A Fuzzy Decision Tree for Bus Network Management

Marcos E. Cintra, Odacir A. Neves

Universidade Federal Rural do Semi-Arido, Brazil
{cintra,odacir}@ufersa.edu.br

Abstract. The planning and managing of bus networks as an urban public transport system is a relevant problem. Several proposals focus the planning of new bus networks. In this paper, we focus the specific task of reallocating buses from different lines of a previously planned bus network in case of restrictions, such as vehicle breakdowns and drivers absences. Our proposal uses a fuzzy decision tree, which combines the graphical representation and high interpretability of decision trees with the use of linguistic variables of fuzzy systems. The fuzzy decision tree is intended to support human experts that have to make decisions regarding reallocating buses from predefined lines in order to optimize the average time that users have to wait for a bus service in any bus stop. The model was induced by the FUZZYDT algorithm using real data collected from a public bus system with 26 bus lines. 16 variables related to characteristics of the bus lines and users were considered as inputs to the system. The output variable of the system presents an evaluation of bus lines when changes in the lines occur. The induced fuzzy decision tree includes only 7 key-variables forming 15 rules. The final model is simple, highly interpretable, and has received positive feedback from experts.

Keywords: bus network management, decision trees, fuzzy decision trees, rule-based fuzzy systems

1. INTRODUCTION

The planning and management of urban public transport systems [Fan and Machemehl 2008; Han and S.; Kim 2011], such as bus networks [Ghatee and Hashemi 2008], is a relevant problem that can be decomposed in a sequence of tasks, including the network design, frequency setting, timetable development, as well as bus and drivers scheduling [Ceder and Wilson 1986], among others.

Regarding bus network planning, from scratch, a strong restriction is the number of available vehicles. In this sense, researchers usually focus on the scenario of distributing a previously defined set of vehicles into a given number of bus lines in order to define a bus network. Some proposals work on the optimization of a set of conflicting objectives [Fonseca and Fleming 1993], usually by adopting methods that generate a Pareto front, which allows to choose solutions from a set of non-dominated options. Clustering algorithms [Oliveira et al. 2009] have also been used for this specific task of bus network planning.

A main issue for the maintenance of a previously planned bus network system is the redistribution of the buses in case of mechanical problems in vehicles and/or absent drivers. Human experts usually tackle the problem by reallocating buses or drivers from non-affected lines, instead of replanning the whole bus network. Our proposal aims at supporting the decision of the human expert on such task.

Decision trees are popular models in machine learning, especially for classification problems, due to the fact that they produce graphical models, as well as text rules, that are easily understandable for final users. Moreover, their induction process is usually fast, requiring low computational power. Fuzzy systems, on the other hand, provide mechanisms to handle imprecision and uncertainty in data, based on the fuzzy logic and fuzzy sets theories [Zadeh 1994]. The combination of fuzzy systems and decision trees [Chandra and Varghese 2009; Cintra et al. 2012; Huang et al. 2008] has produced fuzzy decision tree models, which benefit from both techniques to provide simple, accurate, and highly interpretable models at low computational cost.

In this sense, we use a fuzzy decision tree to provide a human planner with a support system to evaluate possible options in case of bus or driver problems. Such evaluation is focused on the impact of reallocating...
M. E. Cintra and O. A. Neves

buses from lines based on the characteristics of the involved lines and the human experience, whose knowledge is stored in the form of rules in the decision tree.

In order to induce the fuzzy decision tree we use data collected from 26 bus lines. The fuzzy decision tree algorithm adopted to induce the model is named FUZZYDT [Cintra et al. 2012]. The induced decision tree contains a reduced subset of variables, which are defined as linguistic variables, increasing its interpretability.

The remaining of this paper is organized as follows. Section 2 discusses the task of bus networks and public transport planning. Section 3 presents the basic concepts of fuzzy classification systems. Section 4 discusses decision trees and presents the FUZZYDT algorithm. Section 5 presents the induced fuzzy decision tree and the definition of the input attributes for the problem, followed by the conclusions and future work in Section 6.

2. BUS NETWORKS IN PUBLIC TRANSPORT SYSTEMS

The task of planning public transport systems can be divided into the following subtasks [Ceder and Wilson 1986]: i) network design; ii) frequency setting; iii) timetable development; iv) bus scheduling; and v) drivers scheduling.

