The 30th Annual Conference of the IEEE Industrial Electronics Society, November 2 - 6, 2004, Busan, Korea

Application of Data Mining and Artificial Modeling for Coagulant Dosage of Water Treatment Plants Corresponding to Water Quality

Hyeon Bae[1], Student Member, IEEE, Dae-Won Choi[1], Seng-Tai Lee[1], Yejin Kim[2], Sungshin Kim[1], Member, IEEE

[1] School of Electrical and Computer Engineering, Pusan National University, Busan, Korea
[2] Department of Environmental Engineering, Pusan National University, Busan, Korea
e-mail : baehyeon@pusan.ac.kr, kadanza@pusan.ac.kr, youandi@pusan.ac.kr, yjkim@pusan.ac.kr, sskim@pusan.ac.kr

Abstract—**Shortage of water is gradually accelerated because a high standard of living is required and water resources are rapidly run dry. Therefore, effective water treatment is necessary to retain the required quality and amount of water. The general treatment includes coagulation, flocculation, filtering, and disinfections. Coagulation, flocculation, and disinfections are major parts of the water treatment processes. In this paper, new automatic algorithm is proposed for coagulation that is one of the water treatment processes. The proposed method is show how to determine the coagulant sort and amounts using data mining techniques.**

I. Introduction

Water treatment involves physical, chemical and biological changes that transform raw water into potable water. The treatment process used depends on the quality and nature of the raw water. Water treatment processes can be simple, as in sedimentation, or may involve complex physicochemical changes, such as coagulation.

Several types of chemical are applied for the coagulation in the water treatment process. In the applied field, three coagulants are usually used such as PAC, PASS, and PSO-M. The type and dosage are determined based on Jar-test and then the test result is analyzed by expert's knowledge. But this method has a disadvantage that this test needs more two hours to get the result, so it is very difficult to adapt the frequent variation of the water quality. Namely, it is not easy to make an on-line system for coagulation [1].

The coagulation effect depends on turbidity, specific ion, pH, alkalinity, and others [2]. In this paper, coagulant selection is accomplished by a decision tree model using water data of the Duksan water treatment plant and coagulant dosage is estimated by neural network models that perform mapping between the water quality such as pH, turbidity, alkalinity, water temperature, and others and coagulants such as PAC, PASS, and PSO-M.

In past studies, artificial intelligence was imported for coagulant dosage in the water treatment process. Sugeno et al. compared and validated the control performance based on the operator's experience and fuzzy logic [3]. Baba et al. designed the visual function to analyze the coagulation result with a fuzzy controller that was a operation support system [4], Enbutsu et al. used the training algorithm of neural networks and proposed the advanced fuzzy rule generation method to improve the performance [5]. As shown in researches corresponding to coagulation of Korea, neuro-fuzzy model or neural network model was employed to design the models [6-8]. But in the past studies, only one chemical was dealt for coagulation, that is, there are no fluent studies. The equipment and system of the water treatment plant is developed much and the several coagulants are applied. And the high efficient control of coagulants is required for the high quality water supply. And the traditional method of coagulant selection has the other weak point that the Jar-test for determining the type and dosage of coagulants requires at least over 2 hours, so it is difficult to build an automatic dosage system with the traditional technique. To avoid this problem, data mining and artificial intelligence are employed to construct the coagulant dosage system that can adapt the timely variable system and control the coagulants automatically.

II. Water Treatment Process

A. Concept of Water Treatment Process

A water treatment plant contains several processes that are from water resource to custom for water feeding. The water treatment plant should have ability to purify water for standard quality and supply to customs even though the quality of the water source gets worse. The major processes of the water treatment plant consist of following unit processes as shown in Fig. 1.

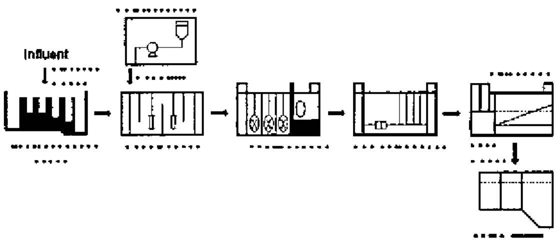


Fig. 1. Concept of water treatment processes.

