Personalized Music Emotion Classification via Active Learning

Dan Su
Dept. of Electronic & Computer Engineering
Hong Kong University of Science & Technology
dsu@ust.hk

Pascale Fung
Dept. of Electronic & Computer Engineering
Hong Kong University of Science & Technology
pascale@ee.ust.hk

ABSTRACT

We propose using active learning in a personalized music emotion classification framework to solve subjectivity, one of the most challenging issues in music emotion recognition (MER). Personalization is the most direct method to tackle subjectivity in MER. However, almost all of the state-of-the-art personalized MER systems require a huge amount user participation, which is a non-neglegible problem in real systems. Active learning seeks to reduce human annotation efforts, by automatically selecting the most informative instances for human relabeling to train the classifier. Experimental results on a Chinese music dataset demonstrate that our method can effectively reduce as much as 80% of the requirement of human annotation without decreasing F-measure. Different query selection criteria of active learning were also investigated, and we found that informativeness criterion which selects the most uncertain instances performed best in general. We finally show the condition of successful active learning in personalized MER is that label consistency from the same user.

Categories and Subject Descriptors

H3.1 [**Content Analysis and Indexing**]: Retrieval models; H5.5 [**Sound and Music Computing**]: Modeling

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Personalization, music emotion classification, active learning

1. INTRODUCTION

One of the most challenging issues in MER is subjectivity. Any individual differences, such as gender, personality, and cultural background of the subject, may lead to different emotion recognition results for the same music piece [15]. This situation is more obvious when dealing with real world music data, when many emotions are not clearly or directly expressed. Common agreement on the perceived emotions is even harder to achieve[16, 14, 3].

MIRUM'12, November 2, 2012, Nara, Japan.

A lot of previous work has been conducted to take subjectivity into account. In the categorical-based music emotion classification area, [12] try to deal with subjectivity by formulating their task into multi-label classification problems , and [15] do so through building a fuzzy classifier, which assigns soft labels to each song with a soft value for each category indicating the strength of that emotion. In the dimensional-based music emotion regression community, E. Schmidt et al. in [7] introduce the idea of modeling the collected dynamic human response labels in the V-A plane as a parameterized stochastic distribution instead of a singular label or value.

However, one common limitation in all of the above mentioned works is that they discard the individuality of casual listeners, which is actually the fundamental cause of subjectivity in MER, as early as at the ground truth collecting stage. Whether the ground truth was collected from experts or non-experts [15, 16, 14], or from a collaborative annotation game based method [7], either expert opinion or a huge amount of user data is needed to obtain a reliable average. As [16] points out, a simple combination of the annotator's recognition results as the ground truth to build a generalized MER system for all users is unlikely to have satisfictory performances in practice, when the perceived emotion of a song differs a lot for different individuals.

Yang et al. in [16], for the first time demonstrate the effect of personalization, to confront subjectivity from its fundamental cause. In [14] they proposed two regression methods for building a personalized MER system; however, both their Bag-of-Users (BoU) models and residual modeling (RM) methods have huge requirements for user participation to train a sufficient personalized model, while heavy user participation is definitely an unnegligible factor in a real world system. On the other hand, their model is for dimensional music emotion regression only. In [3], the author further addresses the subjectivity issue by proposing a computational model that predicts the emotion distribution on the emotion plane instead of a single point; however, their training process still involves lots of human annotation.

Based on the above, we propose active learning, to alleviate user burden for personalized music emotion classification. Active learning aims to minimize user participation without decreasing the results, by selecting the most informative instances to the training data. It has shown its effectiveness in speech summarization [17], text categorization [11], as well as image retrieval [10] .

In summary, the main contributions of the paper are as follows.

- We propose active learning in personalized music emotion classification to solve subjectivity, one of the most challenging issues in MER. It will reduce, by as much as 80%, individual participation, which is one of the most important and unneglected factors that personalized music emotion classification systems should take into account.

- We prove that informativeness criterion of query instances selection in active learning, in general gives the best performance.
- We investigate the conditions to make active learning work effectively. Emotion labels should be annotated consistently in personalized MER system, since each non-expert individual will entirely decide the emotion label for each music.

