JOURNAL OF THE KOREAN SOCIETY OF AGRICULTURAL ENGINEERS DEC. 2005, pp. 13-22

JOURNAL OF THE KOREAN SOCIETY
OF AGRICULTURAL ENGINEERS,
Vol. 47, No. 7

Reservoir Classification using Data Mining Technology for Survivor Function

Park Mee Jeong* · Lee Joon Gu · Lee Jeong Jae*****

Abstract

Main purpose of this article is to classify reservoirs corresponding to their physical characteristics, for example, dam height, dam width, age, repair-works history. First of all, data set of 13,976 reservoirs was analyzed using k means and self organized maps. As a result of these analysis, lots of reservoirs have been classified into four clusters. Factors and their critical values to classify the reservoirs into four groups have been founded by generating a decision tree. The path rules to each group seem reasonable since their survivor function showed unique pattern.

Keywords : Classification, Cluster, Life time, Probability

I. Introduction

South Korea has 17,882 reservoirs in rural area and more than half of them are very old because they were constructed before 1940s. Thus, it's difficult to examine all the reservoirs whether they need repairs or not. A summarized examination have been regularly conducted to all reservoirs and a few reservoirs are examined thoroughly even if thorough examination is sophisticated and reliable. It needs to set a criterion to decide which reservoirs are examined thoroughly. Some researchers have suggested life cycle cost analysis as a systematical management strategy, and Kim (2003) applied survivor function into irrigation system to evaluate serviceability of the agricultural irrigation systems. However, that survival function did not respect irrigation systems characteristics, for example, its location, benefitted area, capacity and so on, and a unique survival function was proposed even though that can be varied according to their physical characteristics and management history and so on.

Reservoir classification is useful to make a reliable life cycle cost analysis. Reservoirs currently have been classified according to service, dam height and material. According to

* Research Institute for Agriculture and Life Sciences
** KARICO
*** Department of Agricultural Engineering, Seoul National University
*** Corresponding author. Tel.: +82-2-880-4592
Fax: +82-2-873-2087
E-mail address: ljj@snu.ac.kr

service, they are divided into four classes; stage dam, storage dam, multipurpose dam and barrier dam. And they are classified into seven groups by material and into 2 groups by dam height of 1.5 m. However, their survivor function can differ in accordance with how frequently they do service even if the group is within same purpose. Since drainage basin, crops, rainfall frequency, drought and so on determine reservoir service frequency. Finally, several survivor functions that depend on the characteristics of a reservoir and service area should be suggested. It helps to make their survivor functions more reliable and manage them discriminately.

This article is aimed to propose classification parameters and their value of irrigation system. It focuses on reservoir as agricultural irrigation systems and shows classification reasonability by the significance of survivor functions of classified group.

II. Fundamental Theories and Clustering Algorithm

1. Cluster Analysis

Classification is the process of finding out parameters and their values where the class label of each object is known. Unlike classification, clustering is the process of grouping the data whose class label is not known into classes or clusters so that objects within a cluster have high similarity in comparison with one another, but are dissimilar to objects in other clusters. Therefore, clustering is proper to group the reservoirs because we don't know even how many classes are suitable to divide.

The most well known and commonly used portioning method is k means. The k means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low. Cluster similarity is measured with regard to the mean value of the objects in a cluster, which can be viewed as the cluster's center of gravity. The k means algorithm proceeds as follows. It randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, the squared error criterion is used, defined as eq. (1). This process is summarized in Fig. 1.

$$E=\sum_{i=1}^{k}\sum\nolimits_{p\in C_i}|p-m_i|^2 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots (1)$$

where, P: a data point in space representing a given object, m_i: mean of cluster C_i

There is one problem on cluster number. If the data set can be represented in two dimensional coordinates, analyzer can decide clusters number as shown in Fig. 2, but most data set comprises more than two parameters. So it is impossible to find out clusters number intuitively with graph in a coordinate system. Self Organizing Maps (SOM) as another clustering method can solve this problem with a feature map. This method has advantage on projecting n dimensional data set into one or two dimensional data set. It is useful