B. Water Treatment Process: Coagulation

In an influent basin, several inspection of water that is from primary chlorination is performed with an automatic inspection technique. In a mixing basin, water is treated by optimal amounts of coagulants (PAC, PASS, PSO-M, etc.) and distributed according to basin capacity of the settling basin. And multi-resource water is mixed.

Coagulation is the destabilization of colloidal particles brought about by the addition of a chemical reagent called as coagulant. Flocculation is the agglomeration of destabilized particles into microfloc and after into bulky floccules which

0-7803-8730-9/04/$20.00 ©2004 IEEE

can be settled called floc. The addition of another reagent called flocculant or a flocculant aid may promote the formation of the floc. Coagulation and flocculation are defined as follows [9].

TABLE I. DEFINITION OF COAGULATION AND FL●CCULATION.

Items	Coagulation	Flocculation
Coagulant	Polymeric salt	Organic polymer
Destabilization	High	Very low
Floc	Not formed	Growing
Mixing condition	Faster	Slower
Zeta potential	Big change	Very small

C. Affecting Factors to the coagul●tion

1) Alk●linity

Coagulants have two primary actions in water treatment: they neutralize negative particles in raw water and they react with alkalinity present to form a metal hydroxide floc. Traditional coagulants are acidic and so decrease system alkalinity. The coagulant used can affect the alkalinity adjustment in low alkalinity water. The alkalinity consumed by different coagulants can vary greatly. For example, ferric chloride (dry basis) consumes 1 ppm alkalinity/ppm coagulant, alum (dry basis) consumes ●.5 ppm alkalinity/ppm coagulant, while a PAC requires only ●.088 ppm alkalinity/ppm coagulant.

2) pH

The primary coagulants used in potable water treatment hydrolyze in water and ev●lve aluminum or ferric hydroxide floc. Hydrolysis speed and efficiency depend on pH, alkalinity and water temperature. Speed varies from less than 0.001 second to more than 1 second. Efficiency increases if more insoluble hydroxides form for entrapment and if the hydrolyzed products have a higher charge that enhances neutralization. Alum removes turbidity best between pH 6.0 and 7.5.

3) Chl-a

Appearance of algae increase pH of water and the algae are one of factors that makes difficult to treat water. If there is algae appearance in water resources, dosage of chlorine and coagulants should be increased. When the algae appear, pH is increased and functional group COO- creates complex compounds with Al ions. Thus, dosage Al coagulants cannot perform coagulation well.

4) Temper●ture

The temperature dependence of most chemical reactions stems from the activation energy associated with them. In general, the rates of chemical reactions decrease with decreasing temperature. Turbidity and color are indirectly related to temperature, because temperature affects coagulation. The efficiency of coagulation is strongly temperature dependent, and the optimum pH for coagulation decreases as temperature increases. This shift is probably important only when coagulant doses are close to the experimentally determined minimum. As a longer settling time is not available in a plant with a fixed flow rate and basin capacity, the efficiency of color and turbidity removal by coagulation and sedimentation may be less in winter than in summer.

5) Turbidity

Turbidity is more than an optical parameter. Although it is defined by light scattering, it relates broadly to the nature of the particles present. The form of turbidity present affects coagulant selection and dosage. Plants that have trouble removing turbidity must determine if the cause rests with the form of turbidity or with plant factors. This can involve a degree of detective work. Consider a common problem in which turbidity that is easily treated at one coagulant dose at one time does not respond well to multiple doses of the same coagulant at another time. If pH is much above 3.5 to 4.0, the coagulant it contains will hydrolyze and floc with the carrier water. The solution added to the raw water will be partially spent and will lose a part of its ability to remove turbidity.

D. Co●gulant Types

Recently, inorganic coagulants of Al family are employed in most of the water treatment plant in Korea. These coagulants show the very good performance for coagulation in the water treatment plants. The typical inorganic coagulants of Al family are as follows.

1) PAC

PAC (Poly Aluminum Chloride) is the second-generation coagulant. This coagulant also shares high market portion in Korea. For example, 100% of PAC is applying in the water treatment plant in Seoul. But, there is a problem that the stability of the product is very short. If the product is manufactured incorrectly, the product will make sedimentation and formation of sludge in the pipes less than six months. This can be cause the reduction of use. Because the coagulation capacity is very good, PAC is broadly used in the coagulation process in spite of this problem.