The rest of the papers is organized as follows. First, we introduce the preliminary subjective experiments to verify the tenability of our starting point for building personalized music emotion classifiers using active learning. Then the proposed active learning algorithm will be explained in section 3. Experimental setups, results and further discussions are shown in section 4. Finally we give our conclusion and discuss future work.

2. PRELIMINARY SUBJECTIVE TEST

At the starting point, to measure the tenability of applying active learning for personalized music emotion classification, we conducted preliminary subjective tests to measure the inter-labeler agreement between participants on different emotion categories on a subset of our dataset. It is a Chinese dataset with randomly selected songs and no label information. We picked up 96 Chinese songs from the dataset. Participants in our subjectivity tests are all casual music listeners with different gender and cultural backgrounds. To avoid the effect of cultural differences on our measurement, we divided participants either into the Chinese or the Western group, according to their culture background. Inter-labeler agreement was measured within the group. They performed the preliminary music emotion evaluation tasks individually through an online website where they listened to 30 second music clips provided under different emotion categories. They then indicated whether their perceived emotion of that music clip belonged to that emotion category or not by clicking yes or no.

There is no expert-agreed label information provided for our Chinese music dataset. However, instead of asking participants for their emotion label after listening to each Chinese music clip, which is a multiple-choice question, emotion labels were assigned to each Chinese music clip beforehand. Participants' cognitive burdens were alleviated by the binary choice of yes or no. The assigned labels for the Chinese music dataset are obtained by binary classifiers trained from a Western music dataset, which has an overall cross-validation accuracy of 91%. Of course, the cross-language music emotion classifiers method for our Chinese music emotion labeling might not be accurate since no adaptation work was performed during the procedure. However, we want to measure the inter-labeler agreement of the labels between participants. Whether the labels are correct or not will not affect our measurement actually. On the other hand, this method should be more accurate than random assignment, since, intuitively, there are some commonalities in emotion expression and perception in music across languages via audio content.

Cohen's Kappa statistic was adopted to measure the inter-labeler agreement between each of two participants. It has been utilized a lot in speech emotion annotation areas [8]. Cohen's Kappa κ_c is calculated as

$$\kappa_c = \frac{p_A - p_R}{1 - p_R} \quad (1)$$

where p_A is the relative observed agreement between the two labelers and p_R is the expected agreement if both labelers would annotate on a random basis. According to [5], the interpretation of k can be interpreted as: $<$0: No agreement; 0.0-0.2: Slightly agreement; 0.21-0.40: Fair agreement; 0.41-0.60: Moderate agreement; 0.61-0.80: Substantial agreement; 0.81-1.00: Almost perfect agreement.

Table 1: Kappa values between three Chinese annotators on a subset of the Chinese music dataset

Labelers	1&2	1&3	2&3	**Avg**	Agreement
angry	0.307	0.536	0.462	0.435	moderate
calm	0.267	-0.067	-0.114	-0.029	no
happy	0.6	0.067	-0.069	0.199	slightly
sad	0.093	0.253	0.831	0.392	fair

Table 2: Kappa values between three Western annotators on a subset of the Chinese music dataset

Labelers	1&2	1&3	2&3	**Avg**	Agreement
angry	0.397	0.235	0.549	0.394	fair
calm	0.170	0.170	0.327	0.222	fair
happy	0.285	0.121	0.378	0.261	fair
sad	0.683	0.516	0.806	0.669	substantial

Table 3: Kappa values between three Western annotators on a subset of the Western music dataset

Labelers	1&2	1&3	2&3	**Avg**	Agreement
angry	0.533	0.583	0.861	0.659	substantial
calm	0.384	0.509	0.524	0.472	moderate
happy	0.552	0.528	0.609	0.563	moderate
sad	0.571	0.400	0.400	0.457	moderate

Table 4: Kappa values between three Chinese annotators on a subset of the Western music dataset

Labelers	1&2	1&3	2&3	**Avg**	Agreement
angry	0.340	0.661	0.463	0.484	moderate
calm	0.216	0.296	0.254	0.255	fair
happy	0.5	0.372	0.203	0.358	fair
sad	0.577	0.493	0.561	0.544	moderate

Table 5: Self-to-Self Kappa values of the same Chinese annotators on the same subset of Chinese music dataset

