/(ASCE)TE.1943-5436.0000992

CE Database subject headings: Pavement management; Data collection; Geographic information systems; Decision making.

Author keywords: Pavement management; Data mining; GIS; Pavement distress.

Introduction

A pavement management system (PMS) is a valuable tool and one of the critical elements of the highway transportation infrastructure (Tsai et al. 2004; Kulkarni and Miller 2003). The earliest PMS concept can be traced back to the 1960s. With rapid increase of advanced information technology, many investigators have successfully integrated the Geographic Information System (GIS) into PMS for storing, retrieving, analyzing, and reporting information needed to support pavement-related decision making. Such an integration system is thus called G-PMS (Lee et al. 1996). The main characteristic of a G-PMS is that it links data/information to its geographical location (e.g., latitude/longitude or state plane coordinates) instead of the milepost or reference-point system traditionally used in transportation. Moreover, the GIS can describe and analyze the topological relationship of the real world using the topological data structure and model (Goulias 2002; Lee et al. 1996). GIS technology is also capable of rapidly retrieving data from a database and can automatically generate customized maps to meet specific needs such as identifying maintenance locations. Therefore, a G-PMS can be enhanced with features and functionality by using a GIS to perform pavement management operations, create maps of pavement condition, provide cost analysis for the recommended maintenance strategies, and long-term pavement budget programming.

With the increasing amount of pavement data collected, updated and exchanged due to deteriorating road conditions, increasing traffic loading, and shrinking funds, many knowledge-based expert systems (KBESs) have been developed to solve various transportation problems (e.g., Abkowitz et al. 1990; Nasser 2007; Spring and Hummer 1995). A comprehensive survey of...
KBESs in transportation is summarized and discussed by Cohn and Harris (1992). However, only a few scholars have investigated applying data mining and knowledge discovery (DMKD) to PMSs. For example, Attok-Okin (2002, 1997) presented application of rough set theory to enhance the decision support of the pavement rehabilitation and maintenance. Prechavarakul and Hadipriono (1995) applied KBES and fuzzy logic for minor rehabilitation projects in Ohio. Wang et al. (2003) discussed the decision-making problem of pavement maintenance and rehabilitation. Leu et al. (2001) investigated the applicability of data mining in the prediction of tunnel support stability using an artificial neural network algorithm. Sarasua and Jia (1995) explored an integration of GIS technology with knowledge discovery and expert system for pavement management. Ferreira et al. (2002) explored the application of probabilistic segment-linked pavement management optimization model. Chan et al. (1994) applied the genetic algorithm for road maintenance planning. Soibelman and Kim (2000) discussed the data preparation process for construction knowledge generation through knowledge discovery in databases, as well as construction knowledge generation and dissemination.

This paper presents a research effort undertaken to explore the applicability of data mining in combination with GIS technology to pavement management to better decide maintenance strategies, set rehabilitation priorities, and make investment decisions. A brief overview of data mining techniques as applied to pavement management is presented in “Data Mining and Inductive Learning.” Data collected from the Institute for Transportation Research and Education (ITRE) of North Carolina State University and pavement condition rating are presented in “Pavement Performance Measures.” Examples using data mining to reveal unknown patterns and rules in the database are presented in “Data Mining-Based Maintenance and Repair (M&R).” Finally, the findings and conclusions are discussed.

Data Mining and Inductive Learning

The goal of applying DMKD technology to a PMS is to discover any pertinent information in the pavement management database by creating a decision tree and/or inducing rules. Some information is stored in an obvious location of the database, and thus can be obtained by traditional database query operation, such as maximum and minimum width of the roads. Other knowledge is hidden in a more remote area of the database (e.g., pavement condition patterns), and thus only can be obtained by intelligent technology, such as data mining. Several data mining techniques have been developed over the last few decades by artificial intelligence community. Generally, the data mining techniques can be categorized in four categories, depending on their functionality: (1) classification; (2) clustering; (3) numeric prediction; and (4) association rules (Michalski 1983).