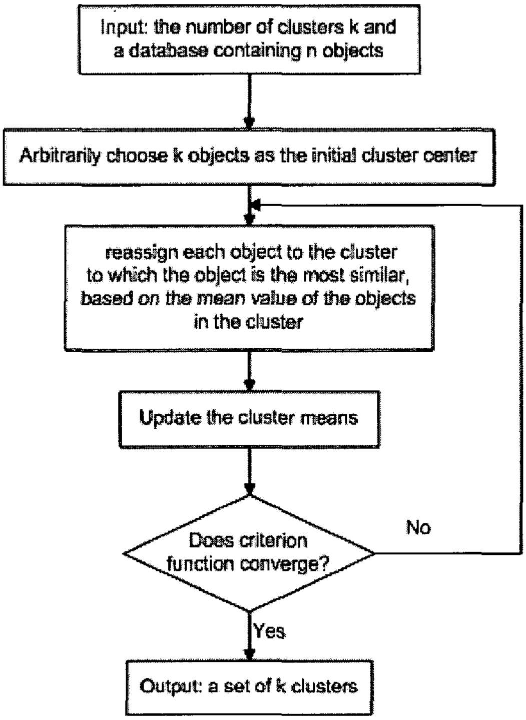

Input: the number of clusters k and a database containing n objects
Arbitrarily choose k objects as the initial cluster center
reassign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster
Update the cluster means
Does criterion function converge?
No
Yes
Output: a set of k clusters

Fig. 1 k means procedure

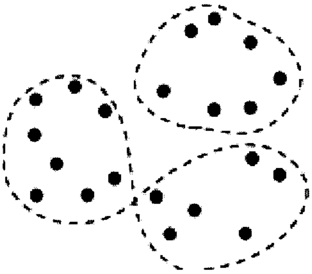


Fig. 2 Class number of a set of objects

to decide clusters number, but clustering process takes long time. So we only use SOMs to determine k as input of k means with a feature map and then employ k means as a grouping process.

2. Classification Analysis

Data classification is for describing a predetermined set of data classes or concepts. Common method for classification is a decision tree induction. A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute. The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top–down recursive divide and conquer manner. The algorithm is summarized in Fig. 3. The tree starts as a single node representing the training samples. If the samples are all of the same class, then the node becomes a leaf and is labeled with that class. Otherwise, the algorithm uses an entropy based measure known as information gain as a heuristic for selecting the attribute that will best separate the samples into individual classes. In this version of the algorithm, all attributes are categorical, that is discrete valued. Continuous valued attributes must be discretized. A branch is created for each known value of the test

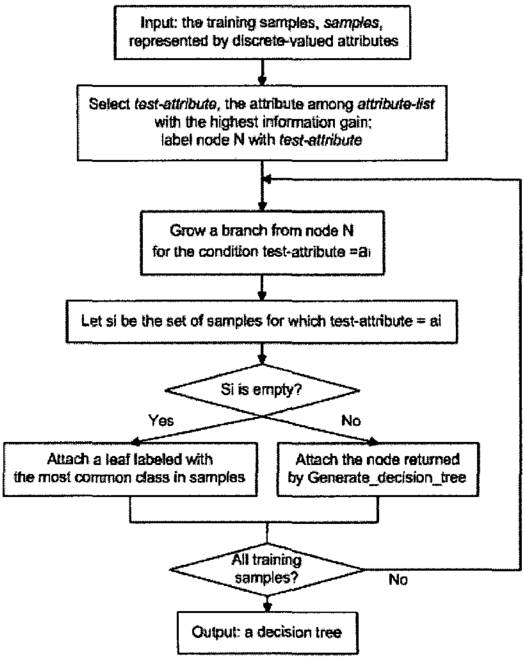

Input: the training samples, samples, represented by discrete-valued attributes
Select test-attribute, the attribute among attribute-list with the highest information gain; label node N with test-attribute
Grow a branch from node N for the condition test-attribute =ai
Let si be the set of samples for which test-attribute = ai
Si is empty?
Yes
No
Attach a leaf labeled with the most common class in samples
Attach the node returned by Generate_decision_tree
All training samples?
No
Output: a decision tree

