

COMPARITIVE ANALYSIS OF FUZZY DECISION TREE AND LOGISTIC REGRESSION METHODS FOR PAVEMENT TREATMENT PREDICTION

DEVINDER KAUR, HARICHARAN PULUGURTA

Department of Electrical and Computer Sciences, Department of Civil Engineering
University of Toledo
2801 W Bancroft St
USA

devinder.kaur@utoledo.edu, hpulugur@eng.utoledo.edu

Abstract: - Data mining is the process of extraction of hidden predictive information from large databases and expressing them in a simple and meaningful manner. This paper explains the use of Fuzzy logic as a data mining process to generate decision trees from a pavement (road) database obtained from Ohio Department of Transportation containing historical pavement information from 1985 to 2006. Generally there are many attributes in the pavement database and often it is a complicated process to develop a mathematical model to classify the data. This study demonstrates the use of fuzzy logic to generate decision tree to classify the pavement data. Further, the fuzzy decision tree is then converted to fuzzy rules. These fuzzy rules will assist decision-making process for selecting a particular type of repair on a pavement based on its current condition. The fuzzy decision tree induction method used is based on minimizing the measure of classification ambiguity for different attributes. These models overcome the sharp boundary problems, providing soft controller surface and good accuracy dealing with continuous attributes and prediction problems. This method was compared with common logistic regression model for predicting the pavement treatment. The results show that the fuzzy decision method outperforms the logistic regression model by 10%. The fuzzy decision tree method generates the rules, which gives the better understanding of the relationship between the parameters and the pavement treatment prediction.

Key-Words: - Pavement Management, Classification Ambiguity, fuzzy ID3, Logistic Regression

1 Introduction

Pavement Management is a topic of great significance in civil engineering. It is essential to develop reliable pavement management systems, which have the ability to estimate the overall pavement condition and the ability to forecast when and what kind of repair will be needed on certain pavements. The models of the pavement performance prediction are developed using the past pavement performance data. Thus Pavement Performance Prediction models are integrated into the decision making process and help to schedule the repairs and enable to estimate the budgets.

Historically, statistical formulae mostly based on regression analysis have been used for developing models for pavement performance prediction [11]. The regression equations illustrate the effects of various factors on the performance of pavements.

These regression equations are valid only under certain conditions and should not be used if actual conditions are different. This approach is very cumbersome and time consuming in terms of the calculations and in terms of the acquisition of the data required for doing such calculations.

Therefore, there have been attempts by many researchers in the field to use other alternative techniques based on soft computing i.e. fuzzy logic, neural network and evolutionary computation and hybrid techniques based on these three methods to develop models for Pavement Management Systems [16, 17, 18, 19, 21]. Fuzzy logic approach has been used to develop pavement performance model for the asphalt pavement [14]. In this paper author Kaur and Tekkedil built a fuzzy model for the asphalt pavements. It is known from expert opinion that sub grade material plays a very crucial role in the pavement performance. They built fuzzy models based on three different types of sub grades viz.,

clay (code 51), sandy clay (code 52) and sand (code 53). Fig. 1 depicts the layered construction of asphalt pavement.

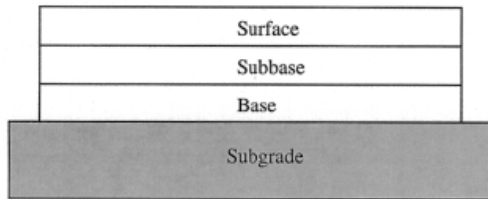


Fig. 1: Layered Construction of roads

The other parameters of the fuzzy model were surface thickness, age of the road, and total traffic count.

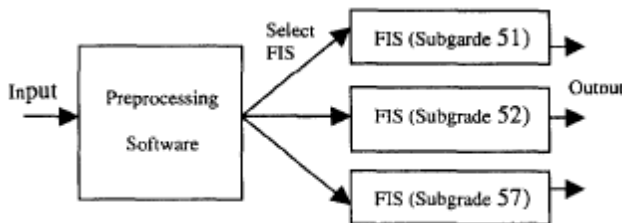


Fig 2: Design of the fuzzy software

Fig. 2 shows the fuzzy inference system which was developed corresponding to three sub grades types [14]. Pre-processing software was written to select the appropriate fuzzy inference system depending on the sub grade type. The software predicts the Rut depth. Fig. 3 represents the graphical user interface of the software developed for predicting the Rut depth. The result of the fuzzy approach was compared with the regression analysis and it was established that fuzzy inference based model outperformed the regression model. Kaur and Datta [12] also used Neuro- Fuzzy techniques to model the pavement performance prediction of asphalt pavements.