2) PASS

PASS (Poly Aluminum Sulfate Silicate) is developed in Canada that is inorganic polymer coagulant including *Si* and whose molecular weight is from 100,00● to 3●●,000. Because it contains *Si*, sedimentation ability and removing effect of micro turbid materials is very excellent. Especially, the coagulation capacity of PASS is improved in the winter. Adding Si can improve the capability of organic removal and effect of sedimentation.

3) PSO-M

PSO-M (Poly Organic Aluminum Magnesium Sulfate) has 3.0 of pH and specific gravity is 1.22. This coagulant is used under the general condition of water source because its price is cheaper than other coagulants. This coagulant has depression effect of increased pH that is caused by the propagation of algae at the dry season. This coagulant is not easily frozen and suitable for rapid variation of the water quality because the variation of coagulation corresponding to pH variation is very wide.

III. Data Mining

A. Basic concept of data mining

Data mining is searching and analysis processing that seeks ill-defined rules or relationship using automatic approaches from large database [10]. This method reduces the human's favor and uses the artificial intelligence to extract the rules and relationship mechanically that is called as machine learning technique [11].

The data analysis has three groups in general. The first group includes logistic regression analysis, multivariate discriminant analysis and others that are traditional statistical approaches, the second group contains decision tree, decision rules, and others that are machine learning techniques, and the final group consists of back-propagation, Kohonen self-organizing network, and others that are artificial neural network methods. The neural network and decision tree are generally applied for classification in data mining.

1) Decision tree

A decision tree algorithm was applied to select the coagulant type. The decision tree is a map of the reasoning process. It can be used to explain why a question is being asked. The following decision tree assumes that questions are answered with a certain yes or no. A tree that allows answering with a partial yes or no would have a much larger number of end nodes. In real world problems, the intuition of a human expert, or expert system software, is necessary to determine the likely end node. Each end node represents a situation with known effective and efficient leadership styles.

2) Neural network

A neural network model was employed to design the prediction model of amounts of coagulant. The neural network is an information-processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. Neural networks, like people, learn by example.

IV. Experimental Results

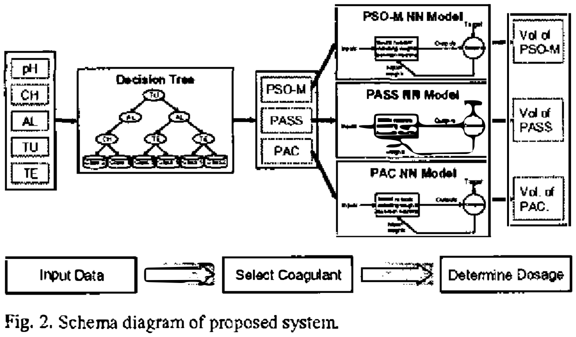


Fig. 2. Schema diagram of proposed system.

A. Characteristics of Applied Water Data

Fig. 3 is physical data that is from the Duksan water treatment plant. The data consist of temperature, pH, turbidity, alkalinity, and Chl–a that can be rapidly changed respect to the seasons. Especially, the variation of the turbidity and alkalinity is very severe between 106 and 200. This is caused by heavy rain and typhoon in summer. The all of inputs are very important for data modeling but some input variables show the closer relationship with coagulant selection. For example, the turbidity and water temperature have strong relationship with selected coagulants.

Fig. 4 shows the dosed coagulants. As shown in Fig. 4, the type and amount of coagulants are dependent on the season. The variation of coagulants between 100-day and 200-day is significantly different. The designed decision tree classifies the patterns. The model generates the If-Then rules indicating the relationship between inputs and outputs. This modeling is one of major goals in this paper.

The applied data is 2003-year data of the Duksan water treatment plant in Busan. These data were collected by Jar-test manually because the type and amount of coagulants are determined by the Jar-test in the fields. Coagulant selection is achieved by considering pH, turbidity (Tu), and alkalinity (Al) of daily data of the water quality. In the selection, the pH and alkalinity are sub-index variables. On the other hand, the turbidity affects strongly coagulant dosage.

