Application of Classifier Integration Model with Confusion Table to Audio Data Classification

Miso Jang and Dong-Chul Park

Abstract—**A novel and efficient method for the Classifier Integration Model(CIM) by adopting the concept of confusion table is extensively studied and reported in this paper. The method considers not only the diagonal elements of each confusion table but also the non-diagonal elements obtained from the existing confusion tables for local classifiers in CIM for more accurate classification performance. The CIM with Confusion Table (CIM-CT) method is applied to two different data sets, Iris data set and an audio signal data set for evaluation of the CNN-CT scheme. The experimental results show that the CIM-CT method outperforms a conventional classifier and the classifier with utilizing only the diagonal elements of confusion tables in CIM in terms of classification accuracy.**

Index Terms—**Classifier, features, confusion table, machine learning.**

I. INTRODUCTION

Most of pattern classification tasks usually include a feature extraction procedure as the first step of the whole scheme. Extracting proper features for a given pattern classification problem is a vital task for designing an accurate pattern classifier. The importance of extracting proper features goes even higher when the classification problem has to work with multimedia data. For audio signal data, acoustical elements including bandwidth, loudness, pitch, and harmony are widely used along with music-dependent features including rhythmic content, pitch content, and timbre texture [1]. Most widely used features for audio signal data classification problems include the energy features and timbre texture [2], MFCC (Mel-scaled Frequency Cepstral Coefficients) [3], and DWT (Discrete Wavelet Transform) [4]. Note that there exist various features for audio signals and each of these available features has its own advantage that can characterize each audio signal. However, only some of the most widely used features are adopted as inputs to the classifiers in this study, since the scope of this paper is more on efficient classifier design.

Once proper features are extracted for a given audio signal classification task, the next step is to design an efficient classifier for the given set of audio signals. When more features among all the available features are used for the given audio signal classifier, more accurate classification results can be expected as long as the available features can be handled properly in the given classifier. Handling different feature sets as inputs to a given audio signal classifier, however, is not a simple task. One of the reasons why this task is rather difficult is that each feature of the given audio signal is independent of the others and each feature generally has a different dimensionality and magnitude from other features. When using different features in order to characterize different audio signals, a normalization process for each feature is usually required first so that all different features can be combined. The normalization process usually requires to find the minimum and maximum values for all the feature values involved. The use of the same minimum and maximum values for the feature normalization process, however, does not have any justification in addition to its simplicity in use. This non-homogeneity in properties of different features makes it hard to use various features as a combined form for the audio signal classifier.

Conventionally, only some of the features considered as important for a given task among all the available features are used as input data by concatenating the chosen features as shown in Fig. 1. Note that the concatenation of different features makes the dimension of the input very high and the resulting high dimensionality of input data to a classifier can cause problems in training the classifier. In order to train the classifier efficiently, it generally requires more training data when the dimensionality of input data goes higher. This problem is often called as Curse of Dimensionality [5].
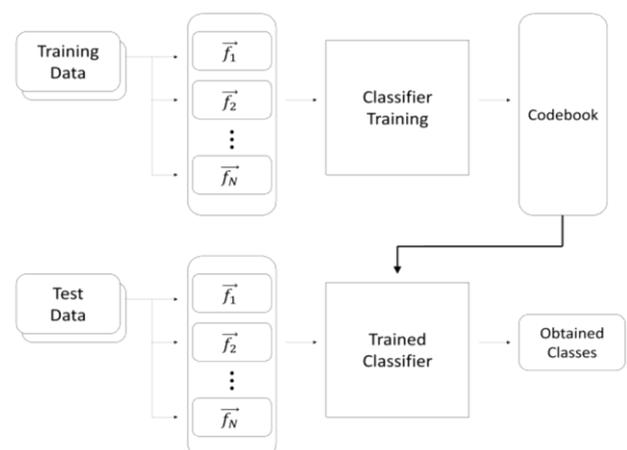
Fig. 1. Conventional classifier model [6].

In order to facilitate utilizing different sets of features in

designing efficient classifiers, classifier models such as PFC (Partitioned Feature-based Classifier)[6] and CIM (Classifier Integration Model) [7] have been proposed.

