Active Class Selection for Arousal Classification

Dongrui Wu[1] and Thomas D. Parsons[2]

[1] Machine Learning Laboratory, GE Global Research Center
One Research Circle, Niskayuna, NY 12309 USA
wud@ge.com
[2] Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Playa Vista, CA 90094 USA
tparsons@ict.usc.edu

Abstract. Active class selection (ACS) studies how to optimally select the classes to obtain training examples so that a good classifier can be constructed from a small number of training examples. It is very useful in situations where the class labels need to be determined before the training examples and features can be obtained. For example, in many emotion classification problems, the emotion (class label) needs to be specified before the corresponding responses can be generated and recorded. However, there has been very limited research on ACS, and to the best knowledge of the authors, ACS has not been introduced to the affective computing community. In this paper, we compare two ACS approaches in an arousal classification application. Experimental results using a kNN classifier show that one of them almost always results in higher classification accuracy than a uniform sampling approach. We expect that ACS, together with transfer learning, will greatly reduce the data acquisition effort to customize an affective computing system.

Keywords: Active class selection, active learning, affective computing, arousal classification, nearest neighbors classification, transfer learning.

1 Introduction

Active learning [11, 13, 21] has been attracting a great deal of research interest recently. It addresses the following problem: Suppose that we have lots of unlabeled training examples and the labels are very difficult, time-consuming, or expensive to obtain; then, which training examples should be selected for labeling so that the maximum learning (classification or prediction) performance can be obtained from the minimum labeling effort? For example, in speech emotion estimation [30, 6], the utterances and their features can be easily obtained; however, it is difficult to evaluate the emotions they express. In this case, active learning can be used to select the most informative utterances to label so that a good classifier or predictor can be trained based on them. Many different approaches have been proposed for active learning [21] so far, e.g., uncertainty sampling [9], query-by-committee [22, 12], expected model change [20], expected error reduction [18], variance reduction [2], and density-weighted methods [32].

S. D´Mello et al. (Eds.): ACII 2011, Part II, LNCS 6975, pp. 132–141, 2011.

One fundamental assumption in active learning is that the training examples can be obtained without knowing the classes (i.e., features can be obtained without knowing the labels). However, in practice there may be situations that the class label needs to be determined first before the training examples can be obtained. For example, in the arousal classification experiment reported in [28], where a Virtual Reality Stroop Test (VRST) was used to obtain training examples, one needs to select a level of arousal (which is the class label) first, and then displays the appropriate test to elicit the corresponding physiological responses, from which the features can be extracted [16]. A classifier is then constructed to estimate a subject's arousal level from physiological responses. So, the problem becomes how to optimally select the classes to obtain training examples so that a good classifier can be constructed from a small number of training examples.

Unlike the rich literature on active learning, there has been limited research on active class selection (ACS). Weiss and Provost [27] studied a closely-related problem: if only n training examples can be selected, in what proportion should the classes be represented? However, as suggested and verified by Lomasky et al. [10], if one can control the classes from which training examples are generated, then utilizing feedback during learning to guide the generation of new training data may yield better performance than learning from any *a priori* fixed class distributions. They proposed several ACS approaches to iteratively select classes for new training instances based on the existing performance of the classifier, and showed that ACS may result in better classification accuracy.

In this paper we apply ACS to the arousal classification problem [28], where three arousal levels need to be distinguished using physiological responses. We implement two of Lomasky et al.'s ACS approaches and perform extensive experiments to compare their performance. To the authors' best knowledge, this is the first time that ACS has been introduced to the affective computing community.

The remainder of this paper is organized as follows: Section 2 introduces the ACS algorithms. Section 3 presents the experimental results of ACS in arousal classification. Section 4 draws conclusions and points out some future research directions.

2 Active Class Selection (ACS)

This section introduces two iterative ACS algorithms. For simplicity we use the k-nearest neighbors (kNN) classifier; however, the algorithms can also be extended to more advanced classifiers like the support vector machine (SVM) [25].