If the turbidity become over 50NTU, PAC is used, if algae is severe, PASS is applied, and besides previous two cases, PSO-M is employed. The price of PASS and PSO-M is similar, but PAC is more expensive than both. The both coagulants are separately used corresponding to the algae or alkalinity. But above knowledge is not exactly correct in the physical field. Therefore, the data-based analysis is necessary in the rule generation for coagulation.

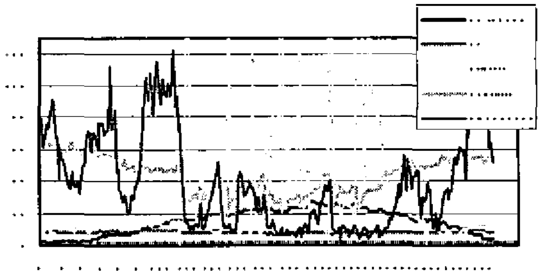


Fig. 3. Data trend of influent water for 1 year (1 data/day).

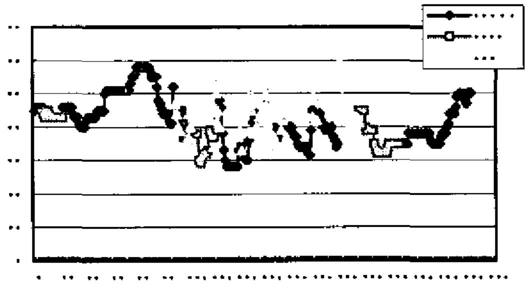


Fig. 4. Dosing coagulant corresponding to input conditions.

B. Application of Decision Tree for Rule Extraction

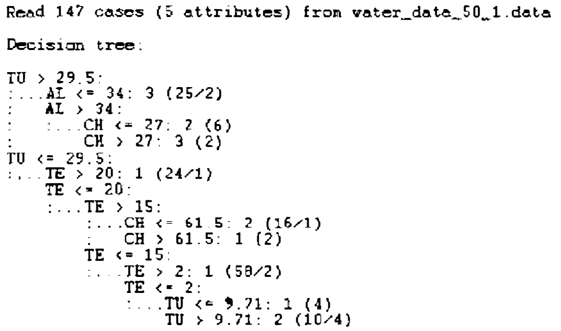
pH, turbidity, alkalinity, cha-a, and temperature are used for inputs of decision tree models. The inputs affect the coagulant selection in the field. In this paper, inputs are five and outputs are three for the models. Three outputs contain PSO-M, PASS, and PAC that are applied coagulant in the Duksan water treatment plant in Busan.

The general-purpose model needs to be considered importantly when construct the rule and model. That is, over-fitted trained model can make bad performance with respect to the unseen data that are not applied in modeling. Therefore, in this paper, training data and testing data are divided by two data sets from the original data set and the model is constructed using these two sets.

1) Training: 50%, Testing: 50%, Pruning: 25%

50% data of 2003-year data were applied in the model training and the decision tree model was constructed by 25% pruning. The remaining 50% data were employed for the model validation and error values were calculated for the model performance. Fig. 5 shows the generated rules and Fig. 6 shows the testing results.

As shown in Table 2, the testing result is much worse than that of training. This is called as insufficient training that is caused by leakage of training cases.



```
Read 147 cases (5 attributes) from water_date_50_1.data

Decision tree:

TU > 29.5:
:...AL <= 34: 3 (25/2)
:   AL > 34:
:   :...CH <= 27: 2 (6)
:       CH > 27: 3 (2)
TU <= 29.5:
:...TE > 20: 1 (24/1)
    TE <= 20:
    :...TE > 15:
        :...CH <= 61.5: 2 (16/1)
        :   CH > 61.5: 1 (2)
        TE <= 15:
        :...TE > 2: 1 (58/2)
            TE <= 2:
            :...TU <= 9.71: 1 (4)
                TU > 9.71: 2 (10/4)
```

Fig. 5. 50% of training data (pruning 25%).

