Multimedia Systems (2017) 23:251–264
DOI 10.1007/s00530-015-0489-y

REGULAR PAPER

Feature selection and feature learning in arousal dimension of music emotion by using shrinkage methods

Jiang Long Zhang[1] · Xiang Lin Huang[1] · Li Fang Yang[1] · Ye Xu[1] · Shu Tao Sun[1]

Received: 1 January 2015 / Accepted: 3 October 2015 / Published online: 20 October 2015

Abstract Music emotion recognition is an important topic in music information retrieval area. A lot of acoustic features are used to train a music classification or regression emotion model. However, these existing features may not be efficient for classification or regression task. Furthermore, most works do not explain why these features do work for classification. In our work, eight features are extracted to represent the arousal dimension of music emotion, and various commonly used statistical learning methods such as Logistic Regression, and tree-based methods are applied to interpret important features. Then the shrinkage methods are applied to feature selection and classification in music emotion recognition for the first time. Our tests show that the proposed approaches are efficient for feature selection just as entropy-based filter methods, and better than wrapper methods. The shrinkage methods can produce more continuous and low variance model than wrapper methods. Then, we discover that the most useful features are low specific loudness sensation coefficients (low-SONE), root mean square and loudness-flux. Moreover, the shrinkage methods apply in logistic regression perform better for classification than most of other methods. We get an average accuracy rate of 83.8 %.

Keywords Features selection · Features learning · Music arousal dimension classification · Statistical learning · Shrinkage method

1 Introduction

Due to the explosion of vast and easily accessible digital music libraries, automatic recognition of emotion in musical audio has gained increasing attention in recently years. As a result, it is necessary to find an effective way to retrieve or classify them. But music, as a complex acoustic and temporal structure, is rich in content and expression, which can be highly subjective and hard to quantify [1]. There have much progress in machine learning method for estimating human emotional response to music [2], while little improvement has been made in terms of compact or interpret features. In general, most of the classification or regression methods use as many acoustic feature domains (e.g., loudness, timbre, rhythm) as possible, and perform dimensionality reduction techniques such as principal component analysis (PCA) or Karhunen–Loeve (K–L) transform, in order to remove the correlation among features. However, these methods may produce an increase in prediction accuracy; the model interpretability is decreased. Hence, most of the previous works do not explain the relationship between emotional associations and acoustic features clearly.

In our work we focus on the feature selection and feature learning in arousal dimension, which is based on Thayer's

Communicated by B. Prabhakaran.

✉ Jiang Long Zhang
zhangjianglong135@126.com

Xiang Lin Huang
huangxl@cuc.edu.cn

Li Fang Yang
yanglifang@cuc.edu.cn

Ye Xu
Xuye1031@163.com

Shu Tao Sun
stsun@cuc.edu.cn

[1] School of Computer, Communication University of China, Beijing, China

two-dimensional (energy and stress) model of mood [3]. The energy dimension corresponds to the arousal, and the stress corresponds to valence. Arousal is a physiological and psychological state of being awake or reactive to stimuli (music signal in our experiment). Thayer [3] classifies emotions in terms of high arousal (arousing emotion) or low arousal (calming emotion). Arousing emotions contain: joy, happiness, anger, frustration, hate, excitement and so on. Calming emotions can be expressed by contentment, sadness, confusion, shame, guilt and satisfaction. Lu [4] extracts intensity features, which are composed of the spectrum sum of the signal and the spectrum distribution in each sub-band. Then, the intensity features are used to classify the low arousal (Contentment and Depression) and high arousal (Exuberance and Anxious/Frantic) music. In his experiment, K–L transform is applied to remove the correlation among these raw features. Hence, we attend to make an assumption that the arousal dimension only depends on the intensity features in our model just as what Lu has done. In this paper, we will extract several intensity features without PCA or K–L transform being applied to raw features. Then various ordinary statistical learning methods such as logistic regression and tree-based methods are applied to build the training model and learn the important features. What is more, the shrinkage methods [5] are applied to select the important feature, and classify the music signals. We also compare other feature selection methods to learning the relationship between the emotions and features. To our best knowledge, at the time of writing there are no other shrinkage methods work that have applied in music emotion classification.

The rest of this paper is organized as follows. A review of literature on feature selection and feature learning in music emotion recognition is given in Sect. 2. Next, we extract the intensity features to characterize the arousal dimension music clip in Sect. 3. Then feature selection algorithms are applied to select important features in Sect. 4. Meanwhile, we learn features in various statistical models, and select important features by using shrinkage methods in Sect. 5. Several classification algorithms are applied to classify the feature vector. The result and analysis are reported in Sect. 6. Finally, the conclusions and future work are given in Sect. 7.

2 Related work

2.1 Feature selection

The general approaches to implement emotion classification or regression of audio are using supervised machine to train statistical models based on the different level of music features. But, these existing features may not be efficient for classification. Moreover, irrelevant or redundant features may lead to inaccurate conclusion. Hence feature selection is an important topic in machine learning.

There are three main approaches to select important features: filter, wrapper and dimension reduction. Filter methods select features use a model-independent heuristic such as Regressional Relief-F (RReliefF) [6], information gain evaluator and Correlation-based Feature Selection (CFS) [7]. Wrapper methods invoke the target learning algorithm to evaluate a feature set, such as: Forward Selection (FS), Backward Elimination (BE) and Hybrid forward and backward stepwise Selection (HS). Dimension reduction methods involve projecting the p predictors into a M-dimensional subspace, where $M < p$.

Miyoshi [8] uses correlation coefficients between features and music mood scores to select important features. He believes that the features with high correlation against the mood scores are suitable for detecting the mood score. Meanwhile, Yang [9] believes that RReliefF takes feature interrelationship into account, it is better than other statistical measures such as correlation coefficient in information gaining. He adopts RReliefF for each arousal-valence data space, and ranks the features by importance from 114 dimensions of features. Then, the top-18 and top-15 features are selected as the best features dimensions. He eventually comes to conclusion that top features for arousal are related to spectral shape and pitch. The top features for valence are more closely related to rhythmic (beat and tempo) and pitch. Although he gets a slight improvement from using RReliefF, it may be subject to Subset Selection Bias (SSB) [10]. In Arefin Huq's [11] work, feature selection algorithms including filter methods (RReliefF and CFS) and wrapper methods (biased FS and biased BE) are applied to select a subset of the 160 features to improve regression performance. He discovers that above feature selection algorithms do not improve the performance in his system due to the phenomenon of SSB, which make train model tend to over-fitting. Our propose methods (shrinkage methods) provide a good solution of over-fitting. Saari [12] presents a framework for obtaining realistic performance estimate of wrapper selection by taking into account the simplicity and generalizability of the classification models. Only four features, which contain mode major and key clarity, combine with dynamical, rhythmical, and structural features are selected by using backward elimination algorithm. And, the best classification method is k-nearest neighbors. Ruxanda [13] labels each song into eight emotional categories according to Hevner's emotional terms, and performs six dimensionality reduction algorithms: maximum likelihood common factor analysis (FA), information gain (infoGain) evaluator algorithm, genetic-search method (GA), K–N–Match algorithm (KNM), PCA and pivot-based algorithm. Then, he gets the best result by

Table 1 Feature description

Index	Feature	Description
1	Loudness-Mean	Mean of total loudness
2	Loudness-Std	Standard deviation of loudness
3	Low-SONE	The ratio of first ten critical-band loudness to total loudness
4	RMS	Average energy of a signal
5	Low-energy	Low energy rate
6	Loudness-flux	Amount of loudness change
7	Loudness-centroid	Center of mass of the loudness
8	Loudness-flatness	Smooth or spiky of loudness

using infoGain evaluator algorithm with six dimensions of feature: perceptual tempo, rhythm motion, spectral-flux, roughness, articulation and pitch density. Schmidt [14] searches the most informative features for mood detection by investigating multiple sets of acoustic features, including psychoacoustic [mel-cepstrum (MFCC) and statistical frequency spectrum descriptors] and music-theoretic (estimated pitch chroma). The support machine regression with MFCCs feature set resulting in the highest performance for time varying musical emotion regression.