Fig. 3 Classification algorithm

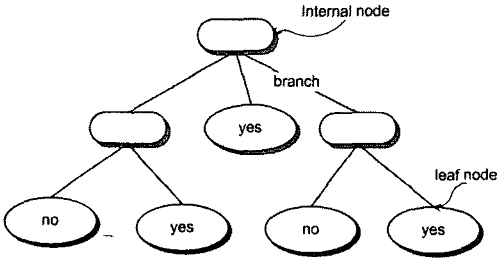


Fig. 4 Decision tree

attribute, and the samples are partitioned accordingly. The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once an attribute has occurred at a node, it need not consider in any of the node's descendents. The recursive partitioning stops only when any one of the following conditions is true:

(a) All samples for a given node belong to the same class, or

(b) There are no remaining attributes on which the samples may be further partitioned. In this case, majority voting is employed. This involves converting the given node into a leaf and labeling it with the class in majority among samples. Alternatively, the class distribution of the node samples may be stored.

A decision tree is widely used to predict and classify data since it's easy to understand rules to show that with trees. There are many algorithms to build a tree, out of them, CART (classification and regression trees) and CHAID (chi squared automatic interaction detection) are most popular. C4.5 algorithm is widely used because this builds more understandable rules without so many computations. This algorithm can treat with categorized variables as well as continuous variables. We employed this algorithm to find out classification rules building decision tree.

3. Reservoir Clustering and Classification

As a data set of reservoir objects is unlabeled, clustering process should be conducted. The number of classes in a data set heavily influences the performance of a cluster analysis. So it's more important to decide suitable class numbers. There are two kinds of methods, data observation and evaluating the clustering performance. Self Organizing Maps enables data observation. SOMs are an analytical tool that provides a topological mapping from input space to the clusters. In SOMs, the clusters are organized into a grid that is usually two dimension. A feature map through topological mapping show how many cluster exist in the data set. On the contrary, inter-intra ratio of k-means is numerical value. This value represents the ratio of a inter cluster dissimilarity to intra cluster dissimilarity computing. The lower this value is, the better performance is. After we decide parameter, *k*, of *k*-means using these all methods, we cluster a data set. Then, a decision tree is generated to describe a predetermined set of data classes or concepts. This process is summarized as following.

① Exporting input variables with a correlation analysis

② Preprocessing such as transformation, imputation of selected input data

③ Making a feature map using SOMs to decide clusters number k, of k means

④ Compute inter/ intra ratio of k means

⑤ Deciding k comparing feature map of 3rd step and inter/intra ratio of 4th step.

⑥ Clustering a data set of reservoir objects with k means

⑦ Making path rules of each group by building a decision tree.

4. Data Investigation and Preprocessing

1) Data Investigation

The data for clustering and classification of reservoirs comes from database of KARICO in the year of 2000. That comprises repair management history which composed of 19 fields such as hydrosphere name and location, benefitted area, reservoir capacity, surface area, embankment height and length, constructed year, and so on with 13,976 records and physical characteristics which are 12 fields, 3,312 records and through examination data which composed of 13 fields, and 975 records. Those data gathered from 3,358 reservoirs which KARICO has managed. In this paper, 21 variables were selected in three databases.