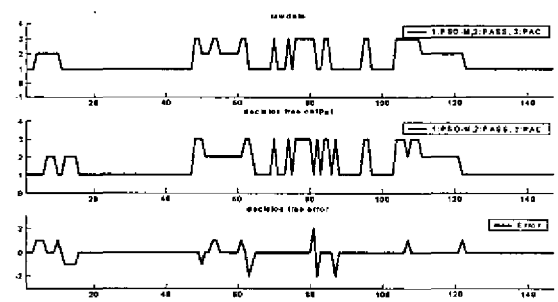


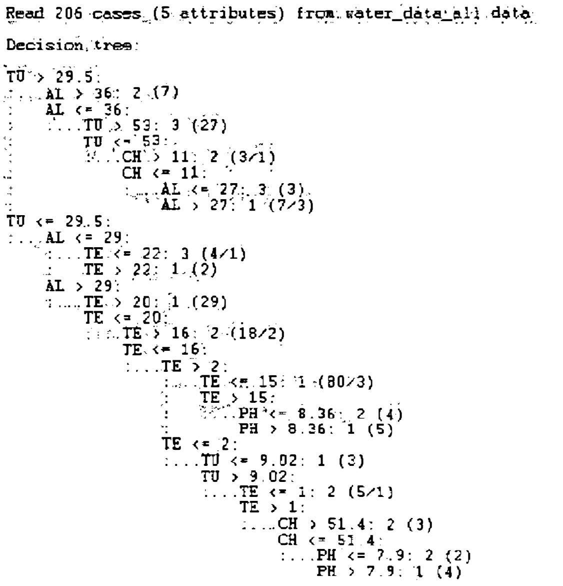
Fig. 6. Raw data, result graph of decision tree, and error graph.

2) Training: 70%, Testing: 30%, Pruning: 70%

70% data of 2003-year data were applied in the model training and the decision tree model was constructed by 70% pruning. The remaining 30% data were employed for the model validation and error values were calculated for the model performance. Fig. 7 shows the generated rules and Fig. 8 shows the testing results.

As shown in results, the performance of the trained model with 70%:30% data ratio is better than that of trained model with 50%:50% data ratio. And in this research, we considered pruning ratio of the decision tree that influences the model performance and controls the over-fitting. In the first model, 25% of pruning was performed but in this model, 70% of pruning was achieved that means the first model is simpler than the second model.

As shown in Table 2, the result of 25% pruning is worse than that of 70% pruning. From this result, we can infer that the proper number of rules is able to design a good model for the nice performance.



```
Read 206 cases (5 attributes) from water_data_all.data

Decision tree:

TU > 29.5:
:...AL > 36: 2 (7)
:   AL <= 36:
:   :...TU > 53: 3 (27)
:       TU <= 53:
:       :...CH > 11: 2 (3/1)
:           CH <= 11:
:           :...AL <= 27: 3 (3)
:               AL > 27: 1 (7/3)
TU <= 29.5:
:...AL <= 29:
    :...TE <= 22: 3 (4/1)
    :   TE > 22: 1 (2)
    AL > 29:
    :...TE > 20: 1 (29)
        TE <= 20:
        :...TE > 16: 2 (18/2)
            TE <= 16:
            :...TE > 2:
                :...TE <= 15: 1 (80/3)
                :   TE > 15:
                :   :...PH <= 8.36: 2 (4)
                :       PH > 8.36: 1 (5)
                TE <= 2:
                :...TU <= 9.02: 1 (3)
                    TU > 9.02:
                    :...TE <= 1: 2 (5/1)
                        TE > 1:
                        :...CH > 51.4: 2 (3)
                            CH <= 51.4:
                            :...PH <= 7.9: 2 (2)
                                PH > 7.9: 1 (4)
```

Fig. 7. 70% of training data (pruning 70%).

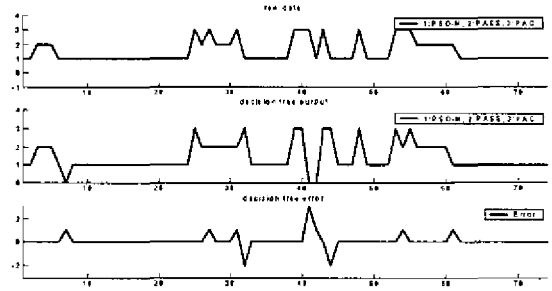


Fig. 8. Raw data, result graph of decision tree, and error graph..

TABLE 2. RESULTS OF DECISION TREE.

