Fusion of Structure Adaptive Self-Organizing Maps Using Fuzzy Integral

Kyung-Joong Kim
Yonsei University
Department of Computer Science
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
uribyul@candy.yonsei.ac.kr

Sung-Bae Cho
Yonsei University
Department of Computer Science
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
sbcho@cs.yonsei.ac.kr

Abstract- **Recently, many researchers attempt to develop an effective SOM-based pattern recognizer for high performance classification. Structure adaptive self-organizing map (SASOM) is a variant of SOM that is useful to pattern recognition and visualization. Fusion of classifiers can overcome the limitation of a single classifier by complementing each other. Fuzzy integral is a combination scheme that uses subjectively defined relevance of classifiers. In this paper, fusion of SASOM's using fuzzy integral is proposed for web mining problem. User profile represents different aspects of user's characteristics and needs an ensemble of classifiers that estimate user's preference using web content labeled by user as "like" or "dislike." The proposed method estimates the user profile using subsets of important features extracted from user-rated web documents. Using UCI Syskill & Webert data, the method is tested and compared with other classifiers including ID3, BP and naïve Bayes classifier. Experimental results show that the fusion of SASOM's using fuzzy integral can perform better than not only previous studies but also majority voting of SASOM's.**

I. INTRODUCTION

Self-organizing map (SOM) is a very useful neural network to visualize large-dimensional data for mining knowledge and an efficient tool to cluster data [1,2]. Some researchers attempt to apply SOM to pattern classification [3,4]. Like other models of neural networks, one of the shortcomings is a difficulty to determine the size and structure of the network. In the previous work [5] we also proposed an efficient pattern recognizer based on a dynamic node splitting scheme for the SOM, which shapes the general structure-adaptive SOM (SASOM) into a pattern recognizer by splitting a node representing more than one class into a submap (composed of four nodes). Ensemble of SASOM's trained independently using different feature sets provides high performance in digit recognition problem [6].

In this paper, we focus on web content mining for creating a user profile from HTML documents and user's preference record for them. Estimating the user profile needs non-linear function because it has the properties that are not easily captured by simple guess. Also, it contains many aspects of user's preference such as "I like a web page that has a funny story," "I like a web page that is likely related to sports," and "I like a web page that has easily understandable words." It is difficult to estimate such properties as a single machine learning model and needs to combine a number of models that complement each other with different expertise.

We have adopted the ensemble of SASOM's to estimate the user profile and each SASOM is trained independently using different feature sets. Three different feature-ranking methods are used for the problem of web mining. They are information gain, TFIDF, and odds ratio [7]. These three methods are representative feature ranking methods for text and simply implemented. There are many combination methods such as Borda count, Condorect function, voting, weighted voting, Bayesian averaging, Dempster-Shafer theory, and behavior knowledge space [8]. However, these methods cannot insert user's subjective preference on classifier into combination procedure and have little flexibility. Fuzzy integral is a combination method to aggregate evidence from multiple sources using fuzzy measure and user's subjective evaluation on classifiers' relevance [9].

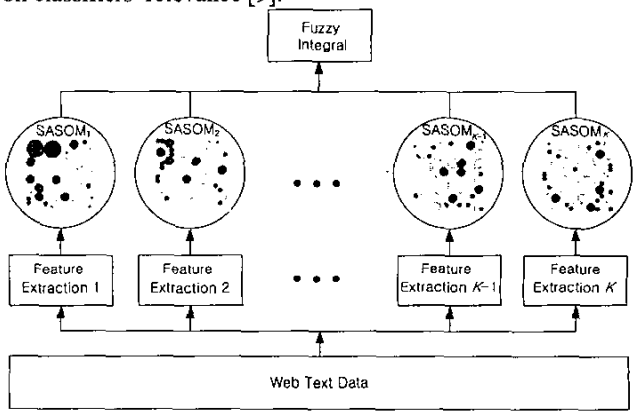


Fig. 1. Overview of the proposed method for web mining

Fig. 1 shows the overview of the proposed scheme. From a preprocessed feature vector for each web text, each feature extraction method selects relevant feature sets for training: Each feature set is used to train one SASOM. After training, each SASOM has different topology as shown in this figure. Fuzzy integral aggregates evidence from multiple sources, at the end. This ensemble classifier can be used to predict user's preference on unknown web documents as a user profile. To evaluate the proposed method, UCI KDD Syskill & Webert data set is adopted [10]. The data set contains four different topics and related web pages with user's preference record that is "hot," "medium," or "cold." Problem is to estimate user's preference on unknown web page as "hot," or "cold" ("medium" and "cold" is aggregated because "medium" is a few). Pazzani reported that naïve Bayes classifier performed better than other methods such as neural network and ID3 for the data [11]. For comparison, majority voting is adopted.