- **Classification:** this technique is designed essentially to generate predictive models by analyzing an existing database to determine categorical divisions or patterns in the database. Thus, this method focuses on identifying the characteristics of the group or class to which each record belongs in the database;
- **Clustering:** the purpose of this data mining technique is to group or classify items that seem to fall naturally together in the database when there is no preidentified classification or group;
- **Numeric prediction:** this is essentially a classification learning technique with the intended outcome of a numeric value (numeric quantity) rather than a category (discrete class). Thus, the numeric values are used for prediction; and
- **Association rule:** the purpose of this data mining technique is to find useful associations and/or correlation relationships among large sets of data items. Association rules, expressed by in the form of “if-then” statements, shows attribute value conditions that occur frequently together in a given data set. These rules are computed from the data, unlike the if-then logic rules, and thus are probabilistic in nature.

The main differences among the earlier discussed techniques lie in which algorithms and methods will be used to extract knowledge and how the mined knowledge and discovery rules are expressed. The data mining technique used in this paper is an association/inductive learning rule, so that relevant and implicit knowledge, such as road condition distress pattern, can be extracted. Many inductive learning algorithms, which mainly come from the machine learning, have been presented, such as AQ11 (Michalski and Larson 1975), AQ15 (Michalski et al. 1986; Hong et al. 1995) and AQ19 (Kaufman and Michalski 1999), AE1 and AE9 (Hong et al. 1995), concept learning system (CLS) (Hunt 1966), ID3, C4.5, and C5.0 (Quinlan 1979, 1986, 1993), and CN2 (Clark and Niblett 1989), and so on. CLS (Hunt et al. 1966) used a heuristic look ahead method to construct decision trees. Quinlan extended CLS method by using information content in the heuristic function, called ID3. ID3 method adopts a strategy, called “divide and conquer,” and selects classification attributes recursively on the basis of information entropy. Quinlan (1993) further developed his method, called C4.5, which only dealt with strings of constant length. C4.5 not only can create a decision tree, but induce equivalent production rules as well, and deals with multi-class problems with continuous attributes. C5.0 is an upgraded version of C4.5. This method requires the training data, which are usually constructed by several tuples, each of which has several attributes. Note that if the records in the database are taken as tuples and fields as attributes, the C5.0 algorithm is easily realized in a pavement database. Thus, the knowledge discovered by C5.0 algorithm is a group of classification rules and a default class, and furthermore, with each rule there is a confidence value (between 0 and 1). Moreover, the C5.0 method can generate a decision tree and associated decision rule.

**Decision Tree**

The inductive learning C5.0 method is a top-down induction algorithm that is capable of generating a decision tree model to classify data in a database. The decision tree is capable of partitioning the data at each level based on a particular feature of the data. The processes of C5.0 algorithm are: (1) a decision tree is built by choosing an attribute that best separates the classes in the database and (2) a pruning strategy is used to prune a branch when the introduced error is one standard error of the existing errors. The implementation of the C5.0 algorithm requires recursive iteration to the partitions until partitions only have a single class, or the partition becomes too small. With such recursive iteration, data with higher partitioning ability will be toward the root, and iterative classification occurs at the leaves in a decision tree. At each iteration, the data are fed into a filtering method, which reduces the number of features. In our project, the decision tree categorizes the entire subject according to whether or not they are likely to have hypertension.
**Association Rule**

In addition to the decision tree generation, the C5.0 algorithm also induces association rules. An association rule differs from a classification rule because the associate rule can be used to predict any attribute (not just the group or class), and more than one attribute’s value at a time. An associate rule also gives an occurrence relationship among factors. A typical association rule is if-then sentence, i.e.,

```
IF Cause_1, AND/OR Cause_2
 THEN Result (or consequence)
```

This rule means “if Cause_1 and/or Cause_2 are true, and then Result (the association rule) would be generated at cases of x% with a confidence of x%.” With such a rule, a piece of knowledge can be discovered from data set. The rules induced by C5.0 algorithm also provide with support, confidence level, and capture using statistical tests to correct the decision tree model (Chae et al. 2001).

