A LDA-Based Approach to Lyric Emotion Regression

Deshun Yang, Xiaoou Chen, and Yongkai Zhao

Institute of Computer Science & Technology, Peking University,
100871, Beijing, China
{yangdeshun,chenxiaoou,zhaoyongkai}@icst.pku.edu.cn

Abstract. Lyrics can be used to predict the emotions of songs, and if combined with methods based on audio, better predictions can be achieved. In this paper, we present a new approach to lyric emotion regression. We first build a Latent Dirichlet Allocation (LDA) model from a large corpus of unlabeled lyrics. Based on the model, we can infer the latent topic probabilities of lyrics. Based on the latent topic probabilities of labeled lyrics, we devise a scheme for training and integrating emotion regression models, in which separate models are trained for latent topics and the outputs of those models are combined to get the final regression result. Experimental results show that this scheme can effectively improve the emotion regression accuracy.

Keywords: Latent Dirichlet Allocation, lyric emotion regression.

1 Introduction

The explosion of digital music calls for new ways of music retrieval. Organizing and searching music according to its emotional content is such a new way. In supporting emotion-based retrieval, automatic music emotion recognition plays an important role, in that it can act as a means of automatic music emotion annotation.

Due to the emotion representation adopted, music emotion recognition can take either the form of emotion classification where categorical emotion models are adopted, or the form of emotion regression where emotion is scaled and measured by a continuum of two or more real-valued dimensions.

Compared to other related tasks of music concept recognition, such as genre recognition, emotion recognition is still in its early stage, though it is attracting more and more attention from the research community. As an indicator, MIREX run Audio Music Mood Classification contest for the first time in 2007. Since then, the contest is run each year and the performance of the best algorithm improves year on year, but is still far behind that of the recognition algorithms for other music concepts.

Until recently, most of the research work on music emotion recognition is based on music audio analysis, where only audio data is used to predict music emotions. However, there appears a few research papers on music emotion recognition through lyrics. In fact, for a song, lyric is an essential component, and contains information useful for predicting the emotion of the song. In addition, studies show that lyrics are complementary to audio in predicting song emotions.

Y. Wang and T. Li (Eds.): Knowledge Engineering and Management, AISC 123, pp. 331–340.
springerlink.com

Lyric emotion recognition is mostly carried out as an supervised machine learning task, where training lyrics are labeled with emotions and represented by vectors of features and a learning model, e.g., SVM, is employed on the training examples. N-gram features, having been well explored in other natural language processing tasks, are naturally introduced to the task of lyric emotion recognition. In addition to n-gram features, semantic model based features have also been tried for lyric emotion recognition.

LDA (Latent Dirichlet Allocation) [1] was proposed by David M. Blei in 2003. It is a generative graphical model that can be used to model and discover the underlying topic structures of any kind of discrete data. LDA has exhibited superiority on latent topic modeling of text data in the research works of recent years. Ralf Krestel et al [2] introduced a method based on LDA for recommending tags of resources. István Bíró et al [3] applied an extension of LDA for web spam classification, in which topics are propagated along links in such a way that the linked document directly influences the words in the linking document.

In this paper, we try to exploit both LDA model based features and n-gram features to build a better-performing lyric emotion regression model. For this purpose, we propose an emotion regressor training and integration scheme. First of all, we build a LDA model from a large corpus of unlabelled lyrics. The model learns a range of readily meaningful topics relating to emotion. Then, based on the model, we infer the latent topic probabilities of labeled lyrics to be used as training examples. According to the topic probabilities of the lyrics, we distribute them among subsets of examples, with each subset dedicated to a unique topic. We then train a regressor for each topic on its subset of training lyrics which are represented by n-gram features. To compute the emotion of a given lyric, we first call the individual regressors with the lyric's n-gram features and then combine the multiple results into a final value according to the lyric's LDA topic probabilities.

The rest of the paper is organized as follows: Section 2 will review some related works. Section 3 gives the emotion model we adopt. Section 4 describes in detail our scheme. Section 5 explains the details of experiments we did, including datasets and features. Section 6 shows evaluation experiments for proposed method. Finally, Section 7 gives conclusions and future work directions for this research.