Labelers	1&1	2&2	3&3	**Avg**	Agreement
angry	0.769	0.693	0.769	0.75	substantial
calm	0.670	0.783	0.426	0.638	substantial
happy	0.634	0.605	0.672	0.637	substantial
sad	1	0.466	0.801	0.817	almost

Table 1 shows the kappa values of the three Chinese annotators on a subset of the Chinese music dataset on four basic emotion categories. Labelers 1, 2 and 3 are all native Chinese speakers. It can be seen that it is harder to reach an agreement between the non-experts for the real world music data. We also asked three non-Chinese speakers to do the same annotation tasks in order to filter out the effect of lyrics. The results are shown in Table 2. Again, it shows us that little common agreement has been reached.

The majority voting results on the Chinese music clips were calculated and are shown in Table 6. From Table 1 and Table 6, it can be concluded that even though common agreement of the perceived emotion for each music clip can be achieved through majority voting, the low kappa values on different emotion categories between

Table 6: Statistical results on agreements of the subset of the Chinese music dataset

	# of clips with 3 or 2 agreements	# of clips with no agreement	Total
angry	14(53.8%)	12(46.1%)	26
calm	20 (66.7%)	14 (46.7%)	30
happy	16 (53.5%)	14 (46.7%)	30
sad	9 (69.2%)	4 (30.1%)	13

different participants indicate that common agreement between different individuals is actually hard to reach. So we therefore doubt the performances of the system trained by using ground-truth data built from non-experts by simply averaging opionions. If agreement between the ground-truth builders can not even be achieved, how can it satisfy other users?

We also conducted human evaluation experiments on a Western dataset, with labels provided as the agreed label from four experts, and pre-selected by these experts to best represent each emotion. We picked up 120 Western songs with equally distribution on the four emotion categories. Table 3 and Table 4 show the kappa values on a subset of the Western music dataset by three Chinese and three Western participants. We can see that for the Western music dataset, which has been filtered by experts, the inter-label agreements on the four basic emotions between different participants are still not as high as we expected.

A further experiment was conducted by asking the three Chinese annotators to do the evaluation again, on another days on the same Chinese real world music dataset whose emotion labels did not reach common agreement between the three labelers. The kappa value results are shown in Table 5. The results in Table 5 and Table 1 indicate that, even though it is not easy to reach an agreement between different annotators on the perceived emotion of the real world music, the perceived emotion in music is relatively stable for each annotator him/herself, at least in a certain period.

Based on the above results and discussions, we may conclude that in order to build satisfactory music emotion classifiers for all users, personalized music emotion classifiers should be considered, since training sets built from non-experts (Table 6) have low inter-labeler user agreements(Table 1) while relatively high self-to-self consistency can be reached for each user him/herself (Table 5). On the other hand, even a training set annotated by experts, inter-labeler user agreements on the emotion labels are not so high as we expected (see Table 3 and Table 4).

For the personalized music emotion classifiers, user participation in the training process is an important and unneglectable issue. So, how to reduce human annotation efforts will become the key point.

3. ACTIVE LEARNING

Yang et al. [14] made an assumption that high user participation is not an issue when proposing their two-layer personalized MER system, which obviously can not hold true in practice. Naturally, active learning becomes the solution.

3.1 Algorithm Description

The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer labeled instances if it is allowed to selectively choose the data from which it learns. An active learner selectively chooses the most informative instances for soliciting labels to be labeled by an oracle (e.g., a human annotator) such that the number of instances requiring labeling could be reduced dramatically.

The proposed active learning algorithm for personalized music emotion classification is shown in Algorithm 1. We randomly select N0 instances (both positive and negative instances need to be included) to train the initial classifier model for each emotion category. At the active learning stage, the classifier is used to select the most informative N instances as queries from a pool of unlabeled instance to be labeled by the user u. Then the labeled instances will be removed from the unlabeled set, and added to the training set to retrain the classifier. Iteration goes on until the there are no unlabeled instances or the classification performance satisfies our expectation.

3.2 Query Instances Selection Criteria

One of the major issues in active learning frameworks is the query instances selection strategy. There are mainly two approaches for selecting the most informative query instances: certainty-based methods [11], and committee-based methods.

Certainty-based approaches are often straightforward for probabilistic learning models. The following two strategies for certainty-based approach will be employed.