- **Support**: this measures the percentages of the training data that support the rule; i.e., how many data generates the left hand side of the rule. The support is the number of cases in which the rule is found and a pattern is defined by the rule. Once this pattern is identified, it can be used to predict the similar patterns in database;
- **Confidence**: the confidence indicates a probability of a certain rule. It can be used to measure the accuracy of rule; and
- **Capture**: this is used to measure the percentage of records that are correctly captured by this rule. For example, if a rule with capture is close to 100%, it means that all observations with this class are closely clustered.

In this paper, the initial association rules will be induced by using the C5.0 algorithm. The final rules are obtained by postprocessing of the initial rules. The attributes for deductive learning are the same as that in inductive learning except the class label attribute. The final rules are used to identify the occurrence relationship between pavement rehabilitation and various road distress conditions.

**Pavement Performance Measures**

**Data Sources**

In 1983, ITRE of North Carolina State University began working with the Division of Highways of the North Carolina DOT (NCDOT) to develop and implement a PMS for its 60,000 mi of paved state highways. At the request of several municipalities, the NCDOT has made the PMS available for North Carolina municipalities. The ITRE modified the system for municipal streets and has used it in more than 100 municipalities in North and South Carolina. The data sources for this experiment are provided by the ITRE, the following nine common types of distress are considered to assess the necessity for road rehabilitation (Table 1). The rating was determined by visual evaluation/inspection at each section of roadway.

**Distress Rating**

All the analyses and pavement condition evaluations presented in this paper are based on the pavement performance measures. A common acceptable pavement performance measure is pavement condition index (PCI), which was first defined by the U.S. Army. Under the PCI, the pavement condition is related to factors such as structural integrity, structural capacity, roughness, skid resistance, and rate of distress. These factors are quantified in the evaluation worksheet that field inspectors used to assess and express the local pavement condition and severity of damage. Note that inspectors primarily relied on their own judgment to assess the distress conditions. Usually, the PCI is quantified into seven levels, corresponding to from Excellent (over 85) to Failed 0.

Table 1 presents eight types of distresses that are evaluated for asphaltic concrete pavements. The severity of distresses in Table 1 is rated in four categories ranging from very slight to very severe. Extent (or density) is also classified in five categories ranging from few (less than 10%) to throughout (more than 80%). The identification and description of types of distress, severity, and density are as follow, respectively.

- The road conditions of alligator cracking are rated as a percentage of the section that falls under the categories of none, light, moderate, and severe. Percentages are shown as 1=10%, 2=30%, 3=60%, up to 10=100%. The appropriate percentages were placed under none, light, moderate, and severe. These percentages should always add up to 100%;
- The severity levels of distresses for block cracking, transverse cracking, bleeding, rutting, utility cut patching, patching deterioration, and raveling are rated four levels: none (N), light (L), moderate (M), and severe (S), respectively; and
- The severity levels of ride quality are classified as: average (L), slightly rough (M), and rough (S).
Potential Rehabilitation Strategies

Based on the knowledge obtained from ITRE in the field, the ITRE has classified flexible pavement rehabilitation needs into three main categories according to the type of the problem to be corrected: (1) cracking; (2) surface defect problems; and (3) structural problems. These problems can be treated using crack treatment, surface treatment, and nonstructural overlay (one- and two-course overlay), respectively. To select an appropriate treatment for rehabilitation and maintenance, seven potential rehabilitation and maintenance strategies have been proposed by the NCDOT (see Table 2). Which treatment strategies will be carried out for a pavement segment is traditionally determined by the ITRE who comprehensively evaluate all types of distresses. This paper seeks to replace this process of decision making for rehabilitation strategies with data mining technology. Thus, one of the eight potential rehabilitation strategies will be determined based upon the association rule in data mining, which will be generated by the data mining technology.