As shown in Fig. 2, the each of the available features extracted from independent feature extractors is separately used as an input feature vector to an independent local classifier of PFC. There is no need to go through any preprocessing procedure involved with other features. This independent use of each feature can preserve the individuality of each feature for a specific classifier. Each of the local classifiers in PFC is independently trained with a specific feature extracted from different feature extractor. A local classifier with a specific feature does not interfere with other classifiers in PFC. A local classifier with a specific feature in PFC does not interfere with other classifiers. Note that the separation of input features allows us to train each local classifier efficiently with smaller number of training data when compared with the case of combining all the feature vectors in one feature as is the case of conventional classifiers. This idea of separated features can alleviate the problem of training data size in practical applications.
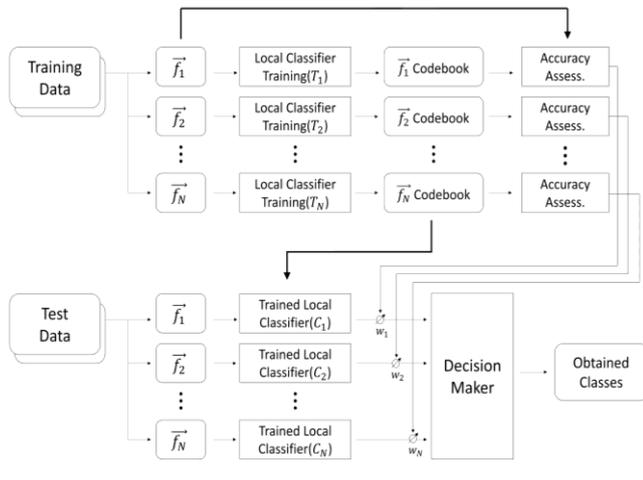
Fig. 2. Partitioned feature-based classifier model [6].

In FCM and CIM, all the available feature vectors can be utilized effectively for the classifier. Each feature extracted by different feature extraction procedure is considered independently as the input to a local classifier that uses the specific feature as the only input vector. The outputs of these local classifiers, where each of the local classifiers utilizes a different feature for each classifier, are then combined to produce the final classification decision of the input data in an optimal fashion.

In order to enhance the classification accuracy for CIM, this paper addresses an efficient method to combine the outputs of all the local classifiers. While CIM uses only the diagonal elements of each confusion table of the local classifier, the CIM with Confusion Table(CIM-CT) method is designed to use the non-diagonal elements in addition to the diagonal elements of confusion tables.

The remaining portion of this paper is organized as follows: a brief summary of the CIM is first given and CIM enhancement by using the confusion table, CIM-CT, is then introduced in Section II. Experiments on two sets of data for CIM and CIM-CT and the evaluation results are reported in

Section III. Finally, Section IV concludes this paper.

II. CLASSIFIER MODELS

A. Classifier Integration Model

This model was first proposed to cope with the problems experienced when training data with various features [7]. Fig. 3 shows a schematic diagram of CIM. The CIM can utilize all the available N features extracted from data by utilizing a number of local classifiers k, $1 \leq k \leq N$, while conventional classifiers use only selected features at once by concatenating these features. While training FCM, each classifier k produces w_k that shows the accuracy information of the corresponding classifier k. The PFC produces the class results based on the probability of correct classification of each local classifier. However, if we have information on the existing tendency of each local classifier's classification results, it would be better to use the tendency in producing classification results for the whole system. For example, C_i produces a high classification accuracy result on class j, of data while C_i has a tendency to produce a very low classification accuracy result on another class k, it would be more reasonable to give more credit when C_i produces a result of Class j on a given data than when this classifier C_i assesses the classification result as Class k for the given data[6] .
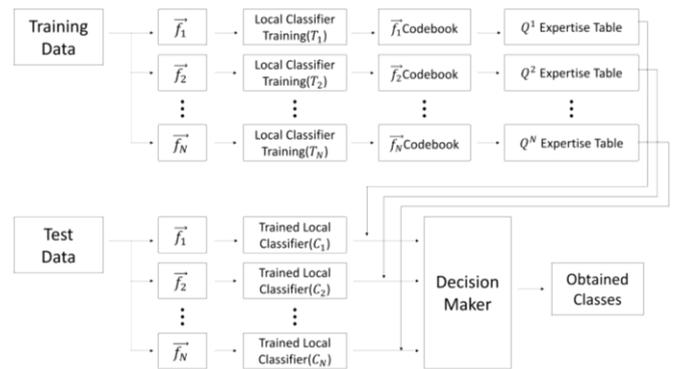
Fig. 3. Multiple feature-based classifier model [7].