2 Related Works

In the beginning, most research works of this field focus on acoustic analysis, finding new features or new classification models to improve the accuracy. Liu et al [4] presented a mood recognition system with a fuzzy classifier. In this system, all the music file were translated into MIDI form. Some music-related features, like tempo, loudness, timbre etc were extracted, by which they build a fuzzy classifier to recognize the mood of music. Lie Lu et al [5] presented a framework for automatic mood detection from audio data. The framework has the advantage of emphasizing suitable features for different tasks. Three feature sets, including intensity, timbre, and rhythm are extracted for classification.

At the same time, there appears a group of researchers who begin to notice the importance of lyrics. Some researchers make an effort to construct effective emotion

lexicon and use it to compute the emotion of a piece of lyric [6,7]. Others try to use machine learning methods to solve this problem.

Hu et al [8] made a detailed comparative analysis between audio and lyrics. Among the 18 mood categories, lyric features outperformed audio spectral features in seven categories. Only in one category, audio outperformed all lyric features.

Xia et al [9] proposed a model called sentiment vector space model (s-VSM) to represent song lyric documents which uses only sentiment related words. Their experiments prove that the s-VSM model outperforms the VSM model in the lyric-based song sentiment classification task.

Y. Yang et al [10] exploited both audio and lyrics for song emotion classification. They used bi-gram bag-of-words and Probabilistic Latent Sentiment Analysis(PLSA) to extract lyric features. The results show that the inclusion of lyric features significantly enhances the classification accuracy of valence.

3 Emotion Model

Roughly speaking, emotion models can be classified into categorical models and dimensional models.

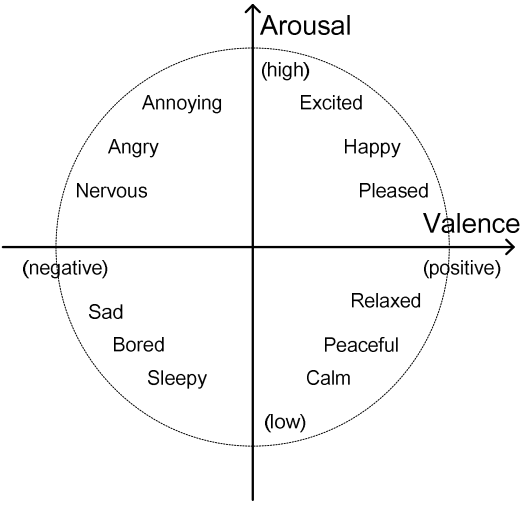

Fig. 1. Thayer's arousal-valence emotion plane

In a categorical model, a number of distinct emotion classes are identified, each of which is described by a group of related adjectives or nouns. Hevner's Adjective Circle [11] is a typical categorical model. Dimensional models reduce emotion to two or more dimensions of continuous values. Thayer's model[12] (see Fig. 1) is a dimensional model. It consists of two dimensions: Arousal and Valence. The Arousal dimension (the horizontal axis in Fig. 1) reflects the degree of stress and Valence (the vertical axis) the degree of energy.

Dimensional models are thought to be superior to categorical models in that they can represent the fine details of emotion and therefore don't have the inherent ambiguities between emotion categories in categorical models. When using a dimensional model, human annotators label a song in a continuous multi-dimensional emotion space freely without being confined to predefined emotion tags.

In the presented work, we adopt Thayer's two dimensional emotion model and we focus only on the problem of *Valence* value regression in this paper.

4 Regression Model Training and Integration Process

The process is shown in Fig. 2.