Table 1. Nine Common Types of Distresses to Be Used in This Study

<table>
<thead>
<tr>
<th>Distress</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alligator cracking (four types of rates are given)</td>
<td>Alligator none (AN) Alligator light (AL) Alligator moderate (AM) Alligator severe (AS)</td>
</tr>
<tr>
<td>Block/transverse cracking (BK)</td>
<td>This indicates the overall condition of the section as follows: • N—none; • L—light; • M—moderate; and • S—severe</td>
</tr>
<tr>
<td>Reflective cracking (RF)</td>
<td>The same rating as BK’s</td>
</tr>
<tr>
<td>Rutting (RT)</td>
<td>The same rating as BK’s</td>
</tr>
<tr>
<td>Raveling (RV)</td>
<td>The same rating as BK’s</td>
</tr>
<tr>
<td>Bleeding (BL)</td>
<td>The same rating as BK’s</td>
</tr>
<tr>
<td>Patching (PA)</td>
<td>The same rating as BK’s</td>
</tr>
<tr>
<td>Utility cut patching</td>
<td>The condition is designated as follows: • L—average; • M—slightly rough; and • S—rough</td>
</tr>
<tr>
<td>Ride quality (RQ)</td>
<td></td>
</tr>
</tbody>
</table>
Data Mining-Based Maintenance and Repair (M&R)

Decision Trees of M&R

Implementation of data mining-based maintenance and rehabilitation usually involves the following five basic steps: (1) problem identification; (2) knowledge acquisition; (3) knowledge representation; (4) implementation; and (5) validation and extension. The first step of problem identification has already been discussed, which is to reveal and extract the relevant information hidden in the pavement database to predict potential pavement conditions and plan the rehabilitation. As to the second step, it has been obtained a significant amount of knowledge from the ITRE in the field, such as distress ratings, PCI ratings, and potential rehabilitation strategies. As for the third step, the decision tree and induced rule to be derived by the C5.0 algorithm constitute the knowledge representation. The theory behind the decision tree and inductive rules has already been described earlier (see Data Mining and Inductive Learning, supra.). The experimental results will now be presented in this section, which constitutes the fourth step of implementation. Finally, Step 5, validation, will be discussed next as well.

Experiments

The CTree for Excel tool (Saha 2003) was used to create the decision tree. This tool is based on the C5.0 algorithm, which lets user build a tree-based classification model. The classification tree can generate the association rules. The basic steps are as follows.

Step 1: Load Pavement Database. It is first loaded the pavement database into the data worksheet, in which the observations were placed in rows and the variables were placed in columns. In addition, for each column in the data worksheet, it is chosen an appropriate type, i.e., Omit, Class, Cont, and Cat. The implications of these types are:
• Omit to drop a column from model;
• Cat to treat a column as categorical predictor;
• Cont to treat a column as continuous predictor; and
• Output to treat a column as class variable.

Because the type, Cont, in CTree only recognizes the digital number, while the distresses in Table 1 were measured by letter, N, L, M, and S, these letters have to be quantified into 100, 75, 50, and 25. In the CTree tool, the maximum predictor variables are 50, and only a class variable is allowed. Application will treat the class variable as categorical, and each of categorical predictors (including class variable) is limited to 20. After the data are uploaded into software, the class variables and data do not have blank rows and blank columns, which are treated as missing values in the CTree tool.

