J Inf Process Syst, Vol.13, No.2, pp.400~416, April 2017
https://doi.org/10.3745/JIPS.04.0032

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Rough Set-Based Approach for Automatic Emotion Classification of Music

Babu Kaji Baniya* and Joonwhoan Lee*

Abstract

Music emotion is an important component in the field of music information retrieval and computational musicology. This paper proposes an approach for automatic emotion classification, based on rough set (RS) theory. In the proposed approach, four different sets of music features are extracted, representing dynamics, rhythm, spectral, and harmony. From the features, five different statistical parameters are considered as attributes, including up to the 4^{th} order central moments of each feature, and covariance components of mutual ones. The large number of attributes is controlled by RS-based approach, in which superfluous features are removed, to obtain indispensable ones. In addition, RS-based approach makes it possible to visualize which attributes play a significant role in the generated rules, and also determine the strength of each rule for classification. The experiments have been performed to find out which audio features and which of the different statistical parameters derived from them are important for emotion classification. Also, the resulting indispensable attributes and the usefulness of covariance components have been discussed. The overall classification accuracy with all statistical parameters has recorded comparatively better than currently existing methods on a pair of datasets.

Keywords

Attributes, Covariance, Discretize, Rough Set, Rules

1. Introduction

The automatic emotion classification of music has gained increasing attention in the field of music information retrieval. The research activities in this field are not only highly diversified, but also constantly growing. The diversity comes from the fact that emotions manifest in humans in a variety of ways. Automatic emotion classification establishes certain relationships between music and its effect in human emotional state like angry, happy, sad, etc. In addition, the growth is inevitable nowadays, to increase the accessibility to music databases. As the amount of music content continues to explode, the searching time is unexpectedly increasing. The solution of widespread music grouped under different emotions could lead to a reduction in the information retrieval time on an online system.

Even though music emotion recognition (MER) research has received increased attention in recent years, it faces many limitations and open problems. In fact, the current accuracy and feature selection of MER system show that there is plenty of room for improvement. For example, in Music Information

Manuscript received August 24, 2015; accepted July 5, 2016.
Corresponding Author: Joonwhoan Lee (chlee@jbnu.ac.kr)
* Dept. of Computer Science and Engineering, Chonbuk National University, Jeonju, Korea ({everwith_7, chlee}@jbnu.ac.kr)

Retrieval Evaluation eXchange (MIREX), the highest recorded classification accuracy in the mood classification task was around 65%.

Some of the major difficulties in MER are related to the fact that the perception of emotions evoked by a song is inherently subjective. Besides, there is still much ambiguity regarding its description [1]. In general, there are two models to describe emotions, i.e., a categorical and a dimensional one [2]. The categorical model focuses on the characteristics that distinguish emotions from one another. The essentiality of this model is the concepts of basic emotions, such as happiness, sadness, anger, fear, disgust, and surprise. On the other hand, the dimensional model focuses on identifying emotions based on their position in a continuous dimensional emotion space with a small number of axes. In the space, each bipolar axis usually has its own meaning i.e. valence or arousal.

In addition to the description model of emotion, recognition schemes must be addressed with a set of features involved in MER as in other recognition problems. Several schemes have been put forward for MER, including regression [1], ranking-based [3], and multi-modal [4]. The regression scheme represents each song as a point on an emotion space, so that it can be categorized as an approach that uses the dimensional model. The ranking-based scheme determines the coordinates of a song on a 2-D emotion space in a dimensional model, by the relative emotion of the song with respect to other songs, rather than directly computing the exact emotion of the song. The multi-modal approach is to combine information from distinct sources, namely audio, MIDI, and the lyrics of a song. In this approach, the information on music emotion is extracted separately from each modality, and later is combined, for improving the classification accuracy.

Usually, features for MER can be obtained using an audio toolbox, such as PsySound [5], Marsyas [6], MIR [7] or MA [8], which includes many feature extractors. Frequently, features that can be extracted by the toolbox are categorized in several groups, like dynamics, rhythmic, spectral, harmonic, and so on. All the above mentioned schemes use diverse audio features, and various machine learning algorithms. For example, the regression scheme used 114 features, and considered three different kinds of classifier, i.e., multiple linear regression (MLR) [9], Adaboost.RT [10], and support vector regression (SVR) [11]. In the ranking-based [3] scheme, it used 157 features for evaluation, and also considered two different classifiers for performance measurement, including radial basis function (RBF) [12] and SVR [11].

Although those learning approaches could give good results in terms of accuracy, it is not easy to clearly interpret the learned classifier. Originally, rough set (RS) theory was invented in 1982 by Zdzislaw Pawlak, and is concerned with the classification and analysis of imprecise, uncertain or incomplete information and knowledge [13]. Different from the conventional machine learning algorithm, the RS-based approach represents the classification rules with "IF-THEN" rules. In addition, the attributes to check the firing condition of each rule are well preserved without any transformation, as in other learning algorithms. Therefore, the RS-based approach make it easy to visualize which attributes are important, after removing dispensable ones in the rule generation process. Furthermore, the set of logical decision rules, with their strength for classification of a certain class has been obtained.

In this paper, a RS-based approach for the classification of music emotion has been proposed. Since the output after rule generation process is classification rules, this approach could be inherently characterized as a kind of categorical approach in emotion recognition.