The 1-norm shrinkage (regularization/penalty) methods can also perform feature selection, as it shrinks the coefficient estimates towards zero. However, there are no relative works in music emotion recognition area. It is likely because 1-norm shrinkage method is younger than other feature selection methods. One of the most popular shrinkage methods is Least Absolute Shrinkage and Selection Operator (Lasso), which is first introduced by Tibshirani [15]. In signal processing area, it is also called basis pursuit (BP), which finds signal representations in over-complete dictionaries [16]. Hence it can also be used to learn the feature representation. The shrinkage can delete noise features and select important features on one hand; on the other hand, it has effect of controlling the variances of estimated coefficients, and hence prevents overfitting and possibly improves the fitted model's prediction accuracy.

2.2 Feature learning

Feature learning is receiving much attention when deep learning make great success in machine learning field. Deep learning algorithm tries to learn simple features in the lower layers and more complex ones in the higher layers. Smith's [17] work reveals that learning a sparse representation of auditory signals is similar in early audio processing in mammals. Sparse learning not only can be used to select important features, but also provide insight into direct relationship between emotion and acoustic content. Schmidt [18, 19] use deep belief networks (DBNs) [20] with three hidden layers to learn emotion-based acoustic representations directly from magnitude spectra. Their results reveal that the second layer of DBN performs better than other layers in terms of mean error.

3 Feature extraction

In this section, several intensity features are extracted. A total of eight dimension features will be analyzed in our study.

3.1 Pre-processing

Each music clip is first transformed into a uniform format: 22,050 Hz, 16 bits, mono channel, PCM signals and the volume is normalized to a standard value.

3.2 Intensity

Intensity is defined as sound power per unit area, and can be estimated using the amplitude of the music signals. We extract eight features (detail in Table 1): mean of total loudness (loudness-Mean), standard deviation of loudness (loudness-Std), low specific loudness sensation coefficients (low-SONE) rate, mean of root mean square (RMS), low energy rate (low-energy), the flux of loudness (loudness-flux), centroid of loudness (loudness-centroid) and flatness of loudness (loudness-flatness). The extraction of loudness feature and SONE are based on MA toolbox [21], which includes an outer ear model [22], the Bark critical-band [23] rate scale, and spectral masking. Therefore, it better reflects human loudness sensation. Then, MIR toolbox [24] is applied to extract low-energy feature.

In order to study the relationship between feature and arousal response, we apply Box and whisker plots to display variation in samples of a statistical population without making any assumptions. Figure 1 shows the distribution of loudness-Mean which is spilt by the binary arousal variable. It displays a very pronounced relationship between the predictor loudness-Mean and the response arousal. The high arousal has a larger median value of loudness-Mean

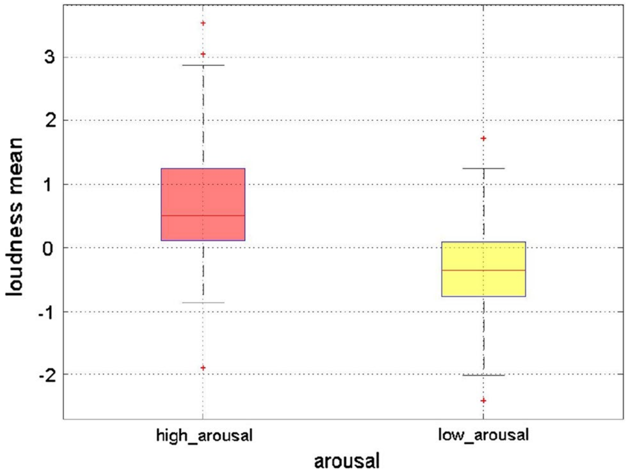

Fig. 1 Boxplot of loudness-mean as a function of arousal

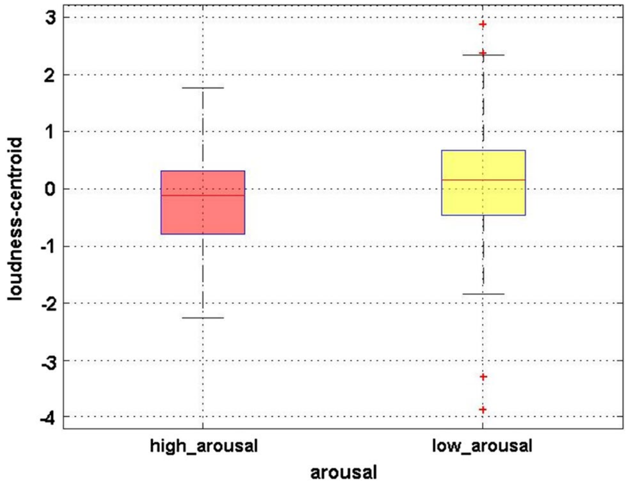

Fig. 2 Boxplot of loudness-centroid as a function of arousal

and inter-quartile-range (IQR) than the low arousal. While, Fig. 2 shows the left box overlaps most with the right box. Hence, the loudness-centroid may provide little information for arousal response.

3.3 Feature representation of a music clip

A total of eight dimension features are used to represent the music clip. Since, the characteristics and dynamics of these feature components are very different. Therefore, a normalization process is performed on each feature component to make their scale similar.

4 Feature selection algorithm

In this section, we introduce several feature selection algorithms including filter methods and wrapper methods in great detail.

4.1 Filter methods

We investigate several filter methods: Relief-F, CFS, Correlation filters (Pearson's correlation, Spearman's correlation) and Entropy-based filters (information gain, gain-ratio).

Relief-F ranks each feature based on how well it separates points of varying output values. The distance function can be Manhattan distance or Euclidean distance. Šikonja [6] finds that there is little significant difference in the estimations using these two metrics. But, Relief-F is sensitive to the number of nearest neighbors and sample size, which make hard to tune the parameters.

CFS attempts to discover sets of features that have low correlation with each other but high correlation with the output. It calculates a matrix of feature-class and feature–feature correlations from the training data and then searches the feature subset space using a best first search. Hall [7] compares CFS to Relief-F, and the result shows that CFS performs more feature selection than Relief-F does.