Input:
D_u: a set of unlabeled Chinese song lyrics;
D_l: a set of Chinese song lyrics which have been labeled with emotion values;
K: the number of latent topics in the LDA model built;
$\overline{P}$: topic probability threshold for example selection;
$D_l[k]$: subset of labeled lyrics selected for topic k;
$\widehat{R}$: a basic regression algorithm;
Build a LDA model M_{LDA} **with** K **topics from** D_u
Infer the topic probabilities of lyrics in D_l **based on** M_{LDA}
Extract uni-gram features of lyrics in D_l
Train regressors for topics, do for $k = 1,2, \dots, K$
1) For each lyric in D_l, if its probability to generate topic k is larger than $\overline{P}$, then put it into $D_l[k]$;
2) Call $\widehat{R}$ to train a regression model for topic k on $D_l[k]$;
Predict the emotion value of each lyric in D_l
1) Call $\widehat{R}$ with each topic's regression model and the lyric's uni-gram features to get an emotion value;
2) Combine the values got above into a final value for the lyric;

Fig. 2. Regression model training and integration process

The process works on two sets of examples; one is used for fitting the LDA model, and the other for training and testing the emotion regressors. A basic regression algorithm, such as SVM(Support Vector Machine) regression algorithm, is used in the process.

First of all, We fit a LDA model from a corpus of unlabeled lyrics and, with the model, infer the topic probability distributions of labeled lyrics. Instead of using the topic probabilities to represent lyrics in the training instances, we use them to base a framework for training and integrating emotion regression models. We assume that there exist more explicit relationships between the latent topics of the LDA model and the two dimensional emotion space of lyrics. And, lyrics which have higher probabilities to generate the same topic express emotions in similar ways, depending on the topic. So, for each latent topic, we choose those labeled lyrics whose probability to

generate the topic is greater than a threshold to compose a training set, and train a regression model on the set of selected training lyrics.

To predict the emotion value of a given lyric, all the regressors are called and provided with the lyric's uni-gram features. Then the outputs of the regressors are integrated to get the final value. We devised and experimented with two alternative mechanisms for combining the multiple values to get the final emotion value.

1) ***Mechanism A*** (MA)
   For a lyric, we choose the topic on which the lyric has the largest generation probability and call the topic's regressor with the lyric's uni-gram features to get the emotion value. This mechanism is based on the assumption that a lyric is only talk about one topic, on which the lyric has the largest generation probability. Of cause this assumption is sometimes not true, because there exit lyrics which generate multiple topics with almost the same high probabilities. So we propose the other mechanism.
2) ***Mechanism B*** (MB)
   For a lyric, we first choose those topics on which the lyric has a probability higher than a pre-set threshold. Then we call the regressors corresponding to the topics, and provide them the lyric's uni-gram features. To get the final emotion value, we adopt a weighted-voting method to integrate the multiple values returned by the regressors. For a lyric, the weight of a regressor is computed as in formula (1):

$$w_l[\hat{r}_k] = \frac{p(z_k|l)}{\sum_{j:p(z_j|l)\geq\bar{p}} p(z_j|l)} \ . \tag{1}$$

   where $w_l[\hat{r}_k]$ denotes the weight of regressor k for lyric l. Regressor k corresponds to topic k. $p(z_k|l)$ denotes the probability of lyric l generating topic k. This mechanism may select multiple topics that the lyric most probably talks, not only the single most probable topic. It doesn't have the weakness of the first one.

Our scheme for regression model integration differs from the Adaboost-style methods in that different lyrics have different topic probability distributions and therefore, different weights for the regressors. That is to say, the weights for the regressors vary with lyrics whose emotion value needs to be figured out.

It needs to be pointed out that the LDA model data consists of two parts: document-topic probability distribution and topic-word probability distribution. The document-topic distribution is usually taken as features of documents. In our work, we only use the document-topic distribution of the LDA model.

5 Experiments

5.1 Data Sets

We employed two sets of lyric examples. One is for building the LDA model and the other for training the emotion regression model. We downloaded about 35,000 song

lyrics from Internet, covering most of the contemporary Chinese popular songs, and use them to build the LDA model.

The training set we use to train the emotion regression models is the same as that used in [7]. We had downloaded 981 Chinese songs(including both the waveform data and lyric text data) from www.koook.com, which was then one of the most popular Chinese music websites on Internet. The songs were labeled manually with VA(Valence and Arousal) values. The values of VA emotion dimensions were confined in the range of [-4.0,+40]. Seven people took part in the annotation work, and everyone labeled all of the songs. From the original 981 songs, we selected only those songs which had been labeled with similar VA values by at least six people. At last, we got 500 songs labeled with VA values and they will be used to train the emotion regression models. More than 270 artists appear in our final collection and the genres of the songs in the collection include pop, rock & roll and rap.