In the literature, it is possible to find different approaches for planning public transport systems. The tabu metaheuristic search is used in [Pacheco et al. 2001]. Genetic Algorithms (GAs) are used for this task in [Goldberg 1989] aiming at improving an existing service by redistributing and reducing the number of transfers, as well as the travel time of each user. GAs are also used in [Kidwai et al. 2005] to bus scheduling by performing two steps: in the first one, the minimum frequency of buses required on each route, with the guarantee of load feasibility, is determined by considering each route individually; in the second step, the fleet of buses is taken as upper bound and fleet size is again minimized by considering all routes together.

In [Yu et al. 2011], the authors combine the tabu search and genetic algorithms to define a solution by compromising the quality of the services to the users and the total cost of the services. In other words, in [Yu et al. 2011], the authors try to maximize the quality of the services and minimize their operational cost. In [Fan and Machemehl 2008], the authors use the tabu search to solve a non linear model and redistribute urban buses, presenting a sensitivity analysis showing that the tabu search obtained better results than the genetic algorithm approach. Three computational programs were developed by Han and Kim [Han and S.; Kim 2011] based on simulated annealing, tabu search, and genetic algorithms in order to solve the redistribution of urban buses problem, presenting a comparison of the results. In [Tsubouchi 2009], instead, the authors work with the on-demand bus scheduling problem by executing two main algorithms: a vehicle-choosing algorithm and a routing algorithm. Both algorithms in [Tsubouchi 2009] are based on expert heuristics.

Although several proposals can be found in the literature for public transportation systems, the task we focus in our work, the redistribution of buses in case of bus breakdowns and problems with drivers absences, is usually performed by human experts. Once a problem is detected, the manager of the transport system can simply ignore it, when the impact on the system is considered tolerable, or decide to reallocate buses, in case no extra buses or drivers exist, if the impact on the whole system is considerable. In order to decide which bus line will be least affected by a reallocation, the human operator must analyse all existing information regarding the lines.

Due to human limitations to process information in reduced time, we propose the use of a fuzzy decision tree system to support the human expert. The fuzzy decision tree includes previous knowledge extracted from human experience in the form of rules.

3. FUZZY CLASSIFICATION SYSTEMS

Classification is a relevant task of machine learning that can be applied to pattern recognition, decision making, and data mining, among others. The classification task can be roughly described as: given a set of objects
A Fuzzy Decision Tree for Bus Network Management

\[ E = \{e_1, e_2, ..., e_n\}, \text{also named examples or cases, which are described by } m \text{ features, also named variables or attributes, assign a class } c_i \text{ from a set of classes } C = \{C_1, C_2, ..., C_j\} \text{ to an object } e_p, e_p = (a_{p_1}, a_{p_2}, ..., a_{p_m}). \]

Fuzzy classification systems are rule-based fuzzy systems that require the granulation of the features domain by means of fuzzy sets and partitions. The linguistic variables in the antecedent part of the rules represent features, or attributes, and the consequent part represents a class. A typical fuzzy classification rule can be expressed by

\[ R_k : \text{IF } X_1 \text{ is } A_{1l_1} \text{ AND } ... \text{ AND } X_m \text{ is } A_{ml_m} \text{ THEN Class } = C_i \]

where \( R_k \) is the rule identifier, \( X_1, ..., X_m \) are the features of the example considered in the problem (represented by linguistic variables), \( A_{1l_1}, ..., A_{ml_m} \) are the linguistic values used to represent the feature values, and \( C_i \in C \) is the class. The inference mechanism compares the example to the rules in the fuzzy rule base in order to assign a class to the example.

The classic and general fuzzy reasoning methods [Cordon et al. 1999] are widely used in the literature. Given a set of fuzzy rules (fuzzy rule base) and an input example, the classic fuzzy reasoning method classifies this input example using the class of the rule with maximum compatibility to the input example, while the general fuzzy reasoning method calculates the sum of compatibility degrees for each class and uses the class with highest sum to classify the input example. The classic fuzzy reasoning method is also known and the best rule method, while the general fuzzy reasoning method is also known as the best class method.

4. DECISION TREES

As previously mentioned, decision trees provide popular and powerful models for machine learning which are easily understandable and intuitive. Decision trees also require low computational power and usually produce competitive models that can be expressed graphically or as a set of rules.

C4.5 [Quinlan 1996] is one of the most relevant and well-known decision tree algorithms. C4.5 uses the information gain and entropy measures to decide on the importance of the features, which can be numerical and/or categorical. C4.5 recursively creates branches corresponding to the values of the selected features until a class is assigned as a terminal node. Each branch of the tree can be seen as a rule, whose conditions are formed by their attributes and respective tests.

In order to avoid overfitting, C4.5, as well as most decision tree algorithms, includes a pruning process. Specifically, C4.5 adopts a post-pruning strategy. This way, the pruning process takes place after the tree is completely induced. The pruning process of C4.5 basically assesses the error rates of the tree and its components directly on the set of training examples [Quinlan 1993].

