TAMPERE UNIVERSITY OF TECHNOLOGY

An active learning method using clustering and committee-based sample selection for sound event classification

Citation
Shuyang, Z., Heittola, T., & Virtanen, T. (2018). An active learning method using clustering and committee-based sample selection for sound event classification. In *16th International Workshop on Acoustic Signal Enhancement, IWAENC 2018* (pp. 116-120). IEEE. https://doi.org/10.1109/IWAENC.2018.8521336

Year
2018

Version
Peer reviewed version (post-print)

Link to publication
TUTCRIS Portal (http://www.tut.fi/tutcris)

Published in
16th International Workshop on Acoustic Signal Enhancement, IWAENC 2018

DOI
10.1109/IWAENC.2018.8521336

License
Other

Take down policy
If you believe that this document breaches copyright, please contact cris.tau@tuni.fi, and we will remove access to the work immediately and investigate your claim.

Download date:22.01.2021

AN ACTIVE LEARNING METHOD USING CLUSTERING AND COMMITTEE-BASED SAMPLE SELECTION FOR SOUND EVENT CLASSIFICATION

Zhao Shuyang, Toni Heittola, Tuomas Virtanen

Tampere University of Technology

ABSTRACT

This paper proposes an active learning method to control a labeling process for efficient annotation of acoustic training material, which is used for training sound event classifiers. The proposed method performs K-medoids clustering over an initially unlabeled dataset, and medoids as local representatives, are presented to an annotator for manual annotation. The annotated label on a medoid propagates to other samples in its cluster for label prediction. After annotating the medoids, the annotation continues to the unexamined sounds with mismatched prediction results from two classifiers, a nearest-neighbor classifier and a model-based classifier, both trained with annotated data. The annotation on the segments with mismatched predictions are ordered by the distance to the nearest annotated sample, farthest first. The evaluation is made on a public environmental sound dataset. The labels obtained through a labeling process controlled by the proposed method are used to train a classifier, using supervised learning. Only 20% of the data needs to be manually annotated with the proposed method, to achieve the accuracy with all the data annotated. In addition, the proposed method clearly outperforms other active learning algorithms proposed for sound event classification through all the experiments, simulating varying fraction of data that is manually labeled.

Index Terms: active learning, K-medoids clustering, committee-based sample selection, sound event classification

1. INTRODUCTION

Sound event classification [1, 2] has many applications such as environmental noise monitoring [3], road surveillance [4] and remote health care [5]. Nowadays, the majority of sound event classification systems [6, 7] are based on supervised learning, which depends on annotated recordings as training material. Preparing the training material is commonly the most time-consuming part in developing a sound event classifier and annotating audio typically costs much more time than recording it. Similar situation has been faced in other applications such as speech recognition [8] and recommendation systems [9], where unlabeled data is abundant but manual labels are expensive to obtain.

The maximum number of labels can be manually assigned is commonly called a *labeling budget*. In order to optimize the classification accuracy with a limited labeling budget, three techniques have been established, including transfer learning [10], active learning [11, 12] and semi-supervised learning [13]. Transfer learning utilizes an audio representation learned from other tasks, where more labeled data is available. Active learning controls which samples will be annotated in order to efficiently utilize the labeling budget. Semi-supervised learning predicts labels for unlabeled data and use them as training material. The three techniques are not mutual exclusive, and can be combined. There are two previous active learning studies on sound event classification, semi-supervised active learning (SSAL) [11] and medoid-based active learning (MAL)[12]. Both of them involves a sample selection mechanism to control the labeling process, and a label prediction mechanism for unlabeled data.

SSAL performs sample selection and label prediction based on a classifier trained with previously labeled data. Samples with low classification confidence are selected for annotation, whereas samples with high confidence, are assigned with the predicted labels. The classifier relies on a decent amount of annotated data to achieve reliable label prediction and confidence estimation. Thus it can hardly optimize a labeling process at the very early stage when little annotated data is available. A solution to this drawback is to utilize the similarities between data points, which rely on no annotation.

MAL completely relies on the similarities between unlabeled data points. It structures unlabeled data into small clusters using K-medoids clustering. Each medoid, as a local representative, is selected for annotation. The label of an annotated medoid is propagated to the whole cluster. After all the medoids are annotated, MAL repeats the whole process on the data that has not been annotated, clustering the data again and presenting the medoids for annotation. However, repeating the process does not utilize previously annotated data, which is important for optimizing the labeling process, after a decent amount of annotated labels are collected.