Data condition (Learning:Testing)	Pruning	Error	
		Learning	Test
70:30	25%	8.3	9.1
	70%	5.3	8.0
50:50	25%	6.8	17.0
	70%	5.4	13.6

C. Model Construction for Coagulant Prediction

Variables of the water quality are water temperature, pH, turbidity, alkalinity, and Chl-a of influent water in the mixing basin. Applied dosage data of coagulant are determined by the expert with Jar-test. The water temperature is ranged from 1 to 27 that is varied corresponding to the season and pH is changed from 6.7 to 9.7. The turbidity is from 5NTU to 15NTU on average but the value become over 609NTU at the storming day. The alkalinity is 21 to 64mg/L and the Chl-a is 4 to 30mg/L in summer or autumn but 30 to 126 in spring or winter.

PAC, PASS, and PSO-M were applied for the coagulants. Under general condition, PSO-M was usually employed; PAC was applied under the case that the turbidity became over 5●NTU; and PASS was used on Jan., May, and Oct. The applied coagulants in this research are based on the dosed in the physical field that is the Duksan water treatment plant. Therefore, the performance of the system needs to be considered by some errors.

The prediction models are constructed with respect to the three coagulants. The decision tree model in the prior step selects one of them. The neural network models predict the coagulant amount for the economic treatment.

1) Prediction Model for PSO-M

The prediction model of PSO-M consisted of water temperature, pH, turbidity, alkalinity, and Chl-a. These input variables were normalized by 0.1 to 0.9 values. Total 186-day data of the water treatment plant were separated by 50% of training data and 50% of testing data. The test is for model validation. Training epochs were 1,000. The number of hidden layers is shown in Table 3.

Fig. 9 shows the testing result with the PSO-M prediction model. As shown in Fig. 9, the estimation result (dash line) has very small error with the raw data of the PSO-M amount (hard line). This means that the estimation model for PSO-M is reasonable for the water treatment system.

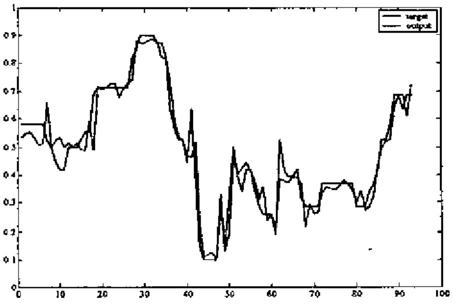


Fig. 9. Simulation result of PSO-M (training 5●%: testing 5●%).

2) Prediction Model for PASS

The prediction model of PASS consisted of water temperature, pH, turbidity, alkalinity, and Chl-a. These input variables were normalized by 0.1 to 0.9 values. Total 58-day data of the water treatment plant were separated by 50% of training data and 50% of testing data. The test is for model validation. The number of hidden layers is shown in Table 3.

Fig. 10 shows the testing result with the PASS prediction model. The performance of PASS is worse than that of PSO-M because the PSO-M data is not continues. PSO-M is usually dosed on Jan., May, and Oct.

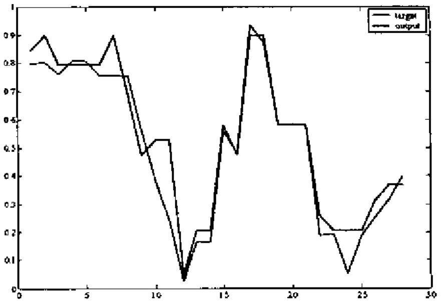


Fig. 10. Simulation result of PASS (training 50%: testing 50%).

3) Prediction Model for PAC

The prediction model of PAC consisted of water temperature, pH, turbidity, alkalinity, and Chl-a. These input variables were normalized by ●.1 to 0.9 values. Total 50-day data of the water treatment plant were separated by 50% of training data and 50% of testing data. The test is for model validation. The number of hidden layers is shown in Table 3.

Fig. 11 shows the testing result with the PAC prediction model.

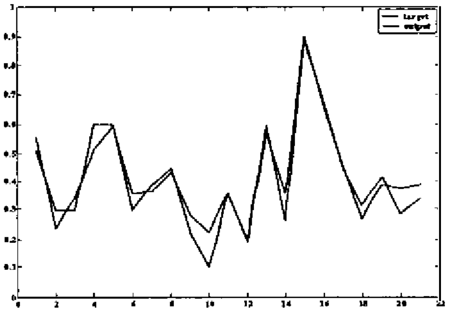


Fig. 11. Simulation result of PAC (training 50%: testing 50%).