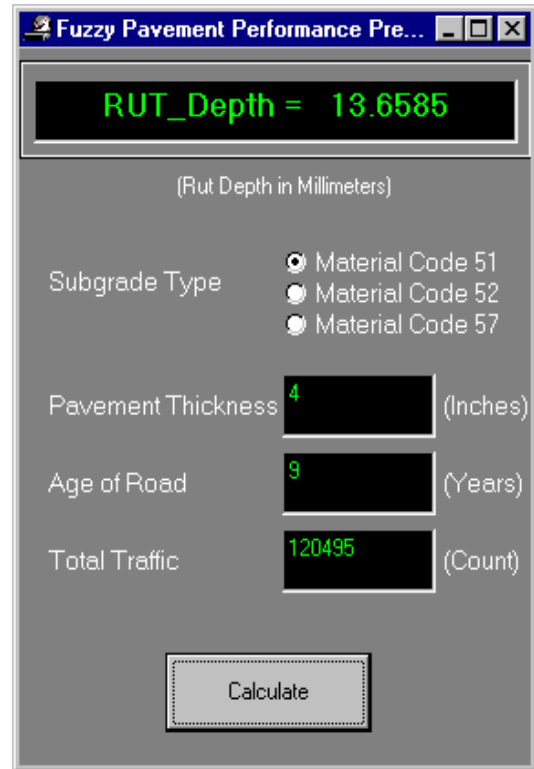


Fig. 3: User Interface of the software

In this paper Adaptive Neuro Fuzzy Inference system (ANFIS) was used to develop the pavement performance prediction model. The initial membership functions and the fuzzy rules were generated from the data using both grid partitioning and subtractive clustering pattern recognition methods. This was followed by training the two different models separately by back propagation (BP) learning algorithm. Then these two different models were validated using different pavement condition dataset.

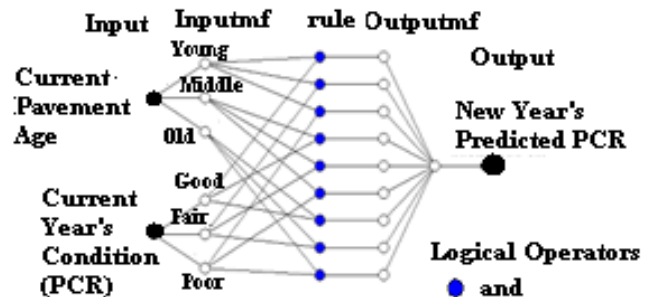


Fig 4: ANFIS Structure of the Model

The asphalt pavement condition database for the interstates and US routes in Ohio, which is available from the Ohio Department of Transportation (ODOT), was used to train, check and evaluate the model. As asphalt pavement sections built prior to 1997 had the same design specification, therefore, 1991 – 1996 pavement condition data was used to train and check the model and 1986 – 1990 condition data was used to evaluate the model.

In that paper authors show that ANFIS (Adaptive Neuro- Fuzzy Model) can be used as a better approach in building pavement deterioration model. ANFIS model can capture many uncertainties that even probabilistic approach used in Markov chain process cannot identify and take care-of [12]. ANFIS based on grid portioning outperforms both ANFIS model based on subtractive clustering and Markov chain prediction model.

It is important to note that the model should be build considering the homogeneity of data set, i.e. the data should be collected from the same population. In other words, different model should be developed and applied for different type of road network, i.e. Interstates/US routes and state routes and for different type of pavement sections.

In the current paper authors explore alternative techniques based on data mining to predict the performance of pavement.

Data mining is the process of extracting hidden information from large databases. Data mining models search databases for hidden patterns, finding classification and predictive information that experts miss because it lies outside their expectations. The classification and prediction problems, where the target attribute is respectively discrete (nominal) and continuous (numerical), are two main issues in data mining and machine learning fields. General methods for these two problems discover rules and models from a database of examples. IF ... THEN ... rules, neural nets, Bayesian nets, and decision trees are examples of such models [2, 8]. There are different ways of representing the patterns discovered by machine learning; each one of the ways dictates the kind of technique to be used to generate that output [15]. General methods of decision trees and classification rules are basic knowledge representation styles that machine learning methods use. Quinlan [7] with ID3 that stands for Interactive Dichotomizer 3 popularized the concept of decision trees. Systems based on this approach use an information theoretic measure of

entropy for assessing the discriminatory power of each attribute.

We are trying to enhance the technique of conventional decision tree with fuzzy logic and designing an algorithm based on fuzzy decision tree for the pavement performance prediction model.

The fusion of fuzzy sets with decision trees enables one to combine the uncertainty handling and approximate reasoning capabilities of the former with the comprehensibility and ease of application of the latter [1]. These models overcome the sharp boundary problems, providing soft controller surface and good accuracy dealing with continuous attributes and prediction problems. The information measure described by ID3 use to split a node is modified to introduce the Fuzzy concept. Sushmita et al. [1] discretize continuous attributes based on the distribution of pattern points in the feature space in linguistic terms using quantiles and use of fuzzy entropy and tree evaluation concept, in terms of compactness and performance. M. J. Kim et al. [3] describe hybrid knowledge integration mechanism using fuzzy genetic algorithm for the optimized integration of knowledge from several sources such as machine knowledge, expert knowledge and user knowledge. Baldwin and Xie [4] describe use of expected entropy and renormalized branch probability in modified fuzzy ID3 algorithm. Olaru and Wehenkel [6] introduce a new method of fuzzy decision trees called soft decision trees (SDT). This method combines tree growing and pruning, to determine the structure of the soft decision tree, with refitting and back fitting, to improve its generalization capabilities. Yuan and Shaw [2] induce a fuzzy decision tree by reducing classification ambiguity with fuzzy evidence. The input data is fuzzified using triangular membership functions around cluster centres obtained using Kohonen's feature map [2]. This study follows the approach proposed by Yuan and Shaw [2] and incorporation of fuzziness at the input by Sushmita et al. [1]. The system is then applied on the pavement management database. A pavement management database stores historical data about pavements (roads) of a network such as its present condition, past condition, geographical location, length, environmental conditions, etc and the attributes are both continuous and discrete. The main purpose of maintaining a pavement management database is to make informed decisions such as type of repairs to be performed on the pavements based on their condition. Given the number of attributes that are present in the pavement