0-7803-7898-9/03/$17.00 ©2003 IEEE

Experimental result indicates that the proposed method using fuzzy integral shows better performance than previous method based on naive Bayes classifier and ensemble of SASOM's using majority voting.

II. STRUCTURE ADAPTIVE SELF-ORGANIZING MAP

SOM is a neural network model that has property of preserving topology of map and is frequently used to visualize high-dimensional data to low-dimensional space. Fig. 2 shows basic structure of SOM. White node represents input node where input vector is inputted and black node represents neuron. Each neuron competes with other neurons to become the winning node.

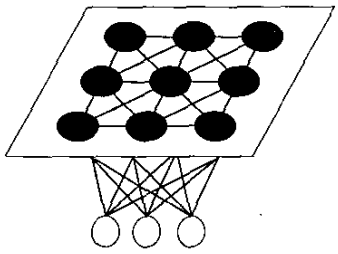


Fig. 2. Basic structure of self-organizing map

Basic SOM fixes the structure of map and shows low performance in classification because each node has data that has different class label. This property is very useful in unsupervised clustering but becomes weak property in classification. When a node has data of different class labels, SASOM divides a node into a submap of 4 nodes.

Dynamic node splitting scheme is able to simultaneously determine a suitable number of nodes and the connection weights between input and output nodes in self-organizing map.

The basic idea is very simple like this.

1. Start with a basic SOM (in our case, a 4×4 map in which each node is fully connected to all input nodes).
2. Train the current network with the Kohonen's algorithm [2].
3. Calibrate the network using known I/O patterns to determine:
   (a) which node should be replaced with a submap of several nodes (in our case, 2×2 map), and
   (b) which node should be deleted. Nodes that have been inactive longer than a specified length of time can be deleted.
4. Unless every node represents a unique class, go to step 2.

A node representing more than one class is replaced with several nodes. Weights of child nodes of parent node are determined as follows:

$$C = \frac{(P \times 2) + \sum N_c}{S} \tag{1}$$

N_c : Neighborhood nodes of child
S : The number of N_c+2

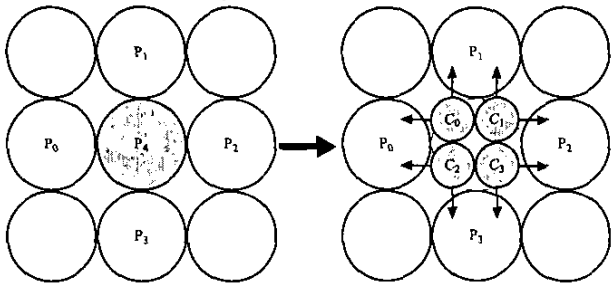


Fig. 3. An example of node splitting

Fig. 3 shows an instance of node splitting. In this case, the weight of C_0 is determined as follows.

$$C_0 = \frac{(P_4 \times 2) + P_0 + P_1}{4} \tag{2}$$

III. FUSION OF SASOM'S USING FUZZY INTEGRAL

A. Feature Selection

Feature selection is the procedure of ranking features based on the information such as frequency and dependency. In text classification, feature is a term in text and has binary value (exist or not). Feature selection procedure is necessary because there are more than 5000 or 6000 features in a collection of 20 web documents. Many features are not useful to improve performance and difficulty arises to learn classifier.

