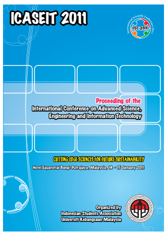
Proceeding of the International Conference on Advanced Science, Engineering and Information Technology 2011

Hotel Equatorial Bangi-Putrajaya, Malaysia, 14 - 15 January 2011

ISBN 978-983-42366-4-9



C4.5 Classifier for Solving the Problem of Water Resources Engineering

Chih-Chiang Wei[#], Jiing-Yun You[*]

[#] Department of Information Management, Toko University
No. 51, Sec. 2, University Rd., Pu-Tzu City, Chia-Yi County 61363, Taiwan, R.O.C.
Tel.: +886 5 3622889, E-mail: d89521007@ntu.edu.tw (Corresponding Author)

[*] Department of Civil Engineering, National Taiwan University
No. 1 Sec. 4, Roosevelt Rd., Taipei 106, Taiwan, R.O.C.
Tel.: +886 2 33664238, E-mail: genejyu@ntu.edu.tw

***Abstract*—The conventional decision-tree algorithm, such as ID3 and C5.0, executes rapidly and can easily be translated into if-then-else rules. This paper introduces popular classifier C4.5 for dealing with water resources engineering problem. The proposed approach was applied to the Shihmen Reservoir, which is one of the largest reservoirs, located upstream of the Tahan River Basin of northern Taiwan. The existing rules, namely M5-C rules, include two main flood stages: the peak-flow-preceding stage and the peak-flow-proceeding stage. The findings show superior performance of the C4.5 rules in contrast to M5-C rules. Accordingly, C4.5 rules can be used to improve the engineering problem.**

***Keywords*—Decision tree, Data mining, Reservoir**

I. Introduction

Decision tree algorithms are widely used in the fields of machine learning and data mining ([1]-[4]). In principle, there are exponentially many decision trees that can be constructed from a given set of attributes. While some of the trees are more accurate than others, finding the optimal tree is computationally infeasible because of the exponential size of the search space [5]. Some well-known decision trees include CART [1], ID3 [5], C4.5 and C5.0 [2], and CHAID [7], which are greed local search algorithms with trees constructed top-down ([8]-[10]).

This paper introduces popular classifier, namely C4.5, for dealing with the reservoir release problem during typhoons. The proposed approach was applied to the Shihmen Reservoir, which is one of the largest reservoirs, located upstream of the Tahan River Basin of northern Taiwan. To derive release rules, steps of the proposed methodology involve: (1) generating the optimal input-output patterns obtained by the flood-control optimization model, (2) deriving the tree-based rules by C4.5, and (3) selecting optimal tree-based rules determined by comparing current rules and tree-based rules.

II. Decision Tree Algorithm

Decision-tree algorithm identifies nuggets of information in bodies of data and extracts information in such a way that it can be used in areas such as decision support, prediction, forecasts, and estimation [11]. C4.5 uses a divide-and-conquer approach to growing decision trees that was pioneered by Hunt and his co-workers [12]. Only a brief description of the method is given here; see Quinlan [2] for a more complete treatment.

The default splitting criterion used by C4.5 is *gain ratio*, an information-based measure that takes into account different numbers (and different probabilities) of test outcomes. Let C denote the number of classes and $p(D, j)$ the proportion of cases in a set D of cases that belong to the jth class. Some test T with mutually exclusive outcomes T_1, $T_2,\ldots, T_k$ is used to partition D into subsets D_1, $D_2,\ldots, D_k$, where D_i contains those cases that have outcome T_i. The tree for D has test T as its root with one subtree for each outcome T_i that is constructed by applying the same procedure recursively to the cases in D_i [13].

The residual uncertainty about the class to which a case in D belongs can be expressed as

$$Info(D) = -\sum_{j=1}^{C} p(D,j) \times \mathrm{lo}\ \ \mathrm{g}\big(p(D,j)\big) \tag{1}$$

and the corresponding information gained by a test T with k outcomes as

$$Gain(D,T) = Info(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} \times Info(D_i) \tag{2}$$

The information gained by a test is strongly affected by the number of outcomes and is maximal when there is one case in each subset D_i. On the other hand, the potential information obtained by partitioning a set of cases is based on knowing the subset D_i into which a case falls; this split information

$$Split(D,T) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \times \log_2\left(\frac{|D_i|}{|D|}\right) \tag{3}$$

tends to increase with the number of outcomes of a test. The gain ratio criterion assesses the desirability of a test as the ratio of its information gain to its split information, as below

$$GainRatio(D,T) = \frac{Gain(D,T)}{Split(D,T)} \tag{4}$$

The gain ratio of every possible test is determined and, among those with at least average gain, the split with maximum gain ratio is selected. In other words the gain ratio expresses the number of bits gained divided by the number of bits consumed by using a certain input variable for partitioning. The variable having the highest gain ratio is considered to be best.

