Neurocomputing 148 (2015) 619–627

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Emotion-driven Chinese folk music-image retrieval based on DE-SVM

Baixi Xing, Kejun Zhang*, Shouqian Sun, Lekai Zhang, Zenggui Gao, Jiaxi Wang, Shi Chen

College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 24 January 2014
Received in revised form
1 August 2014
Accepted 10 August 2014
Communicated by R. Capobianco Guido
Available online 23 August 2014

Keywords:
Music emotion recognition
Image emotion recognition
Differential Evolutionary algorithm
Support vector machine
Back propagation
Cross-media information retrieval

ABSTRACT

In this study, we attempt to explore cross-media retrieval between music and image data based on the emotional correlation. Emotion feature analytic could be the bridge of cross-media retrieval, since emotion represents the user's perspective and effectively meets the user's retrieval need. Currently, there is little research about the emotion correlation of different multimedia data (e.g. image or music). We propose a promising model based on Differential Evolutionary-Support Vector Machine (DE-SVM) to build up the emotion-driven cross-media retrieval system between Chinese folk image and Chinese folk music. In this work, we first build up the Chinese Folk Music Library and Chinese Folk Image Library. Second, we compare Back Propagation(BP), Linear Regression(LR) and Differential Evolutionary-Support Vector Machine (DE-SVM), and find that DE-SVM has the best performance. Then we conduct DE-SVM to build the optimal model for music/image emotion recognition. Finally, an Emotion-driven Chinese Folk Music-Image Exploring System based on DE-SVM is developed and experiment results show our method is effective in terms of retrieval performance.

1. Introduction

Multimedia creation work is always inspired by complicated and rich affection. Emotion feature, being as the high-level feature of multimedia data, comes closer to human cognition [1,2], and it also has contributed to successful multimedia retrieval. Presently, searching by emotion type has been regarded as an effective way and more likely to meet user's need [3]. However, explicit emotion description in specific semantic annotation could not fully express the rich emotion content of the multimedia data (e.g. an image or a song). Considering image and music have the similar feature of emotion, and there is an intricate abstract affective connection between these two types of modality, we would like to discover the possibility of cross-media retrieval on the basis of emotion feature.

Multi-media modality, like music and image, evokes emotions and also expresses emotions [4,5]. The relation between multi-media data and emotion is difficult to analyze for many reasons, however it arouses interest of researchers in many domains and the topic of music/image emotion recognition has been widely studied by different algorithms [6–8].

There are two well-known and dominant models to describe emotions: the discrete model (the categorical emotion model) and the dimensional model (the valence-arousal model). The categorical model describes all emotions as a limited number of universal and innate basic emotions such as Hevner Emotion Ring model Vigorous, Dignified, Sad, Dreaming, Soothing, Graceful, Joyous, and Exciting [9,10]. The valence-arousal model considers all emotions related to the level of valence (a pleasure–displeasure continuum) and arousal (activation–deactivation) [11]. However, both models have limitation, the dimensional model has been criticized for that emotions described in the valence-activation space are lack of differentiation when they are close neighbors, and participants expression term of their responses with only two dimensions also cause limitation. On the other hand, the categorical model has criticism of being inadequate comparing to the richness of emotional content [12]. Both theoretical emotion models have been supported by researchers and applied in music emotion study [13,14]. There is also a study proposed by Eerola about the comparison of these two models [15].

There are many outstanding achievements on music emotion recognition algorithm research. Lin introduced multi-objective algorithm of support vector machine [16], then the AdaBoost method was introduced to improve its performance [17]. Yang has continued his work on music emotion recognition research and application, including research of the concept of emotion intensity for complex music emotion recognition, music emotion classification [18], and video highlights exploration with music emotion recognition [19]. Zhang presented a high effective algorithm called the revised gene expression programming (RGEP) to construct the model for music emotion recognition [20], which proved that evolutionary algorithm is applicable for music emotion research.

On the aspect of image emotion recognition algorithm research, Wang built an image emotion semantic extraction structure and

* Corresponding author.
E-mail address: channy@zju.edu.cn (K. Zhang).

http://dx.doi.org/10.1016/j.neucom.2014.08.007
0925-2312/

pointed out the correlation between image color, texture and shape [21]. Then Wang utilized image color and shape feature to make image emotion labeling work [22]. Jinsub improved the model accuracy of exploring image emotion by comparing Neural Network and Adaptive Fuzzy System [23]. Recent study presented a biologically inspired model of emotion which could generate human-like emotion based on visual stimuli [24]. As a result, although the error exists in the research experiment, the rich varieties of approaches and satisfactory experiment results could support the emotion-driven image retrieval.