TFIDF is multiplication of term frequency and inverse document frequency. This measure is frequently used in text retrieval and very simple.

$$TFIDF = TF \times \log \frac{1}{DF} \tag{3}$$

TFIDF does not use class information of training data to calculate the importance of features. This can degrade the performance of classification. Information gain is the method based on information theory. S is a set of pages and E is expected information gain. $E(W,S)$ means the expectation of term W on the documents set S.

$$\begin{aligned} E(W,S) &= I(S) - (A + B) \\ A &= P(W = present) I(S_{W = present}) \\ B &= P(W = absent) I(S_{W = absent}) \\ I(S) &= \sum_{c \in \{hot, cold\}} - p(S_c) \log_2 (p(S_c)) \end{aligned} \tag{4}$$

The last feature extraction method uses odds ratio. Odds ratio is used when the goal is to make a good prediction for one of the class values.

$$OddsRatio(F) = \log \frac{odds(W = present \mid C_1)}{odds(W = present \mid C_2)} \tag{5}$$

where C_1 and C_2 are class label of binary classification problem. $Odds(X_i)$ is defined as follows.

$$Odds(X_i) = \begin{cases} \dfrac{\frac{1}{n^2}}{1-\frac{1}{n^2}} & P(X_i) = 0 \\ \dfrac{1-\frac{1}{n^2}}{\frac{1}{n^2}} & P(X_i) = 1 \\ \dfrac{P(X_i)}{1-P(X_i)} & P(X_i) \neq 0 \wedge P(X_i) \neq 1 \end{cases} \quad (6)$$

where n is a number of examples.

These three feature selection methods have different properties. TFIDF does not consider class values of documents when calculating the relevance of features but information gain uses class labels of documents. Odds ratio uses class labels of documents but they find features that are useful to classify only one specific class.

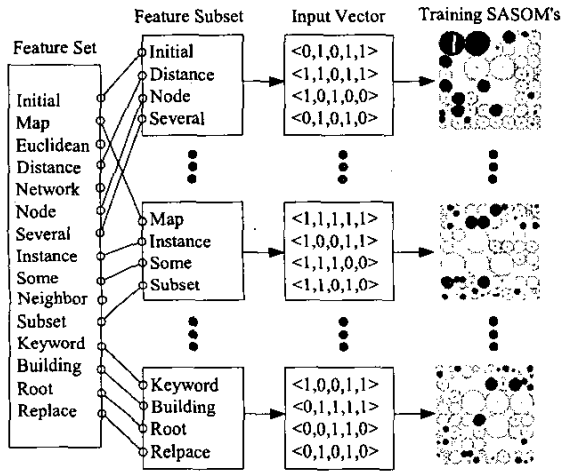


Fig. 4. Training SASOM's using different feature sets

Fig. 4 shows the whole procedure of learning independent classifiers using different feature sets. Feature set is a set of terms in a collection of web documents. Using feature selection methods, we can extract the most relevant features, respectively. For example, the first feature subset selects "Initial," "Distance," "Node" and "Several" as the most relevant features. Using these features, input vectors are constructed. For example, the first web documents using the first feature subset has a vector <0,1,0,1,1> that means "Initial" does not exist, "Distance" exists, "Node" does not exist, "Several" exists, and the class label of this document is "1." Using these input vectors, each SASOM is trained independently.

B. Fuzzy Integral

There are many combination methods that use the decision of multiple classifiers. They assume that each classifier has the same relevance to the problem. Weighted averaging calculates the relevance of each classifier by using objective measure such as classification performance on training data. Fuzzy integral provides the importance of each classifier that is measured subjectively. Final decision is integrated from the evidence of classifiers for each class and the importance of classifiers subjectively defined by users.

The fuzzy integral introduced by Sugeno [12] and the associated fuzzy measures [13], [14] provide a useful way for aggregating information. Fuzzy integral is defined as follows. Fuzzy measure assigns a real value between 0 and 1 for each subset of X.

Definition 1: Let X be a finite set of elements. A set function $g : 2^X \rightarrow [0,1]$ with

1) $g(\varnothing)=0$
2) $g(X)=1$
3) $g(A) \leq g(B)$ if $A \subset B$

is called a fuzzy measure.