management database, it generally requires complex statistical models to model the data. In the current study, fuzzy logic is used to simplify the entire process of decision-making process. Simple rules are generated from fuzzy decision tree which describe the pavement performance prediction.

A separate method based on logistic regression is then developed for comparing the results obtained from fuzzy decision tree. Logistic regression is used when the dependent variable is categorical and the independent variables are numerical and (or) categorical [9]. Similar to fuzzy decision tree methodology developed in the current study, the logistic regression model predicts the probability of a treatment. The treatment with highest probability is chosen as the final prediction. Furthermore, both the methods are applied on a test dataset and the results are compared.

2 Methodology

The Fuzzy logic used in the current study works by measuring the cognitive uncertainty. Cognitive uncertainty is the uncertainty that deals with phenomena arising from human thinking, or reasoning, cognition and perception process, or cognitive information in general [2]. The cognitive uncertainty can be further classified into two subcategories: vagueness and ambiguity. Once the fuzzy sets are introduced, the cognitive uncertainties represented by fuzzy sets can then be measured.

2.1 The Measure of Vagueness

The vagueness or fuzziness of a fuzzy set can be measured by fuzzy entropy [2]. Let A denote a fuzzy set on the universe U with membership function $\mu_A(u)$ for all $u \in U$. If U is a discrete set $U = \{u_1, u_2, \dots, u_n\}$ and $\mu_i = \mu_A(u_i)$, the vagueness or the fuzziness of fuzzy set A is defined by

$$E_V(A) = -\frac{1}{m} \sum_{i=1}^m (\mu_i \ln \mu_i + (1 - \mu_i) \ln(1 - \mu_i)) \quad . \quad (1)$$

where $E_V(A)$ measures the fuzziness or vagueness of a fuzzy set A . The degree of fuzziness expresses the average amount of ambiguity in taking a decision as to whether an element belongs to the set.

2.2 The Measure of Ambiguity

A fuzzy membership function $\mu(x)$ of a variable Y defined on X can also be interpreted as the possibility of taking value x for Y among all elements in X [2]. In this case $\pi(x) = \mu(x)$ for all $x \in X$, can be viewed as a possibility distribution of Y on X . The possibilistic measure of ambiguity or non-specificity is defined as

$$E_a(Y) = g(\pi) = \sum_{i=1}^m (\pi_i^* - \pi_{i+1}^*) \ln(i), \quad (2)$$

where $\pi^* = \{\pi_1^*, \pi_2^*, \dots, \pi_n^*\}$ is the permutation of the possibility distribution $\pi = \{\pi(x_1), \pi(x_2), \dots, \pi(x_n)\}$, sorted so that $\pi_i^* \geq \pi_{i+1}^*$ for all $i = 1, \dots, n$, and $\pi_{n+1}^* = 0$. To measure the ambiguity (overlapping) of an attribute A among its linguistic terms $T(A) = \{T_1, T_2, \dots, T_n\}$, [2] interpret the membership functions $\{\mu_{T_1}(u_i), \mu_{T_2}(u_i), \dots, \mu_{T_n}(u_i)\}$ as a possibility distribution for object u_i to take linguistic term on term label space $T(A) = \{T_1, T_2, \dots, T_n\}$. To normalize the possibility distribution, let

$$\pi_{T_s}(u_i) = \mu_{T_s}(u_i) / \text{Max}_{1 \leq j \leq S} \{\mu_{T_j}(u_i)\}, s = 1, \dots, S. \quad (3)$$

The ambiguity of the attribute A for object u_i therefore can be measured by [2]

$$E_a(A(u_i)) = g(\pi_T(u_i)). \quad (4)$$

The ambiguity of attribute A then is

$$E_a(A) = \sum_{i=1}^m w(E_a(A(u_i))) E_a(A(u_i)), \quad (5)$$

where $w(E_a(A(u_i)))$ is the weight, which represents the relative size. The ambiguity of classes can be measured in the same way as attributes.