For Chinese lyric texts, word segmentation need to be done before any other processing. We do word segmentation for the lyrics by a Chinese NLP tool.

5.2 Lyric Features

For training a regression model with a basic regression algorithm, such as SVM (Support Vector Machine) regression algorithm, we need to represent lyrics by feature vectors. In the field of NLP(Natural Language Processing), n-gram based bag-of-words (BOW) feature model is a commonly used model. In this paper, we focus on the issue of integration of regression models based on the LDA topic probabilities of lyrics, so we simply use uni-gram features to represent lyrics.

The original distinct uni-grams in the lyric corpus count up to tens of thousands. A feature space of this high dimensionality is not feasible for computational processing. To reduce feature dimensions, we divided the real value range of V dimension into four segments as shown in Table 1, and then apply the chi-square feature selection technique. At last we get a feature space of 2000 dimensions.

Table 1. Four segments of the value of V dimension

	$V < -2$	$-2 \leq V < 0$	$0 \leq V < 2$	$V \geq 2$
Class	1	2	3	4

Where n-grams are used as features, there are two commonly used methods of measuring the n-gram weight, tf*idf and Boolean value. Here, tf represents term frequency and idf represents inverse of document frequency. A Boolean value of 1 means that the corresponding n-gram appears in the document, and a value of 0 means that the n-gram doesn't appear. Previous experiments have shown that Boolean values are better than tf*idf, so we use Boolean values in our experiments.

5.3 LDA Model Parameter Estimation Method

It is intractable to compute the parameters of a LDA model exactly. Therefore, all the three parameter estimation methods currently used are approximate solutions. Variational Bayes was proposed by David M. Blei [1], which introduced variational

inference method into standard Expectation Maximization method. Expectation Propagation was proposed by T. Minka [13], which is another approximate solution, also based on variational inference. Gibbs Sampling was proposed by Tomas L. Griffiths [14] which is a approximate method theoretically different from the former two. It is based on the Markov chain Monte Carlo (MCMC) and is a special case of Metropolis-Hasting method. Thomas L. Griffiths and Mark Steyvers compared these three methods and made the conclusion that Gibbs sampling method gives the best result. So, in our experiment we use Gibbs Sampling to estimate the LDA model parameters.

6 Evaluation Results

In our experiments, we use Pearson Correlation Coefficients to measure the performance of the regression models.

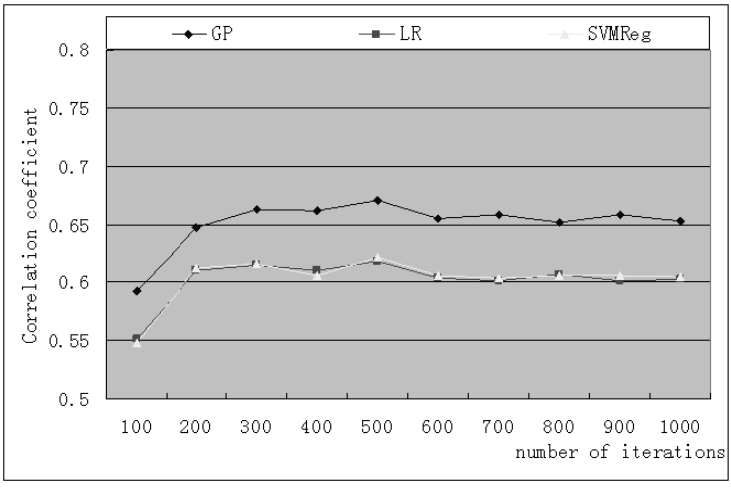

Fig. 3. Test result for different numbers of iterations

The number of iterations the LDA model fitting program runs and the number of latent topics the model contains may affect the performance of the model. We determine these two numbers through experiments. With a series of different values for the two numbers, We build LDA models and infer the topic probabilities of the lyrics based on each model respectively. Then, representing lyrics with their topic probabilities, and employing three basic regression algorithms, SVM (Support Vector Machine) regression, LR (Logistic Regression) and GP (Gaussian Regression), we train emotion regression models and test their performances. We then choose the number of topics and the number of iterations with which the best-performing LDA model has been built.