Definition 2: Let X be a finite set, and $h : X \rightarrow [0,1]$ be a fuzzy subset of X. The fuzzy integral over X of the function h with respect to a fuzzy measure g *is defined by*

$$h(x) \circ g(\cdot) = \max_{E \subseteq X} \left[\min\left(\min_{x \in E} h(x), g(E) \right) \right] \quad (7)$$

The calculation of the fuzzy integral is as follows: Let $Y=\{y_1,y_2,\ldots,y_n\}$ be a finite set and let $h : Y \rightarrow [0,1]$ be a function. Suppose $h(y_1) \geq h(y_2) \geq h(y_3) \geq \ldots \geq h(y_n)$. Then a fuzzy integral, e, with respect to a fuzzy measure g over Y can be computed by

$$e = \max_{i=1}^{n} [\min(h(y_i), g(A_i))] \quad (8)$$

where $A_i=\{y_1,y_2,\ldots,y_i\}$. λ is given by solving the equation

$$\lambda + 1 = \prod_{i=1}^{n} (1 + \lambda g^i) \quad \lambda \in (-1,+\infty) \text{ and } \lambda \neq 0. \quad (9)$$

This equation is derived from following recursive calculation. λ can be easily calculated by using the $(n\text{-}1)$st degree polynomial.

$$\begin{aligned} g(A_1) &= g(\{y_1\}) = g^1 \\ g(A_i) &= g^i + g(A_{i-1}) + \lambda g^i g(A_{i-1}), \quad \text{for } 1 < i \leq n. \end{aligned} \quad (10)$$

Let $C=\{c_1, c_2, c_3,\ldots, c_N\}$ be a set of classes, where binary classification problem has $|C|=2$. Let $Y=\{y_1, y_2,\ldots, y_n\}$ be a set of classifiers, and $h_k : Y \rightarrow [0,1]$ be partial evaluation of the object A (to be classified) for class c_k. $h_k(y_i)$ is an indication of how certain we are in the classification of object A to be in class w_k using the network y_i. The set Y is sorted by the value of $h_k(y_i)$ for each class in descending order. A_{ki} means a set of former i elements in Y for the class k.

$$\text{Final class} = \underset{c_k \in C}{\text{argmax}} \left[\max_{i=1}^{n} [\min(h_k(y_i), g(A_{ki}))] \right] \quad (11)$$

Each SASOM determines the class label of an unknown document as "0" or "1" (in the binary classification problem). If $SASOM_1$ classifies the document as "0", $h_0(\text{sasom}_1)=1.0$ and $h_1(\text{sasom}_1)=0.0$. Supposed that there are three SASOM's, user evaluates classifiers as g^1,g^2 and g^3, respectively. λ is calculated from g^1,g^2, and g^3. It is easily determined from the 2nd degree polynomial based on (9). For each class k, classifiers are sorted by $h_k(\text{sasom}_i)$. By the sorted order, they are labeled as y_1, y_2 and y_3. With $g(y_1)$, $g(y_1,y_2)$ and $g(y_1,y_2,y_3)$, the class label of the unknown document is determined using formula (11).

IV. EXPERIMENTAL RESULTS

The proposed ensemble aims to estimate user profile by learning the data of web documents. From the UCI KDD database, Syskill & Webert data that have a pair of web documents and user's preference value ("hot" or "cold") are accessible. The HTML source of a web page is given. Users looked at each web page and indicated on a 3 points scale (hot medium cold) 50-100 pages per domain (medium is replaced with cold because a few is medium). Syskill & Webert data have four different topics "Bands," "Biomedical," "Goats," and "Sheep." We only use "Goats" and "Bands" data in this paper.

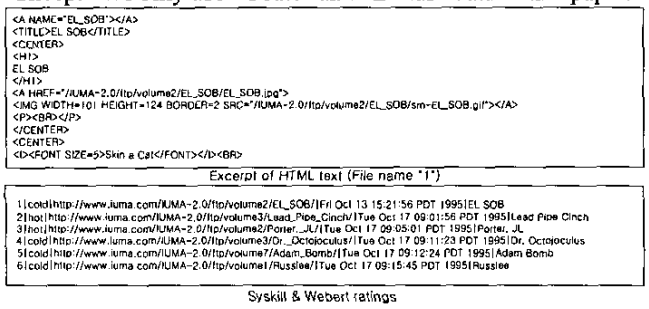


Fig. 5. UCI Syskill & Webert data

"Goats" data have 70 HTML documents and "Bands" 61 HTML documents. Each document has the class label of "hot" or "cold." Fig. 5 shows a HTML file and rating data. Each HTML file contains text related with the topic. Rating file contains file name, rating, URL, date and title orderly. Preprocessing of web documents constructs input vector with selected features and class label. From training data, we extract k important features using three different feature selection methods. Each method ranks all features by different manner. Fig. 6 shows different ranks of features for each method. Using Bands data with 10 training documents, 1200 terms are gathered. In this figure, ranks of a term are different for each method. Document $D=<v_1, v_2, v_3,\ldots, v_{128}, c>$ has three different input vectors that are used to train SASOM's. The procedure of preprocessing of HTML documents is as follows.