2.3 Classification Ambiguity

Knowing single evidence, such as a particular value of an attribute, the classification ambiguity can be defined [2] as follows:

$$G(E) = g(\pi(C/E)), \quad (6)$$

which is measured on the possibility distribution of $\pi(C/E)$ which is defined as [2]

$$\pi(C_i/E) = S(E, C_i) / \text{Max}_j S(E, C_j), \quad (7)$$

where $S(E, C_i)$ represents the *degree of truth* for the classification rule “IF E THEN C_i “, and $\pi(C/E) = \{\pi(C_i/E), i = 1, \dots, L\}$ is a normalized possibility distribution on the no fuzzy label space

$C = \{C_1, C_2, \dots, C_L\}$. Given a fuzzy evidence F and a set of fuzzy evidences $P = \{E_1, E_2, \dots, E_K\}$ defined on object space U , the fuzzy partition P on F is defined [2] as $P/F = \{E_1 \cap F, \dots, E_K \cap F\}$, where each object defined in F is partitioned to E_i with membership $\mu_{E_i \cap F}$. The classification ambiguity of fuzzy partition can be defined as follows [2]:

$$G(P/F) = \sum_{i=1}^K w(E_i/F)G(E_i \cap F), \quad (8)$$

where $G(E_i \cap F)$ is the classification ambiguity with fuzzy evidence $E_i \cap F$, $w(E_i/F)$ is the weight which represents the relative size of subset $E_i \cap F$ in F

$$w(E_i/F) = M(E_i \cap F) / \sum_{j=1}^K M(E_j \cap F). \quad (9)$$

Significant level [2] α for a fuzzy evidence E with membership $\mu_E(u)$, is defined as

$$\mu_{E\alpha}(u) = \begin{cases} \mu_E(u) & \text{if } \mu_E(u) \geq \alpha \\ 0 & \text{if } \mu_E(u) \leq \alpha \end{cases}. \quad (10)$$

i.e., if the membership value of an attribute is less than α , it is not considered for the analysis.

3 Induction of Fuzzy Decision Tree

Yuan and Shaw [2] construct fuzzy decision trees by reducing classification ambiguity with accumulated fuzzy evidences where fuzzy evidence is the knowledge about a particular attribute. The selection of fuzzy evidence is based on its contribution in reducing the classification ambiguity. The method is similar to the non-fuzzy decision tree induction method such as ID3. The fuzzy decision tree induction process suggested in [2] consists of following steps:

- (1) Fuzzifying the training data
- (2) Inducing the fuzzy decision tree
- (3) Converting the decision tree into a set of rules
- (4) Applying fuzzy rules for classification

3.1 Fuzzifying the Training Data

Any input feature value is described in terms of some combination of overlapping membership values in the linguistic property sets *low* (L), *medium* (M) and *high* (H). When input feature is numerical [1] divide it into three partitions (with range [0, 1]) using only two parameters P_{j1} and P_{j2} .

Let F_{jMax} and F_{jMin} denote maximum and minimum

values encountered along feature F_j . The value of P_{j1} is the value of F_j that exceeds one-third of the measurements and less than two-thirds. The value of second quantile P_{j2} is the value of F_j that exceeds two-third of the measurements and less than remaining one-thirds.

3.2 Inducing Fuzzy Decision Tree

With given evidence significant level and truth level induction process consists of following steps [2]:

- 1) Select the attribute with the smallest classification ambiguity as root node.
- 2) Delete all empty branches of decision node. For each non-empty branch of decision node, terminate the branch as leaf if the truth level of classifying into one class is above a given threshold β . Otherwise, investigate if an additional attribute will further partition the branch and further reduce the classification ambiguity. If yes, select the attribute with smallest classification ambiguity as a new decision node from the branch. If not, terminate this branch as leaf.
- 3) Repeat step 2 for all newly generated decision nodes until no further growth is possible, the decision tree is complete.

The input attributes considered for generating a fuzzy decision tree in this study are shown in Table 1. The meanings of each attribute are described in the following Table1. The overall pavement condition is represented by *PCR* (Pavement Condition Rating) 0 represents the worst pavement condition while 100 the perfect condition. *PCR* is calculated from 15 different variables (*Code 1-Code 15*) called distresses or observable faults on a pavement, which are in turn defined by different categories [13].

For example, a code category of “LO” indicates that a particular severity of a particular distress is **Low** on the pavement and it is seen only occasionally. *Half Car Simulation (HCS)* represents the pavement ride condition (rough to smooth). Traffic data is defined by two different attributes that are *ADT* and *ATDT* (average daily traffic and average truck daily traffic).

Functional class represents the type of road (for example, 1=freeway/interstate, 9=local roads etc). *Activity Code-1* is the previous treatment that was performed while *Activity code* represents the repairs performed on the pavements.