Correlation filters are very similar to CFS, but simpler. The Pearson's correlation coefficient and Spearman's correlation coefficient are used to measure the statistical dependence between predictor and response.

Entropy-based filters evaluate the worth of feature based on the information gain. In tree-based model, *Gini* index or cross-entropy is typically used to evaluate the quality of a particular split. We will learn the important feature by using entropy-based filters in next section.

4.2 Wrapper methods

In our work, three wrapper methods will be investigated: forward stepwise selection, backward stepwise selection and hybrid forward and backward stepwise selection.

4.2.1 Forward stepwise selection (FS)

Forward stepwise selection begins with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable (if any) that improves the model the most, and repeating this process until none improves the model.

4.2.2 Backward stepwise selection (BS)

Backward stepwise selection involves starting with all candidate variables, testing the deletion of each variable using a chosen model comparison criterion, deleting the variable (if any) that improves the model the most by being deleted, and repeating this process until no further improvement is possible.

4.2.3 Hybrid forward and backward stepwise selection (HS)

Hybrid versions of forward and backward stepwise selection are available, in which variables are added to the model sequentially, in analogy to forward selection. However, after adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.

5 Feature learning in various model

In this section, we learn the features in various statistical models, including ordinary model—logistic regression model, tree-based model and our proposed model—logistic regression, support vector machine and linear discriminant analysis by using shrinkage method.

In K-class classification problems, we are given a set of training data ($\mathbf{x_1}$, c_1) ($\mathbf{x_2}$, c_2), …($\mathbf{x_n}$, c_n), where the input $\mathbf{x}$ is a p-vector $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots x_{ip})^T$, the output c_i is qualitative and assumes values in a finite set $\{1, 2, \ldots, K\}$. We wish to find a classification rule from the training data, so that when given a new input $\mathbf{x}$, we can assign a class label k from $\{1, 2, \ldots, K\}$ to it.

5.1 Logistic regression (LR)

The logistic regression model arises from the desire to model the posterior probabilities of the K classes via linear functions in predictors $\mathbf{x}$, while at the same they time ensuring that sum to one and remain in [0, 1]. The model has the form

$$p(c_i|\mathbf{x}) = \frac{e^{f_i(\mathbf{x})}}{\sum_{k=1}^{K} e^{f_k(\mathbf{x})}}, \tag{1}$$

where

$$f_k(\mathbf{x}) = b_{k0} + \sum_{j=1}^{p} b_{kj}\mathbf{x_j}, \quad k = 1, 2, \ldots, K \tag{2}$$

LR models are usually fit by maximum likelihood, using the conditional likelihood of c given $\mathbf{x}$. The log-likelihood for N observations is

$$\ell(\beta) = \sum_{i=1}^{N} \log p(c_i|\mathbf{x_i}; \beta), \tag{3}$$

where β is the estimated parameter. It is convenient to code the two class c via a 0/1 response y_i. The log-likelihood can be written as

$$\ell(\beta) = \sum_{i=1}^{N} \left\{ y_i \beta^T x_i - \log(1 + e^{\beta^T x_i}) \right\}. \tag{4}$$

To maximize the log-likelihood, we set its derivatives to zero

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i \left(y_i - \frac{\exp(\beta^T x_i)}{1 + \exp(\beta^T x_i)} \right) = 0 \tag{5}$$

Equation (5) is a smooth convex optimization problem, and can be solved by a wide variety of methods, such as gradient descent, steepest descent and Newton method.

LR models can be used to understand the role of input variables in explaining the outcome. It uses z-statistic to perform the hypothesis tests on the coefficients. A large (absolute) value of the z-statistic indicates evidence against the null hypothesis H_0: $\beta = 0$, where β is the coefficient. So, the LR ranks the feature by the value of z-statistic: the larger value of the z-statistic, the more important of the feature is.

5.2 Decision tree

Decision tree builds classification or regression model in the form of a tree structure. The tree-based methods are simple and useful for interpretation, although they are not competitive with the best supervised learning approaches [5], such as support vector machine.

For a classification tree, we use *Gini* index to measure the total variance across the K classes. The *Gini* index is defined by

$$G = \sum_{k=1}^{K} \widehat{p}_{mk}(1 - \widehat{p}_{mk}), \tag{6}$$

where $\widehat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class.

In order to avoid the overfitting phenomenon, we take tenfold cross-validation to determine the optimal level of tree complexity. Then the cost complexity pruning is used to select a sequence of trees for consideration. Figure 3 shows the tree with 5 terminal nodes results in the lowest cross-validation error rate. In Fig. 4, we start with the loudness-flux node, and look for the binary distinction which gives us most information about the class. The top split assigns

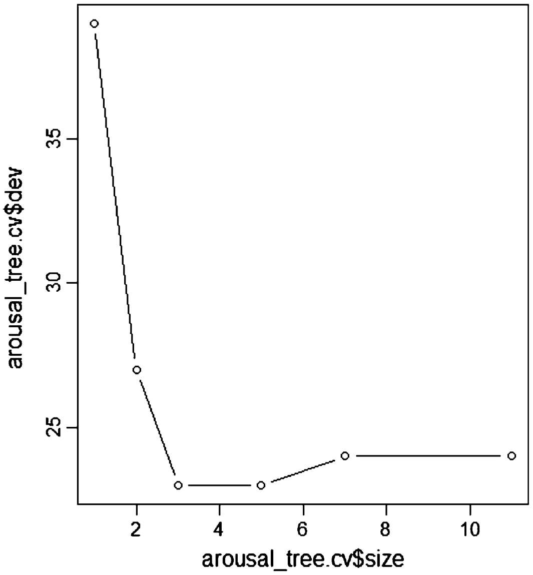

Fig. 3 Tenfold cross-validation error rate as a function of the value of the cost complexity parameter

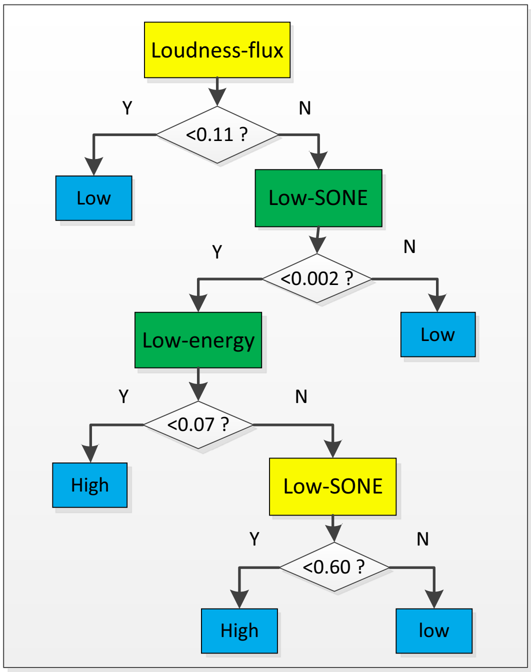

Fig. 4 The pruned tree corresponding to the minimal cross-validation error

observations having loudness-flux <0.11 to the left branch. It means that the low arousal music pieces have a lower value of loudness-flux than the high arousal pieces have. We then take low-SONE to new nodes, and repeat the process until some stopping criterion is met. In our decision tree we use three features: loudness-flux, low-SONE and low-energy.