1. Eliminate non-letters including tags
2. Change capital letter to small one; Stop list is a set of following features
   a) Sort terms by the frequency
   b) Select 600 terms that are highly ranked as stop list
3. Eliminate stop list
4. Make index table <feature, list of documents that have the feature>
5. Calculate feature relevance using TFIDF, Information gain, and odds ratio
6. Sort features by TFIDF and select k features; Sort features by information gain and select k features; Sort features by odds ratio and select k features, where k is determined as 128 [11]
7. Construct input vector for train and test data

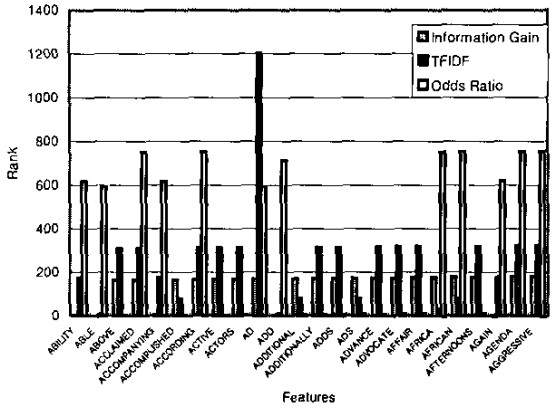


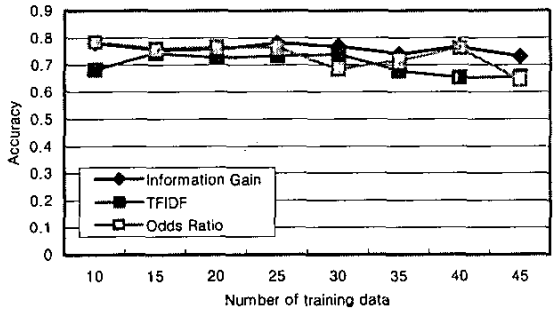
Fig. 6. Rank of features for each feature extraction method

Problem to solve is to predict unknown documents' class using known web documents with ensemble of three different SASOM's trained using the input vectors. For each topic, we have conducted 8 different experiments (each experiment has different number of training data and test data). The number of training data is increased by 5 from 10 to 45. Remaining data are used as test set. Experiments are repeated 10 times and the final result is the average of them. For comparison, Pazzani's results of naïve Bayes classifier, nearest neighbor, ID3, perceptron, Backpropagation, PEBLS, and Rocchio are used [11] (They are summarized in Fig. 10).

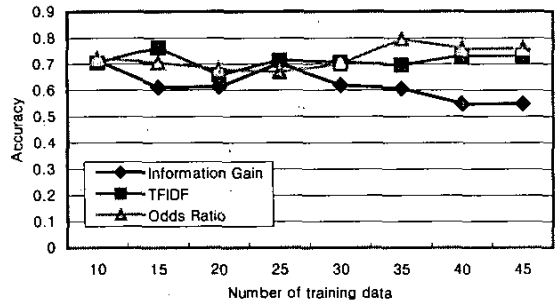
Fig. 7 shows the performance of single classifiers trained using feature subsets. As mentioned before, there are three different feature subsets (information gain, TFIDF and odds ratio). Each classifier shows different performance. In Bands, information gain performs the best. In Goats, odds ratio and TFIDF are good. Accuracy means the ratio of correct predictions.

Fig. 8 shows different maps of SASOM's for three different feature extraction methods. In odds ratio, only nodes with class "1" exist because odds ratio selects the features that are related with specific one class (in this case "1"). User has several reasons why he selects a web page as a "hot." In SOM, neighbor nodes have similar characteristics though they have different class labels. In (a), there are five distinct groups and we can estimate that the user has five different criteria for the decision. Other maps analyze user's criterion with different

manner. By analyzing the criterion, we can capture conceptual features of user's classification and apply them to web mining.