The postpruning method implemented in C4.5 discards subtrees, which are replaced by leaf nodes. The class assigned to a leaf is the most frequent class found among the examples of the training set covered by that leaf. This pruning method analyses the error rate of the tree using just the training examples with which the tree is built. The basic idea is to estimate the real error of a subtree, which, in fact, cannot be determined using only the examples of the training set. If the estimated real error is smaller than the apparent error (error calculated using the set of training examples), the subtree is pruned.

Another important aspect of decision trees is the fact that their induction process selects only the relevant variables for the definition of the final model. Thus, the process of inducing the decision tree model performs an embedded attribute selection process, which simplifies the final model, improving its interpretability.

In this work, we adopt FUZZYDT [Cintra et al. 2012] to generate the fuzzy decision tree in order to support the task of bus rescheduling. FUZZYDT \(^1\) uses the same measures of the classic C4.5 algorithm (entropy and

\(^1\) A test version of FUZZYDT [Cintra et al. 2012] including instructions on how to execute it, is available for download at http://dl.dropbox.com/u/16102646/FuzzyDT.zip

Symposium on Knowledge Discovery, Mining and Learning – KDMiLe 2013
information gain) to decide on the importance of the features. It also uses the same induction strategy to recursively partition the feature space creating branches until a class is assigned to each branch. However, for FUZZYDT, continuous features are defined in terms of fuzzy sets before the induction of the tree. This way, the process of inducing a tree using FUZZYDT takes a set of “discretized” features, since the continuous features are defined in terms of fuzzy sets and the training set is fuzzyfied before the decision tree induction takes place.

The FUZZYDT algorithm uses a previously defined fuzzy data base, i.e., the definition of the fuzzy granulation for the domains of the continuous features in order to replace the continuous attributes of the training set by linguistic labels: the linguistic labels of the fuzzy sets with highest compatibility with the input values are used. In the sequel, FUZZYDT calculates the entropy and information gain of each feature to split the training set and define the test nodes of its tree until all features are used or all training examples are classified. Once the model is induced, a post-pruning process, similarly to C4.5, using 25% confidence limits as default, is performed.

Figure 1 illustrates the process of data fuzzyfication and tree induction for a toy dataset with \( n \) examples, 3 attributes (\( At_1, At_2, \) and \( At_3 \)), and 3 classes (\( C_a, C_b, \) and \( C_c \)).

1. **A FUZZY DECISION TREE MODEL FOR BUS NETWORK MANAGEMENT**

   The generated decision tree is based on real data collected for 26 bus lines in the city of Grenoble, France. Table I presents the names of the 17 variables included in the data, as well as their minimum (Min.), maximum (Max.), and average (Avg.) rates.

   Most of the variables are self-explanatory. Variable Load refers to the number of travellers in the busiest section of the line. The Average Time variable refers to the time taken to travel from the first to the last stop of a line. The Average Waiting Time variable refers to the average time a traveller has to wait for a bus to arrive at any bus stop of the given bus line.

   Each original variable was used to define a linguistic variable, according to the fuzzy logic theory. The linguistic variables were defined by three triangular shaped fuzzy sets, evenly distributed in the domains, according to the equalized universe method [Chen and Wang 1999].

   The choice of 3 fuzzy sets for each linguistic variable was heuristically defined. For this purpose, different decision trees were induced using 3, 5, and 7 fuzzy sets defining the input attributes. Since the best results were obtained using 3 fuzzy sets, the input variables were defined by 3 triangular-shaped fuzzy sets.
A Fuzzy Decision Tree for Bus Network Management