III. Experiment

A. Study site

Shihmen Reservoir, located upstream of the Tahan River (see Fig. 1), is a multipurpose reservoir for irrigation, hydroelectric energy generation, public water supply, flood control, and tourism. In Taiwan, the major operation objective of the Shihmen Reservoir is to mitigate hazard during typhoon invasion.

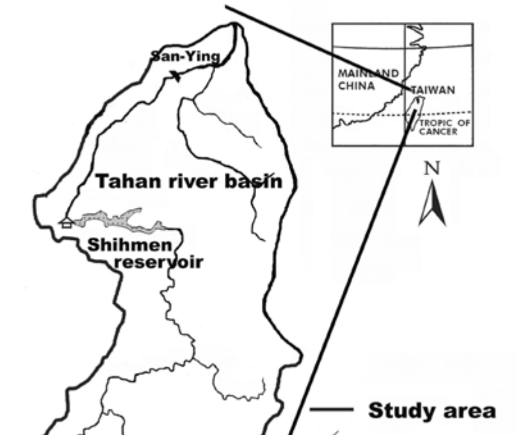


Fig. 1 Map of Tahan River Basin and Shihmen Reservoir

In order to mitigate flood damage, WRA [14] stipulated flood-control operation rules for the Shihmen Reservoir. The existing rules, namely M5-C rules, include two main flood stages: the peak-flow-preceding stage and the peak-flow-proceeding stage. The flood control operation rules adopt the release look-up tables to standardize the water releases during flood periods. These types of release rules are graded by total forecast rainfall, the observed storage level, and the reservoir inflow during flood periods. The details can be seen in WRA [14].

B. Reservoir operation optimization model

In order to generate the input-output patterns, the reservoir operation optimization model for flood control, built by Hsu and Wei [15], is applied to Shihmen reservoir system. This optimization model can identify the best amount of water released at each flood period.

The objectives of this model include: (1) maximizing the peak flow reduction at selected downstream control points; and (2) meeting reservoir storage at the end of the flood. This model follows the reservoir release policies, which present the general flood operation norms for two flood stages. The constraints include continuity equations, release policies, and capacity limitations. The model is summarized as below:

- Objective function

$$\text{Minimize } \left\{C_1 \cdot x^{Q_{\max}} + C_2 \cdot \left(S^{normal} - x_T^S\right)\right\} \tag{5}$$

where C_1 is the coefficient of the peak flow at the downstream control point; C_2 is the coefficient of the storage at the end of the flood; $x^{Q_{\max}}$ is the peak runoff at the selected control point during flood period; S^{normal} is the reservoir maximum storage volume for normal operation; x_T^S is the reservoir storage at the end of the flood; and T is flood periods.

- Constraints

1. Reservoir continuity equations

$$\frac{1}{2}\left(I_{t-1} + I_t\right) - \frac{1}{2}\left(x_{t-1}^R + x_t^R\right) = x_t^S - x_{t-1}^S \qquad 1 \le t \le T \tag{6}$$

2. Downstream flow touting equations

$$x_t^Q = c_0 \cdot x_t^R + c_1 \cdot x_{t-1}^R + c_2 \cdot x_{t-1}^Q \qquad 1 \le t \le T \tag{7}$$

3. Reservoir release policies

(1) Peak-flow-preceding stage

$$x_t^R \le I_t \qquad 1 \le t \le t_\alpha \tag{8}$$

$$0 \le x_t^R - x_{t-1}^R \le \max\left\{\left(I_{t'} - I_{t'-1}\right), t' = 1, \ldots, t\right\} \qquad 1 \le t \le t_\alpha \tag{9}$$