Step 2: Fill-Up Model Inputs. In addition, the tool requires inputting some parameters to optimize the process of decision tree generation. These parameters include:

1. Adjust for number categories of a categorical predictor: while growing the tree, child nodes are created by splitting parent nodes. Which predictor is used for this split is decided by certain criterion. Because this criterion has an inherent bias toward choosing predictors with more categories, input of an adjustment factor will be used to adjust this bias;
2. Minimum node size criterion: while growing the tree, the decision as to whether to stop splitting a node and declare the node as a leaf node will be determined by some criteria which need to be chosen. These criteria are:
   • Minimum node size: a valid minimum node size is between 0 and 100;
   • Maximum purity: an effective value is between 0 and 100. The higher the value of this, the larger will be the tree. Stop splitting a node if its purity is 95% or more (e.g., purity is 90% means) Also, stop splitting a node if number of records in that node is 1% or less of total number of records; and
   • Maximum depth: a valid maximum depth is greater than 1 and less than 20. The higher the value, the larger will be the tree. Stop splitting a node if its depth is 6 or more (depth of root node is 1); any node’s depth is that its parent’s depth +1.
3. Pruning option: this option allows us to determine whether or not to prune the tree when tree is growing, which can help us to study the effect of pruning; and
4. Training and test data: in this research, a subset of data are used to build the model and the rest to study the performance of the model. Also, the tool is required to randomly select the test set at a ratio of 10%.

Experimental Results

With the aforementioned data input, a decision tree is generated. The parameters of the decision tree are listed in Table 3. As seen from Table 3, the total number of nodes is 72, in which the number of leaf nodes is 37. The details on each node can be obtained from the NodeView sheet, which shows the class distribution on each node. Moreover, the decision tree consists of 20 levels. In this test, the number of training observations achieved is 555, the number of predictors is 9. The percentage of misclassified data reaches 61.2% for training data and 60.0% for test data. The number for different types of rehabilitation strategies, including nothing, crack pouring (CP), full-depth patch (FDP), 1-in. plant mix resurfacing (PM1), 2-in. plant mix resurfacing (PM2), skin patch (SKP), and short overlay (SO) in the study area is calculated for training data and test data (Table 4).

In addition, the C5.0 algorithm also generates rules. These associated rules discovered rehabilitation strategies that will be carried out if the conditions are met (Fig. 2). Since only seven rehabilitation strategies in the study area were suggested by the ITRE at the NCDOT, the generated rules by the C5.0 algorithms are clustered manually because of the fact too many rules were generated by the CTree software tool. The method of clustering

<table>
<thead>
<tr>
<th>ID</th>
<th>Rehabilitation strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Nothing</td>
</tr>
<tr>
<td>1</td>
<td>Crack pouring (CP)</td>
</tr>
<tr>
<td>2</td>
<td>Full-depth patch (FDP)</td>
</tr>
<tr>
<td>3</td>
<td>1-in. plant mix resurfacing (PM1)</td>
</tr>
<tr>
<td>4</td>
<td>2-in. plant mix resurfacing (PM2)</td>
</tr>
<tr>
<td>5</td>
<td>Skin patch (SKP)</td>
</tr>
<tr>
<td>6</td>
<td>Short overlay (SO)</td>
</tr>
</tbody>
</table>

Table 2. Potential Rehabilitation Strategies Proposed by Engineers of NCDOT

<table>
<thead>
<tr>
<th>Tree information</th>
<th>Percentage for misclassified data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of nodes</td>
<td>72</td>
</tr>
<tr>
<td>Number of leaf nodes</td>
<td>37</td>
</tr>
<tr>
<td>Number of levels</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 3. Decision Tree Model
the rules are as follows: merging the similar rules in each leaf and node together. As a result, their support and confidence and capture are merged as well. Finally, seven rules, associated their SUPPORT and confidence and capture, are listed in Table 5. As observed from Table 5, The confidence for rehabilitation strategies, CP and PM2 are 100%, and support for rehabilitation strategies are 100% as well. in addition, the support, confidence, and capture for rehabilitation, FDP, are lowest. This means that the decision made for FDP rehabilitation has a low reliability.