We assume that there are C classes and no limits on generating instances of a particular class. All methods begin with a small set of l_0 labeled training examples, where l_i is the number of instances to generate in Iteration i. ACS is used to determine p_i^c ($0 \leq p_i^c \leq 1$), the portion of the l_i instances that should be generated from Class c. We compare the following three approaches, of which the first is our baseline and the latter two are ACS schemes proposed in [10]:

1. *Uniform*: All classes are uniformly sampled, i.e., $p_i^c = \frac{1}{C}$. This is also the baseline method used in [10]. Uniform sampling is the most intuitive and

frequently used method if there is no *a priori* knowledge on how the sampling should be better done.

2. *Inverse* (ACS_1): This method relies on the assumption that poor class accuracy is due to not having observed enough training examples. It requires internal cross-validation to evaluate the performance of the current classifier so that the poor class can be identified. Leave-one-out cross-validation was used in this paper. In Iteration i, we record the classification accuracy (in the leave-one-out cross-validation) for each class, a_i^c, $c = 1, 2, ..., C$. Then, the probability of generating a new instance from Class c is proportional to the inverse of a_i^c, i.e.,

$$p_i^c = \frac{\frac{1}{a_i^c}}{\sum_{c=1}^{C} \frac{1}{a_i^c}} \tag{1}$$

3. *Accuracy Improvement* (ACS_2): This method is based on the intuition that the accuracy of classes that have been well learned will not change with the addition of new data and thus we should focus on classes that can be improved. Again, it requires internal cross-validation to evaluate the performance of the classifier in the current iteration so that its accuracy can be compared with the classifier in the previous iteration. Leave-one-out cross-validation was used in this paper. In Iteration i, we record the classification accuracy (in the leave-one-out cross-validation) for each class, a_i^c, $c = 1, 2, ..., C$. Then, the probability of generating a new instance from Class c is

$$p_i^c = \max\left(0, \frac{a_i^c - a_{i-1}^c}{\sum_{c=1}^{C}(a_i^c - a_{i-1}^c)}\right) \tag{2}$$

The detailed algorithms for ACS_1 and ACS_2 are given below.

Algorithm 1. The algorithm for ACS_1

Input: $N = \sum_{i=0}^{i-1} l_i$ initial training examples in Iteration i; l_i, the number of new instances to generate in Iteration i; k, the size of the neighborhood in the kNN classifier
Output: The l_i new instances generated in Iteration i
foreach j *in* $[1, N]$ **do**
  Compute the kNN classification result using the jth training example in validation and the rest $N - 1$ examples in training;
end
Compute the per-class classification accuracy in the internal leave-one-out cross-validation a_i^c, $c = 1, 2, ..., C$;
Generate l_i new training examples according to (1)

Algorithm 2. The algorithm for ACS_2

Input: $N = \sum_{i=0}^{i-1} l_i$ initial training examples in Iteration i; l_i, the number of new instances to generate in Iteration i; a_{i-1}^c, $c = 1, 2, ..., C$, the per-class classification accuracy in the internal leave-one-out cross-validation from Iteration $i-1$; k, the size of the neighborhood in the kNN classifier;

Output: The l_i new instances generated in Iteration i

foreach j *in* $[1, N]$ **do**

  Compute the kNN classification result using the jth training example in validation and the rest $N-1$ examples in training;

end

Compute the per-class classification accuracy in the internal leave-one-out cross-validation a_i^c, $c = 1, 2, ..., C$;

Generate l_i training examples according to (2)

3 Experiment

This section presents our experimental results on comparing the three sampling approaches, with application to the arousal classification problem introduced in [28].

3.1 Data Acquisition

The use of psychophysiological measures in studies of persons immersed in high-fidelity virtual environment scenarios offers the potential to develop current physiological computing approaches [1] into affective computing [17] scenarios. Affective computing has been gaining popularity rapidly in the last decade because it has great potential in the next generation of human-computer interfaces [17, 24, 26]. An important task in implementing an affective computing system is affect recognition, which recognizes the user's affect from various signals, e.g., speech [5, 8, 19, 30], facial expressions [15, 4], physiological signals [3, 7, 28], etc.

The Virtual Reality Stroop Task (VRST) [16, 28] utilized in this paper involves the subject being immersed into a virtual Humvee as it travels down the center of a road, during which Stroop stimuli [23] appear on the windshield, as shown in Fig. 1. The VRST stimuli are presented within both "safe" (low threat) and "ambush" (high threat) settings. Low threat zones consist of little activity aside from driving down a desert road, while the more stressful high threat zones include gunfire, explosions, and shouting amongst other stressors. Psychophysiological measures of skin conductance level (SCL), respiration (RSP), vertical electrooculograph (VEOG), electrocardiographic activity (ECG), and electroencephalographic activity (EEG) are recorded continuously throughout exposure to the virtual environment.