Table 1 Fuzzified Attributes from ODOT database

Attribute	Type	Representation
PCR	Numerical	0 – 100
HCS	Numerical	>0
AvgADT, AvgTADT	Numerical	> 0
Functional Class	Categorical	1,2,6,7,8,9,11,12,14, 16,17,19
Code 1-15	Categorical	LO,LF,LE,MO,MF, ME,HO,HF,HE, NO(NULL)
Activity Code-1	Categorical	10,20,30,35,40,45,50,55,60,70,77,90,95, 100,110,120
Activity Code	Categorical	10,20,30,35,40,45,50,55,60,70,77,90,95, 100,110,120

Activity code and Activity code-1 are categorical where 10 to 45 represent maintenance activities (small repairs to the pavement), 50 to 60 represent minor activities (minor repairs on the pavements) and 70 to 120 represent major activities (reconstruction of the pavement). Table 2 shows a sample of the data set used in this research. The data are divided into training set and test set (ensuring that the training set contains all the attribute ranges found in the data set).

The significance level and degree of truth are considered as 0.5 and 0.6. Significance level of 0.5 indicates the membership values less than 0.5 are not considered. If the truth level for an attribute at a

branch exceeds 0.6, it becomes leaf [2]. A training set of 329 cases is selected to generate the decision tree. The decision tree is then applied to test data. The process of generating decision tree is done through a user interface as shown in Fig. 5. The user interface provides an option to force the decision tree with a selected root node. In the present study, PCR represents the overall pavement condition and hence PCR is forced as root node in case the Fuzzy ID3 algorithm does not identify it as a root node. Fig. 6 shows the decision tree generated with significance level of 0.5 and truth level of 0.6. Fig.6 shows the decision tree generated with significance level of 0.5 and truth level of 0.6.

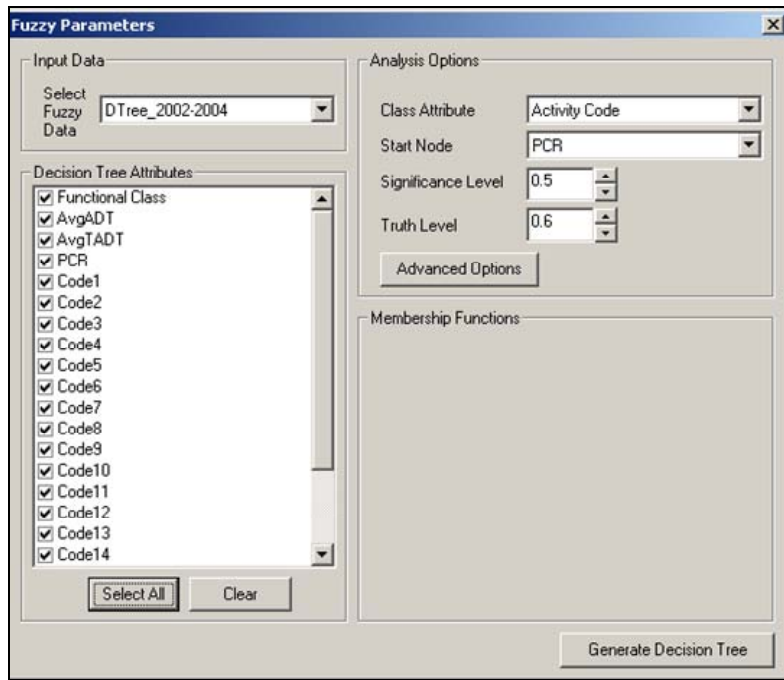


Fig. 5 User Interface to Generate Decision Trees

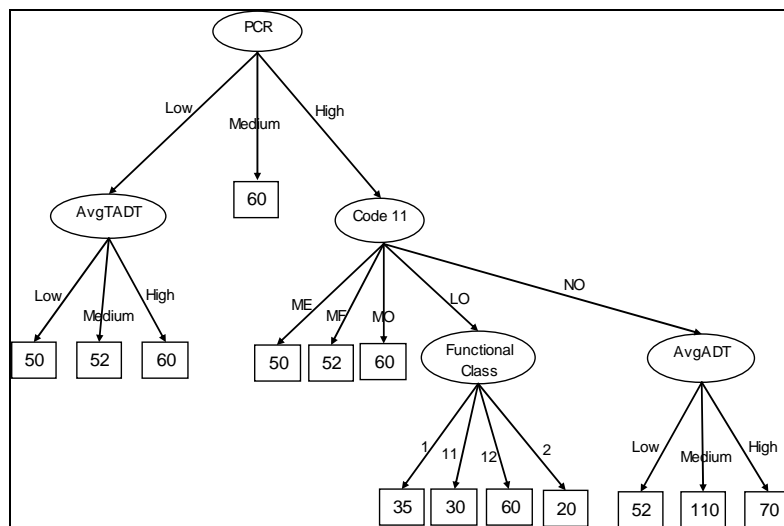


Fig. 6: Fuzzy Decision Tree

Each path of the branches from root to leaf can be converted into a rule with condition part represents the attributes on the passing branches from root to the leaf and the conclusion part represents the class at the leaf with the highest classification truth level [2]. Fig. 7 shows the 14 rules from the decision tree.