In the paper, four different groups of audio features have been selected, including dynamics, rhythmic, spectral, and harmony. From each of the four frame-based features, up to the 4^{th} order central

moments for a music object have been extracted, such as mean, variance, skewness and kurtosis [14,15]. Because it has recently been reported that the higher order moments, like skewness and kurtosis are useful for the genre/mood classification of music [16,17]. In addition, covariance components of pairwise features within the same group of features have been included. For feature extraction, the MIR Toolbox [7] has been taken. Also, the rough set toolbox Rose2 has been used for the experiment of rule generation and classification [18]. Our experiment in Section 3 shows that only a limited number of attributes (statistical parameters) from several audio features turn out to be important and indispensable, which are attack slope, attack time, spectral flux, irregularity, spectral roughness, some of the mel-frequency cepstral coefficients (MFCCs), key clarity, harmonic change detection function (HCDF), and so on.

Another experiment shows that a group of harmony features could give the highest classification accuracy (58%) for emotion classification, as compared to others, i.e., spectral (52%), rhythm (46%), and dynamic (28%) ones. Also, the experiment shows that additional covariance components can increase the performance of accuracy.

To evaluate the overall classification accuracy of RS-based approach, two different datasets are used. The first one was obtained from music tracks of the University of Jyvaskyla (Jyu), Finland [19], and the other from the Centre of Informatics and Systems of the University of Coimbra, Portugal [20]. The former has 50 pieces of music for 4 different classes, and the latter 903 pieces for 5 classes. When including all features and statistical parameters with covariance components, the proposed approach provides comparatively better performance than the others [13,21]. An overall accuracy of around 72% has been achieved for a small formal dataset. With additional covariance components, the accuracy has been increased by 4.5, which indicates that the additional covariance components are helpful.

The rest of this paper is organized as follows. Section 2 describes the proposed RS-based approach, including the audio features that are used. Section 3 describes the experimental results, with discussion. Finally, Section 4 describes the conclusion of the proposed method, and suggests future work for the emotion classification of music.

2. Rough Set Approach Extraction

2.1 Overview of Proposed Approach

One of the important components of emotion classification is the emotion model. In the previous section, there are two emotion models, the categorical and the dimensional. The RS-based approach cannot be a dimensional model, but is a categorical model, because of the inherent nature of the RS approach. Basically, the decision rules from a rough set approach provide a class label so that it cannot be used for a regression problem.

In the dimensional model, Russell's two-dimensional emotion space [22] has been widely used, in which each axis corresponds to valence and arousal, respectively, as shown in Fig. 1. One can roughly partition such continuous space into several categories for classifying music. For example the dataset obtained from Jyu has four classes of emotion including "happy", "angry/fear", "sad", and "tender", which occupy each quadrant from one to four respectively, as shown in Fig. 1. Therefore, the distinction between the two models seems to be caused by the granularity of classes in a continuous space. In

addition, it is noted that any finite number of emotion classes can be allowed, in the RS-based scheme. For example, the dataset in the experiment has 5 classes, which may be allocated in Russel's emotion space.

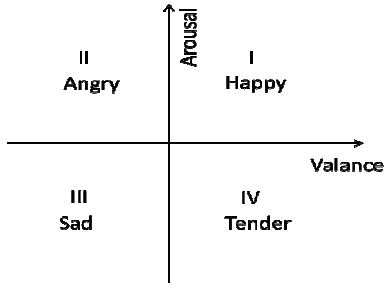

Fig. 1. Emotion space and selected categories.

An overall block diagram of the RS-based scheme is shown in Fig. 2. Feature extraction, normalization and discretization process are shared with both training and classification processes. The goal of training is rule generation with minimal attributes. The generated rules are used to classify the emotion of a music object.

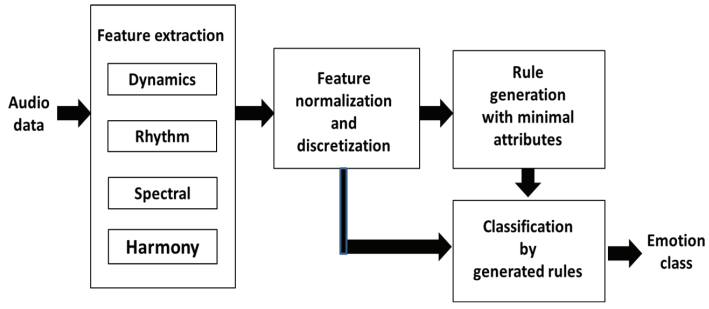

Fig. 2. Block diagram of the proposed rough set-based scheme.

After feature extraction from audio data, a set of statistical parameters up to the 4th order central moments for each frame-based feature is extracted, and a covariance value of pair-wise features. The parameters and covariance values are treated as condition attributes in the rough set approach. Because the ranges of attribute value are diverse, they require compulsory normalization. Thereafter, discretization [23] of the decision table is required, to obtain a decision table, because the rough set approach is based on discrete attribute values. Without a discretized table, there is little chance that a new music object is recognized by a rule directly generated from the normalized table. Therefore, a discretization table is essential, for obtaining high quality classification rules. The next important step is to remove all the dispensable attributes, keeping only the indispensable ones, in order to obtain simplified classification rules in a suboptimal way. In this process, one can choose strong rules, for further simplification of decision rules.

The final stage is to classify the music emotion, based on the obtained rules, for unlabeled music objects. In the research rough set toolbox, Rose2 has been used for training and classification.

2.2 Audio Features

Feature extraction encompasses the analysis and extraction of meaningful information from audio data, in order to obtain a compact and concise description. The features in MIR Toolbox [7] have been chosen in our research, which are divided into four groups, of dynamics, rhythmic, spectral, and harmony, as shown in Table 1.

Frequently, the frame-based features are categorized with respect to the frame length; the feature extracted with shorter (longer) frame is called low (high)-level. In the research, the frame length for the low-level and high-level features were 46 ms and 2 s, respectively, with both having 50% overlap. Dynamics and spectral, and rhythm and harmony features correspond to low and high-level, respectively, as denoted in the first column of Table 1 in parenthesis.