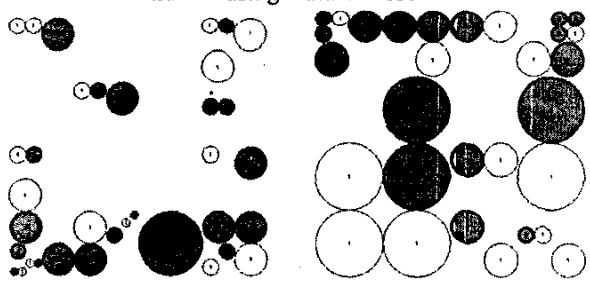


(a) Bands

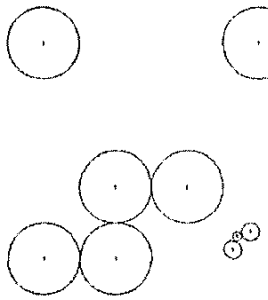


(b) Goats

Fig. 7. Performance of single SASOM classifier that is trained using feature subset



(a) Information Gain (b) TFIDF



(c) Odds Ratio

Fig. 8. Topology of maps (The number of training data is 45 in Goats data)

Table 1 shows the fuzzy measures of classifiers that are determined subjectively. Each classifier has different importance with the different number of training data. Fig. 9 shows the performance of ensemble of SASOM's using fuzzy integral and majority voting. For comparison, the results of Pazzani's naïve Bayes classifier are used. In Bands, fuzzy integral performs better than naive Bayes classifier and majority voting. In Goats, fuzzy integral performs better than the other two methods. Majority voting of SASOM's show better performance than naive Bayes classifier but worse performance than fuzzy integral.

Fig. 10 shows comparison with other seven classifiers including ID3, BP, and naïve Bayes. Figure 10 is based on the results published by Pazzani [11]. Twenty examples are chosen as a reasonable intermediate number of examples [11]. Training set size is 20 and the remaining data are used as a test set. Fuzzy integral performs better than other classifiers. The importance of each classifier in fuzzy integral is described in Table 1 (The number of training data is 20). The value of g does not mean the performance of the classifier but it means the relative favor of the classifier for specific user.

TABLE I
Fuzzy measure of classifiers (1: information gain, 2: TFIDF, 3: Odds ratio)

	Bands			Goats		
Number of training data	g^1	g^2	g^3	g^1	g^2	g^3
10	0.99	0.99	0.99	0.99	0.01	0.99
15	0.99	0.99	0.99	0.99	0.99	0.01
20	0.99	0.99	0.99	0.99	0.99	0.01
25	0.99	0.99	0.99	0.99	0.01	0.99
30	0.99	0.99	0.01	0.99	0.99	0.99
35	0.99	0.01	0.99	0.99	0.99	0.99
40	0.99	0.99	0.99	0.99	0.99	0.99
45	0.99	0.99	0.00	0.99	0.01	0.99

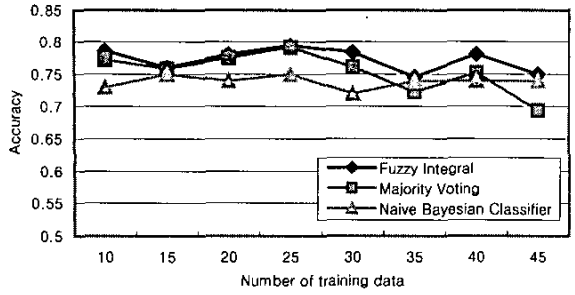
V. CONCLUDING REMARKS

In this paper, we have proposed an ensemble of SASOM's using fuzzy integral to classify web documents based on user's preference. Experimental results show that the proposed method performs better than previous studies and majority voting of SASOM's. Fuzzy integral provides the scheme of measuring the importance of classifiers subjectively. SASOM can classify documents with high performance and visualize its map to understand internal mechanism. The proposed method can be effectively applied to web content mining for predicting user's preference as a user profile.

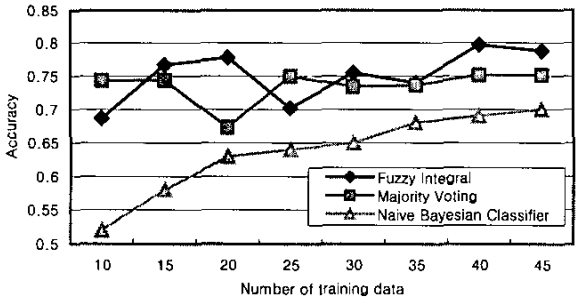
ACKNOWLEDGEMENTS

This work was supported by Biometrics Engineering Research Center, and Brain Science and Engineering Research Program

sponsored by Korean Ministry of Science and Technology.