- *Informativeness* The basic idea of the informativeness strategy is that instances that the current model uncertain about are selected for labeling. For binary classification, instances with a class posterior probability closest to 0.5 will be most uncertainty.

- *Representativeness* Another general active learning criterion is querying the instances that would impart the greatest change to the current model if we knew its label [9]. For a binary classfier, instances with classification probability in the range $[0, \alpha]$ are believed to be most likely negative, while instances with classification probability in the $[\beta, 1]$ are believed to be most likely positive. If instances from the preceding range are misclassified, greatest change will be made to the current model. Adding those samples to the training dataset for labeling will improve the robustness of the model.

4. EXPERIMENTS

4.1 Dataset

The personalized music emotion classificatioin experiment was conducted on a Chinese popular music dataset, which consists of 657 Chinese popular songs from randomly picked artists, ranging from the 80s to now, without pre-selection and with no emotion labels provided.

We selected the 60s to 90s period in each music piece and converted it to 22,050 Hz and 16 bits mono channel PCM WAV file format.

4.2 Feature Sets

We use three toolkits including Marsyas [13], PsySound [1] to extract audio-based, psychoacoustic-based, and speech emotion-based feature sets, respectively [1].

We directly concatenate the three sets of features into a complete feature vector for each music piece.

4.3 Participants and Annotations

Eight Chinese participants were involved in the personalized active learning procedure, four men and four women. They come

[1]http://www.ece.ust.hk/∼dsu/musicemotion/data

Algorithm 1 Active Learning for Personalized Music Emotion Classfication

for $u \leftarrow user1, user2, user3, ...$ **do**
  for $k \leftarrow angry, calm, happy, sad, ...$ **do**
    Initialization:
    For an unlabeled data set U, $i = 0$;
1. Randomly select $N0$ instances of data $Q\{i,k\}$ from U;
   $U_{left}\{i,k\} = U - Q\{i,k\}$;
   $X\{i,k\} = Q\{i,k\}$;
2. Ask user u to label each instance in $X\{i,k\}$, if it belongs to emotion category k or not, store the binary results in a vector $L\{i,k\}$

    Active Learning & Evaluation:
    while $U_{left} \neq O$ **do**

1. $i = i + 1$;
2. Train classfier $C\{i,k\}$ using $X\{i,k\}$ and $L\{i,k\}$;
3. Test $U_{left}\{i,k\}$ by classifier $C\{i,k\}$, selecting N most informative instances of data $Q\{i+1,k\}$ from $U_{left}\{i,k\}$;
   $U_{left}\{i+1,k\} = U_{left}\{i,k\} - Q\{i+1,k\}$;
   $X\{i+1,k\} = X\{i,k\} + Q\{i+1,k\}$;
4. Ask user u to label each instance in $Q\{i+1,k\}$ if it belongs to emotion category k or not, add the binary results into vector $L\{i+1,k\}$;
5. Evaluate the classifier $C\{i,k\}$ on test set $E\{u,k\}$;

    end while
  end for
end for

from different universities in China, and all of them are casual, everyday music listeners.

For time and other considerations, instead of asking participants to do annotations to query instances after each active learning iteration, we simplified the procedure. They were asked to evaluate all of the music emotion beforehand, with labels pool provided by us, e.g., angry, calm, happy, sad and others. The assumption here, that the perceived emotion of a certain song will not change for the same user during short periods, has been verified by the preliminary experiment in Table 5. We only focus on the four emotion categories because they are the four most basic emotions from psychological theories [4].

Labeling instructions were distributed to each participant beforehand during the annotation process. In the instructions, we introduced the purpose of the labeling experiment, the difference between peceived emotion and felt emotion [14], the importance of ignoring the effect of lyrics when labeling since here we only focused on audio-content in music, and so on. They were allowed to do the labeling several times since labeling 657 30s Chinese music clips is a very time-consuming and labor-intensive task. However, we strongly suggested participants to re-listen to some of the music clips and showed the related labels they previously assigned to help them with recall and to keep consistency during the repeated labeling process. The whole procedure for each participants was finished in a maximum of three days.