Table 1. Audio features extraction for rough set-based music mood classification

Feature category	No. of feature/no. of attributes	Name of feature	Statistical parameters
Dynamics	3/15	RMS energy	M, SD, S, K, PC
		Slope	"
		Attack	"
Rhythm	1/4	Tempo	"
Spectral	11/99	Spectral centroid	"
		Brightness	"
		Spread	"
		Rolloff85	"
		Rolloff95	"
		Spectral entropy	"
		Flatness	"
		Roughness	"
		Irregularity	"
		Zero-crossing	"
		Spectral flux	"
	1(13 comps/78)	MFCC (13)	"
Harmony	5/30	Chromagram peak	"
		Chromagram-centrod	"
		Key clarity	"
		Key mode	"
		HCDF	"

Because each frame-based feature provides a sequence of frame-wise values for an audio data, that helps to find out a set of a small number of meaningful scalar values. Usually, for the representative scalar attributes, the sample mean and standard deviation over all frames are taken.The mean (μ) and standard deviation (σ) for a N-frame sequence X_n are calculated by Eqs. (1) and (2), respectively.

$$Mean(\mu) = \frac{1}{N}\sum_{n=1}^{N} X_n \quad (1)$$

$$std(\sigma) = \frac{1}{N}\sqrt{\sum_{n=1}^{N}(X_n - \mu)^2} \tag{2}$$

In addition, the high-order central moments such as skewness and kurtosis over all frames have been taken. Recently, it has been reported that the higher order statistical parameters including skewness and kurtosis are helpful in classifying the genre of music. That is why a feature represented by a random variable may not be Gaussian distributed [24]. The skewness is a measure of asymmetry of a feature sequence, which can be calculated by

$$Kurtosis = \frac{\sum_{n=1}^{N}(X_n - \mu)^4}{(N-1)\sigma^4} - 3 \tag{4}$$

In addition, the inclusion of covariance components of pair-wise features has been proposed in the same group. It is noted that the features in the same group in Table 1 have the same number of frames for an audio data, because they are calculated using the frame length and overlap. Covariance is useful to grasp the polarity, and the degree of relationship between two features. The covariance of two feature sequences X_n and Y_n of an audio data is calculated by

$$Cov(X_n, Y_n) = \frac{1}{N}\sum_{n=1}^{N}(X_n - \mu_X)(Y_n - \mu_Y) \tag{5}$$

where μ_X and μ_Y are the means of X_n and Y_n, respectively.

Therefore, $4n$ attributes of central moments, and $n(n\text{-}1)/2$ attributes of covariance components for n features have been considered in a group. The number of features in each group is represented in the second column of Table 1, with the corresponding number of statistical parameters in parenthesis that are used for condition attributes. Note that for covariance computation MFCC components are separated from other spectral features.

2.3 RS-Based Decision Rule with Minimal Attributes

The selection of suitable attributes is an important problem. A RS-based scheme can significantly reduce the attributes while keeping meaningful information intact [25,26] without using prior knowledge. During the rule generation process in rough set theory, the dispensable attributes are removed so that the final decision table can contain only minimal or indispensable attributes in a suboptimal way.

For the sake of simplicity, in this section the attributes of only two statistical parameters (mean and standard deviation) has been considered, from the features in Table 1.

In rough set theory, a decision table is represented by $T = (U, Q, V, f)$, where U is the universe, with a set of finite objects. For classification problem, a set of attributes $Q = (A, D)$ usually consists of a finite set of condition attributes A, and a decision attributes D. Let

$$V = \bigcup_{q \in Q} V_q \ , \tag{6}$$

where V_q is the domain of the attribute q, and f denotes the total decision function as $f : U \times A \rightarrow V$.

The normalized decision table after feature extraction is shown in Table 2. For the normalization, we simply take the maximum of the absolute value of each attribute to divide the attribute values, so that the range is confined in [-1, 1]. Each row of the table represents a music object, and each column represents an attribute. Therefore, there are 66 condition attributes (A_1–A_{66}) derived from audio features in Table 1, and four different emotion classes for a decision attribute.

If the table has a large number of attribute values, i.e., $card(V_q)$ is very large for some $q \in Q$, then there is little chance of classifying a new music object by the rules generated from the table. Here, *card*() means the number of elements in a set. Therefore, discretization of the data table is essential, by proper partitioning of the attribute values. There are two methods of discretization: local and global. The local method is characterized by the operation being focused on only one attribute at a time, similar to a decision tree. But the global method is characterized by considering all attributes, to determine interval break points. Because the global method provides more efficient partitions than the local, based on that the global method is preferred, even though it takes a larger computation time. Several efficient discretization algorithms that use maximal discernible (MD) heuristics are discussed in detail in [25]. binary discretization has been used for each attribute, because it not only takes a smaller amount of time, but it also makes the structure of final rules and attributes more apparent, for easy interpretation. The part of the decision table discretized into binary values is shown in Table 3.