R1 IF (PCR=Low AND AvgTADT=Low) THEN Class=50
R2 IF (PCR=Low AND AvgTADT=Med) THEN Class=52
R3 IF (PCR=Low AND AvgTADT=High) THEN Class=60
R4 IF (PCR=Med) THEN Class=60
R5 IF (PCR=High AND Code11=ME) THEN Class=60
R6 IF (PCR=High AND Code11=MF) THEN Class=60
R7 IF (PCR=High AND Code11=MO) THEN Class=30
R8 IF (PCR=High AND Code11=LO AND Functional Class=1) THEN Class=35
R9 IF (PCR=High AND Code11=LO AND Functional Class=11) THEN Class=60
R10 IF (PCR=High AND Code11=LO AND Functional Class=12) THEN Class=60
R11 IF (PCR=High AND Code11=LO AND Functional Class=2) THEN Class=52
R12 IF (PCR=High AND Code11=NO AND AvgADT=Low) THEN Class=52
R13 IF (PCR=High AND Code11=NO AND AvgADT=Med) THEN Class=110
R14 IF (PCR=High AND Code11=NO AND AvgADT=High) THEN Class=70

Fig. 7: Fuzzy Rules

The membership of conclusion (classification of each class) is set equal to the membership of condition [2]. The rule with maximum membership value is considered as correct rule and the object is assigned to the class of the outcome of the rule.

4 Logistic Regression

In this paper, the fuzzy decision tree is compared with a different classification method based on Logistic Regression [8]. The comparison was necessary to determine how accurate the fuzzy decision tree was and also to determine whether the results were comparable. Logistic regression is a data mining method that can be used to classify a given dataset. Logistic regression builds a linear model based on a transformed variable [10] often referred to as logit variable.

Logistic regression is a statistical method that can be used for the analysis when the dependent variable is categorical. Logistic regression is a variant of multiple regression [5], which is used to assess the relation between one dependent variable (binary, categorical or ordinal) and several predictor variables (continuous or categorical). Several forms of logistic regression models exist and the choice of a model depends on the type of the classification variable.

The common forms of logistic regression model are binomial, multinomial and ordinal [11]. Binomial logistic regression is used when the independent variable is dichotomous while ordinal logistic

regression is used when ordering within the dependent variable categories is important. However, in the current study since the ordering within the dependent variable has no meaning, multinomial logistic is used.

Multinomial logistic regression is a complicated process. In the current paper since the prediction categories are eleven, the number of logistic regression equations obtained are ten because in logistic regression one of the prediction categories is chosen as a reference category. Similar to fuzzy decision tree, logistic regression equation gives the probability that the object belongs to a certain

class. The general form of a logistic regression is:

$$\frac{P_i}{P_n} = e^{\beta_{i0} + \beta_{i1}X_1 + \beta_{i2}X_2 + \dots + \beta_{ik}X_k} = e^{u_i}, \quad (11)$$

where $i = 1, 2, \dots, n-1$ are number of classifications. As explained earlier, the logistic regression gives the probability of the class, which is calculated as:

$$P_i = \frac{e^{u_i}}{1 + e^{u_1} + e^{u_2} + \dots + e^{u_{n-1}}}. \quad (12)$$

The probability of the reference category n is calculated as:

$$P_n = 1 - \sum_{i=1}^{n-1} P_i. \quad (13)$$

The maximum likelihood estimation procedure is used to obtain the estimates of the coefficients. Maximum likelihood procedure is used to maximize the value of a function called log-likelihood function [8]. In the current paper, the logistic regression model was developed for the same dataset that was used to develop the fuzzy decision tree. As seen from Equation's 12 and 13, there are 10 separate equations that are obtained from logistic regression model. Hence, probability values are obtained as output for each treatment category. The logit value for the treatment categories (see Table 1 for treatment categories) can be calculated as:

$$e^{u(20)} = e^{87.94 - 1.59(STRD) - 0.001(AvgADT) + 0.005(AvgTADT)}, \quad (14)$$

$$e^{u(30)} = e^{74.31 - 0.89(STRD) - 0.001(AvgADT) + 0.005(AvgTADT)}, \quad (15)$$

Table 2 Data Sample

Pave Section	Functional Class	Avg ADT	Avg TADT	PCR	Code										HCS	STRD	Activity Code-1	Class
					1	2	3	.	.	.	13	14	15					
1	1	42310	14380	62	LE	NO	MF					MO	E	NO	86	21.07	0	100
2	12	31250	1660	76	ME	NO	NO					LO	E	NO	113	9.04	0	100
3	12	31250	1660	67	ME	NO	HF					NO	E	NO	153	14.59	0	100
4	2	29150	3180	84	LF	NO	LO					NO	O	NO	93	10.11	0	20
5	2	29150	3180	85	LF	NO	NO					NO	F	NO	0	8.24	0	20
6	11	84590	12710	80	LE	NO	LO					LO	O	NO	73	9.12	60	30
7	11	84590	12710	80	LE	NO	LO					LO	O	NO	73	9.12	60	30
8	1	35050	13640	84	LF	NO	NO					NO	F	NO	47	6.96	0	35
.																		
.																		
.																		
322	11	32870	5010	70	MF	NO	HO					MO	E	NO	62	11.6	0	52
323	11	32870	5010	70	MF	NO	HO					MO	E	NO	62	11.6	0	52
324	12	68540	3860	72	ME	NO	MO					NO	E	NO	110	10.2	0	52
325	12	68540	3860	65	ME	NO	MO					NO	E	NO	80	17.92	0	52
326	12	19890	1370	62	MF	NO	HO					LO	E	NO	176	19.84	0	52
327	12	19890	1370	62	MF	NO	HO					LO	E	NO	176	19.84	77	52
328	12	19890	1370	62	MF	NO	HO					LO	E	NO	120	19.84	77	52