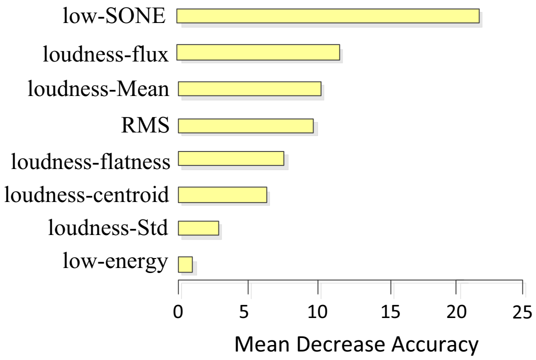

Fig. 5 Variable importance plot by using mean decrease in accuracy for arousal dimension

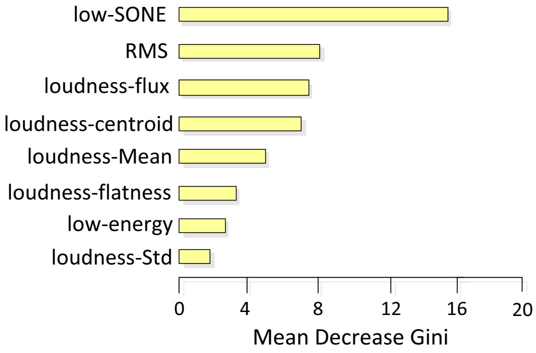

Fig. 6 Variable importance plot by using mean decrease in node impurity for arousal dimension

5.3 Bagging

When we split the training data into two parts at random and fit a decision tree to both halves, the result could be quite different. So, it may suffer from high variance. However, bagging is a general-purpose procedure for reducing the variance. The key to bagging is that trees are repeatedly fit to bootstrapped subsets of the observations. The bagging typically results in improved accuracy over prediction using a single tree, but, it can be difficult to interpret the resulting model. However, there are two measures to obtain an overall summary of the importance of each feature.

5.3.1 Variable importance measure using mean decrease in accuracy

The first measure is based upon the mean decrease of accuracy in predictions on the out of bag samples when a given variable is excluded from the model. In *b*th bootstrapped classification decision tree, it is measured by classification error rate:

$$E_b = 1 - \max_k(\widehat{p}_{mk}), \tag{7}$$

where $\widehat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class. Finally, we average all classification error rates on the out of bag samples.

Figure 5 indicates that among all of the trees considered in bagging, the level of low-SONE is by far the most important variable. What is more, loudness-flux, loudness-Mean and RMS are almost equally important.

5.3.2 *Variable importance measure using mean decrease in node impurity*

The second measure is the total decrease in node impurities from splitting on the variable. In the case of bagging classification tree, we add up the total amount that the *Gini* index is decreased by splits over a given predictor, and average all trees.

A graphical representation of the variable importance for arousal dimension is shown in Fig. 6—the mean decrease in *Gini* index for each feature, relative to the smallest. The features with large mean decrease in *Gini* index are low-SONE, RMS, loudness-flux and loudness-centroid.

5.4 Shrinkage methods

In order to give an interpretable model, we make hypothesis that the response is dependent on small number of variables (features). There are various forms of norm shrinkage methods: l_1-norm (lasso), l_2-norm (ridge), l_∞-norm, hybrid l_1, l_2-norm and so on. We only focus on lasso and ridge penalty in work. Lasso is a regression method that minimizes the sum of squared error loss subject to an l_1-norm constraint on the coefficients. Because of the nature of this constraint it tends to produce some coefficients that are exactly zero and hence selects important features and gives interpretable models. Moreover, the regularization method can prevent over-fitting, while the simple linear models are easily leading to. The Lasso estimate is defined by

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2, \qquad \text{subject to } \sum_{j=1}^{p}\beta_j \le t \tag{8}$$

where N is the sample size, p stands for the number of feature, and the parameters (β_0, β) indicate estimate coefficients.

The Lasso problem can also be written in the equivalent Lagrangian form

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j \right. \tag{9}$$

Here, $\lambda \ge 0$ is a complexity parameter that controls the amount of shrinkage: the larger value of λ, the greater amount of shrinkage.

5.4.1 *Regularized logistic regression*

Linear regression uses least squares (LS) approach to estimate the unknown linear regression coefficients, while to fit the logistic regression model, the maximum likelihood is preferred. So the optimization problem of L1-regularized logistic regression (L1-LR) can be given by:

$$\widehat{\theta} = \arg\min_{\beta}\ \ell(\beta) + \lambda\|\beta\|_1, \tag{10}$$

where $l(\beta)$ is defined in Eq. (4). The objective function in L1-LR is convex, but not differentiable, so the solution of the L1-LR must exist, but it need not to be unique. It can be solved by a generalized Lasso method [25], least angle regression paths (LARS) algorithm [26] or coordinate descent methods [27].

L2-regularized logistic regression (L2-LR) is similar to L1-LR, unless the penalized function. The optimization problem of L2-regularized logistic regression [28] can be given by:

$$\widehat{\beta} = \arg\min \ell(\beta) + \lambda\|\beta\|_2^2 \tag{11}$$

The coefficients are regularized in the same manner as in ridge regression. And the objective function in Eq. (11) is convex and smooth, so it can not produce sparse model. The solution can be solved by repeating the Newton–Raphson steps.

5.4.2 *Regularized support vector machine*

The SVM is a powerful classification tool and has a great success in many applications. In standard two-class classification problem, we encode response $y_i \in \{-1, 1\}$ rather than $\{0, -1\}$ in LR model. The standard 2-norm SVM (L2-SVM) is equivalent to

$$\widehat{\beta} = \arg\min_{\beta_0,\beta_j} \sum_{i=1}^{n}\left[1 - y_i\left(\beta_0 + \sum_{j=1}^{p}\beta_j h_j(x_i)\right)\right]_+ + \lambda\|\beta\|_2^2, \tag{12}$$

where $\{h_1(x), \ldots, h_p(x)\}$ is a dictionary of basic functions, and the loss $(1 - yf)_+$ is hinge loss. The 2-norm penalty shrinks the coefficients and helps to stabilize the solution.

However, just as ridge regression, it does not produce sparse coefficient.

SVM with 1-norm penalty (L1-SVM) can be used to select variable [29]. Zhu et al. [30] propose a solution path algorithm for the 1-norm SVM. L1-SVM is equivalent to fit a model that

$$\widehat{\beta} = \underset{\beta_0,\beta_j}{\arg\min} \sum_{i=1}^{n}\left[1 - y_i\left(\beta_0 + \sum_{j=1}^{p}\beta_j h_j(x_i)\right)\right]_+ + \lambda\|\beta\|_1 \quad (13)$$

In this paper, we assume the model is linearly separable. So, we will concentrate on basic representation rather than a kernel representation.