Table 2. An example of decision table for rule generation

Objects	Condition attributes								Decision attributes
	A_1	A_2	A_3	A_4	A_5	A_6	..	A_{66}	
1	0.792034	0.207966	0.570637	0.486101	0.513899	0.758504	..	0.365720	1
2	0.485027	0.514973	0.43195	0.569337	0.430663	0.732195	..	0.405596	1
3	0.768222	0.231778	0.616885	0.510221	0.489779	0.850356	..	0.293740	1
:	:	:	:	:	:	:		:	:
21	0.704359	0.295641	0.609538	0.390462	0.527423	0.472577	..	0.376042	2
22	0.722238	0.277762	0.581326	0.418674	0.590535	0.409465	..	0.326697	2
23	0.733627	0.266373	0.719684	0.280316	0.578858	0.421142	..	0.384288	2
:	:	:	:	:	:	:		:	:
31	0.679982	0.320018	0.503371	0.496629	0.464181	0.535819	..	0.439244	3
32	0.764347	0.235653	0.593865	0.406135	0.480905	0.519095	..	0.334624	3
33	0.753793	0.246207	0.509861	0.490139	0.480077	0.519923	..	0.400061	3
:	:	:	:	:	:	:		:	:
48	0.821101	0.178899	0.565414	0.434586	0.497072	0.502928	..	0.392448	4
49	0.778716	0.221284	0.650779	0.349221	0.559507	0.440493	..	0.406685	4
50	0.697921	0.302079	0.551555	0.448445	0.534628	0.465372	..	0.476635	4

1=anger/fear, 2=happy, 3=sad, 4=tender in decision attribute.

Table 3. Discretized decision table by global approach

Objects	Condition attributes						Decision attributes
	A_5	A_{25}	A_{48}	A_{55}	A_{61}	A_{63}	
1	0	1	0	0	0	0	1
2	1	1	0	0	1	0	1
3	1	0	1	1	1	0	1
:	:	:	:	:	:	:	:
21	1	1	1	1	1	1	2
22	1	0	1	0	1	1	2
23	1	1	1	1	1	0	2
:	:	:	:	:	:	:	:
31	0	0	0	0	0	1	3
32	0	0	1	1	1	1	3
33	0	1	1	0	0	1	3
:	:	:	:	:	:	:	:
48	0	1	1	0	1	1	4
49	1	1	0	1	1	1	4
50	1	1	0	0	0	1	4

Let $R \subseteq Q$ and $y_a, y_b \in U$. Then, y_a and y_b are indiscernible with respect to the set of attributes R in U, if and only if $f(y_a, q) = f(y_b, q)$, $\forall q \in R$. An elementary set is defined as a set of all indiscernible objects, with respect to a set of specific attributes. An equivalence relation on U for $R \subseteq Q$ is called an R-indiscernibility relation, and denoted as I_R. For example, in I_{A48} from Table 3, the attributes A_{48} generate I_{A48} ={{3,…, 21, 22, 23,,… 32, 33,…, 48},{1,2,…, 31,…, 49, 50}}, because all the elements in each class have the same attribute value. Therefore, the elements in a class are indiscernible, in terms of A_{48}.

Pawlak [27] introduced the concept of approximation in rough set theory. If X is any rough set then $\underline{R}X$ and $\overline{R}X$ are called the lower and upper approximation of X and are defined as $\underline{R}X = \{y \in X \mid I_R(y) \subseteq X\}$ and $\overline{R}X = \{y \in X \mid I_R(y) \cap X \neq \phi\}$, respectively. The concept of lower approximation plays an important role in finding the indispensable attributes.

In RS theory, the reduct and core of attributes are used for reduction of the decision table. A reduct is defined as a minimal set of attributes that can classify the domain of objects as unambiguously as the original set of attributes. A set of attributes occurring in every reduct is called a core. The decision rules can be made based on the attribute in a reduct.

For a given subset of attributes $R \subseteq Q$, an attributes $q \in R$ is dispensable in the set R if and only if $I_R = I_{R-\{q\}}$; otherwise q is indispensable. Let $R \subseteq Q$, and $D \subseteq Q$ have the equivalence relation in U. The R-positive region of D is defined as

$$POS_R(D) = \cup_{z \in I_D} \underline{R}Z, \tag{7}$$

which indicates the set of objects that can be correctly classified into D-elementary set generated by I_D, using the knowledge expressed by I_R.

If $q \in R$ and $POS_R(D) = POS_{R-\{q\}}(D)$, then q is D-dispensable in R, otherwise q is D-indispensable in R. If the set of attributes $G(G \subseteq R)$ is D-indispensable in R and $POS_G(D) = POS_R(D)$, then G is called the

D-reduct of R. To explain the above definitions in detail, let us consider a part of Table 3, as shown in Table 4. Suppose that there are a set of condition attributes $R = \{A_5, A_{25}, A_{48}\}$ and a decision attribute D. Then the corresponding elementary sets are $I_R = \{\{1\}, \{2\}, \{22\}, \{23\}, \{31\}\}$, and $I_D = \{\{1,2\}, \{22,23\}, \{31\}\}$. If an attribute A_5 is removed from R, then $POS_{R-\{A_5\}}(D) = \{1,2,22,23,31\}$.

This indicates that, $POS_{R-\{5\}}(D) = POS_R(D)$, so that the attribute A_5 is D-dispensable in R. Similarly for attribute A_{25}, is D-indispensable in R. Attribute A_{48} again gives $POS_{R-\{A_{48}\}}(D) = \{1,22,31\} \neq POS_R(D)$. Therefore attribute A_{48} is D-indispensable in R. Thus the set $\{A_{25}, A_{48}\}$ is the D-reduct of R.

Table 4. Truncated decision table from Table 3 for illustration

Objects	Condition attributes			Decision attributes
	A_5	A_{25}	A_{48}	
1	0	1	0	1
2	1	1	0	1
22	1	0	1	2
23	1	1	1	2
31	0	0	0	3

By taking the reduct Table 4 can be reduced to Table 5, where '-' symbol indicates the dispensable attribute by taking a reduct. The attribute values (A_{25}=1 Λ A_{48}=1), (A_{48}=1), and (A_{25}=0) indicate the characteristics of decision class 1, 2, and 3, respectively. Note that A_{25} and A_{48} are meaningless to identify class 2 and class 3, respectively, in this specific example. Symbols 'Λ' and 'V' represent conjunction and disjunction operators, respectively. These attribute values are called reduct values. Intersections of the reduct values for each of the decision class yield the core values for the class. In this example, the decision classes 1, 2, and 3 have the same core values as their reduct values, respectively.