Our data may be highly susceptible to noisy, missing and inconsistent due to its very huge size. Data cleaning may be applied to remove

Table 1 Problem formulation: 21 fields 3,358 records

Name	Model role	Measurement	Type	Description
CODE	id	Interval	num	code
ATTACHN	input	binary	num	number of affiliated facilities
COEYR	input	interval	num	constructed year
NEIGHBOR	input	interval	num	dwelling place area
INIRRIA	input	interval	num	irrigated area
NETIRRIA	input	interval	num	net-irrigated area
CHANNELA	input	ordinal	num	channel area
OUTIRRIA	input	interval	num	outside area
DROUGHT	input	ordinal	num	dıought frequency
WATSH	input	interval	num	drainage area
VALISTO	input	interval	num	active storage
FILLSTO	input	interval	num	surface area of reservoir
HEIGHT	input	interval	num	dam height
LENGTH	input	interval	num	length of dam
REPAIRT	input	interval	num	rehabilitation time
REPAIRB	input	binary	num	whether it remains as it is or not
ATTACHT	input	unary	char	kind of a affiliated facility
SOURCE	input	nominal	char	type of hydrosphere:
COSFUND	input	nominal	char	kind of financial resources
STRUCT	input	nominal	char	structural type of dam
REPFUND	input	nominal	char	financial resources for rehabilitation
REPSP	input	nominal	char	rehabilitation history

Table 2 Distribution of values for each of the interval variables

Variable	COEYR	NEIGHBOR	INIRRIA	NETIRRIA	CHANNELA	OURIRRIA
Min.	0	0	0	0	0	0
Max.	99	4,300	15,143	15143	973.1	280.4
Mean	57.44	235.38	154.18	154.18	2.6168	0.8913
Std. Dev	15.533	1,677.3	698.3	695.19	40.736	8.54123
Skewness	0.5249	19.563	12.846	12.945	18.825	22.576
Kurtosis	0.7435	452.77	206.38	209.06	383.5	638.86
Variable	WATSH	VALISTO	FILLSTO	HEIGHT	LENGTH	REPAIRT
Min.	1	0	0	0	0	1
Max.	982,800	180,900	3,650	83	7,807	99
Mean	1,204.6	843.71	16.705	9.9857	191.77	85.264
Std. Dev	24,820	6,255.1	129.26	8.7287	313.5	21.249
Skewness	36.604	19.473	23.353	2.1113	12.602	-3.431
Kurtosis	1,408.5	458.48	605.43	7.5156	237.98	10.968

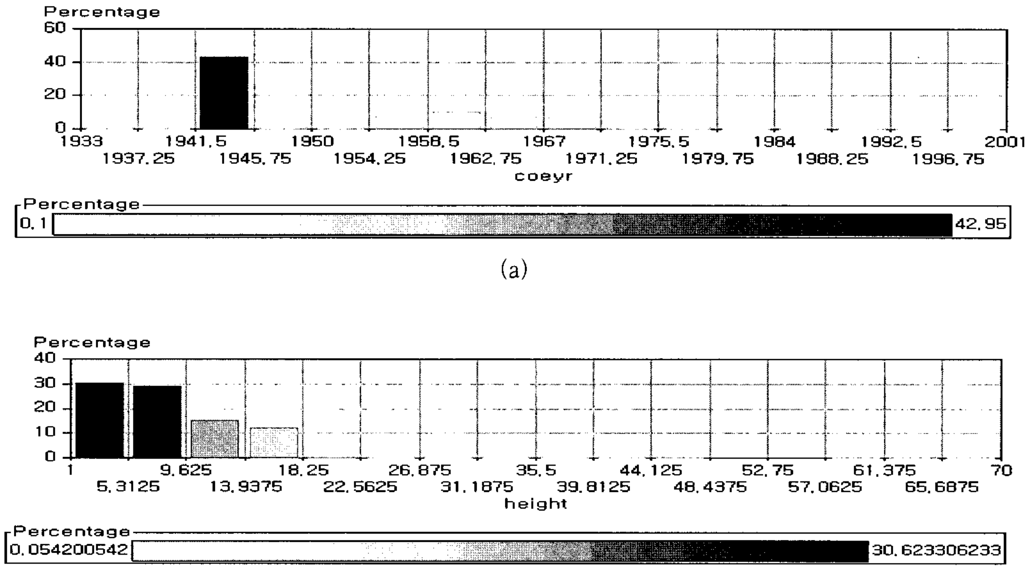
Table 3 Distribution of values for each of the class variables