However, the CIM considers how each local classifier with different feature vector as its input performs on each class whether it assesses correct class or not. The accuracy information, or tendency, of each local classifier during the training procedure is recorded as a form of the confusion table in CIM. The confusion table for the local classifier k can be formulated as shown in Eq. (1):

$$Q_k = \begin{bmatrix} q_{11}^k & q_{12}^k & \cdots & q_{1M}^k \\ q_{21}^k & q_{22}^k & \cdots & q_{1M}^k \\ \vdots & \vdots & \ddots & \vdots \\ q_{M1}^k & q_{M2}^k & \cdots & q_{MM}^k \end{bmatrix} \tag{1}$$

where q_{ij}^k represents the probability that the classifier k, classifies the data as Class j when the data is from of Class i and M denotes the number of classes.

Note that each local classifier in CIM is independently trained with an independent set of features and each local classifier can show different characteristics in its

classification performance in terms of its confusion table. Each local classifier does not have any relation with other local classifiers because it utilizes a specific feature extracted from the data by using an independent feature extractor as its input. Note also that q^k_{ij} is a result of evaluation process after the local classifier is trained with the given training data and each local classifier can utilize any classifier scheme as long as the classifier scheme performs optimally for the given set of features and training data [8].

B. Classifier Integration Model with Confusion Table

When N local classifiers, $\{C_1, C_2, \cdots, C_N\}$ and Q_k, the expertise table, for each classifier, k, are given as results of training with given set of training data, we want to find a proper class for a given data x. Assume that the feature vector f for a data x be as follows :

$$f = [_1{}^T : f_2{}^T : \cdots : f_N{}^T]^T \qquad (2)$$

where T denotes the transpose operator.

Assume that the classifiers are based on unsupervised learning algorithms [8] and the resultant center of the j-th cluster, $\mathbf{P}_j$, $j = 1, 2, \cdots, M$, on the i-th local classifier is obtained by $\mathbf{p}_{ij}$, $i = 1, 2, \cdots, N$. When f_i, $i = 1, 2, \cdots, N$, is shown to C_i, $i = 1, 2, \cdots, N$, the distance between the data x and $\mathbf{P}_j$, $\mathrm{D}(x, \mathbf{p}_{ij})$, is obtained as shown in Eq. (3):

$$\{\mathrm{D}(f_1(x), \mathbf{p}_{1j}), \mathrm{D}(f_2(x), \mathbf{p}_{2j}), \cdots, \mathrm{D}(f_N(x), \mathbf{p}_{Nj})\} \qquad (3)$$

By combining the distance measure in Eq.(3) and the expertise table in Eq. (1), we can compute the $P(C^k_j / x)$ and each local classifier yields a probability that the data x belongs to the class j, C_j, as follows:

$$P(C_j \mid x) = \sum_{k=1}^{N} P(C^K_j \mid x) q^k_{jj} \qquad (4)$$

In Eq. (4), each local classifier considers only the probability that the classifier made correct classifications in the past. That is why only the diagonal components of the confusion table are utilized in Eq. (1). However, in addition to the probability that each specific local classifier made correct classifications in the past, the tendency how each specific local classifier made incorrect classification can give us valuable information when the assessment of class for a given data is made. This idea of utilizing the tendency for each local classifier to make misclassifications can be formulated as follows:

$$P(C_j \mid x) = \sum_{k=1}^{N} \sum_{i=1}^{M} P(C^k_j \mid x) \; q^k_{ij} \qquad (5)$$

The assessment of class for a given data x can be made as follows:

$$\mathrm{Class}\,(x) = \underset{j}{\mathrm{argmax}} \; P(C_j \mid x) \qquad (6)$$

A confusion table Q_k can be updated during the online practice process whenever the ground truth class for a given input is available as well as during training process. Note that the process for calculating the classification accuracy for each local classifier shown in Eq.(3) is for the unsupervised learning case and it will be easily changed to another form when supervised learning scheme is used for local classifiers.