Table I. General characteristics of the involved variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min.</th>
<th>Max.</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Length (m)</td>
<td>7,620.00</td>
<td>30,890.00</td>
<td>16,563.14</td>
</tr>
<tr>
<td>2 Number of vehicles</td>
<td>2.00</td>
<td>14.00</td>
<td>6.22</td>
</tr>
<tr>
<td>3 Buses per Km</td>
<td>81.60</td>
<td>524.50</td>
<td>198.73</td>
</tr>
<tr>
<td>4 Interval between buses (seconds)</td>
<td>318.00</td>
<td>1,680.00</td>
<td>846.39</td>
</tr>
<tr>
<td>5 Rotation time (seconds)</td>
<td>2,832.00</td>
<td>7,388.00</td>
<td>4,594.51</td>
</tr>
<tr>
<td>6 Average speed (km/h)</td>
<td>9.70</td>
<td>18.00</td>
<td>12.79</td>
</tr>
<tr>
<td>7 Capacity</td>
<td>90.00</td>
<td>135.00</td>
<td>95.67</td>
</tr>
<tr>
<td>8 Number of travellers</td>
<td>235.00</td>
<td>2,706.00</td>
<td>1,137.73</td>
</tr>
<tr>
<td>9 Travellers per km</td>
<td>690.00</td>
<td>13,173.00</td>
<td>4,186.63</td>
</tr>
<tr>
<td>10 Average occupation rate</td>
<td>0.00</td>
<td>0.00</td>
<td>0.22</td>
</tr>
<tr>
<td>11 Maximum occupation rate</td>
<td>0.19</td>
<td>0.68</td>
<td>0.41</td>
</tr>
<tr>
<td>12 Commuting travellers</td>
<td>62.00</td>
<td>861.00</td>
<td>339.80</td>
</tr>
<tr>
<td>13 Average length of a route (km)</td>
<td>66.90</td>
<td>4,889.00</td>
<td>3,333.70</td>
</tr>
<tr>
<td>14 Average time (seconds)</td>
<td>497.00</td>
<td>1,121.00</td>
<td>864.51</td>
</tr>
<tr>
<td>15 Buses per km per travellers</td>
<td>0.10</td>
<td>0.39</td>
<td>0.21</td>
</tr>
<tr>
<td>16 Load</td>
<td>125.00</td>
<td>967.00</td>
<td>468.71</td>
</tr>
<tr>
<td>17 Average waiting time (seconds)</td>
<td>300.00</td>
<td>730.00</td>
<td>520.25</td>
</tr>
</tbody>
</table>

The input variables shown in Table I were used in order to induce a decision tree model that classifies the resulting Average Waiting Time for a given bus service.

Figure 2 presents the definition of the input variables in terms of fuzzy sets (see Table I).

Fig. 2. Definition of the input variables using 3 triangular fuzzy sets, evenly distributed in their domains.

Specifically for the definition of the output variable of the fuzzy decision tree, the Average Waiting Time, we adopted 5 evenly distributed triangular fuzzy sets. The linguistic values were chosen with the assistance of an expert in public transportation and aim to reflect the adequacy of the waiting time of a passenger for a bus service. The set of linguistic values is composed of Excellent, Good, Regular, Long, and Very Long. Figure 3 presents the definition of the output variable, Average Waiting Time, using 5 triangular fuzzy sets, evenly distributed in its domain.

After the definition of the original variables into linguistic variables, a set of real world examples, formed by real values recorded for a set of 26 bus lines, were used to induce the fuzzy decision tree. Besides the real values for each of the attributes listed in Table I, the examples also contain the resulting Average Waiting Time values. In total, 51 examples were used in the process of the fuzzy decision tree induction.

Figure 4 presents the induced fuzzy decision tree model. Notice that Figure 4 uses E for Excellent, G for Good, R for Regular, L for long, and V for Very Long.

The accuracy of the induced fuzzy decision tree (Figure 4), in relation to the set of examples, was 94.12%. Once the induced model has linguistic variables and values, the inference process becomes quite straightforward and fast. Thus, several changes and lines can be tested analysed in a short time.
Fig. 3. Definition of the *Average Waiting Time* linguistic variable using 5 triangular fuzzy sets, evenly distributed in the domain, which ranges from 300 to 730 seconds, or from 5 to 12 minutes.

Fig. 4. Induced fuzzy decision tree to support the task of bus reallocation.

Some desirable characteristics regarding the variables that are considered relevant for the problem can be found in the induced model. For instance, only 7 of the 16 input variables are used in the model. These 7 variables are the relevant ones for the FuzzyDT algorithm. This reduction in the set of variables produces a simpler and more interpretable model, also contributing for a fast inference.

Regarding the variables in the fuzzy decision tree, the first three ones, *Interval Between Buses*, *Average Speed*, and *Number of Vehicles*, are, intuitively, considered as the deciding factors for the reallocation of vehicles. The remaining attributes are related to the number of travellers (*Capacity*, *Occupancy Rate*, and *Number of Travellers*). The *Average Time* is also considered in the model. Notice that although 7 variables are used in the induced fuzzy decision tree, in the worst case, only 5 of the 16 available variables have to be evaluated in order to estimate the *Average Waiting Time*.