Variable	Values	Missing rate	Variable	Values	Missing rate	Variable	Values	Missing rate
SOURCE	3	0	COSFUND	8	8	REPFUND	6	81
ATTACHT	1	0	DROUGHT	5	2	REPSP	19	81
ATTACHN	2	0	STRUCT	6	8	REPAIRB	2	2

noise and correct inconsistencies in the data and data transformations, such as normalization, may be applied to improve the accuracy and efficiency of mining algorithms involving distance measurements. In the numerical data, we investigated the data with minimum, maximum, mean value, standard deviation, skewness and kurtosis..Table 2 and 3 show the investigation results and a Fig. 5 shows data distributions, for example, of variables 'COEYR' and 'HEIGHT'. The minimum value of length and height variables are all zero. Since it means they have missing data, data cleaning process is necessary and most of kurtosis and skewness is extremely high. Thus transformation is requested. data units are various so scaling precess is also required.

2) Data Preprocessing

As shown in Table 2, minimum value of each variables is zero means that most variables have missed. There are many methods for filling in the missing values. ignoring the tuple, filling in manually, using the attribute mean, using the attribute mean for all samples belonging to the same class, and using the most probable value. The last method is a popular strategy because it uses the most information from the present data to predict missing values with regression, inference-based tools, or decision tree induction. We instead used the attribute modes to fill



(a)

(b)

Fig. 5 Distribution of interval values; (a) COEYR, (b) HEIGHT

in the missing value when missed values are less than thirty percent. Otherwise, we reduced the variable to prohibit changing significantlly the values distribution. The highly correlated fields also was deleted from correlation analysis. As a result of correlation analysis, for example, variable 'repairt', 'repairsp', 'repfund' have a close correlation with each other. Thus, two new variables were regenerated. One variable is computed as eq. (2) named as 'usetime'. Another variable means whether once reservoir have been repaired in its life time or not. It is a binary data computing with eq. (3)

$$'usetime' = present(year) - constructedyear - repairedyear \quad (2)$$

$$'repairb' = \begin{cases} o & \text{if}(repairdt > 0) \\ = 1 & else \end{cases} \quad \cdots\cdots\cdots (3)$$

In k-means, data measurement depend on distance computation. So the data was transformed using normalization where the attribute data are scaled so as to fall within a small specified range such as −1.0 to 1.0 They are min-max normalization and z-score normalization.

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \quad (4)$$

$$v' = \frac{v - \overline{A}}{\sigma_A} \cdots\cdots\cdots (5)$$

Min-max normalization preserves the relationships among the original data values. It performs a linear transformation on the original data by computing eq. (4). In z-score normalization, the values for an attribute are normalized based on the mean and standard deviation of the attribute and it is useful when there is outliers.

5. Modeling

In a proposed algorithm, input variables are selected and transformation and imputation of the selected variables are done. Next step is to decide parameter, k, of k-means. At first, twelve dimensional variable values have been mapped into two-dimensional topology with a self organized maps. Fig. 6 shows that input data set may be clustered into 4 or 5 groups. To clarify a class number, we also computed the ratio of an inter cluster dissimilarity to intra cluster dissimilarity when varying k means cluster parameter from 2 to 6. Fig. 7 shows that the ratio decreases steeply within 4 clusters but after that, decreases smoothly. So the graph says that this data set should be divided into four small groups. From the two results, we concluded that four classes exist in this data set.

We clustered a data set using k-means with a parameter, k, of 4. Our model have labeled 54 reservoirs class 1, 617 reservoirs class 2, 763 reservoirs class 3 and the rest class 4. As mentioned in section 2, cluster analysis can divide unknown data set into several classes, but can not describe pattern or rule of each classes. It needs data classification using decision tree induction.