There are many different scenarios eliciting different levels of arousal in VRST. In this study we chose the following three of them to affect different arousal levels, which had been used in [28]: 1) Scenario I: Low threat, color naming; 2) Scenario II: High threat, color naming; and, 3) Scenario III: High threat,

(a) (b) (c)

Fig. 1. The Humvee Stroop Scenarios. (a) Color Naming; (b) Word Reading; and, (c) Interference.

interference. Each scenario consisted of 50 tests. Three colors (Blue, Green, and Red) were used, and they were displayed with equal probability. In Scenario I, 50 colored numbers were displayed at random locations on the windshield one by one while the subject was driving through a safe zone. Scenario II was similar to Scenario I, except that the subject was driving through an ambush zone. Scenario III was similar to Scenario II, except that Stroop tests instead of color naming tests were used. In terms of arousal, the three scenarios are in the order of I < II < III.

A total of 19 college aged students participated in this experiment. Strict exclusion criteria were enforced so as to minimize the possible confounding effects of additional factors known to adversely impact a person's ability to process information, including psychiatric (e.g., mental retardation, psychotic disorders, diagnosed learning disabilities, Attention-Deficit/Hyperactivity Disorder, and Bipolar Disorders, as well as substance-related disorders within two years of evaluation) and neurologic (e.g., seizure disorders, closed head injuries with loss of consciousness greater than 15 minutes, and neoplastic diseases) conditions. The University of Southern California's Institutional Review Board approved the study. After informed consent was obtained, basic demographic information was obtained.

3.2 Comparative Study

One of the 19 subjects did not respond at all in one of the three scenarios, and was excluded as an outlier. Only the remaining 18 subjects were studied. Each subject had 150 responses (50 for each arousal level). The same 29 features as those in [28] were used. In the comparative study $k = \{1, 2, 3, 4\}$, since we want to examine whether the performance of ACS is consistent for different k. We studied each subject separately, and for each subject $l_0 = k + 1$ (so that we can run leave-one-out cross-validation using the kNN classifier). Only one new instance was generated in each iteration. After Iteration i, the kNN classification performance was evaluated using the rest $150 - (k + i + 1)$ responses from the same subject. We repeated the experiment 100 times (each time the l_0 initial training examples were chosen randomly) for each subject and k and then report

the average performance of the three class-selection approaches. It is necessary to repeat the experiment many times to make the results statistically meaningful because there are two forms of randomness: 1) a subject generally had different responses at the same arousal level (class label), so for the same sequence of class labels the training examples were different; and, 2) the new class label was generated according to a probability distribution instead of deterministically.

Experimental results for $k = \{1, 2, 3, 4\}$ are shown in Fig. 2. Each of the first 18 sub-figures in (a)-(d) represents a different subject, and the last sub-figure shows the average performance of the three class-selection approaches over the 18 subjects. Observe that:

1. Generally ACS_1 (Inverse) always outperformed the uniform sampling approach. To show that the performance difference is statistically significant, we performed paired t-tests to compare the average performances of ACS_1 and the uniform sampling approach for $k = \{1, 2, 3, 4\}$ and $\alpha = 0.05$. When $k = 1$, $t(17) = 4.66$, $p = 0.0002$. When $k = 2$, $t(17) = 9.06$, $p < 0.0001$. When $k = 3, t(16) = 8.27, p < 0.0001$. When $k = 4, t(15) = 7.97, p < 0.0001$. Clearly, the performance difference is always statistically significant.
   Interestingly, in [10] Lomasky et al. pointed out that the inverse method did not work well. We think this is because the performance of an ACS approach is highly application-dependent, and the inverse approach is particularly suitable for the arousal classification problem. This was partially supported by the fact that Lomasky et al. compared five different sampling approaches in [10] on two datasets, and none of them seemed to be universally better than others. However, more experiments and analysis are needed to better understand the underlying reasons and also the stability of the inverse method.
2. ACS_2 (Accuracy improvement) always had the worst performance among the three sampling approaches. This is because in each iteration only one new instance is generated (assume it belongs to Class c'), and hence very probably in the next iteration only the classification accuracy of Class c' is improved; as a result, ACS_2 keeps generating new instances from Class c' and makes the class distribution very imbalanced. Some typical trajectories of selected training example classes for Subject 1 are shown in Fig. 3. Clearly, ACS_2 tends to stick to a particular class.
3. The overall classification performance decreased as k increased. The exact reason is still under investigation. However, the performance downgrade of ACS_1 was smaller than the uniform sampling approach. This suggests that ACS_1 is less sensitive to k, which is good when it is difficult to determine the optimal k.