$$e^{u(31)} = e^{79.20-0.95(STRD)-0.001(AvgADT)+0.004(AvgTADT)}, \quad (16)$$

$$e^{u(35)} = e^{96.60-9.72(STRD)-0.019(AvgADT)+0.054(AvgTADT)}, \quad (17)$$

$$e^{u(38)} = e^{111.71-3.55(STRD)-0.002(AvgADT)+0.005(AvgTADT)}, \quad (18)$$

$$e^{u(50)} = e^{70.62-0.21(STRD)-0.001(AvgADT)+0.004(AvgTADT)}, \quad (19)$$

$$e^{u(52)} = e^{74.46-0.41(STRD)-0.001(AvgADT)+0.004(AvgTADT)}, \quad (20)$$

$$e^{u(60)} = e^{66.366-0.18(STRD)-0.001(AvgADT)+0.005(AvgTADT)}, \quad (21)$$

$$e^{u(100)} = e^{70.19-0.32(STRD)-0.001(AvgADT)+0.004(AvgTADT)}, \quad (23)$$

$$e^{u(110)} = e^{75.25-0.86(STRD)-0.001(AvgADT)+0.004(AvgTADT)}, \quad (25)$$

It is important to note that the independent variables shown in the Equations 14 to 25 are identified by the model logistic regression methodology. In other words, *STRD*, *AvgADT* and *AvgTADT* are the only significant variables. The other independent variables have no or very small power in predicting the treatment categories.

The probability of treatment categories are then calculated using Equation's 12 and 13. The final prediction (treatment category) is chosen as the treatment category with highest probability value. For example consider a pavement section with independent variables *STRD*, *AvgADT* and *AvgTADT* values as 8.8, 11498 and 4583 respectively. According to Equation 20, the e^{u_i} for treatment category 20 can be calculated as:

$$e^{u(20)} = e^{87.94-1.59(8.8)-0.001(11498)+0.005(4583)}. \quad (26)$$

$$e^{u(20)} = e^{80.18} = 6.6 \times 10^{34}. \quad (27)$$

In a similar fashion, e^{u_i} is calculated for all the treatment categories (denominator of Equation 12) as 8.9×10^{36} .

Equation 12 is then used to determine the probability that the pavement section needs a certain category. In the current example, the probability that treatment category 20 needs to be performed on the pavement section is calculated as:

$$P_{20} = \frac{6.6 \times 10^{34}}{8.8 \times 10^{36}} = 0.01. \quad (28)$$

The probability value in the Equation 20 indicates that the probability that the pavement section requires a treatment 20 to be performed is only 0.01. In a similar fashion Equations 15 to 25 are used to determine the probabilities for all the treatment categories. In the current example, the probability of treatment category 38 is highest with a value of 0.98. Based on this maximum probability, the treatment category predicted for the pavement section is 38.

5 Model Comparison

Before developing the fuzzy decision tree and the logistic regression equations, the data were divided into test set and a training set. The models were developed using the training set. The developed models are then applied on the test data set (the test data set was not included in the model development stage).

Table 3 shows the results obtained by fuzzy decision tree and logistic regression. The test dataset contained 100 pavement sections. The fuzzy decision tree was able to classify 70% of the pavement sections correctly and the logistic model was able to classify 60% of the pavement sections correctly.

6 Conclusion

Pavement management database consists of many different attributes that are both continuous and categorical in nature. It is often required in pavement management to determine the type of repair needed for a pavement. This decision is based on the condition of the pavement whether it is in good condition or fair condition and also with respect to different attributes such as traffic, weather conditions etc.

It is a complicated process to develop a statistical model based on all these attributes. In this study a more straightforward approach is used and is demonstrated using actual data. A fuzzy decision tree is generated which is then converted to simple rules. The rules are then tested on a test data set and the results showed that the accuracy of the tree was approximately 70%.

Furthermore, a multinomial logistic regression was used to classify the dataset and the result from the logistic regression model was compared to the fuzzy decision tree. The accuracy of the logistic regression model was 60%.

Table 3 Results

Pave Section	Functional Class	Avg ADT	Avg TADT	PCR	Code										Activity Code	Fuzzy Tree	Logistic Model			
					1	2	3	4	5	6	7	8	9	10				11	12	13
1	2	10965	3470	61	ME													50	50	52
2	2	10910	3500	53	ME													50	50	50
3	2	4690	350	80	LE													60	52	38
4	2	9773	1300	78	LE													35	52	38
5	12	112708	6353	65	ME													60	60	70
.
99	12	112708	6353	65	ME													60	60	70
100	1	15978	3932	72	MF													50	30	52

The results show that the fuzzy decision tree is more accurate than the logistic model. Furthermore, fuzzy decision tree is easy to develop than the logistic model. In the current study logistic model has ten separate equations.