(2) Peak-flow-proceeding stage

$$x_t^R \le \max\left\{I_{t'}, t' = 1, \ldots, T\right\} \qquad t_\beta \le t \le T \tag{10}$$

$$x_t^R \le x_{t-1}^R \qquad t_\beta \le t \le T \tag{11}$$

4. Physical limitations

$$x_t^S = \sum_{i=1}^{V+1} \eta_{i,t} \cdot S_{l_i} \qquad 1 \le t \le T \tag{12}$$

$$S^{dead} \le x_t^S \le S^{max} \qquad 1 \le t \le T \tag{13}$$

$$x_t^Q \le x^{Q_{\max}} \qquad 1 \le t \le T \tag{14}$$

where t is the time-step (hour) index; I_t is the reservoir inflow at time t; x_t^R is the reservoir release variable at time t; and x_t^S is the reservoir storage variable at time t; x_t^Q is the downstream runoff at selected control points at time t; c_0, c_1, and c_2 are channel routing coefficients; t_α is the end time of peak-flow-preceding stage; t_β is the starting of peak-flow-proceeding stage; S_{l_i} is the linear interval boundary of reservoir storage volume; S_{l_1} is the volume of dead storage (S^{dead}); $S_{l_{(V+1)}}$ is the volume of full storage (S^{max}); and $\eta_{i,t}$ is the variable that lies between 0 and 1, and is associated with a proportion factor between adjacent bound values of storage at time period t.

C. Data and analysis

Data of 36 typhoons (1987–2004) are available. The reservoir flood control problem is formulated as a mixed-integer linear programming (MILP) model and also carried out by the optimization solver LINGO. The generated input-output patterns of the total 1,438 hourly data, including 335 records of the peak-flow-preceding stage and 1,103 records of the peak-flow-proceeding stage, can then serve as the training and testing datasets of C4.5. The patterns are split into two halves. The first half was used for training and the remaining data was for testing.

The attributes Level, Inflow, Storage, and Time were selected; meanwhile, Release was as the target. The class level was discretized into 10% steps, corresponding to the ratio of release (= Release / Peak inflow). Because the flood duration (T) per typhoon is quite various, generally ranging from 12 hours to 5 days, thus the time periods need to be dimensionless. The dimensionless formulas of the attribute Time are as follows:

$$\text{Time}(\%) = \frac{t}{2t_\alpha} \times 100 \qquad 1 \le t \le t_\alpha \tag{15}$$

$$\text{Time}(\%) = \left(\frac{t - t_\alpha}{2(T - t_\alpha)} + 0.5 \right) \times 100 \qquad t_\alpha < t \le T \tag{16}$$

where t_α is the end time of the peak-flow-preceding stage.

From Eqs. (15) and (16), the two stages of peak-flow-preceding and peak-flow-proceeding are bounded in the range of [0, 0.5] and [0.5, 1], respectively. Table 1 demonstrates a typical typhoon for the Shihmen Reservoir under optimized flood control operation.

D. Results

The C4.5 algorithm analyzed is performed using the Clementine software [11]. Results show that in the peak-flow-preceding stage the accuracy is 84.71% and 71.51% for the training and testing phases, respectively; and in the peak-flow-proceeding stage the accuracy is 83.72% and 72.73% for the training and testing phases, respectively.

Figs. 2 and 3 compare the target and prediction in the testing phase. As can be seen, the C4.5 solution can make good classification. Moreover, Figs. 4 and 5 depict the scattered plots of desired target values vs. predicted values for Stages I and II in the testing phase.

TABLE I
EXAMPLE OF TYPICAL TYPHOON FOR FLOOD CONTROL

Flood stage	Time period (hour)	Attributes				Release (mcm/hr)	Release Ratio (%)	Target (Class)
		Time (%)	Inflow (10^6 m^3)	Storage (10^6 m^3)	Level (m)			
Peak-flow-preceding	1	7.1	2.4	275.2	247.4	2.4	10	1
	2	14.3	3.8	275.2	247.4	3.8	16	2
	3	21.4	6.2	275.2	247.4	6.2	27	3
	4	28.6	11.2	275.2	247.4	11.2	49	5
	5	35.7	17.9	275.2	247.4	17.9	78	8
	6	42.9	20.1	275.7	247.5	19.0	83	9
	7	50.0	22.9	278.2	247.7	19.0	83	9
Peak-flow-proceeding	8	54.2	20.3	280.8	248.0	19.0	83	9
	9	58.3	16.6	280.3	247.9	19.0	83	9
	10	62.5	11.1	276.7	247.6	16.0	70	7
	11	66.7	8.2	272.3	247.1	12.0	52	6
	12	70.8	4.3	267.5	246.5	10.0	44	5
	13	75.0	3.6	262.5	246.0	8.0	35	4
	14	79.2	2.8	258.7	245.5	6.0	26	3
	15	83.3	2.3	256.3	245.3	4.0	17	2
	16	87.5	2.0	255.0	245.1	3.0	13	2
	17	91.7	1.8	254.4	245.0	1.9	8	1
	18	95.8	1.4	254.2	245.0	1.6	7	1
	19	100	1.3	254.0	245.0	1.5	7	1