(a) Bands



(b) Goats

Fig. 9. Combination of SASOM's using fuzzy integral

REFERENCES

[1] J. Vesanto, "SOM-based data visualization methods," *Intelligent Data Analysis*, vol. 3, no. 2, pp. 111-126, August 1999.

[2] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464-1480, 1990.

[3] P. N. Suganthan, "Pattern classification using multiple hierarchical overlapped self-organising maps," *Pattern Recognition*, vol. 34, no. 11, pp. 2173-2179, Nov 2001.

[4] S.-B. Cho, "Neural-network classifiers for recognizing totally unconstrained handwritten numerals," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 43-53, Jan 1997.

[5] S.-B. Cho, "Self-organizing map with dynamical node splitting: Application to handwritten digit recognition," *Neural Computation*, vol. 9, no. 6, pp. 1343-1353, 1997.

[6] S.-B. Cho, "Ensemble of structure-adaptive self-organizing maps for high performance classification," *Information Sciences*, vol. 123, no. 1-2, pp. 103-114, March 2000.

[7] D. Mladenic and M. Grobelnik, "Feature selection on hierarchy of web documents," *Decision Support Systems*, vol. 35, pp. 45-87, 2003.

[8] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, and A. Gelzinis, "Soft combination of neural classifiers: A comparative study," *Pattern Recognition Letters*, vol. 20, pp. 429-444, 1999.

[9] S.-B. Cho, and J.-H. Kim, "Combining multiple neural networks by fuzzy integral for robust classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 2, pp. 380-384, February 1995.

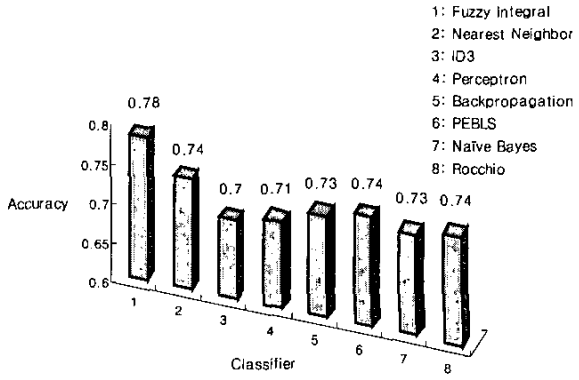
[10] S. Hettich and S. D. Bay, The UCI KDD Archive, http://kdd.ics.uci.edu.

[11] M. Pazzani and D. Billsus, "Learning and revising user profiles: The identification of interesting web sites," *Machine Learning*, vol. 27, pp. 313-331, 1997.

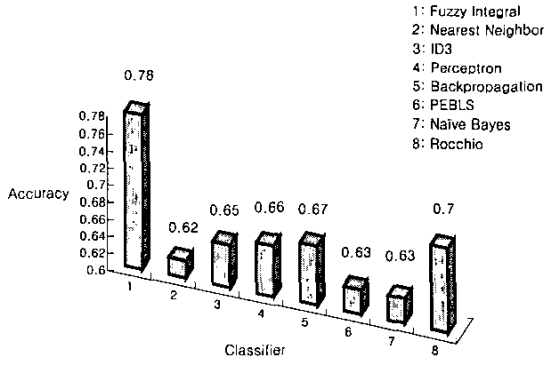
[12] M. Sugeno, "Fuzzy measures and fuzzy integrals: A survey," *Fuzzy Automata and Decision Processes*, Amsterdam: North Holland, pp. 89-102, 1977.

[13] K. Leszeynski, P. Penczek and W. Grochulskki, "Sugeno's fuzzy measures and fuzzy clustering," *Fuzzy Sets and Systems*, vol. 15, pp. 147-158, 1985.

[14] R. R. Yager, "Element selection from a fuzzy subset using the fuzzy integral," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, pp. 467-477, 1993.



(a) Bands



(b) Goats

Fig. 10. Performance comparison with other classifiers