III. EXPERIMENTS AND RESULTS

In order to evaluate the CIM with Confusion Table (CIM-CT), a benchmark data set called Iris data set from UCI Machine Learning Repository [10] and an audio data set are used for the experiments. The audio data set consists of 10 different classes of audio signals: speech, pop, rock, country, folk, classic, hip-hop, blues, metal and jazz music signals. Each of the class has 300 of 30s long excerpts. Fig. 4 shows examples of data. These excerpts are converted into 44KHz and 16 bits mono audio files for feature extraction process. The frame size and texture window size are 23ms (512 samples) and 1,000ms (43 frames), respectively. The feature extraction procedure for the audio data set is performed by using jAudio that is an open source software for audio feature extraction [11]. The features extracted from audio signals are 30-dimensional Timbral texture, 17-dimensional Pitch contents, 9-dimensional FFT coefficients, 10-dimensioanl MFCC, and 24-dimendional DWT(Discrete Wavelet Transform) coefficients. The timbral texture features are extracted by using SSFT (Short Time Fourier Transform) and include means and variances of zero-crossings, roll-off, spectral centroid, flux, and low energy. Because of its ability to represent one of the most important characteristics of a given audio signal, MFCC has been widely accepted as an important feature in speech recognition tasks for representing frequency elements in audio signals. MFCC can be obtained for each frame after dividing audio signals into predetermined frames. The rhythmic content features are basically obtained by using DWT and include the mean and variances of relative amplitude of peaks, beat histogram sum. When we concatenate these features, it becomes a 90-dimensional feature. Note that the Iris data set includes the feature information already and no feature extraction procedure is necessary.

For evaluating the performance of CIM-CT and other conventional classifiers, each data set is divided into two parts: 1) training data: randomly selected data samples, 90% of the total data, and 2) test data: the remaining 10% data. Experiments with 10 different combinations of training and test data sets are carried out for the evaluation of different classifier schemes. The classification accuracies are shown in terms of mean and standard deviation of the test results after these 10 trials of training and evaluating the trained classifiers.

Experiments on Iris data set are carried out first in order to show how the CIM-CT and conventional classifier can be compared. The Iris data set consists of 3 classes with 50 instances each class. These three classes represent three species of iris flowers (Iris setosa, Iris virginica, and Iris versicolor). Each data has 4 features and these features represent the length and the width of the sepals and petals for a given iris flowers. Iris data classification problem has been widely used as a benchmark problem in pattern recognition literature. Note that one class in this data set is linearly separable from the other two classes while the two classes are not linearly separable from each other [12].

The all possible combinations of four input variables are applied to CIM-CT. For CIM-CT, we can have 12 different combinations ($_4C_1 + {}_4C_2 + {}_4C_3 + {}_4C_4 = 12$) of input features