Table 5. Reduced table from Table 4

Objects	Condition attributes			Decision attributes
	A_5	A_{25}	A_{48}	
1	-	1	0	1
2	-	1	0	1
22	-	0	1	2
23	-	1	1	2
31	-	0	0	3

Table 6. Reduced table from Table 5

Decision rule no.	Statement of the rule	
	IF	THEN
1	(A_{25}=1 Λ A_{48}=0)	1
2	(A_{48}=1)	2
3	(A_{25}=0 Λ A_{48}=0)	3

This reduced Table 5 can be implemented to generate the decision rules in "IF-THEN" format. Table 6 shows an example of the decision rules obtained from a part of the whole decision table.

The rule generation from real data is not such a simple task as in the above example, because there

can be insignificant rules from noisy data. Therefore, after removing dispensable attributes only significant rules can be chosen for further simplification. Usually the rule section process is based on the evaluation of goodness for each derived rule, after removing the dispensable attributes.

The strength and coverage can be used as the measure of goodness for selecting a decision rule. These are defined as follows.

The rule generation from real data is not such a simple task as in the above example, because there can be insignificant rules from noisy data. Therefore, after removing dispensable attributes only significant rules can be chosen for further simplification. Usually the rule section process is based on the evaluation of goodness for each derived rule, after removing the dispensable attributes.

The strength and coverage can be used as the measure of goodness for selecting a decision rule. These are defined as follows.

$$Strength = \frac{card\left(A(S) \cap f(S)\right)}{card\left(U\right)} \tag{8}$$

$$Coverage = \frac{card\left(A(S) \cap f(S)\right)}{card\left(f(S)\right)} \tag{9}$$

In the equations $A(S)$ and $f(S)$ denote the number of cases (objects) that can be captured by a rule S in all decision classes, and that in the corresponding decision class, respectively. From Table 2, Table 7 has been obtained, after removing the dispensable attributes. The last column of the table represents the number of captured cases (objects), with corresponding strength/coverage values in parenthesis.

Table 7. Simplified form of the decision table

Rule no.	Condition attributes						Decision attributes	No. of captured cases (strength/coverage)
	A_5	A_{25}	A_{48}	A_{55}	A_{61}	A_{63}		
1	0	1	1	0	0	0	1	15 (0.30/0.75)
2	1	0	1	0	1	0	1	2 (0.04/0.10)
3	1	0	1	1	1	0	1	2 (0.04/0.10)
4	1	0	0	1	1	1	1	1 (0.02/0.05)
5	1	1	1	1	1	1	2	6 (0.12/0.60)
6	1	0	1	0	1	1	2	2 (0.04/0.20)
7	1	1	1	1	1	0	2	2 (0.04/0.20)
8	0	0	0	0	0	1	3	7 (0.14/0.70)
9	0	0	1	1	1	1	3	2 (0.04/0.20)
10	0	1	1	0	0	1	3	1 (0.02/0.10)
11	1	1	0	1	1	1	4	6 (0.12/0.60)
12	1	1	0	1	1	1	4	3 (0.06/0.30)
13	1	1	0	0	0	1	4	1 (0.02/0.10)

According to Table 7, the first rule is represented as: "*IF* (A_5=0 Λ A_{25}=1 Λ A_{48}=1 Λ A_{55}=0 Λ A_{61}=0 Λ A_{63}=0) *THEN* music emotion belongs to class 1". Similarly, rule two states that: "IF (A_5=1 Λ A_{25}=0 Λ A_{48}=1 Λ A_{55}=0 Λ A_{61}=1 Λ A_{63}=0) *THEN* music emotion belongs to class 1". There are four decision rules

for class 1. Hence, the corresponding emotion (angry/fear) of music can be described by four different rules. Other decision classes can also have multiple decision rules.

Table 8. Reduced values only considering four strong rule in Table 7

Rule no.	Condition attributes						Decision attributes
	A_5	A_{25}	A_{48}	A_{55}	A_{61}	A_{63}	
1	-	1	-	0	-	0	1
2	1	1	-	1	1	1	2
3	0	-	-	-	0	1	3
4	0	-	1	0	1	-	4

As mentioned before, one can consider only the strong decision rules with higher strength or coverage in each class, for further simplification of the rules. For example, suppose a rule that has been the strongest coverage value selected in each class in Table 7. Then there are four decision rules. Table 8 shows the corresponding reduct values for the selected four strongest rules from Table 7. The final decision rules can be generated using the reduct values in the "IF-THEN" format, as shown in Table 9.

Here, it has to note that there can be the case (object) that does not satisfy any of the "IF" conditions of simplified rules. Therefore, an additional "ELSE" condition is included. The input object that meets this "ELSE" condition can be treated as an "unknown" class.

Table 9. Decision rules obtained from reduced values in Table 8

Decision rule no.	Statement of the rule	
	IF	THEN
1	$(A_{25}=1 \Lambda A_{55}=0 \Lambda A_{63}=0)$	1
2	$(A_5=1 \Lambda A_{25}=1 \Lambda A_{55}=1 \Lambda A_{61}=1 \Lambda A_{63}=1)$	2
3	$(A_5=0 \Lambda A_{61}=0 \Lambda A_{63}=1)$	3
4	$(A_5=0 \Lambda A_{48}=1 \Lambda A_{55}=0 \Lambda A_{61}=1)$	4
5	ELSE	Unknown

3. Experimental Results and Discussion

In order to validate the proposed rough set approach, there are several experimental results included in this section. In all the experiments, 5-fold cross-validation has been used to evaluate the performance of the classification accuracy. Also, the strong rules have been selected, for further simplification of the decision rules. The rules are removed in a class if they have less than 1/10 of the strongest rule, with respect to coverage value. For classification of music, a pair of datasets has been used. One was obtained from the music tracks of Jyu, Finland [19], and the other from the University of Coimbra, Portugal [20].