Fig. 3 shows the results of the performance test of the LDA models got with different numbers of iterations and a fixed number of topics, 15. From Fig. 3, we can see that when the number of iterations exceeds 300, the performance of the LDA model becomes stable. So, in the following experiments, we set the number of iteration to 500.

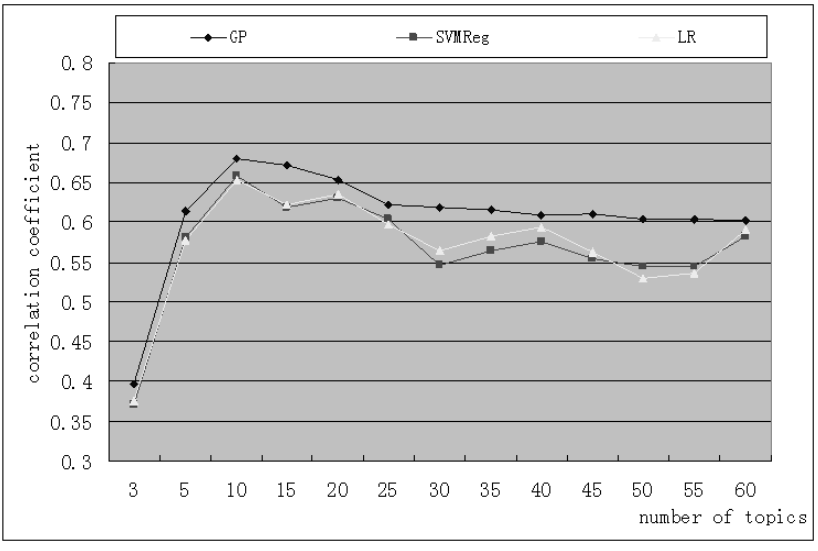

Fig. 4. Test result for different numbers of topics

Fig. 4 shows the results of the performance test of the LDA models got with different numbers of topics. We can see from Fig. 4 that, the optimal number of topics is around 10, much smaller than that for texts in many other domains. This shows that in lyrics, although the word vocabulary of lyrics commonly used is large, but these words can be clustered semantically into a few categories. In following experiments, we set the number of topics to 10.

In our regression model training and integration experiment, the uni-gram features we use to represent lyrics consist of 2000 dimensions. It is not practical for Gaussian Regression and Logistic Regression programs to work in a feature space with this high dimensionality. So, we only adopt SVM regression program in the experiment.

The value of parameter $\overline{P}$ in Fig. 2 affects the selection of training examples. So it has influences on the performance of regression models trained. We do experiments with the value set to 0.03, 0.05, 0.1, 0.15, 0.2, 0.25, and 0.3, respectively and find that at first, with the increase of the value, the performance gets better, which is in our expectation. But when the value becomes greater than 0.15, the performance starts to get worse. The reason for the worse regression models may be that the training set gets too small. In fact, if the value is too large, few lyrics will be chosen as training examples.

The last parameter to be set is $\bar{p}$ in formula (1). We set its value to 0.1, which is the average probability a lyric generates each topic (1/10).

To get a reliable evaluation result, we took 5-fold cross-validation on the set of labeled lyrics. The final result is shown in table 2. The baseline SVM regression model has been trained on the set of all labeled lyrics. It can be seen from the table that the two model-integration mechanisms both outperform the baseline SVM model. MA raises the correlation coefficient by 3.3% and MB increases it by 4.1%.

Table 2. Final results

	SVM Model (baseline)	Topic SVM Models-MA	Topic SVM Models-MB
CF	0.731	0.764	0.772

We can get the conclusion that, for a specific lyric to be recognized, if we select only those examples that are similar to the test example to train a regression model, the resulted model will be better than the model trained on all examples. In other words, an example which talks about topics totally different from that talked about by the test example is considered to be noise for training the regression model, and we should remove it from the training set.