For the 657 Chinese songs, Figure 1 shows each participant's annotation distribution on the categories angry, calm, happy, sad and others. It can be seen that the labeling by each participants on the emotion categories differs a lot for individuals. We calculated the inter-labeler agreement kappa values between those participants on the four basic emotion categories on a subset of the Chinese

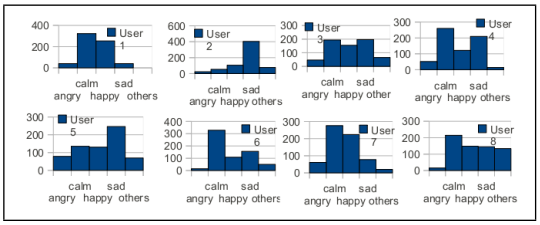

Figure 1: Participants' evaluation results on whole dataset

dataset, and self-to-self inter-labeler measurement by asking the participants to re-do the annotation again several days later. Again the results are consistent with our preliminary experiment results in section II.

4.4 Machine Learning Classifier

SVMs was adopted as our base classifier, whose effectiveness with active learning has been demonstrated in [11, 10, 17]. We adopted libSVM [2] with linear kernel since the dimensions of our feature sets are almost the same order of magnitude with the size of our dataset. Non-linear kernel may cause the model over-fitting. For each emotion category, binary SVMs classifier was trained, and the posterier probabilities were estimated by Platt's equation in [6].

Feature selection was left to SVMs itself. Since active learning is a dynamic process, specific feature sets selected by some extra feature selection methods may not work well as the number of instances increases when iteration goes on.

4.5 Performance Measurement

F-measure for the positive instances was chosen as a performance measurement, since it is a combination of the precision and recall. It is not fair to measure the performance by the overall accuracy. Since the negative samples always hold the largest portion of the whole dataset, it will bias the real performances of our methods.

$$F - measure = 2 * \frac{Precision * Recall}{(Precision + Recall)} \quad (2)$$

4.6 Active Learning Parameters Setting

Ten instances were randomly picked at the initialization stage, with at least two positive instances included.

As for the query instance number N at each iteration, fewer instances each iteration with more iteration rounds usually performs better than more instances each iteration with fewer iteration rounds[10] for a certain number of labeled instances. But considering that more iteration rounds will make participants bored, we selected 10 queries each rounds, which is the best number to balance the classifier performance and the user endurance, according to our user survey.

We employed the following instance selection strategies in our experiments:

- *Random Selection Baseline* A baseline system that will randomly select the query instances from the unlabeled pool at each round.
- *Informativeness* We choose 10 instances at each iteration round with classification probability that is nearest to 0.5.
- *Representativeness* We choose the most negative 5 instances with classification probability nearest to 0 and the most positive 5 instances with classification probability nearest to 1.

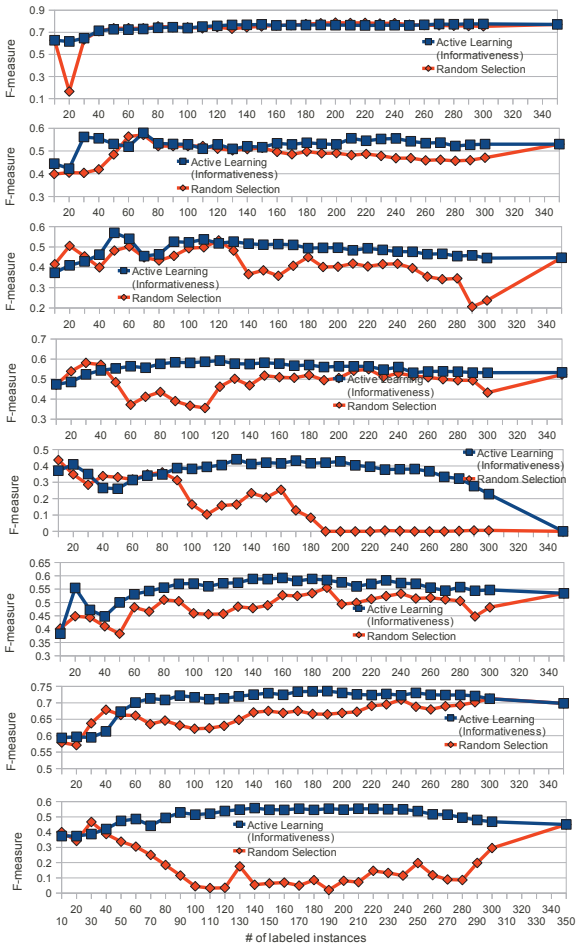

Figure 2: F-measure of personalized music emotion classification on happy emotion by all participants

- *Hybrid* We combine the informativeness and representativeness strategies, to select the 5 most informative instances and 5 most representative instances at each iteration round.