TABLE 3. NUMBER ●F HIDDEN LAYERS.

Coagulant	Number of 1^{st} hidden layer	Number of 2^{nd} hidden layer
PSO-M	3	4
PASS	2	3
PAC	1	2

RMSE (Ro●t Mean Square Error) was employed to evaluate the performance of simulation. Table 4 shows the prediction results of each coagulant using RMSE. As shown in Table 4, the prediction results for three coagulants are reasonable for the proposed system.

The result of the PSO-M model shows the best performance because training data of PSO-M are much more than these of PASS and PAC. If there is some more data of PAC or PASS, the performance of the model should be better. But the model performance is not only determined by the data amount but also by data characteristics. PAC data are less than PASS data but the performance of PAC is better than that of PASS. This is caused by the data characteristics. In other words, the pattern of PAC data is more significant than that of PASS to be estimated by the prediction models.

$$RMSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - o_i\right)^2 \qquad (1)$$

TABLE 4. RESULTS ●F MODEL PREDICTI●N.

Coagulant	Data conditions (Training: Testing)	RMSE
PSO-M	50%:50%	0.0●58
	70%:30%	0.0100
PASS	50%:50%	0.0157
	70%:30%	0.0153
PAC	50%:50%	0.0118
	70%:30%	0.0156

V. CONCLUSIONS

In this paper, the decision tree was applied to select the proper coagulant corresponding to the input conditions and neural network models were employed to predict the amount of each coagulant. This proposed system can rapidly adapt to the variation of the water factor of influent.

The result of decision tree sh●wed the good performance in coagulant selection but over-fitting could occur because of leakage of training data. That is a weak point of the decision tree modeling. Therefore, the cross validation or pruning is necessary to avoid the over-fitting. In this paper, also the prediction models were designed to estimate the amount of coagulants. The model performance was calculated by the RMSE. As considering this RMSE, the built models showed the reasonable results. That is, the proposed modeling can be suitable to estimate the coagulant volume.

VI. REFERENCES

[1] Korea Institute of Water and Environment, Development of an Automatic Decision System of Coagulant Dosage (the second year), Technical Report, Korea Water Resources Co., pp. 14-17, 1977.

[2] A. P. Black S. A. Hamnah, "Electrophoretic Studies of Turbidity removal Coagulant with Aluminum Sulfate," J. AWWA, vol. 53, p. 438, 1961.

[3] Yagishita, O., Itoh, O., Sugeno, M., "Application of fuzzy reasoning to the water purification process," Industrial Application of Fuzzy Control, pp. 19-39, 1985.

[4] K. Baba, I. Enbutu, H. Matuzki, and S. Nogita, "Inteligent suport system for water and sewage treatment plants which includes a past history learning function-coagulant injection guidance system using neuralnet algorithm," Instrumentation, Control and Automation of Water and Wastewater Treatment and Transport systems (IAWPRC), pp. 227-234, 1990.

[5] I. Enbutsu, K. Baba, N. Hara, K. Waseda, and S. Nogita, "Integration of multi AI paradigms for intelligent operation support systems-fuzzy rule extraction from a neural network," Instrumentation, Control and Automation of Water and Wastewater Treatment and Transport systems (IAWQ), pp. 333-34●, 1993.

[6] B. J. Lee, "A Study on Determination Model of Coagulant PAC Dosing Rate using a Neural Network Theory," M.S.' thesis, Chonnam National University, pp. 34-41, 1992.

[7] B. Y. Park, "A Study on Determination Model of Coagulant Dosing Rate in Water Treatment Process using a Neural Network Theory," M.S.' thesis, Chonnam National University, pp. 45-52, 20●0.

[8] H. S. Kim and S. H. Kim, "A Study on ●ptimum Coagulant Feeding Rate using Jar-test," Journal of the Korean Society of Water and Wastewater, vol. 2, pp. 39-45, 1993.

[9] J. ●. Kwag, Principle and Application of Physical and Chemical Water Treatment. Gisam publisher, pp. 192-2●9, 1998.

[10] M. Berry and G. Linoff, Mastering Data Mining. NY: Wiley, 200●.

[11] N. S. Jang, S. W. Hong, and J. H. Jang, Data Mining. Daechung Media, pp. 25-27, 1999.