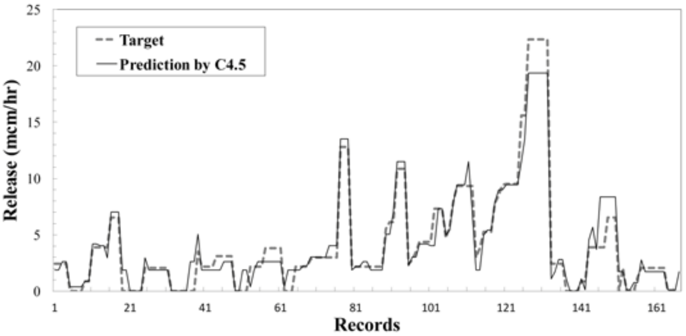


Fig. 2 Comparisons of target values vs. predicted values deducted in peak-flow-preceding stage of testing phase

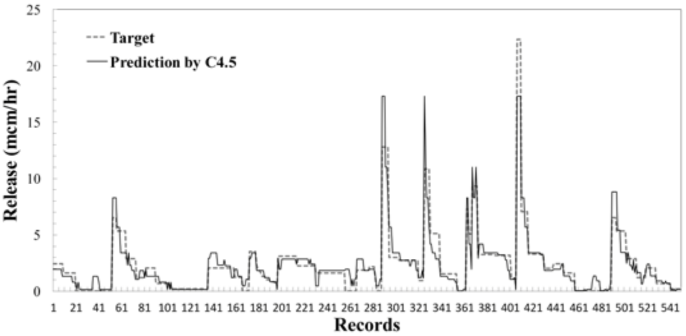


Fig. 3 Comparisons of target values vs. predicted values deducted in peak-flow-proceeding stage of testing phase

For comparison C4.5 rules with M5-C rules, the criteria used was Relative Mean Square Error (RMSE), defined as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{j=1}^{n} \left(y^{\mathrm{predict}}(j) - y^{\mathrm{target}}(j)\right)^2}{n}} \qquad (17)$$

where $y^{\mathrm{predict}}(j)$ is the predicted value deducted from tree-based rules at record j, $y^{\mathrm{target}}(j)$ is the target value at record j, and n denotes the number of hourly records.

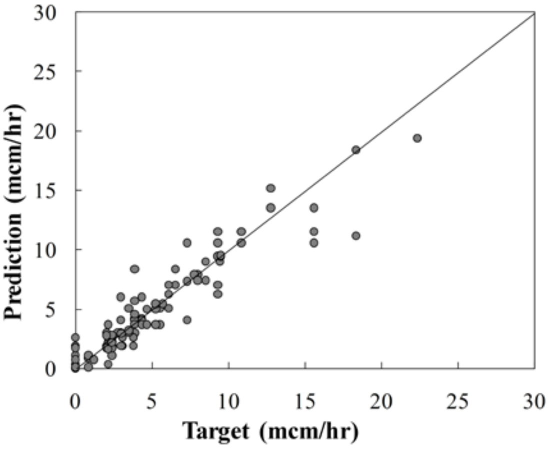


Fig. 4 Scattered plots of target values vs. predicted values in peak-flow-proceeding stage of testing phase