The Jyu dataset are divided into four different classes, for 50 music data. Each class has a different number of musical items: anger/fear (20), sad (10), happy (10), and tender (10), where the value in each parenthesis represents the number of musical items. There are four distinct groups of features in Table 1. In order to find out the degree of influence of each group, the Jyu dataset with each separate group of features have been tried to classify emotion. In this experiment all the statistical parameters as condition

attributes have used including mean, standard deviation, skewness, kurtosis and covariance components. Among them, harmony features provided the highest classification accuracy (58%). The spectral category has a larger number of features, but it gave a classification accuracy lower than that of the harmony category. The rhythmic feature, which has the least number of features among the four groups, showed good classification accuracy (46%). The result is summarized in Fig. 3.

In the second experiment tried to find out which pair or set of statistical parameters gives the better accuracy. For the experiment all four groups of features have been used in the Jyu dataset. The results of the experiment are shown in Table 10. The overall classification accuracy of the two statistical parameters of mean and standard deviation were 61.5%. Adding high order moments like skewness and kurtosis, to the mean and standard deviation increased the classification accuracy up to 4.5% and 4.0% respectively. When the four statistical parameters of mean, standard deviation, skewness, and kurtosis were combined together, the classification accuracy reached 67.5%. The overall classification rate, after adding the covariance components with four statistical parameters, was increased by 4.5%. Thus it found that the covariance components also contributed to improving the performance of music emotion classification.

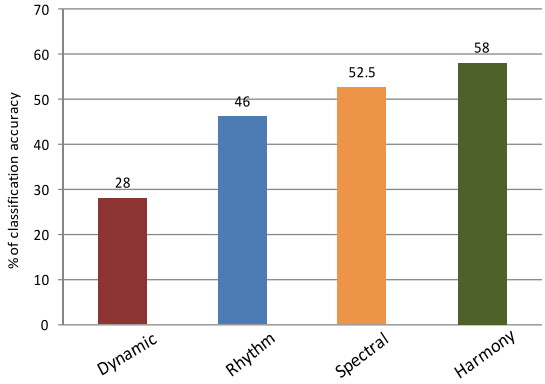

Fig. 3. Classification accuracy of music emotion based on different feature categories.

Table 10. Emotion classification considering with different statistical parameters

Conditional attributes	No. of final rules/ no. of attributes in final rules	Attributes in final rules	Overall classification accuracy (%)
M, SD	13/6	I = {mean of AS, SR, the 8th comp. of MFCC, KC, HCDF, SD of 1st comp. of MFCC}	61.5
M, SD, S	17/16	II = IU{skewness of AS, SR, 1st and 8th comp. of MFCC, HCDF, T, ST, F, SF}	66.0
M, SD, K	16/15	III = IU{Kurtosis of AS, SR, 1st and 8th comp. of MFCC, T, ST, F, SF, IR}	65.5
M, SD, S, K	19/25	IV = I U II U III	67.5
M, SD, S, K, PC	23/31	V = IV U{6 covariance components}	72.0

AS=attack slope, SR=spectral roughness, KC=key clarity, T=tempo, ST=slope time, F=flatness, SF=spectral flatness, IR=irregularity, MFCC=mel-frequency cepstral coefficient, HCDF=harmony change detection function.

In addition, it was observed that in constructing final classification rules, only several features and the statistical parameters were effective. The effective features are listed in the last column in Table 10, which are AS (attack slope), AT (attack time), T (tempo), SF (spectral flux), SR (spectral roughness), IR (irregularity), the 1st component MFCC, the 8th component MFCC, KC (key clarity), and HCDF. Again, note that there are 6 indispensable attributes from covariance components remaining in the final classification rules, as shown in the last row of Table 10.

Another notable thing is that the effective attributes in the final classification rules in the 3rd column of Table 10, when increasing the statistical parameters. The effective attributes in the final rules with additional statistical parameters include all the attributes in the previous trial, without the additional parameters. For example, when all 5 statistical parameters were considered, the effective attributes in the final rules included all the attributes obtained from the trial with 4 statistical parameters. This fact can be verified, by comparing every entry of the third column in Table 10. This means that the attributes taken to be indispensable in a trial with a smaller number of condition attributes are always selected to be indispensable, even though the number of condition attributes is increased. In this opinion, this might happen due to the orthogonal nature of tessellation in global discretization.