As previously stated, decision trees can be seen as a set of rules. The fuzzy decision tree induced for the bus rescheduling problem (Figure 4) can be seen as the following set of rules.
A Fuzzy Decision Tree for Bus Network Management

1 – If INTERVAL BETWEEN BUSES is High then Average Waiting Time is Very High

2 – If INTERVAL BETWEEN BUSES is Low and AVERAGE SPEED is Medium or High then Average Waiting Time is Good

3 – If INTERVAL BETWEEN BUSES is Low and AVERAGE SPEED is Low and NUMBER OF VEHICLES is Low then Average Waiting Time is Regular

4 – If INTERVAL BETWEEN BUSES is Low and AVERAGE SPEED is Low and NUMBER OF VEHICLES is High then Average Waiting Time is Excellent

5 – If INTERVAL BETWEEN BUSES is Low and AVERAGE SPEED is Low and NUMBER OF VEHICLES is Medium and CAPACITY is High then Average Waiting Time is Good

6 – If INTERVAL BETWEEN BUSES is Low and AVERAGE SPEED is Low and NUMBER OF VEHICLES is Medium and CAPACITY is Low or Medium then Average Waiting Time is Regular

7 – If INTERVAL BETWEEN BUSES is Medium and AVERAGE SPEED is Low then Average Waiting Time is Long

8 – If INTERVAL BETWEEN BUSES is Medium and AVERAGE SPEED is High then Average Waiting Time is Very Long

9 – If INTERVAL BETWEEN BUSES is Medium and AVERAGE SPEED is Medium and NUMBER OF VEHICLES is High then Average Waiting Time is Very Long

10 – If INTERVAL BETWEEN BUSES is Medium and AVERAGE SPEED is Medium and NUMBER OF VEHICLES is Low and AVERAGE TIME is Low or Medium then Average Waiting Time is Very Long

11 – If INTERVAL BETWEEN BUSES is Medium and AVERAGE SPEED is Medium and NUMBER OF VEHICLES is Low and AVERAGE TIME is High then Average Waiting Time is Long

12 – If INTERVAL BETWEEN BUSES is Medium and AVERAGE SPEED is Medium and NUMBER OF VEHICLES is Medium and OCCUPANCY RATE is Low then Average Waiting Time is Regular

13 – If INTERVAL BETWEEN BUSES is Medium and AVERAGE SPEED is Medium and NUMBER OF VEHICLES is Medium and OCCUPANCY RATE is High then Average Waiting Time is Long

14 – If INTERVAL BETWEEN BUSES is Medium and AVERAGE SPEED is Medium and NUMBER OF VEHICLES is Medium and OCCUPANCY RATE is Medium and NUMBER OF TRAVELLERS is Low then Average Waiting Time is Regular

15 – If INTERVAL BETWEEN BUSES is Medium and AVERAGE SPEED is Medium and NUMBER OF VEHICLES is Medium and OCCUPANCY RATE is Medium and NUMBER OF TRAVELLERS is Medium or High then Average Waiting Time is Long

Although the previous set of rules can be used to evaluate the reallocation of buses, its graphical representation as a decision tree (Figure 4) is easier and more intuitive to make inferences about a real situation and, thus, take decisions. The model was analysed by experts in urban transport and received positive feedback.

6. CONCLUSIONS AND FUTURE WORK

The planning and managing bus networks in urban public transport systems is a relevant problem. The maintenance of existing bus networks is usually carried out by human experts. In this work, we focus on the task of reallocating buses from a previously planned bus network. More specifically, we tackle the problem of restrictions in the number of extra vehicles and/or drivers in case of vehicle breakdowns or in case of absent drivers.

The combination of decision trees and fuzzy systems has generated fuzzy decision trees, which benefit from both techniques to provide simple, accurate, and highly interpretable models at low computational costs.

In this paper, we present a fuzzy decision tree model to support human experts on the task of reallocation of buses in case of buses breakdowns or absence of drivers. The model was induced using real data from a bus system with 26 bus lines. The examples contain 16 variables related to characteristics of the bus lines, such as their length and interval between buses, as well as characteristics of the travellers and the usage of the bus services, such as the average number of travellers, average number of commuting travellers, among others.
M. E. Cintra and O. A. Neves

The FUZZYDT algorithm was used to induce the fuzzy decision tree. The induced model presents a reduced set of variables (7 of the original 16 available variables). The variables used in the model include the important ones, according to experts, for the evaluation of bus lines. The final fuzzy decision tree presents five levels and a total of 15 terminal nodes (15 rules).

The model presented high accuracy rate, regarding the training example set, and has also received positive feedback from experts in urban transport.

As future work, we intend to compare the proposed model to other interpretable models, such as, classic decision trees. We also intend to further investigate the definition of the fuzzy sets used for the variables of the system.

REFERENCES