4.7 Experimental Results

We applied 2-fold stratified sampling methods to divide each participant labeled dataset into training set and testing set. For each classifier, 10 groups of experiments were conducted based on different sets of randomly picked initialization sets. Results shown on the curve are averaged over the 10 groups of results.

Fig. 3, shows F-measure curves of the personalized music emotion classification results on the four basic emotions by one of the participants using the four strategies. The reason for the randomness of curves for sadness will be analysed in the next discussion section. For the other three emotion categories, active learning using the informativeness strategy generally gives the best performances. The curves always converge fast, and the F-measures increase stably. The hybrid strategy also performs well. We can see that with only 17% (60 instances), the F-measure almost reached its best performance, which even better than the F-measure results when all of the training data is added in.

Fig. 2 is the F-measure curves of the personalized music emotion classification by all of the 8 participants on happiness, by comparing the best performed active learning strategy using the informa-

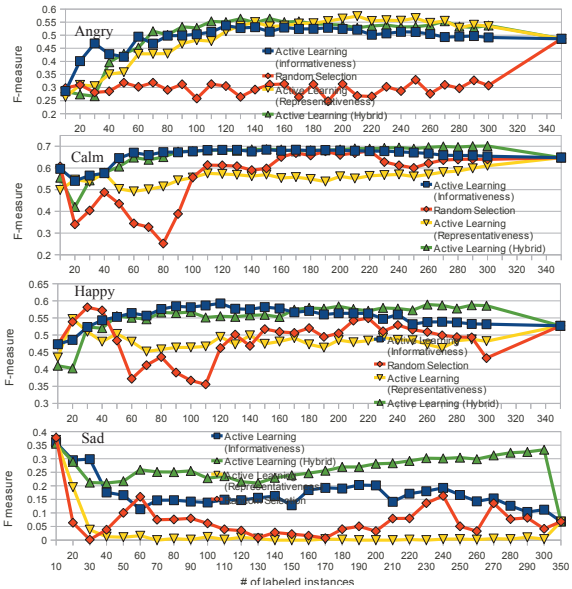

Figure 3: F-measure of participant 4 on different emotion categories

tiveness and random selection criterion. Also, the failure of participant 5 on happy emotion will be analysed in the following section.

4.8 Error Analysis: The Noisy Oracles Issue

One of the strongest assumptions in most active learning work is that the quality of labeled data is high. For our personalized Chinese music emotion classification experiment, labels are entirely the opinion of the single, individual participant who is non-expert. This may cause the quality of the labeled data to not be reliable. First, since the Chinese songs are randomly drawn from the real-world, there are certain portions of the music whose emotions are implicitly difficult to label. Second, it is natural that participants become distracted or fatigued over time. Even though we considered these issues when designing the personalized experiment to reduce participants' labeling variabililty, by allowing them to choose 'others' when it was hard to decide the emotion labels using our provided adjectives set, and suggesting them to re-listen to their previous music-label pairs when they came back to the annotation task after a break, etc., we still could not eliminate variability in the labeled data from each individual participant, especially for our annotation tasks which has 657 music clips in total. This may explain why some of the participant's F-measure curves are not what we expected just like participant 4's sad emotion performed in Fig. 3.