Table 11. Classification rules and coverage

Emotion class	Decision rules	Coverage (%)
Anger/fear	$(A_{25}=0) \Lambda (A_{55}=1) \Lambda (A_{63}=0)$	75
	$(A_5=0) \Lambda (A_{25}=1) \Lambda (A_{63}=0)$	10
	$(A_5=1) \Lambda (A_{25}=0) \Lambda (A_{61}=1) \Lambda (A_{63}=0)$	10
	$(A_5=1) \Lambda (A_{48}=0) \Lambda (A_{55}=1) \Lambda (A_{61}=0)$	5
Happy	$(A_5=1) \Lambda (A_{25}=0) \Lambda (A_{55}=0) \Lambda (A_{61}=1) \Lambda (A_{63}=1)$	60
	$(A_5=1) \Lambda (A_{25}=1) \Lambda (A_{48}=1) \Lambda (A_{55}=1)$	20
	$(A_5=0) \Lambda (A_{48}=0) \Lambda (A_{55}=0) \Lambda (A_{61}=1)$	20
Sad	$(A_5=0) \Lambda (A_{61}=0) \Lambda (A_{63}=1)$	70
	$(A_5=1) \Lambda (A_{25}=1) \Lambda (A_{55}=0) \Lambda (A_{61}=0) \Lambda (A_{63}=0)$	20
	$(A_5=1) \Lambda (A_{48}=0) \Lambda (A_{55}=1) \Lambda (A_{61}=0)$	10
Tender	$(A_5=0) \Lambda (A_{48}=1) \Lambda (A_{55}=0) \Lambda (A_{61}=1)$	60
	$(A_5=0) \Lambda (A_{25}=1) \Lambda (A_{61}=1) \Lambda (A_{63}=1)$	30
	$(A_{25}=1) \Lambda (A_{48}=0) \Lambda (A_{61}=1)$	10

The result of the rough set approach can show the strength of the classification rule, as well as the indispensable attributes. Table 11 shows the final 13 rules with coverage, only considered the statistical parameters, including the mean and standard deviation. Because the table is intended to just show the resulted rules with coverage, all music objects in the Jyu dataset were used for rule generation, without held-out data.

In this experiment the common attributes has retained, even though the final decision rules and their strengths change for every round of cross validation. This is due to the fact that 80% of the data for training is enough to capture the overall behavior of the dataset, for preserving the attributes and rules.

In general, the comparison of emotion classification is quite difficult, because the features and methods are diverging. This approach has tried to compare with others for the Jyu dataset that have been recently published. As shown in Table 12, our approach provided higher classification accuracy

than the others. The reason for the higher accuracy is partially due to the various features and additional statistical parameters, such as covariance components that have not been used before. The corresponding confusion matrix is given in Table 13.

Again, the same experiment has been performed with the attributes that have been found to be indispensable in the last row of Table 10. Even though a slightly different accuracy has been obtained for each class, the overall accuracy has not changed, as shown in Table 14. This means the rules consisting of selected indispensable attributes are close to the optimal solution so that there is little room to change the classification performance.

Because all rules did not choose after removing insignificant attributes, the "ELSE" condition is necessary to take care of uncovered objects, as mentioned in Table 4. Note that there was no such music object in Tables 13 and 14 that can be treated as "unknown". This means that the final "IF-THEN" rules from the rough set approach were so general, that they could classify all the unseen music objects in the dataset.

Table 12. Comparison of classification accuracy with other approaches of the same datasets

References	Classification accuracy (%)
Our approach	72
[18]	66
[19]	56.5

Table 13. Confusion matrix of Jyu datasets

	Anger/fear	Happy	Sad	Tender	Unknown
Anger/fear	**17**	2	1	1	0
Happy	1	**7**	0	0	0
Sad	0	1	**6**	4	0
Tender	0	2	2	**6**	0

Table 14. Confusion matrix of Jyu datasets of classification accuracy using limited number of attributes

	Anger/fear	Happy	Sad	Tender	Unknown
Anger/fear	**17**	1	2	0	0
Happy	1	**5**	2	2	0
Sad	3	0	**6**	1	0
Tender	0	2	0	**8**	0

Because the Jyu dataset has only 50 music objects, a similar experiment has been conducted with the Coimbra dataset, which consists of five different classes, for 903 music objects. In the dataset, each class contains a set of adjectives. The class, its corresponding adjectives, and the number of musical items are summarized in Table 15.

For the classification example, we obtained 53.82% of overall accuracy, which is better than the 46.3% in the research results that Panda et al. [13] obtained for the same dataset. Here, all 5 statistical parameters as attributes have been used for all 4 categories of features. The confusion matrix is shown in Table 16. In the table, one can see that there is no music object that is not captured by the final "IF-

THEN" rules, as shown in the last column. Therefore, we can conclude that the common attributes in a pair of results from both datasets are generally influential, in classifying music emotion in the RS-based approach.

Table 15. Coimbra dataset

Class label	Adjectives in class	No. of music
C1	passionate, rousing, confident, boisterous, rowdy	170
C2	rollicking, cheerful, fun, sweet, and amiable/good natured	164
C3	literate, poignant, wistful, bittersweet, autumnal, brooding	215
C4	humorous, silly, campy, quirky, whimsical, witty, wry	191
C5	aggressive, fiery, tense/anxious, intense, volatile, visceral	163

Table 16. Confusion matrix of Coimbra datasets of classification accuracy using rough set theory

	C1	C2	C3	C4	C5	Unknown
C1	**91**	27	8	9	35	0
C2	40	**85**	23	14	2	0
C3	8	43	**129**	21	14	0
C4	23	38	91	**92**	19	0
C5	20	8	26	20	**89**	0

4. Conclusions

The automatic emotion classification of music has gained increasing attention in the field of music information retrieval. In this paper, a RS-based approach for the classification of music emotion has been proposed, which can be characterized as a categorical approach to emotion recognition.

Different from the conventional machine learning algorithm, the RS-based approach represents the classification rules with "IF-THEN" rules. In addition, the RS-based approach makes it easy to visualize which attributes are important, after removing dispensable ones in the rule generation process. Furthermore, a set of logical decision rules has been obtained, with their strength, for classification of a certain class.

In the paper, four different groups of audio features have been implemented: dynamics, rhythm, spectral, and harmony. From each of the four frame-based features, four statistical parameters for condition attributes have been extracted, of mean, variance, skewness and kurtosis. In addition, in the paper, the inclusion of covariance components of pairwise features has been proposed within the same group of features.