In summary, we have demonstrated through a simple kNN classifier the advantages of ACS, which include higher classification accuracy and more robustness to parameter selection.

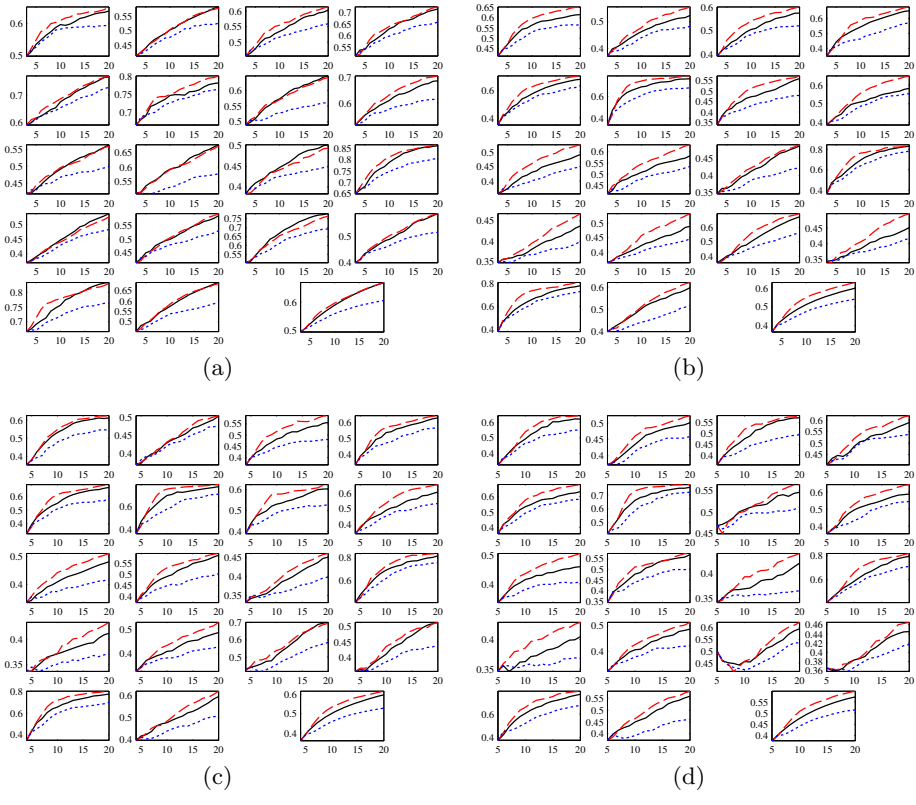

Fig. 2. Performance comparison of the three class-selection approaches on the 18 subjects. (a) $k = 1$; (b) $k = 2$; (c) $k = 3$; (d) $k = 4$. The horizontal axis shows the number of training examples, and the vertical axis shows the testing accuracy on the remaining examples from the same subject. —: Uniform; – – –: ACS_1 (Inverse); - - -: ACS_2 (Accuracy Improvement).

We need to point out that ACS has more computational cost than the uniform sampling approach, because before acquiring each new training example it needs to compute the leave-one-out cross-validation performance and then to determine which class to sample. However, since a person's physiological responses or affective states cannot change very quickly (usually on the order of seconds), and the extra computational cost only occurs during the training process, it does not hinder the applicability of ACS.

Finally, note that the purpose of the experiments is not to show how good a kNN classifier can be in arousal classification; instead, we aim to demonstrate how ACS can improve the performance of an existing classifier. Also, as we have shown in this section, not necessarily all ACS algorithms can always improve the classification performance. For each particular application, a small dataset may be needed to identify the ACS approach.