To further verify the emotion labels' consistency, we ran 10-fold cross-validation classification on all of the binary Chinese datasets for the four emotion categories labeled by each participant. Since we assume that if the data is consistent in terms of emotion labels, classifiers can separate positive and negative instances successfully, at least to some extent, with relatively satisfied performances(e.g., F-measure). Table 7 shows the 10-fold cross validation results in terms of F-measure, on the positive instances of each emotion binary datasets labeld by each participants, as well as active learning's status on those datasets. By using ✓we mean that active learning effectively gives satisfied performances (As iteration goes on, performance increases in a stable way), just like the curve shown in Fig. 2. From the comparison of the two subtables in Table 7, we

Table 7: F-measure VS active learning status on the Chinese dataset. Left table is the 10-fold fmeasure results on the Chinese dataset labeled by each participant, while table on the right shows the active learning status, ✓means active learning works well

UserID	Angry	Calm	Happy	Sad
1	**0.789**	**0.780**	**0.761**	0.000
2	**0.526**	0.000	**0.532**	**0.833**
3	0.138	**0.505**	**0.427**	0.029
4	**0.442**	**0.603**	**0.552**	0.000
5	0.024	0.000	0.179	0.345
6	0.000	**0.689**	**0.596**	0.060
7	0.353	**0.690**	**0.734**	0.000
8	0.111	**0.483**	**0.426**	0.040

UserID	Angry	Calm	Happy	Sad
1	✓	✓	✓	×
2	✓	×	✓	✓
3	×	✓	✓	×
4	✓	✓	✓	×
5	×	×	×	×
6	×	✓	✓	×
7	✓	✓	✓	×
8	×	✓	✓	×

can confirm that in order to make active learning work successfully, the music dataset needs to satisfy a certain consistency in terms of emotion labels.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose active learning to solve the subjectivity issue which is one of the major issues in MER. We first show the tenability of applying active learning in a personalized music emotion classificatioin framwork through our preliminary experiments. We further prove the effectiveness of active learning in a personalized music emotion classification system with randomly picked Chinese song dataset with 8 non-experts, which show that active learning can reduce as much as 80% the requirement of human labeling without decreasing the whole performance in terms of F-measure. Different query instances selection criteria were also investigated, and the F-measure curves show that the informativeness criterion, which select the most uncertain instances at each round, in general performs best, i.e., the F-measure curves converge fast and stably as iteration goes on. Finally, we investigate conditions to make active learning work effectively. We show that emotion labels should be annotated consistently in order to make active learning work well, since each non-expert individual will entirely decide the emotion label for each music.

For future work, we could apply active learning in the two-layer model proposed in [14], or apply it on a middle layer with a user clustered emotion classifier.

6. ACKNOWLEDGMENTS

This research is supported by a grant from Velda Limited VELD01 through the HKUST R&D Corporation Ltd. The author would like to thank Anik Dey and all other student helpers for this project.

7. REFERENCES

[1] D. Cabrera et al. Psysound: A computer program for psychoacoustical analysis. In *Proceedings of the Australian Acoustical Society Conference*, volume 24, pages 47–54, 1999.

[2] C. Chang and C. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[3] H. Chen and Y. Yang. Prediction of the distribution of perceived music emotions using discrete samples. *Audio, Speech, and Language Processing, IEEE Transactions on*, (99):1–1, 2011.

[4] P. Juslin and P. Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238, 2004.

[5] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[6] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[7] E. Schmidt and Y. Kim. Prediction of time-varying musical mood distributions from audio. *ISMIR2010*, pages 465–470, 2010.

[8] A. Schmitt, U. Tschaffon, and W. Minker. Inter-labeler agreement for anger detection in interactive voice response systems. In *Intelligent Environments (IE), 2010 Sixth International Conference on*, pages 112–115. IEEE, 2010.

[9] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

[10] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM, 2001.

[11] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[12] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multilabel classification of music into emotions. In *Proc. 9th International Conference on Music Information Retrieval (ISMIR 2008), Philadelphia, PA, USA*, volume 2008, 2008.

[13] G. Tzanetakis and P. Cook. Marsyas: A framework for audio analysis. *Organised sound*, 4(3):169–175, 1999.

[14] Y. Yang, Y. Lin, and H. Chen. Personalized music emotion recognition. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 748–749. ACM, 2009.

[15] Y. Yang, C. Liu, and H. Chen. Music emotion classification: a fuzzy approach. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 81–84. ACM, 2006.

[16] Y. Yang, Y. Su, Y. Lin, and H. Chen. Music emotion recognition: The role of individuality. In *Proceedings of the international workshop on Human-centered multimedia*, pages 13–22. ACM, 2007.

[17] J. Zhang, R. Chan, and P. Fung. Extractive speech summarization by active learning. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 392–397. IEEE, 2009.