that can be used as an input for a classifier. Note that "# of classifiers 4" in Table I implies all the 4 features are used separately as 4 of one-dimensional features and each one-dimensional input vector is used separately for separate classifier for CIM-CT. Note also that " 01/23" in "# of classifiers 2" in Table I represents the case when the first 2 features (feature #0 and feature #1) are concatenated and used as an input to the Classifier #1 while feature #2 and feature #3 are concatenated and used as another input to the classifier #2. Table I summarizes the results. As can be noticed from Table I, classification accuracies for all different combinations of features are far higher than the conventional classifier that utilizes all the 4 features as the one 4-dimensional feature vector (as shown in the first column of Table I) while CIM-CT can combine various combinations of features CIM-CT shows 67.08% - 95.00% accuracy on average while conventional classifier that concatenates all 4 features shows 57.50% on average. Note that the case of "# of classifiers 2" with 2 inputs that consists of concatenating the first 3 features (feature #0, feature #1, and feature #2) to one input and feature #3 as the other input shows the best classification accuracy, 95.0%. The results show that the accuracy is higher when the feature #0 and the feature #1 are concatenated and used as an input to the local classifier. Note also that the best combination of features and the number of classifiers are not known in advance.
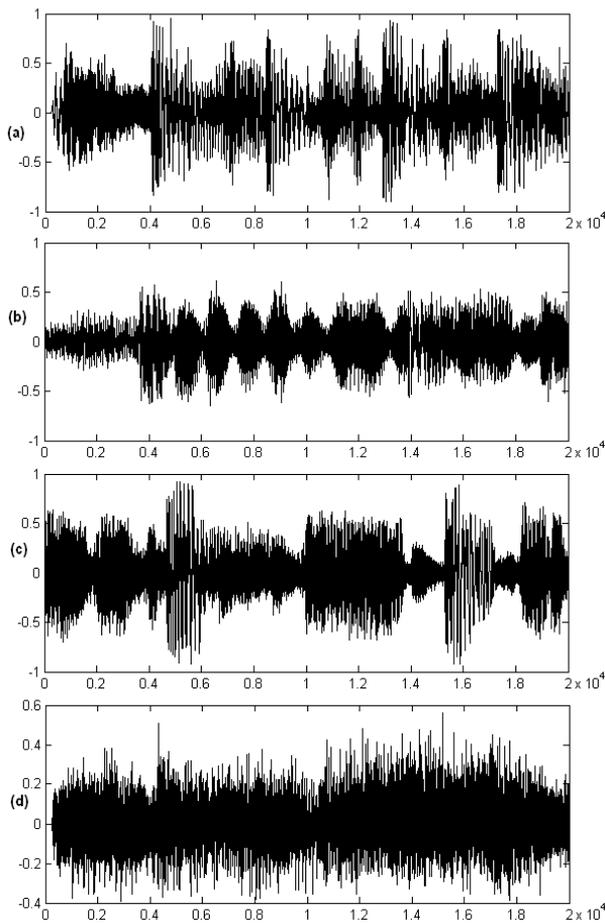
Fig. 4. Examples of audio signals: (a) Hiphop, (b)Folk, (c)Pop, (d)Rock [9].

The classification results for the conventional classifier, CIM with utilizing only q^k_{ii}, and CIM-CT with utilizing q^k_{ij} on audio data are then summarized in Table II. As can be witnessed from Table II, both of the classifiers based on CIM with utilizing only q^k_{ii} and CIM-CT with utilizing q^k_{ij} outperform significantly the conventional classifier that does not utilizes the concept of classifier integration. The results shown in Table II imply that the CIM-CT outperforms the conventional CIM with utilizing only q^k_{ii} when we consider the classification accuracy. Note also that the standard deviation for the classification accuracies of CIM with utilizing q^k_{ij} is 1.2% and 0.75% lower than the conventional classifier and CIM with utilizing only q^k_{ii}, respectively.

TABLE I: Classification Accuracies with Different Combinations of Features on CIM-CTs for Iris Data

# of classifiers	Feature groups	Avg. (Std. Dev.)
1	0123	57.5 (14.5)
2	01/23	89.6 (9.1)
	02/13	89.6 (12.1)
	03/12	67.1 (8.2)
	123/0	90.8 (7.4)
	023/1	73.3 (10.0)
	013/2	91.3 (9.7)
	012/3	95.0 (4.4)
3	01/2/3	93.3 (7.1)
	02/1/3	86.7 (5.8)
	03/1/2	83.3 (5.8)
4	0/1/2/3	85.8 (7.0)

TABLE II: Classification Accuracies with Different Classifier Schemes on Audio Data Classification Problem

All features–in-One	CIM with q^k_{ii} only	CIM with q^k_{ij}
66.2% (5.32)	75.8% (4.87)	91.3% (4.12)

After successful evaluation of the proposed CIM-CT on Iris data set, we perform some experiments on more practical classification problem with audio signal data. In these experiments with audio signal data, there also exist different sets of features obtained from different feature extraction methods on audio signal data. In experiments, the conventional classifier with concatenating all the available features, the conventional CIM with utilizing only q^k_{ii}, and the CIM with utilizing q^k_{ij} are evaluated on on audio signal data. Note that the conventional All features–in-One classifier concatenates all the features extracted from audio signals, 30-dimensional Timbral texture, 17-dimensional Pitch contents, 9-dimensional FFT coefficients, 10-dimensioanl MFCC, and 24-dimendional DWT coefficients and the dimension of its input is 90 (= 30+17+9+17+24).