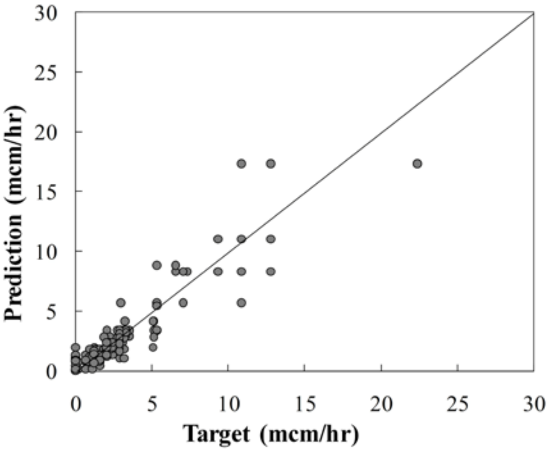


Fig. 5 Scattered plots of target values vs. predicted values in peak-flow-proceeding stage of testing phase

Results show that in the peak-flow-preceding stage the RMSE value is 1.777 mcm/hr and 2.876 mcm/hr for the testing by C4.5 and M5-C rules, respectively; and in the peak-flow-proceeding stage the RMSE value is 1.908 mcm/hr and 3.253 mcm/hr for the testing by C4.5 and M5-C rules, respectively. The findings show superior performance of the C4.5 rules in contrast to M5-C rules. Accordingly, C4.5 rules can be used to improve the engineering problem.

IV. Conclusions

Decision-tree algorithm identifies nuggets of information in bodies of data and extracts information in such a way that it can be used in areas such as decision support, prediction, forecasts, and estimation. This study aimed to develop the operation rules (decision trees) with respect to flood control during typhoons using classical, popular C4.5 algorithm.

The proposed approach was applied to the Shihmen Reservoir. The existing rules, namely M5-C rules, were used to compare with C4.5. The findings show superior performance of the C4.5 rules in contrast to M5-C rules in two main flood stages: the peak-flow-preceding stage and the peak-flow-proceeding stage. This study has successfully developed a methodology incorporated with the C5.0 algorithm for extracting the tree-based rules for flood control in Shihmen Reservoir system.

Acknowledgment

The support under Grant No. NSC 98-2111-M-464-001 from the National Science Council, Taiwan is greatly appreciated. In addition, the authors are indebted to the reviewers for their valuable comments and suggestions.

References

[1] L. Breiman, Friedman, J. H., Olshen, R. A., and Stone, C. J., 1984. Classification and regression trees. Wadsworth, Belmont, CA, U.S.

[2] J. R. Quinlan, (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo.

[3] F. Berzal, J. C. Cubero, N. Marín, and D. Sánchez, “Building multi-way decision trees with numerical attributes,” *Information Sciences*, vol. 165(1-2), pp.73-90, 2004.

[4] B. E. T. H. Twala, M. C. Jones, and D. J. Hand, “Good methods for coping with missing data in decision trees,” *Pattern Recognition Letters*, vol. 29(7), pp. 950-956, 2008.

[5] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*, Boston: Addison Wesley, pp. 193-5, 2006.

[6] J. R. Quinlan, *Discovering rules by induction from large collection of examples*, In: Michie D, ed., Expert Systems in the Micro Electronic Age, Edinburg: Edinburg University Press, 1979.

[7] G. V. Kass, “An exploratory technique for investigating large quantities of categorical data,” *Applied Statistics*, vol. 29, pp. 119-127, 1980.

[8] J. Rissanen, and M. Wax, “Algorithm for constructing tree structured classifier,” U.S. Patent 4719571, 1998.

[9] K. M. Sreerama, “Automatic construction of decision trees from data: A multidisciplinary survey,” *Data Mining and Knowledge Discovery*, vol. 2, pp. 245-389, 1998.

[10] M. Kearns, and Y. Mansour, “On the boosting ability of top-down decision tree learning algorithms,” *Journal of Computer and System Sciences*, vol. 58(1), pp. 109-128, 1999.

[11] SPSS Inc., *Clementine 7.0 user’s guide*, Chicago: SPSS, 2002.

[12] E. B. Hunt, J. Marin, and P. J. Stone, *Experiments in induction*, New York: Academic Press, 1966.

[13] J. R. Quinlan, “Improved use of continuous attributes in C4.5,” *Journal of Artificial Intelligence Research*, vol. 4, pp. 77-90, 1996.

[14] WRA, Guidelines of Shihmen Reservoir operations (in Chinese), Taoyuan: Water Resources Agency Press, 2002.

[15] N. S. Hsu, and C. C. Wei, “A multipurpose reservoir real-time operation model for flood control during typhoon invasion,” *Journal of Hydrology*, vol. 336, pp. 282-293, 2007.