A decision tree is built with recursive partitioning. While partitioning recursively, large or small tree is made if the partitions fits data set excessively or too little. training accuracy and validation accuracy computed as shown in Fig. 8. A training accuracy converges into 0.9986 and validation accuracy converges into 0.9981 when

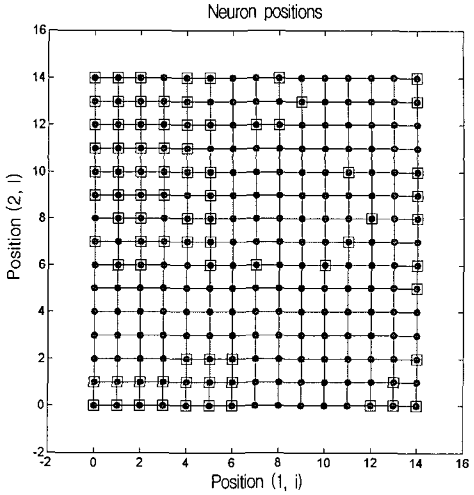


Fig. 6 Feature map of SOMs

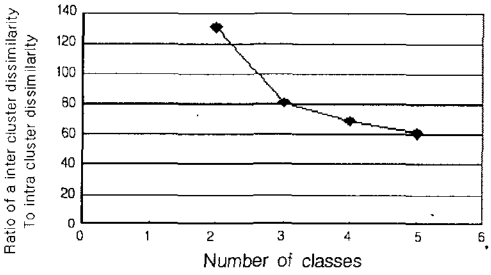


Fig. 7 Ratio of a inter cluster dissimilarity to intra cluster dissimilarity

decision tree comprise five leaf nodes. A decision tree generates 4 internal nodes and 5 leaves with a misclassification rate of 0.001. A classification analysis results that internal nodes are variables of DROUGHT, COSFUND, and STRUCT in a decision tree. When value of DROUGHT is less than 3 and value of COSFUND is less than 4, they are class 3, and if value of DROUGHT is less than 3 and value of COSFUND is less than 7, or value of DROUGHT is more than 3 and value of COSFUND is more than 7, that case is class 4. In the case of class 2, value of DROUGHT is more than 3 and value of COSFUND is less than 2 and value of STRUCT is less than 4. Otherwise, these

Table 7 Characteristics of each group

	Drought	Cosfund	Struct
Class 1	more than 3 years	national or local	etc.
Class 2	more than 3 years	national or local	homegeneous earth dam, concrete gravity dam, zoned embankment dam, embankment with clay core
Class 3	less than 3 years	national, local, Farmland Improvemnet Sssociagion (FIA), or PL480	not considered
Class 4	less than 3 years	etc.	not considered
	more than 3 years	FIA, PL480, private, personal	not considered