In this paper, we pursue the insight of Chinese folk music and image by machine learning approach on the emotion aspect to explore cross-media retrieval possibility. According to the current research that music-image cross-media retrieval by semantics could achieve encouraging performance [3,25]. However, to our knowledge, there is little research about emotion driven cross-media retrieval in the literature. Since emotion feature is regarded as the higher level feature than semantics feature [1], it is a challenge to explore the retrieval performance by emotion feature. We hold the view that emotion within multimedia data is much more profound than words, so any retrieval method by specific text has its limitation, since some emotion feature is complicated and hard to utter. A single picture could convey complex ideas, as the saying "a picture is worth a thousand words". Cross-media retrieval could be a new way to express and convey the user retrieval need.

Therefore the aim of our survey is to explore emotion in music and image, and bridge music and image by their emotion feature. We propose the Chinese Folk Music-Image Exploring System that could search out images by inputting a music piece driven by emotion similarity. It devotes in providing an interesting and effective method to bridge music and image by the emotion meaning.

This paper is organized as follows: Section 2 introduces methods of BP, LR, SVM and DE applied in this study; Section 3 describes the experiments, including building Chinese folk music and image emotion database, experiment of emotion labeling for music and image database, and comparing the model efficiency of BP, LR and DE-SVM modeling for music and image emotion recognition; Section 4 briefly introduces the user interface of the Chinese Folk Music-Image Exploring System; Section 5 concludes with some directions for future research.

2. Methodology

2.1. Back propagation

Back propagation (BP) has the great strength in non-linear solutions to ill-defined problems, while the only layer nervous system could only solve linear questions. Back Propagation Neural Network includes input layer, hidden layer, and output layer. BP learning has two steps in processing, including forward propagation and backward propagation of error. The advantages are that this synergistically developed back-propagation architecture is the most popular, effective model for complex, multi-layered networks.

2.2. Linear regression

Linear regression answers questions about a scalar-dependent variable on one or more predictors, including prediction of future values of a response, discovering predictors importance, and estimating the impact of changing a predictor [26]. Linear regression could be easier to fit the models which depend linearly on their parameters, because comparing to the models of non-linear parameters. Linear regression is extensively used in practical applications.

2.3. Support vector machine

The support vector machine (SVM) is a supervised classification system that minimizes an upper bound on its expected error. It attempts to find the hyper plane separating two classes of data that will generalize best to future data. Such a hyper plane is the so-called maximum margin hyper plane, which maximizes the distance to the closest points from each class. More concretely, given data points $x_0, x_1, \ldots, x_n$ and class labels $y_0, y_1, \ldots, y_n$, $y_i \in -1, 1$, any hyper plane separating the two data classes has the form:

$$y_i(w^T x_i + b) > 0 \forall i \tag{1}$$

Let w_k be the set of all such hyper planes. The maximum margin hyper plane is defined by

$$W = \sum_{i=1}^{n} a_i y_i X_i \tag{2}$$

where the $a_0, a_1, \ldots, a_n$ maximize

$$L_D = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j X_i^T X_j \tag{3}$$

Subject to

$$\sum_{i=1}^{n} a_i y_i, a_i \geq 0 \forall i \tag{4}$$

For linearly separable data, only a subset of the axis will be non-zero. These points are called the support vectors and all classification performed by the SVM depends on only these points and no others. Thus, an identical SVM would result from a training set that omits all the remaining examples. The representation of the data in this feature space need not be explicitly calculated if there is an appropriate Mercer kernel operator for which

$$K(X_i, X_j) = \Phi(X_i) \cdot \Phi(X_j) \tag{5}$$

Data that is not linearly separable in the original space may become separable in this feature space.

$$k(x_i, x_j) = \exp(-\gamma \| x_i - x_j \| 2), \quad \text{for } \gamma > 0 \tag{6}$$

For the data sets which are non-separable, a slack variable g_i is introduced to relax the constraint as

$$y_i(w^T x_i + b) \geq 1 - g_i, \quad g_i \geq 0 \forall i \tag{7}$$

Since the g_i will cause error of classification, so an additional cost $(C/n)\sum_{i=1}^{n} g_i$ for errors is added. Usually an optimal model could be achieved by seeking for the best parameters of C and G.

2.4. Differential Evolutionary-Support Vector Machine algorithm

In this study, we utilize the advantages of Differential Evolutionary algorithm to explore the best parameters of C, G in SVM modeling [27], and propose a promising algorithm of Differential Evolutionary-Support Vector Machine (DE-SVM) in music/image emotion recognition research.

Differential evolution is designed to be a stochastic real parameter optimization algorithm, which has similar operating steps as other evolutionary algorithms, mutation, cross over and selection [28,29], includes following steps [27] (see Fig. 1):

Step 1: Mutation, the parameters is optimized in a given problem initialized in vector.

For target vector $x_{k,g}, k = 1, 2, 3 \ldots, n$, mutant vector is obtained according to

$$v_{k,g+1} = x_{a1,g} + F \cdot (x_{a2,g} - x_{a3,g}) \tag{8}$$

while $a_1, a_2, a_3 \in 1, 2, \ldots, n$ are random indexes, a_1, a_2, a_3 are chosen to be diverse from the running index k, so $n \geq 4$), in order to allow