**M&R Rules Verification**

The aforementioned generated decision tree theoretically organizes the “hidden” knowledge in a logical order. A determination as to whether the tree is used to provide a useful methodology for selecting a feasible and effective treatment strategy for rehabilitation (from the seven predetermined strategies) will require this prototype of rules to be tested and validated, and then modified or extended, if necessary. In the study area, seven rules have been created to handle operations involved in the spatial knowledge of the PMS. Upon carefully checking the rules, it was determined that these rules are not completely correct. Some of rules are redundant. For this reason, the AIRA tool for Excel v1.3.3 is used to verify these rules. This tool is an add-in for MS-Excel and allows user to extract the “hidden information” (i.e., discover rules) right from the spreadsheets for small to midrange database files. With the same data set and the AIRA tool, 41 rules are generated, part of which are listed in Fig. 2. Obviously, many rules will result in misclassification, and thus have to be reduced. To this end, the following schemes are adopted in this research:

1. If the attribute values simultaneously match the condition of the rules induced by both decision tree and AIRA reasoning, the rules will be retained;
2. If attribute values simultaneously match the conditions of several rules, those rules with the maximum confidence will be kept;
3. If attribute values simultaneously match several rules with the same confidence values, those rules with the maximum coverage of learning samples will be kept; and
4. If attribute values do not match any rule, this class of attribute is defined as the rule, nothing treatment.

With the schemes proposed earlier for rule reduction, 14 rules were still retained. However, only seven rehabilitation strategies in the study area were suggested by the ITRE. Thus, the following method is suggested to further reduce the number of rules:

1. Reduce the attribute data sets from alligator cracking family through
   \[ AC = \{[AN], [AL], [AM], [AS]\} \]
2. Reduce the attribute data through checking the PCI values. The principles are,
   a. If AC=M or S reduce other attributes; and
   b. If RT=S, reduce other attributes.

With the aforementioned reduction methods, the seven rules is finally refined (see Fig. 3).

**Mapping of Data Mining-Based Decision of M&R**

With the rules induced earlier, the rehabilitation strategies can be predicted and decided for each road segment in the database. In other words, the operation using D MKD only occurs in the database, and thus the results cannot be visualized and displayed on either map or screen. However, the GIS system is capable of rapidly retrieving data from the database and automatically generating customized maps to meet specific needs such as identifying maintenance locations, visualizing spatial and nonspatial data, and linking data/information to its geographical location. Thus, this paper employed ArcGIS software in combination with the earlier induced results to create the decision-making map for maintenance and rehabilitation. The basic operation consists of taking each of the above rules as a logic query in the ArcGIS
software, and then queried results are displayed in the ArcGIS layout map. The results are listed in Figs. 4–9, corresponding to each of the preset rehabilitation strategies, respectively. The rehabilitations suggested by the ITRE of North Carolina State University are superimposed with the decisions made under the analysis presented in this paper. As seen from Figs. 4–9, each rehabilitation strategy derived in this paper can be located with its geographical coordinates, and visualized with its spatial, nonspatial data, and different colors.

**Comparison Analysis and Discussion**

**Comparison Analysis**

To verify the results of the proposed method for the decision making of road maintenance and rehabilitation, the recommended pavement segment treatments produced in this paper were compared with those suggested by the ITRE at the NCDOT. All pavement treatments derived by this paper and by NCDOT are displayed in Fig. 10. A comparison analysis for both the numbers and the location of each treatment strategy derived by this paper and by NCDOT in the study area is listed in Table 6. As seen from Table 6, the number of the crack pouring treatment derived by this paper and by NCDOT is the same, i.e., 3, but the locations of the three roads are not the same. Specifically, the location of one road derived by the method presented in this paper is different from the one derived by NCDOT. The number of the full-depth patch (FDP) treatments suggested by NCDOT is 34, but 29 under the method employed in this paper. The difference between two methods is five. Moreover, the location of three road segments for FDP treatment is different for two methods. In the 1-in. plant mix resurfacing (PM1) strategy, NCDOT suggested six roads for PM1 treatment, but this paper indicates seven roads using data mining. Moreover, a road location is different for two methods. For the skin patch (SKP) rehabilitation strategy, 65 roads are suggested by NCDOT for treatment, but 56 roads are identified using data mining. Moreover, 13 road locations are different between two methods. For the short overlay rehabilitation strategy, three roads...
are suggested for treatment, but five roads are recommended by the methods employed in this paper, of which the locations of two roads are different between the two methods.