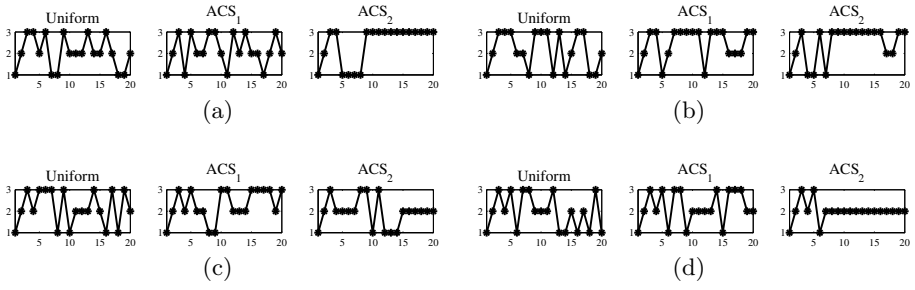

Fig. 3. Typical trajectories of selected training example classes for Subject 1. (a) $k = 1$; (b) $k = 2$; (c) $k = 3$; (d) $k = 4$. The horizontal axis shows the index of training examples, and the vertical axis shows the corresponding class index.

4 Conclusions and Future Research

Active class selection studies how to optimally select the classes to obtain training examples so that a good classifier can be constructed from a small number of training examples. In this paper, we have compared two ACS approaches in an arousal classification application. Experimental results using a kNN classifier showed that the inverse ACS approach generally resulted in higher classification accuracy and more robustness than the uniform sampling approach. To the best knowledge of the authors, this is the first time that ACS has been applied to affective computing problems.

Our future research includes:

1. To compare ACS approaches using more advanced classifiers like SVM, logistic regression, etc, and also on more affective computing datasets, to study whether the performance improvement is consistent and universal.
2. To integrate ACS with feature selection. As it has been shown in [28], many of the 29 features are not useful. However, the useful features are subject-dependent. As the features directly affect the NNs, it is necessary to integrate ACS with feature selection for further performance improvement.
3. To integrate ACS with classifier parameter optimization, e.g., k in the kNN classifier, and C, ϵ and the kernel parameters in the SVM [25].
4. To combine ACS with transfer learning [14,31] in affective computing. A major assumption in many classification and prediction algorithms is that the training and test data are in the same feature space and have the same distribution. However, this does not hold in many real-world applications. For example, in the arousal classification experiment introduced in this paper, a subject's physiological responses at a certain arousal level are generally quite different from another's. This makes it difficult to make use of other subjects' responses. In this paper we ignored other subjects' responses completely in classifying an individual subject's arousal level. However, all subjects' responses should still be similar at some extent, and hence other subjects'

responses may also be useful in classifying an individual subject's arousal levels. Transfer learning is a framework to address this kind of problems by making use of auxiliary training examples. We [29] have applied inductive transfer learning to the above arousal classification problem and showed that the auxiliary training examples can indeed improve the classification performance. We expect that further performance improvement can be obtained by combining transfer learning and ACS, i.e., very few user-specific training examples are needed to obtain satisfactory classification accuracy if ACS and transfer learning are integrated properly. This would make it much easier to customize an affective computing system for individual use.