TABLE III: Classification Accuracies with Different Algorithms on Audio Signal Data for Different Classes

	All features -in-One	CIM with q^k_{ii} only	CIM with q^k_{ij}
Speech	97.3	98.2	98.7
Pop	51.5	62.7	88.4
Rock	67.1	81.0	92.2
Country	90.2	91.2	95.5
Fork	25.3	42.3	82.2
Classical	97.2	98.1	98.2
Hiphop	74.6	81.3	92.6
Blues	86.2	88.4	92.1
Jazz	24.5	47.3	82.3
Metal	48.2	67.5	90.7
Avg.	66.2	75.8	91.3

Table III shows a summary of classification results for

different algorithms on audio signal data for each of the specific classes. The CIM-CT with utilizing q^k_{ij} shows a remarkable improvement on Pop, Rock, Fork, Jazz, and Metal class data over conventional All features–in-One classifier and CIM with utilizing only q^k_{ii}. The conventional classifier with all features–in-One by concatenating all the available 90-dimensional features suffers from confusing Jazz, Fork, Pop, Rock, and Metal significantly. Similar trend is continued with the case for CIM with utilizing only q^k_{ii} even though some improvement over the conventional classifier with All features–in-One is achieved. The improvement is more significant when CIM uses q^k_{ij}. Note that the confusion in classifying Jazz, Fork, and Pop leaves a room for improvement in future research. Research on proper features to discriminate Jazz, Fork, and Pop data might be the next step for improving the classification accuracy.

Table IV shows the confusion matrix of the CIM with q^k_{ij} on our audio data classification problem for each of the 10 class data. The confusion matrix shows that how each of the 10 class data is classified correctly and how each of the class data is confused with other classes. For example, Jazz class data are least accurately classified and confused mostly with Blues class data as Table IV implies. If we have additional local classifiers with more appropriate features that can discriminate the data between Jazz and Blues data to the CIM, the overall classification accuracy of the classifier should be improved without sacrificing accuracies of other classes.

TABLE IV: CONFUSION MATRIX OF THE CLASSIFICATION RESULTS OF CIM-CT WITH QKIJ ON AUDIO DATA CLASSIFICATION PROBLEM

	Spch.	Pop	Rock	Cntry.	Fork	Clas.	Hphp.	Blues	Jazz	Metal
Speech	98.7	0.3	0	0.4	0.1	0	0.1	0.2	0.2	0
Pop	1.5	88.4	0.9	3.7	4.6	0	0.1	0.3	0.4	0.1
Rock	0.3	2.5	92.2	0.1	2.5	0	1.7	0.2	0.5	0
Cntry	0.7	2.2	0.3	95.5	0.8	0	0.1	0.3	0.1	0
Fork	0.6	5.4	4.8	3.6	82.2	0	0.2	1.4	1.8	0
Classic	0	0.1	0	0.4	0	98.2	0	0.7	0.6	0
Hiphop	1.8	0.4	2.3	0.1	0.2	0	92.6	1.2	0.9	0.5
Blues	0.1	0.3	0.1	2.1	1.7	0	0	92.1	3.6	0
Jazz	2.8	1.7	1.2	2.9	1.3	0.1	1.4	6.3	82.3	0
Metal	0	0.4	4.6	0.1	0.1	0.1	3.8	0	0.2	90.7

IV. CONCLUSIONS

In this paper, a classifier integration model that utilizes efficiently the confusion information on local classifiers is applied to audio signal classification problem. This classifier model can reduce the complexity of classifiers that stems from the multi-dimensional feature information extracted through various feature extraction methods from data sets. The CIM with Confusion Table (CIM-CT) also uses each feature vector separately as an input feature for an independent classifier while conventional classifiers use the entire feature vectors extracted from the original data as a whole by concatenating these feature vectors as an input feature for the classifier. The CIM-CT utilizes the tendency how each specific local classifier made incorrect classification in addition to the probability that the classifier made correct classifications in the past. The tendency how a local classifier made incorrect classification can be obtained from the confusion table in terms of q^k_{ij} while q^k_{ii} implies the correct classification. In order to evaluate the CIM-CT with q^k_{ij} information from confusion table, experiments on two data sets are carried out. The performances of different classifier schemes such as a conventional classifier, CIM with utilizing only q^k_{ii} and CIM with q^k_{ij} are measured in terms of classification accuracy. The results summarized in Table I show that the CIM with q^k_{ij} outperforms significantly the conventional All features–in-One scheme in terms of classification accuracy when Iris data set is utilized for experiments. The results summarized in Table II and Table III show that the CIM with q^k_{ij} outperforms the other two classifier schemes significantly when specific class data such as Jazz, Fork, Pop, Rock, and Metal are considered. These 4 class data are mostly confused in the conventional classifier with All features–in-One by by concatenating all the available 90-dimensional features. The improvement of the CIM-CT over the conventional All features–in-One classifier and CIM with utilizing only q^k_{ii} on these 4 class data is significant.