**Discussion**

The decision trees are based on the knowledge acquired from pavement management engineers to be used for rehabilitation strategy selection. A decision tree is used to organize the obtained knowledge in a logical order. Thus, the decision trees can determine a technically feasible rehabilitation strategy for each road segment. Different decision trees can be built to acknowledge changes. For example, the decision trees were based on severity levels of individual distresses in this paper. If the pavement layer thickness and material type are taken as knowledge, or work history, pavement type, and ride data are taken as knowledge for generating decision trees, these decision trees generated by different types of knowledge are different. This means the association rules generated by different knowledge are different as well.

On the other hand, from our experiment, the rules induced from the decision tree are probably inexact, thereby verification of the rules is needed. This project, in fact, tested different tools (e.g., SPSS, ACRT, RSES, CHIAD) to endeavor to generate an “optimum” or “ideal” decision tree and rules. Unfortunately, the decision trees and rules generated by these tools are not completely the same. Therefore, postprocessing for verification of the rules is still needed. Through the project, it has been realized that:

1. The decision tree and rules induced by the data mining method are not completely exact, i.e., the DMKD method is not as exact as individual inspection and observation. Therefore, postprocessing for refining the rules is still essential; and
2. The advantages of applying the DMKD method for road maintenance and rehabilitation are: (a) it can largely increase the speed of choosing a rehabilitation strategy, thus largely save time and money, and shorten the project period and (b) the same treatments proposed by the DMKD method in a
road network level should have similar pavement conditions, thus, this method avoids any human factors such as visual distress data by engineers.

Conclusions

This paper conducts research and analysis related to applying data mining technology in combination with GIS for pavement management. The main purpose of the research is to utilize data mining techniques to find some useful knowledge hidden in the pavement database. The C5.0 algorithm has been employed to generate decision trees and association rules. The induced rules have been used to predict which maintenance and rehabilitation strategy to be selected for each road segment. A pavement database covering four counties in the state of North Carolina, which are provided by the ITRE at NCDOT, has been used to test the proposed method. The comparison of two decisions for rehabilitation treatment suggested by NCDOT and by the methodology presented in this paper has been conducted. From the experimental results, it was found that the rehabilitation strategies derived by the rules, i.e., C5.0 method, are different from those suggested by NCDOT. After combining other technologies, e.g., AIRA method, and postprocessing, seven rules are finally refined. Using the final rules, mapping for different types of pavement rehabilitation strategies is created using ArcGIS v. 9.1. When compared with the results from NCDOT, the number and location of the suggested road rehabilitations are different. The maximum error for the number of suggested road rehabilitations is nine, and for the location is 13 out of 65 (see Table 6).

Through this project, it has been concluded that

1. The DMKD technology can make consistent decisions regarding road maintenance and rehabilitation for the same road conditions with less interference from human factors;
2. The DMKD technology can largely increase the speed of decision making because the technology automatically generates a decision tree and rules if the expert knowledge is given. This can save significant time and money for PMS;
3. Integration of the DMKD and GIS can provide the PMS with the capabilities of graphically displaying treatment decisions, visualize the attribute and nonattribute data, and link data and information to the geographical coordinates; and
4. The DMKD method is not quite as intelligent as individual observation. In other words, the decisions and rules determined by DMKD are not exact, and thus, postprocessing for verification and refinement is necessary.

Acknowledgments

The pavement database was provided by Greg Ferrara, GIS Program Manager at the Institute for Transportation Research and Education (ITRE) of North Carolina State University. John Roklevi at the ITRE provided essential help in data delivery, data interpretation, and explained related the metadata. The writers sincerely thank both of them. The writers also thank the project administrators for granting permission to use their data.

References