References

1. Allanson, J., Fairclough, S.: A research agenda for physiological computing. Interacting With Computers 16, 858–878 (2004)
2. Cohn, D.: Neural network exploration using optimal experiment design. In: Proc. Advances in Neural Information Processing Systems (NIPS), Denver, CO, vol. 6, pp. 679–686 (1994)
3. Fairclough, S.H.: Fundamentals of physiological computing. Interacting with Computers 21, 133–145 (2009)
4. Fasel, B., Luettin, J.: Automatic facial expression analysis: A survey. Pattern Recognition 36(1), 259–275 (2003)
5. Grimm, M., Kroschel, K., Mower, E., Narayanan, S.S.: Primitives-based evaluation and estimation of emotions in speech. Speech Communication 49, 787–800 (2007)
6. Grimm, M., Kroschel, K., Narayanan, S.S.: The Vera Am Mittag German audio-visual emotional speech database. In: Proc. Int'l Conf. on Multimedia & Expo. (ICME), Hannover, German, pp. 865–868 (2008)
7. Kim, J., Andre, E.: Fusion of multichannel biosignals towards automatic emotion recognition. In: Lee, S., Ko, H., Hahn, H. (eds.) Multisensor Fusion and Integration for Intelligent Systems. LNEE, vol. 35, pp. 55–68. Springer, Heidelberg (2009)
8. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. IEEE Trans. on Speech and Audio Processing 13(2), 293–303 (2005)
9. Lewis, D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Proc. Int'l. Conf. on Machine Learning (ICML), New Brunswick, NJ, pp. 148–156 (July 1994)
10. Lomasky, R., Brodley, C.E., Aernecke, M., Walt, D., Friedl, M.: Active class selection. In: Proc. 18th European Conference on Machine Learning, Warsaw, Poland, pp. 640–647 (September 2007)
11. MacKay, D.J.C.: Information-based objective functions for active data selection. Neural Computation 4, 589–603 (1992)
12. McCallum, A., Nigam, K.: Employing EM in pool-based active learning for text classification. In: Proc. Int'l. Conf. on Machine Learning (ICML), Madison, WI, pp. 359–367 (July 1998)
13. Muslea, I., Minton, S., Knoblock, C.A.: Active learning with multiple views. Journal of Artificial Intelligence Research 27, 203–233 (2006)
14. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering 22(10), 1345–1359 (2010)

15. Pantic, M., Rothkrantz, L.: Automatic analysis of facial expressions: The state of the art. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(12), 1424–1445 (2000)
16. Parsons, T., Courtney, C., Arizmendi, B., Dawson, M.: Virtual reality Stroop task for neurocognitive assessment. Studies in Health Technology and Informatics 143, 433–439 (2011)
17. Picard, R.: Affective Computing. The MIT Press, Cambridge (1997)
18. Roy, N., McCallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: Prof. Int'l. Conf. on Machine Learning (ICML), Williamstown, MA, pp. 441–448 (2001)
19. Schuller, B., Lang, M., Rigoll, G.: Recognition of spontaneous emotions by speech within automotive environment. In: Proc. German Annual Conf. on Acoustics, Braunschweig, Germany, pp. 57–58 (March 2006)
20. Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. In: Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, pp. 1289–1296 (December 2008)
21. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
22. Seung, H., Opper, M., Sompolinsky, H.: Query by committee. In: Proc. ACM Workshop on Computational Learning Theory, Pittsburgh, PA, pp. 287–294 (July 1992)
23. Stroop, J.: Studies of interference in serial verbal reactions. Journal of Experimental Psychology 18, 643–661 (1935)
24. Tao, J., Tan, T.: Affective computing: A review. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 981–995. Springer, Heidelberg (2005)
25. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Berlin (1995)
26. Vesterinen, E.: Affective computing. In: Digital Media Research Seminar, Finland (2001)
27. Weiss, G.M., Provost, F.: Learning when training data are costly: The effect of class distribution on tree induction. Journal of Artificial Intelligence Research 19, 315–354 (2003)
28. Wu, D., Courtney, C.G., Lance, B.J., Narayanan, S.S., Dawson, M.E., Oie, K.S., Parsons, T.D.: Optimal arousal identification and classification for affective computing: Virtual Reality Stroop Task. IEEE Trans. on Affective Computing 1(2), 109–118 (2010)
29. Wu, D., Parsons, T.D.: Inductive transfer learning for handling individual differences in affective computing. In: D´Mello, S., et al. (eds.) Affective Computing and Intelligent Interaction, Part II, vol. 6975, pp. 142–151. Springer, Heidelberg (2011)
30. Wu, D., Parsons, T.D., Mower, E., Narayanan, S.S.: Speech emotion estimation in 3D space. In: Proc. IEEE Int'l Conf. on Multimedia & Expo. (ICME), Singapore, pp. 737–742 (July 2010)
31. Wu, P., Dietterich, T.G.: Improving SVM accuracy by training on auxiliary data sources. In: Proc. Int'l Conf. on Machine Learning, Banff, Alberta, Canada, pp. 871–878 (July 2004)
32. Xu, Z., Akella, R., Zhang, Y.: Incorporating diversity and density in active learning for relevance feedback. In: Proc. European Conference on Information Retrieval (ECIR), Rome, Italy, pp. 246–257 (April 2007)