Note that 1-norm penalty is not differentiable at zero. This important singularity property ensures that the L1-SVM is able to delete many noise features by assigning their coefficients to zero [31], so it can select important features.

5.4.3 Shrunken centroids regularized discriminant analysis (SCRDA)

Linear Discriminant Analysis (LDA) assumes each classes have a multivariate normal distribution with common covariance matrix $\sum(p \times p)$ and different mean vectors $\mu_k(p \times 1)$. According the *Bayes' theorem*

$$p(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)} \quad (14)$$

we maximize the posterior probability that the observation belongs to a particular group. Therefore, in two class condition, the *discriminant function* can be given

$$\delta_k(x) = x^T \sum^{-1} \mu_k - \frac{1}{2}\mu_k^T \sum^{-1} \mu_k + \log \pi_k \quad (15)$$

In order to stabilize the sample covariance estimate and solve the singularity problem, some forms of regularization are imported on $\sum$

$$\widehat{\sum} = \alpha \sum + (1-\alpha) I_p \quad (16)$$

ωηερε $0 < \alpha < 1$. And Tibshirani proposes Nearest Shrunken Centroids (NSC) [32] as prototypes for each class to identify the subsets of the variables that best characterize each class. His basic idea is to shrink the class centroids $\overline{x}_{ik} = \sum_{j \in C_k} x_{ij}/n_k$ toward the overall centroid $\overline{x}_i = \sum_{j=1}^{n} x_{ij}/n$, after normalizing within-class standard deviation $s_i^2 = (1/(n-K)) \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \overline{x}_{ik})^2$ for each variable. We can write

$$\overline{x}_{ik} = \overline{x}_i + m_k s_i d_{ik} \quad (17)$$

where $m_k = \sqrt{1/n_k - 1/n}$. Then, the *soft thresholding* is applied for each d_{ik} to reduce by an amount Δ. This is expresses as

$$d_{ik}^{'} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+ \quad (18)$$

if Δ cause d_{ik} to shrink to zero for all classes k, then the variable i does not contribute to response. SCRDA is shrinks the centroids in (15) before calculating the discriminant score [33].

6 Experimental evaluation

In this section, the above approaches are evaluated with database. Then, we present the performance of mood detection in arousal dimension and analyze the result.

6.1 Data set

There are several datasets that are designed with the music emotion detection. Goto's RWC database [34] contains some original recordings for use in research. But the database has no arousal dimension classification. 1000 songs [35] is set up for music emotion recognition research, while it is suitable for temporal dynamic regression model. SOUNDTRACKS [36] is designed by Tuomas Eerola, which including three dimension (valence, energy, tension) quantitative value and five basic emotions (anger, fear, happy, sad, tender) quantitative value. These data sets are not suitable in our experiment.

The perception of emotion may vary not only between the genres of music but also across different language, cultural, so, in our research, we focus on detecting the Chinese popular music emotion and limiting cultural backgrounds of annotators to one cultural group. We collect 200 Chinese popular songs randomly from a popular Commercial website. The annotation can be grouped into *expert-based*

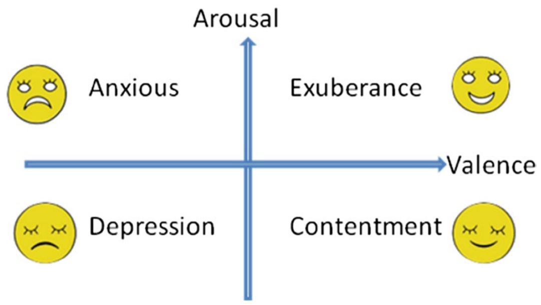

Fig. 7 Music emotion taxonomy in our subjective test

Table 2 Subjective test table

Index	Title	Emotion	Cue
1	…….	A Exuberance √ B Anxious C Depression D Contentment	A Timbre B Rhythm √ C Lyric √ D Background

[4, 34] and *subject-based* [9, 35, 37]. In our experiment, subject-based method is applied, because most of the people do not have the experience of professional music train. We design several rules to enhance the reliability of annotations.

- Reduce the emotion variation within the segment. 30-second segment starting from the 30th second of a song.
- Introduce basic knowledge of music emotion taxonomy in our annotation (in Fig. 7).
- Listen and annotate the music in a quiet and peaceful environment.
- Allow the user to skip songs when none of the candidate emotion is appropriate to describe the affective content of the song.
- Design a user-friendly annotation interface (in Fig. 7)
- Provide the music cue (in Table 2 last column) that may affect the expression of emotion.

Firstly, each song is annotated by ten subjects. Then, we remove the songs which have ambiguous emotion (the songs do not get the consensus among 50 % people). Lastly, 171 pieces of music with the length of 30s are selected as the data set.

6.2 Experimental result analysis

In our experiments, we randomly choose 100 pieces of music for ten-fold cross-validation, and the left 71 pieces of music for testing.

6.2.1 LR model versus L2-LR

Table 3 shows the coefficient estimates and relative information, which are estimated in LR model and L2-LR model. The first column stands for the features used in the model, the second column denotes the estimated coefficients, the third column stands for the estimated standard error, and the last column stands for the *p* value.

Hypothesis test is performed in LR model and L2-LR model. And a *p* value which is the estimated probability of rejecting the null hypothesis can helps us to determine the significance of results. A small *p* value indicates that there is an association between the predictor and the response. Typical *p* value cutoffs for rejecting the null hypothesis are 5 or 1 %. We can see that the *p* value of low-SONE and loudness-flux are less than 5 % in LR model. But we cannot infer the other features are less significant, because, some relationships exist among these features. Notice that the *p* values in L2-LR model associated with low-SONE, RMS and loudness-flux are 0.2, 0.8 and 2.3 %, respectively, which indicates that each of these features is associated with the probability of arousal response. Low-SONE and loudness-flux are significant features, which are estimated both by LR and L2-LR method, while, the importance of RMS can't be detected in LR model. This is because loudness-flux and loudness-Mean are correlated to the variable RMS, which can be seen in Table 4. The correlation between loudness-Mean and RMS is 0.87. What is more, we can see in the Table 3: the loudness-Mean and loudness-flux have a large negative coefficient, which can be canceled by a similarly positive coefficient on its correlated RMS. However, LR model can prevent this phenomenon by using shrinkage methods. Hence the importance of RMS can be detected in L2-LR model while cannot be in LR model.