References:

[1] S. Mitra, K. M. Konwar, and S. K. Pal, "Fuzzy Decision Tree, Linguistic Rules and Fuzzy Knowledge-Based Network: Generation and Evaluation," in *IEEE Trans. Systems Man Cybernet*, Vol. 32, Nov. 2002, pp. 328-339.
 [2] Y. Yuan and M. J. Shaw, "Induction of fuzzy decision trees," *Fuzzy Sets Syst.*, Vol. 69, 1995 pp. 125-139.
 [3] H. J. Kim, I. Han, and K. C. Lee, "Hybrid Knowledge Integration Using the Fuzzy Genetic Algorithm: Prediction of the Korea Stock Price Index," *Intell. Sys. Acc. Fin. Mgmt.* 12: 43-60 2004
 [4] J. F. Baldwin and D. W. Xie, "Simple Fuzzy Logic Rules Based on Fuzzy Decision Tree for Classification and Prediction Problem," *Journal of Intelligent Information Processing (IIP)* 2004 Available at:

<http://eis.bris.ac.uk/~enxdx/files/ICIIPBaldwinandXie.pdf>

[5] J. R. Quinlan, "Induction on decision trees," *Machine Learning*, Vol. 1, pp. 81-106, 1986.
 [6] C. Olaru and L. Wehenkel, "A Complete Fuzzy Decision Tree Technique," *Fuzzy Sets Syst.*, vol. 138, pp. 221-254, 2003.
 [7] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
 [8] S. Mernard, *Applied Logistic Regression*. (Sage University Papers on Quantitative Applications in the Social Sciences, series no. 07-106). Thousand Oaks, CA: Sage, 2002
 [9] F. C. Pampel, *Applied Logistic Regression*, (Sage University Papers on Quantitative Applications in the Social Sciences, series no. 07-132). Thousand Oaks, CA: Sage, 2000
 [10] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edition, CA: Morgan Kaufmann, 2005

- [11] M. J. Norusis, *SPSS 15.0 Advanced Statistical Procedures Companion*, Prentice Hall, New Jersey, 2007
- [12] D. Kaur and D. Datta, "Soft Computing Technique in Prediction of Pavement Condition", *Proceedings of WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics (CIMMACS)*, ISBN: 978-960-6766-21-3, ISSN: 1790-5117, Tenerife, Dec. 14-16, 2007, pp. 88-93.
- [13] D. Kaur and Haricharan Pulugurta, "Fuzzy Decision Tree Based Approach to Predict the Type of Pavement Repair", *Proceedings of WSEAS International Conference on Applied Informatics and Communications*, ISSN: 1790-5117, ISBN: 978-969-8457-96-6, Vouliagmeni Beach, Athens, Aug. 24-26, 2007, pp. 1-6.
- [14] D. Kaur and Dilip Tekkedil, "Fuzzy Expert System for Asphalt Pavement Performance Prediction" *2000 IEEE Intelligent Transportation Systems Conference Proceedings*, Dearborn, MI, USA, Oct. 1-3, 2000. ISBN: 0-7803-5971-2, pp. 428-433.
- [15] Sung Hee Kim and Nakeseok Kim, "Development of performance prediction models in flexible pavement using regression analysis method", *KSCE Journal of Civil Engineering*, Vol. 10, No. 2, ISSN: 1226-7988, March 2006, pages 91-96.
- [16] W-D Yu, Hybrid Soft Computing Approach for Mining of Complex Construction Databases, *Journal of Computing in Civil Engineering*, ASCE, Vol.21, No.5, 2007, pp 343-352.
- [17]. J. J. Ortiz-Garcia, S. B. Costello and M.S. Snaith, Derivation of Transition Probability Matrices for Pavement Deterioration Modeling, *Journal of Transportation Engineering*, ASCE, Vol.132, No.2, 2006, pp 141-161.
- [18]. A.K. Sandra, V.R. Vinayaka Rao, K. S. Raju, and A. K. Sarkar, Prioritization of Pavement Stretches using Fuzzy MCDM Approach – A Case Study, *11th Online World Conference on Soft Computing in Industrial Applications*, September 18-October 6, 2006,
- [19]. Y. Kleiner, R. Sadiq, and B. Rajani, Modeling the deterioration of buried infrastructure as fuzzy Markov process, *Journal of Water Supply Research and Technology: Aqua*, v. 55, No. 2, March 2006, pp 67-80.
- [20]. Ohio Department of Transportation, *Pavement Condition Rating Manual*, <http://www.dot.state.oh.us/pavement/PavementManagement/Manual.htm>
- [21]. J.S.R. Jang, C.-T. Sun and E. Mizutani, *Neuro – Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*, Prentice Hall, Upper Saddle River, NJ 07458, 1997.