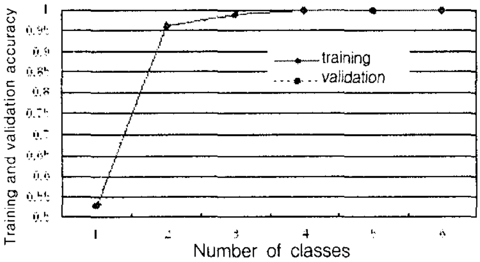


Fig. 8 Training and validation accuracy

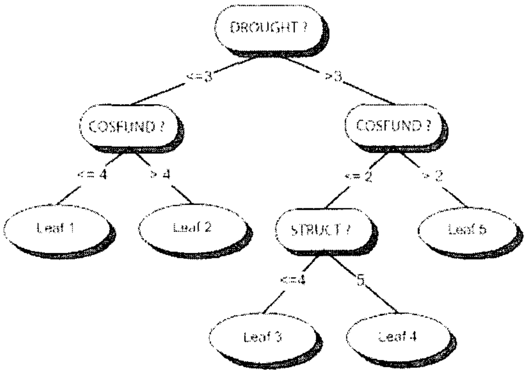
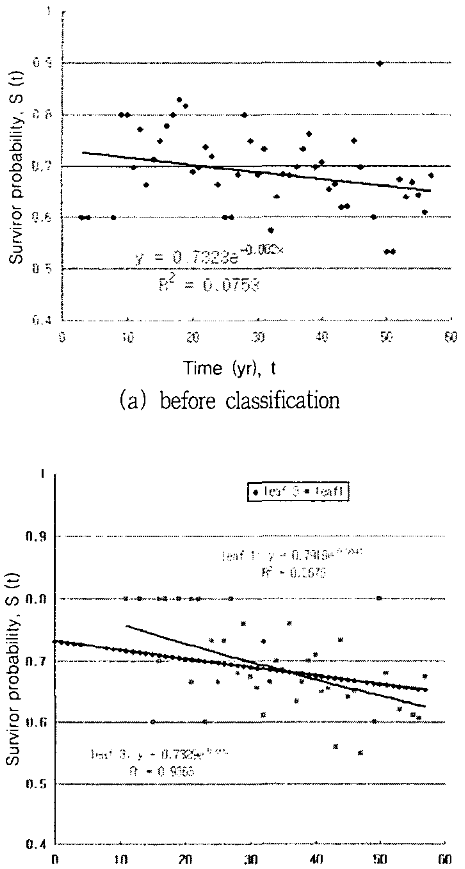


Fig. 9 Decision tree for reservoir classification

data are classified into class 1. The characteristics of each group summarizes as shown in Table 3.

6. Pattern Evaluation

We classified reservoirs to make a significant survivor function. After classification analysis, we divided reservoirs into four groups. In a generated decision tree, leaf 1 and leaf 3 classified well because they compose only one group at the rate of more than 99%. So we made their survivor functions and function of un-classified reservoirs. Survivor function is usually regressed into a Weibull function. Fig. 10 (a) is a survivor function of unclassified reservoirs as a Weibull function. Fig. 10 (b) shows survivor functions of leaf 1 and leaf 3. In Fig 10 (b), reservoirs classified into a leaf 3 should be managed more throughly than reservoirs of a leaf 1 before they have passed about 35 years but, leaf 1 survivor rate is lower than leaf 3 after that time. In other words, it's possible to apply a different management strategy into a reservoir classified with the model. Although survivor function doesn't well regressed, more improved survivor function has been estimated than un-classified data set. Furthermore, It is supposed to estimate more fitted function if data preproce ssing is reasonably improved.



(a) before classification

Time (yr), t

(b) before classification

Fig. 10 Estimated survivor function of (upper) unclassified reservoirs and (lower) reservoir group classified into leaf 1 and 3

7. Conclusions

There are lots of reservoirs in South Korea. They have each other different characteristics in their capacity, usage and so on. It means that different management strategy are requested. It may be unreasonable to use a same survivor function for life cycle analysis so this article purposes to classify reservoirs corresponding to their physical characteristics such as height, width, age, repair works history and so on. First of all, data set of 3,358 reservoirs was analyzed using k means and self organized maps. As a result of that analysis, four clusters have been extracted. Factors and their critical values to classify 3,358 reservoirs into four groups have been founded by Decision Tree. The path rules to each group have feasibility since their each survivor function showed unique pattern.

Acknowledgements

This research was supported by a grant (2003103-03-2) from Agricultural R&D Promotion Center.

References

1. jHan iawei and Micheline Kamber, 2001, Data mining concepts and techniques, Morgan Kaufmann
2. The data of the safety diagnosis for irrigation facilities during 12 years. KARICO
3. Kim, Han Joong, Lee, Jeong Jae, and Im, Sang Joon, 2003, Life reliability analysis of irrigation system, Journal of the Korean Society of Agricultural Engineers 45 (2)
4. Lee, Joon-gu, 2003, Prediction model of remaining service life concrete with carbo-nation, Jounal of the Korea Concrete Institute 15 (4)
5. MAF, KARICO, The annual statistics report of agricultural infrastructure development & improvement project.