Our experiment has shown that only a limited number of attributes from several audio features turned out to be indispensable, which are the attack slope, attack time, spectral flux, irregularity, spectral roughness, some of the MFCCs, key clarity, HCDF, and so on.

Another experiment has shown that a group of harmony features could give the highest classification accuracy among the four groups of features. In addition the experiment has shown that additional covariance components can increase the performance of classification accuracy. When this approach

includes all features and statistical parameters with covariance components, the proposed approach has provided comparatively better accuracy than others on a pair of datasets.

Acknowledgement

This work was partially supported from National Research Foundation of Korea (NRF-2015R1D1 A1A01058062).

References

[1] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448-457, 2008.

[2] D. Cabrera, "PsySound: a computer program for psychoacoustical analysis," in *Proceedings of the Australian Acoustical Society Conference*, Melbourne, 1999, pp. 47-54.

[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, 2002.

[4] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 5-18, 2006.

[5] T. Li and M. Ogihara, "Content-based music similarity search and emotion detection," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, Toulouse, France, 2004, pp. 705-708.

[6] A. Sen and M. Srivastava, *Regression Analysis: Theory, Methods, and Applications*. New York: Springer, 1990.

[7] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199-222, 2004.

[8] Y. H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762-774, 2011.

[9] D. P. Solomatine and D. L. Shrestha, "AdaBoost.RT: a boosting algorithm for regression problems," in *Proceedings of IEEE International Joint Conference on Neural Networks*, Budapest, Hungary, 2004, pp. 1163-1168.

[10] O. Lartillot and P. Toivianen, "MIR in MATLAB (II): a toolbox for musical feature extraction from audio," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR2007)*, Vienna, Austria, 2007, pp. 127-130.

[11] E. Pampalk, "A MATLAB toolbox to compute music similarity from audio," in *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR2004)*, Barcelona, Spain, 2004, pp. 1-4.

[12] F. Pachet and P. Roy, "Improving multilabel analysis of music titles: a large-scale validation of the correction approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 335-343, 2009.

[13] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, "Multi-modal music emotion recognition: a new dataset, methodology and comparative analysis," in *Proceedings of 10th International Symposium on Computer Music Multidisciplinary Research (CMMR'2013)*, Marseille, France, 2013, pp. 1-13.

[14] B. K. Baniya, D. Ghimire, and J. Lee, "Evaluation of different audio features for musical genre classification" in *Proceedings of IEEE Workshop on Signal Processing Systems (SiPS)*, Taipei, Taiwan, 2013, pp. 260-265.

[15] B. K. Baniya, J. Lee, and Z. N. Li, "Audio feature reduction and analysis for automatic music genre classification," in *Proceedings of 2014 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, San Diego, CA, 2004, pp. 457-462.

[16] B. K. Baniya, D. Ghimire, and J. Lee, "Automatic music genre classification using timbral texture and rhythmic content features," in *Proceedings of 17th International Conference on Advanced Communication Technology (ICACT)*, PyeongChang, Korea, 2015, pp. 434-443.

[17] B. K. Baniya and C. S. Hong, "Music mood classification using reduced audio features," in *Proceedings of Korea Computer Congress (KCC)*, Jeju, Korea, 2015, pp. 915-917.

[18] ROSE2 (Rough Sets Data Explorer) [Online]. Available: http://idss.cs.put.poznan.pl/site/rose.html.

[19] Soundtracks datasets for music and emotion [Online]. Available: www.jyu.fi/music/coe/materials/emotion/soundtracks.

[20] MOODetector [Online]. Available: http://mir.dei.uc.pt/downloads.html.

[21] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and simplicity as criteria in feature selection: application to mood classification in music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1802-1812, 2011.

[22] J. A. Russell, A. Weiss, and G. A. Mendelsohn, "Affect grid: a single-item scale of pleasure and arousal," *Journal of Personality and Social Psychology*, vol. 57, no. 3, pp. 493-502, 1989.

[23] H. S. Nguyen, "Discretization of real value attributes: a Boolean reasoning approach," Ph.D. dissertation, Department of Mathematics, Warsaw University, Warsaw, Poland, 1997.

[24] A. Papoulis and S. Unnikrishna Pillai, *Probability, Random Variables, and Stochastic Processes*. Boston, MA: McGraw-Hill, 2002.

[25] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*. Dordrecht: Kluwer Academic Publisher, 1991.

[26] S. Dalai, B. Chatterjee, D. Dey, S. Chakravorti, and K. Bhattacharya, "Rough-set-based feature selection and classification for power quality sensing device employing correlation techniques," *IEEE Sensors Journal*, vol. 13, no. 2, pp. 563-573, 2013.

[27] Z. Pawlak, "Rough set theory and its applications," *Journal of Telecommunications and Information Technology*, vol. 9, no. 3, pp. 7-10, 2002.

Babu Kaji Baniya

He received the B.E. degree in Computer Engineering from Pokhara University, Nepal in 2005, M.E. degree in Electronic Engineering and Ph.D. in Computer Science and Engineering from Chonbuk National University, Korea in 2010 and 2015. Currently, he is pursuing his postdoc in department of Biomedical Engineering, University of South Dakota. His main research interest includes audio signal processing, music information retrieval, pattern recognition, phenotype classification, source separation, etc.

Joonwhoan Lee

He received his B.S. degree in Electronic Engineering from the Hanyang University, Korea in 1980. He received his M.S. degree in Electrical and Electronics Engineering from KAIST, Korea in 1982 and the Ph.D. degree in Electrical and Computer Engineering from University of Missouri, USA, in 1990. He is currently a Professor in Department of Computer Engineering, Chonbuk National University, Korea. His research interests include image processing, computer vision, emotion engineering, etc.