The standard errors estimated in L2-LR model are smaller than in LR model—average 0.29 versus average 0.72. It means that the performance of LR can be improved

Table 3 Estimated coefficient of the LR model and L2-LR model feature

Feature	LR results			L2-LR results		
	Coefficient	Standard error	Pr(>|z|)	Coefficient	Standard error	Pr(>|z|)
Loudness-Mean	−2.25	1.31	*0.085*	−0.30	0.27	0.27
Loudness-Std	0.88	0.57	0.12	0.055	0.29	0.85
Low-SONE	1.48	0.43	*0.0005*	0.94	0.30	*0.002*
RMS	1.44	1.08	0.18	−0.83	0.31	*0.008*
Low-energy	0.28	0.41	0.50	0.14	0.28	0.61
Loudness-flux	−1.66	0.79	*0.046*	0.14	0.28	*0.023*
Loudness-centroid	0.29	0.39	0.46	−0.73	0.32	0.92
Loudness-flatness	−0.77	0.82	0.35	−0.03	0.30	0.20
Absolute average	1.13	0.72	0.22	0.39	0.29	0.36

Typical *p* value cutoffs for rejecting the null hypothesis are 5 or 1 % (in italics)

Table 4 The correlation matrix that contain all of the pair-wise correlations among the predictors in training data

Feature	Loudness-Mean	Loudness-Std	Low-SONE	RMS	Low-energy	Loudness-flux	Loudness-centroid	Loudness-flatness	*Y*
Loudness-Mean	1.0	0.72	−0.52	*0.87*	−0.44	0.69	−0.43	−0.92	−0.55
Loudness-Std		1.0	−0.39	0.63	−0.25	0.65	−0.46	−0.70	−0.40
Low-SONE			1.0	−0.54	0.05	−0.58	0.11	0.41	0.60
RMS				1.0	−0.40	*0.75*	−0.40	−0.84	−0.63
Low-energy					1.0	−0.20	0.36	0.44	0.18
Loudness-flux						1.0	−0.36	−0.63	−0.60
Loudness-centroid							1.0	0.43	0.22
Loudness-flatness								1.0	0.45
Y									1.0

Typical *p* value cutoffs for rejecting the null hypothesis are 5 or 1 % (in italics)

by using shrinkage method. Moreover, the coefficients estimated by L2-LR are shrunk from absolute average 1.13 to absolute average 0.39, which can reduce the variance of the predicted value.

6.2.2 Applying wrapper methods in LR

Then, the wrapper methods are applied in LR model. We use the Akaike information criterion (AIC) [38] as a measure of relative quality of a statistical model. The results are quite similar by using FS, BS and HS methods. FS method selects low-SONE as the most important feature, following by loudness-flux. While, BS and HS method drop out the loudness-centroid features firstly, following by RMS, low-energy, loudness-Mean, loudness-flatness, and loudness-Std, lastly, leaving low-SONE and loudness-flux. The wrapper methods give a more easily interpreted model than LR without using wrapper methods, and show that only two features are enough for LR model classification.

The wrapper methods and shrinkage method with 1-norm produce a model that is interpretable. But the wrapper methods have possibly lower prediction error than the full model. Moreover, because it is a discrete process—variables are either retrained or discarded—it often exhibits high variance. Hence, the shrinkage methods are more continuous than wrapper methods, and do not suffer high variance.

6.2.3 Coefficient path estimated by L1-LR

Figure 8 shows the Lasso estimates coefficients as a function of L1 norm of the coefficients in LR model. The L1-norm of the coefficients forms the *x*-axis, and the vertical breaks indicate where the coefficients are shrunk to zero. When $\lambda = 0$, the L1-LR simply gives LR fit, when λ becomes sufficiently large, the L1-LR gives the null model in which all coefficient estimates equal to zero. Moving from left to right in Fig. 8, each curve represents a coefficient (labeled on the right) as a function of L1 norm of the coefficients. We observe that at first the L1-LR results in a model that contains only the RMS predictor. And the low-SONE enters the model simultaneously, shortly followed by loudness-flux. Eventually, the remaining variables enter the model. The vertical breaks indicate where the active set (labeled on the top) is modified. Hence, depending on the value of λ, the L1-LR can produce a model involving any number of variables. If $\beta \leq 1$, only three features (RMS, low-SONE and loudness-flux) will be selected, the other coefficients are assigned to zero. So, it can produce a sparse model, while L2-LR can not. Moreover, the L1-LR provides more detail than LR and L2-LR model. The RMS and low-SONE are almost equally important, so as low-energy and loudness-flatness.

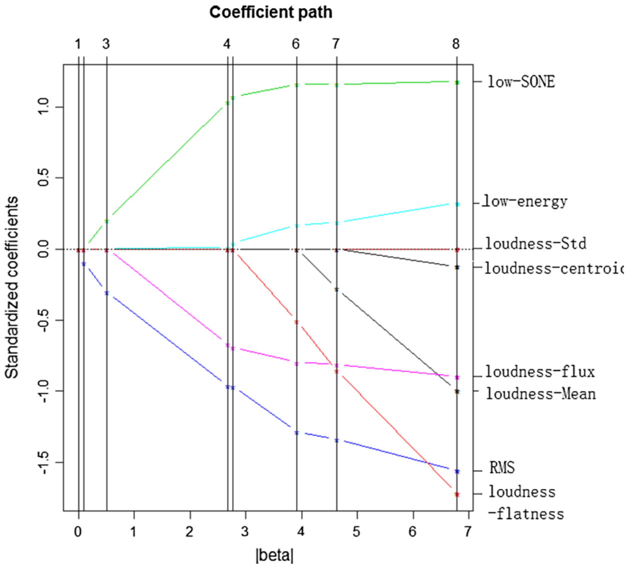

Fig. 8 Profiles of L1-norm coefficients

We can get relationship between features and response from the sign of feature coefficient. The sign of low-SONE coefficient is positive which indicates that for fixed value of other features, the higher value of low-SONE, the more likely the music signal is low arousal. Maybe, less brightness music produce low arousal emotion. In the same way, the high arousal music signal has high value of RMS, loudness-flux and so on. This is because a high volume or fast and large change in the loudness music could lead one to increase heart rate and blood pressure, which make people produce high arousal emotion.

The value of λ is very important for L1-LR model, and we need a method to determine which models is the best. One general approach is to use *K*-fold cross-validation, where the training data is used for both training and testing in an unbiased way. Figure 9 illustrates tenfold cross-validation on training data set. For classification model, the binomial deviance is selected as the measure of risk. The left vertical line corresponds to the minimum error, while the right vertical line the largest value of lambda such that the error is within one standard error of the minimum. Then, we select the tuning parameter value $\lambda = 0.225$ for which the cross-validation error is the smallest.

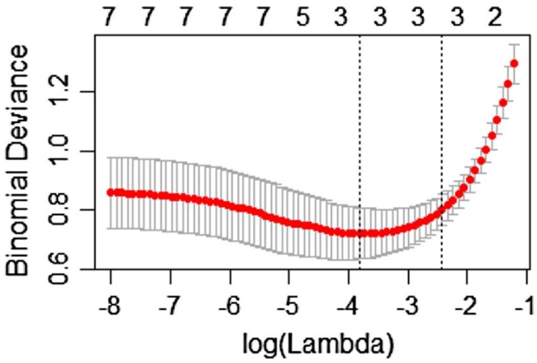

Fig. 9 Binomial deviance as a function of log (Lambda)

Table 5 The results of feature learning in various models, which only show first three features

Method	1st	2nd	3rd
LR	Low-SONE	Loudness-flux	Loudness-Mean
LR-FS	Low-SONE	Loudness-flux	
Bagging-A	Low-SONE	Loudness-flux	Loudness-Mean
Bagging- G	Low-SONE	RMS	Loudness-flux
L1-LR	RMS	Low-SONE	Loudness-flux
L2-LR	Low-SONE	RMS	Loudness-flux