But we can also see that the accuracy does not improve much. This may be caused by the small size of the training set. After the screening process, the set of selected training examples for each topic are all much smaller than the original training set. So, the regression models may be under-trained. If the size of the training set is too small, the hyper plane of the SVM model can not reflect the true information of example distribution.

Mechanism MB is better than mechanism MA. It proves our idea mentioned before. Some lyrics talk about mainly a single theme from the beginning to the end, expressing a happy feeling, for example. For these songs, MA mechanism can give a much better prediction. But there are a number of songs that express multiple themes at the same time.

Overall, the regression model training and integration scheme which uses LDA model based information is effective to improve the final regression accuracy.

7 Conclusion and Future Work

In this paper, we investigated a lyric emotion regression model training and integration scheme which is based on the LDA model based information about lyrics. We train a separate SVM regression model for each latent topic and integrate these regression models based on the latent topic probabilities of lyrics. The experimental results show that this method can effectively improve the regression accuracy.

In the future, we will try LDA models which include both bi-grams and uni-grams. By considering bi-gram patterns of words, we believe the resulted LDA models will be more semantically expressive. In addition, we will do experiments on a larger training set of about 2,000 Chinese songs with emotion annotations, to see if better-performing regression models can be obtained.

Acknowledgments. This work was supported by Beijing Natural Science Foundation(Multimodal Chinese song emotion recognition) and National Development and Reform Commission High-tech Program of China under Grant No. [2010]3044.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. In: NIPS 2002 & JMLR 2003
2. Krestel, R.: Latent Dirichlet allocation for tag recommendation. In: Proceedings of the Third ACM Conference on Recommender Systems, pp. 61–68 (2009)
3. Bíró, I., Siklósi, D., Szabó, J., Benczúr, A.A.: Linked latent Dirichlet allocation in web spam filtering. In: Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web, pp. 37–40 (2009)
4. Liu, D., Zhang, N.Y., Zhu, H.C.: Form and mood recognition of Johann Strauss's waltz centos. Chin. J. Electron. 12(4), 587–593 (2003)
5. Lu, L., Liu, D., Zhang, H.-J.: Automatic Mood Detection and Tracking of Music Audio Signals. In: IEEE Transactions on Audio, Speech and Language Processing (2006)
6. Meyers, O.C.: A mood-based music classification and exploration system. Master's thesis, Massachusetts Institute of Technology (2007)
7. Hu, Y., Chen, X., Yang, D.: Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In: Proc. of the International Society for Music Information Conference, Kobe, Japan (2009)
8. Hu, X., Downie, J.S.: When lyrics outperform audio for music mood classification: a feature analysis. In: Proceedings of 11th International Society for Music Information Retrieval Conference (2010)
9. Xia, X.Y., Wang, L., Wong, K., Xu, M.: Sentiment vector space model for lyric-based song sentiment classification. In: Proc. of the Association for Computational Linguistics, ACL 2008, pp. 133–136. Columbus, Ohio, U.S.A (2008)
10. Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., Chen, H.H.: Toward Multimodal Music Emotion Classification. In: Huang, Y.-M.R., Xu, C., Cheng, K.-S., Yang, J.-F.K., Swamy, M.N.S., Li, S., Ding, J.-W. (eds.) PCM 2008. LNCS, vol. 5353, pp. 70–79. Springer, Heidelberg (2008)
11. Hevner, K.: Experimental studies of the elements of expression in music. American Journal of Psychology 48(2), 246–268 (1936)
12. Thayer, R.E.: The Biopsychology of Mood and Arousal. Oxford Univ. Press, New York (1989)
13. Minka, T.: Expectation propagation for approximate Bayesian inference. In: Proc. 17th Conf. on Uncertainty in Artificial Intelligence (2001)
14. Griffiths, T.L., Steyvers, M.: Finding Scientific Topics. PNAS (101), 5228–5235 (2004)