Furthermore, the CIM with non-diagonal elements in addition to the diagonal elements of confusion tables shows very stable classification accuracy when the standard deviation of the classification accuracy is considered. The results also imply that the CIM-CT scheme can be successfully used for more complicated pattern classification problems that are involved with various multi-dimensional features on data sets. We can conclude that the proposed CIM-CT can alleviate the complexity of the classifiers involved in practical problems.

REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, 2002.
[2] A. Meng, P. Ahrendt, J. Larsen, and L. Hansen, "Temporal feature integration for music genre classification," *IEEE Trans. on Audio Speech and Language Processing*, vol. 15, no. 5, pp. 1654-1664, June 2007.
[3] J. Foote, "Content-based retrieval of music and audio," *Multimedia Storage and Archiving Syst. II*, pp. 138-147, 1997.
[4] N. Junejo, N. Ahmed, M. A. Unar, and A. Q. K Rajeput, "Speech and image compression using discrete wavelet transform," in *Proc. IEEE/ Sarnoff Symp. on Advances in Wired and Wireless Communication*, May 2005, pp. 45-48.
[5] G. V. Trunk, "A problem of dimensionality: A simple example," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 3, pp. 306–307, 1979.
[6] D.-C. Park, "Partitioned feature-based classifier model," in *Proc. IEEE Int. Symposium on Signal Processing and Information Technology (ISSPIT)*, 2009, vol. 1, pp. 412–417.
[7] D.-C. Park, T. Jeong, Y. Lee, and S. Min, "Satellite image classification using a classifier integration model," in *Proc. 9th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*, 2011, pp. 90-94.
[8] D.-C. Park, "Centroid Neural Network for Unsupervised Competitive Learning," *IEEE Trans. on Neural Networks*, vol. 11, no. 2, pp. 520-528, Mar. 2000.

[9] D.-C. Park, D.-H. Nguyen, S. Beack, and S. Park, "Classicfication of audio signals using gradient-based fuzzy c-means algorithm with divergence measure," *Lecture Note on Computer Science*, vol. 3767, pp. 698-708, 2005.

[10] D. Dua and E. K. Taniskidou. (2017). *UCI Machine Learning Repository*. Irvine, CA: University of California. [Online]. Available: http://archive.ics.uci.edu/ml

[11] D. McEnnis, C. McKay, I. Fujinaga, and P. Depalle, "jAudio: A feature extraction library," in *Proc. the International Conference on Music Information Retrieval*, 2005, pp. 600-603.

Miso Jang received the B.S. and M.S. degrees all in electronics engineering from MyongJi University, Rep. of Korea, in 2014 and 2016, respectively. Since 2018, she has been with MindIn Technology Inc. and participated in developing super resolution image processing with machine learning. Her research interests include image and video processing, pattern recognition, computer vision, machine learning algorithms and their applications to image enhancement problems including super resolution and haze removal.

Dong-Chul Park received the B.S. degree in electronics engineering from Sogang University, Seoul, Korea, in 1980, the M.S. degree in electrical and electronics engineering from the Korea Advanced Institute of Science and Technology, Seoul, Korea, in 1982, and the Ph.D. degree in electrical engineering, with a dissertation on system identifications using artificial neural networks, from the University of Washington (UW), Seattle, in 1990. From 1990 to 1994, he was with the Department of Electrical and Computer Engineering, Florida International University, The State University of Florida, Miami. Since 1994, he has been with the Department of Electronics Engineering, MyongJi University, Korea, where he is a professor. He is a pioneer in the area of electrical load forecasting using artificial neural networks. He has published more than 180 papers, including more than 50 archival journals in the area of neural network algorithms and their applications to various engineering problems including electric load forecasting, image compression, speech recognition, time-series prediction, and pattern recognition. Dr. Park was a member of the Editorial Board for the IEEE Transactions on Neural Networks from 2000 to 2002.