The blank entries correspond to null feature

Table 6 The results of feature ranking in various methods, which only show first three features

Method	1st	2nd	3rd
Relief-F-2	Low-energy	Low-SONE	RMS
Relief-F-3	Low-energy	Loudness-centroid	RMS
CFS	Low-SONE		
Pearson	Low-SONE	Loudness-flux	RMS
Spearman	Low-SONE	Loudness-flux	RMS
Information gain	Low-SONE	RMS	Loudness-flux
Gain-ratio	RMS	low-SONE	Loudness-flux

The blank entries correspond to null feature

6.2.4 *Model-based feature rank*

Various training models are used to learn the relationship between features and response, the results of which are shown in Table 5. The LR and L2-LR model rank the feature by *z*-statistic, the descriptions of which are shown in Sect. 5.1. And L1-LR method ranks the features by the time when features enter the model, where λ varies from sufficiently large to zero: the earlier the sequence feature enters the model, the more important the feature will be. The bagging methods use two metric—classification accuracy (Bagging-A) and *Gini* index (Bagging-G). Most methods exclude L1-LR method evaluate that the low-SONE is the most important feature. And in Sect. 6.2.3, we show that low-SONE and RMS are almost equally important. Only L1-LR, L2-LR and Bagging-G methods estimate RMS as another important feature, while the remaining three methods can not. The reason for LR and LR by using forward selection algorithm (LR-FS) is stated in Sect. 6.2.1. And Bagging-A is measured by misclassification rate, which is less sensitive to changes in the node probabilities.

6.2.5 *Filter-based feature rank*

Then, we apply the filter methods which are described in Sect. 4.1 to measure the importance of features and rank the features. The ranking results are shown in Table 6. The number in “Relief-F-2” means the nearest neighbor number is two. The Relief-F algorithms are sensitive to the number of nearest neighbors and sample size. We choose different nearest neighbors but same sample size, and the results are different. CFS only chose one feature—low-SONE, which does not have enough information to build a good model. And the two correlation methods get the same results—loudness-flux is more important than RMS. It is interesting to see that information gain filter gets the same result as Bagging-G method and L2-LR method. Because, in tree-based model, *Gini* index can be replaced by cross-entropy, which bases on information gain. Meanwhile, gain-ratio filter gets the same result as L1-LR method. And

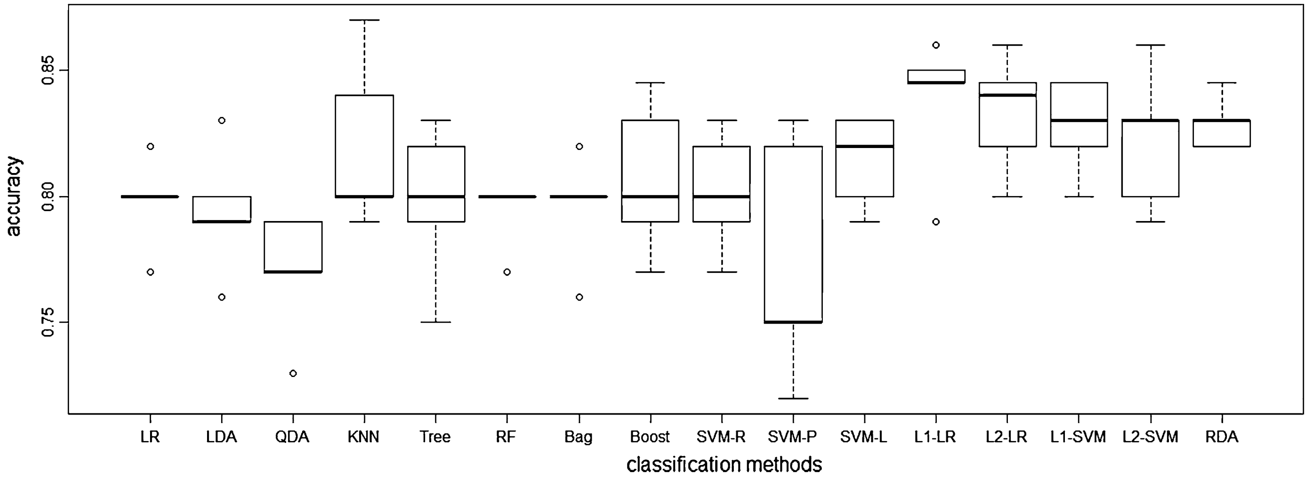

Fig. 10 Performance data for various classification methods

most of filter methods can detect the important of RMS. Furthermore, the features selected by entropy-based filter methods are similar to shrinkage method.

6.2.6 Classification

We imply various classification methods including the shrinkage methods (which present above) in arousal dimension to test the performance of models. The classification methods include: LR, LDA, Quadratic Discriminant Analysis (QDA), K-Nearest Neighbor (KNN), Tree, RandomForest (RF), Bagging (Bag), Boosting (Boost), Support Vector Machine with Radial kernel (SVM-R), Support Vector Machine with Polynomial kernel (SVM-P), Support Vector Machine with Linear kernel (SVM-L), L1-LR, L2-LR, L1-SVM, L2-SVM and Shrunken Centroids Regularized Discriminant Analysis (SCRDA). The parameters in various classification methods are determined by cross-validation. For example, we chose K value as 6 in KNN model. The accuracy rates are shown in Fig. 10, where X-axis denotes the classification methods, and Y-axis denotes the accuracy rate. In each box, the central mark is the median value, and the edges of each box are the 25th and 75th percentiles.

We see that the L1-LR model gets the highest median value, average accuracy rate of 84 %, and its box range is rather narrow. It means that the L1-LR model is quite stable. We find that the results with shrinkage methods are more stable than without having them. Moreover, the L2-LR, L1-SVM and RDA methods also get promising average result of 83 %. While in LR model, the average accuracy is only 80 %. The advantage of shrinkage methods are rooted in the bias-variance trade off. By controlling the tuning parameter, shrinkage can perform well by trading off a small increase in bias for a large decrease in variance.

The best accuracy rate of other no penalized methods is KNN's—82 % on average, and 80 % as median. The tree-based methods including pruning tree, Random-Forest, Bagging and Boosting have nearly the same median accuracy rate, and the Boosting performs a little better than the other tree-based methods. Then, we see that linear kernel may be more suitable than polynomial kernel in our model. SVM-P gets an average accuracy rate of 77.4 %, while SVM-L gets 81.4 %. Moreover, SVM with penalty performs better than SVM without penalty. We find that the performance of LDA is better than QDA, but a litter poorer than LR. And the regularized LDA achieves 5 % higher accuracy than LDA. Because by using regularization not only stabilizes the variance but also reduces the bias of discriminant function. As a result, the prediction accuracy is improved. And linear model may be more suitable than quadratic model in music arousal response.

7 Conclusion and future work directions

In this paper, we focus on the feature selection and learning in arousal dimension by using various methods. We study the features in Logistic Regression model firstly. Then the most commonly used methods—wrapper (forward selection, backward selection, hybrid selection), are applied to select the important features. In order to produce a more interpretable model, the tree-based methods are used to build the model and give the important features. Lastly, the shrinkage methods are applied in LR, SVM and LDA model.

Experimental results show that shrinkage methods are powerful methods for both learning informative feature as well as making classification. We get a more interpretable and continuous model than other wrapper methods, which only three features—low-SONE, RMS and loudness-flux are selected in L1-LR and L2-LR models. What is more, LR using shrinkage methods makes an average promotion of 3–4 % than LR. Lastly, we find that the accuracy performance by shrinkage methods is better than most of the other no-penalty classification methods.

There is still much room for feature improvement of our work. In this paper, we only focus our work on arousal dimension, while the other dimension—valence, is also very important for music emotion recognition. Then we still do not understand why feature selection results by using shrinkage methods are quite similar to use entropy-based filter methods. Lastly, new classification method such as deep learning (deep belief networks) can also be a good direction in our feature work.

Acknowledgments This work is supported by Communication University of China Engineering Project 3132014XNG1429, 2012BAH17B02 And the National key Science & Technology Pillar Program of China under Grant No. 2012-BAH51F02.

References

1. Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G., Richardson, P., Scott, J., Speck, J.A., Turnbull, D.: Music emotion recognition: a state of the art review. ISMIR, Utrecht, Netherlands, pp. 255–266 (2010)
2. Yang, Y.-H., Chen, H.H.: Machine recognition of music emotion: a review. ACM Trans. Intell. Syst. Technol. (TIST) **3**(3), 1–30 (2012). doi:10.1145/2168752.2168754
3. Thayer, R.E.: The Biopsychology of Mood and Arousal. Oxford Univ. Press, New York (1989)
4. Lu, L., Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signals. IEEE Trans. Audio Speech Lang. Process. **14**(1), 5–18 (2006)
5. Hastie, T., Tibshirani, R., Friedman, J.: Elements of Statistical Learning. Springer, New York (2009)
6. Šikonja, M.R., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. **53**, 23–69 (2003)
7. Hall, M.: Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of the 17th International Conference on Machine Learning (ICML), Stanford, CA, USA, pp. 359–366 (2000)
8. Miyoshi, M., et al.: Feature selection method for music mood score detection. IEEE, Modeling, Simulation and Applied Optimization (ICMSAO), 2011 4th International Conference, pp. 1–6 (2011)
9. Yang, Y.H., Lin, Y.C., Su, Y.F., et al.: A regression approach to music emotion recognition. IEEE Trans. Audio Speech Lang. Process. **16**(2), 448–457 (2008)
10. Miller, A.: Subset Selection in Regression. CRC Press, Boca Raton, London (2002)
11. Huq, A., Bello, J.P., Rowe, R.: Automated music emotion recognition: a systematic evaluation. J. New Music Res. **39**(3), 227–244 (2010)
12. Saari, P., Eerola, T., Lartillot, O.: Generalizability and simplicity as criteria in feature selection: application to mood classification in music. IEEE Trans. Audio Speech and Lang. Process. **19**(6), 1802–1812 (2011)
13. Ruxanda, M.M., Chua, B.Y., Nanopoulos, A., Jensen, C.S.: Emotion-based music retrieval on a well-reduced audio feature space. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference, pp. 181–184 (2009)
14. Schmidt, E.M., Turnbull, D., Kim, Y.E.: Feature selection for content-based, time-varying musical emotion regression. International Conference on Multimedia Information Retrieval, ACM, pp. 267–274 (2010)
15. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. **58**(1), 267–288 (1996)
16. Chen, S., Donoho, D., Saunders, M.: Atomic decomposition by basis pursuit. SIAM Rev. **43**(1), 129–159 (2001)
17. Smith, E.C., Lewicki, M.S.: Efficient auditory coding. Nature **439**, 978–982 (2006)
18. Schmidt, E.M., Kim, Y.E.: Learning emotion-based acoustic features with deep belief networks. IEEE, Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop, pp. 65–68 (2011)
19. Schmidt, E.M., Scott, J., Kim, Y.E.: Feature learning in dynamic environments: modeling the acoustic structure of musical emotion. In: ISMIR, pp. 325–330 (2012)
20. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
21. Pampalk, E., Rauber, A., Merkl, D.: A MATLAB toolbox to compute music similarity from audio. In: Proceedings of the ISMIR International Conference on Music Information Retrieval (ISMIR) (2004)
22. Painter, T., Spanias, A.: A review of algorithms for perceptual coding of digital audio signals. In: Digital Signal Processing Proceedings (DSP), pp. 179–208. IEEE (1997)
23. Zwicker, E.: Subdivision of the audible frequency range into critical bands. J. Acoust. Soc. Am. **33**(2), 248 (1961)
24. Lartillot, O., Toiviainen, P.: MIR in Matlab (II): a toolbox for musical feature extraction from audio. International Society for Music Information Retrieval (ISMIR), pp. 127–130 (2007)
25. Roth, V.: The generalized LASSO. IEEE Trans. Neural Netw. **15**(1), 16–28 (2004). doi:10.1109/TNN.2003.809398
26. Efron, B., et al.: Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)
27. Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. Ann. Appl. Stat. **1**(2), 302–332 (2007)
28. Lee, A., Silvapulle, M.: Ridge estimation in logistic regression. Commun. Stat. Simul. Comput. **17**, 1231–1257 (1988)
29. Bradley, P.S., Mangasarian, O.L.: Feature selection via concave minimization and support vector machines. In: Machine Learning Proceedings of the Fifteenth International Conference (ICML'98). Morgan Kaufmann, San Francisco, CA, pp. 82–90 (1998)
30. Zhu, J., Rosset, S., Hastie, T., Tibshirani, R.: 1-norm support vector machines. Adv. Neural Inf. Process. Syst. **16**(1), 49–56 (2003)
31. Friedman, J., Hastie, T., Rosset, S., Tibshirani, R., Zhu, J.: Discussion of boosting papers. Ann. Stat. **32**, 102–107 (2004)
32. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Class prediction by nearest shrunken centroids with applications to dna microarrays. Stat. Sci. **18**(1), 104–117 (2003)
33. Guo, Y., Hastie, T., Tibshirani, R.: Regularized discriminant analysis and its application in microarrays. Biostatistics **1**(1), 1–18 (2005)
34. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: RWC music database: popular, classical and jazz music data-bases. In: Proceedings of the International Conference on Music Information Retrieval (ISMIR), pp. 287–288 (2002)

35. Soleymani, M., et al.: 1000 songs for emotional analysis of music. In: Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia, pp. 1–6 (2013)
36. Eerola, T., Lartillot, O., Toiviainen, P.: Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models. International Society for Music Information Retrieval (ISMIR), pp. 621–626 (2009)
37. Li, T., Ogihara, M.: Detecting emotion in music. In: Proceedings of the ISMIR International Conference on Music Information Retrieval (ISMIR), pp. 239–240 (2003)
38. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csáki, F. (eds.) 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2–8, 1971, Budapest, Akadémiai Kiadó, pp. 267–281 (1973)