In this study, we propose an active learning method that targets on optimizing the whole labeling process, utilizing both the similarities between data points and data annotated previously in the labeling process. The proposed method performs clustering and presents medoids to an annotator similarly to MAL. After annotating all the medoids, the annotation continues to the samples with mismatched prediction results from two classifiers: a nearest-neighbor classifier and a model-based classifier, both trained with annotated data. A segment with mismatched predictions is ranked by the distance to its nearest annotated sample, farthest first. In each iteration, a batch of top ranked samples are selected for annotation, and the rest of the samples update their predicted labels to the labels of their nearest annotated samples.

The structure of the paper is as follows. The problem of optimizing a labeling process is described in Section 2. The proposed active leaning algorithm is introduced in Section 3. The evaluation of the proposed system is presented in Section 4. The conclusion is drawn in Section 5.

2. PROBLEM STATEMENT

We state the problem of optimizing the process of labeling acoustic training material. A set of N sound segments $\mathcal{S} = \{s_1, s_2, ...s_N\}$ is given, initially unlabeled. A set of M sound event classes $\mathcal{C} =$

Funded by European Unions H2020 Framework Programme through ERC Grant Agreement 637422 EVERYSOUND and 737472 SMARTSOUND.

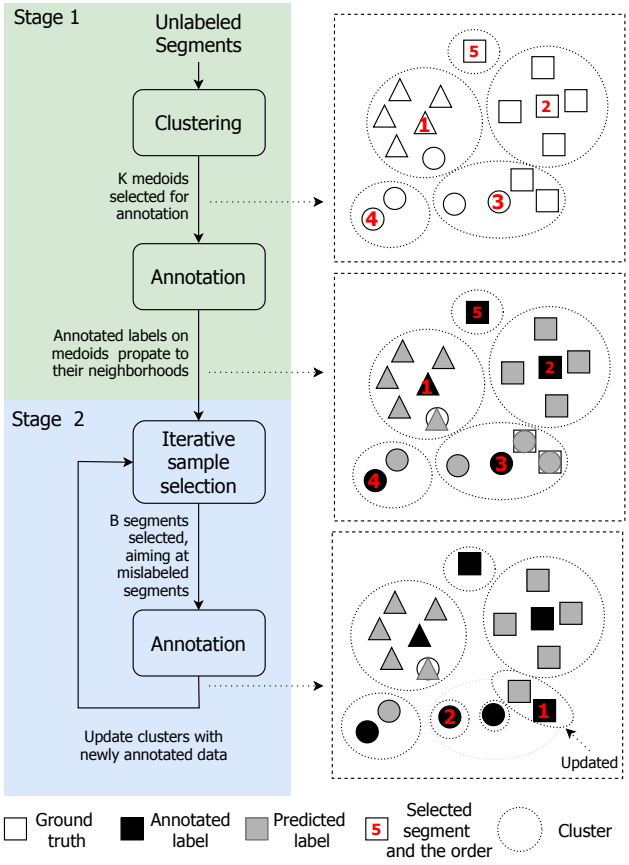

Fig. 1. Illustration of the labeling process, controlled by the proposed method. Each segment is represented with a geometric drawing and the shape represents the class.

$\{c_1, c_2, ...c_M\}$ is pre-defined. A label $l = (s, c) \in \mathcal{S} \times \mathcal{C}$ associates a segment s with a class c.

In a labeling process, an annotator examines sound segments and assigns labels. A label $l = (s, c)$ is added to a label set $\mathcal{L} \subset \mathcal{S} \times \mathcal{C}$, by associating a segment s to a class c. The segments that are manually examined and annotated are called *annotated segments*, denoted as $\mathcal{A}$. The segments that are not examined are called, *unexamined segments*, and denoted as $\mathcal{U} = \mathcal{S} \setminus \mathcal{A}$.

A labeling process produces a label set $\mathcal{L}$, including annotated labels ($\mathcal{L}_\mathcal{A}$) on $\mathcal{A}$ and possibly machine-generated predicted labels ($\mathcal{L}_\mathcal{U}$) on $\mathcal{U}$. The produced label set is used to train a supervised classifier. The problem is to optimize the labeling process that the obtained label set results in the most accurate classifier, under a labeling budget.

3. THE PROPOSED METHOD

The proposed method is illustrated in Figure 1. The input is a set of segments $\mathcal{S}$, initially unlabeled. Sound segments are typically sliced from audio recordings.A set of labels $\mathcal{L}$ is produced through a labeling process, controlled by the proposed method. The labeling process ends when all the segments are annotated or the labeling budget runs out. After the labeling process, $\mathcal{L}$ are used for supervised learning.

The proposed method has two stages. In the first stage, K-medoids clustering is performed and the medoids, as local representatives, are presented to an annotator for manual annotation. An annotated label propagates to the whole cluster as predicted labels. By the end of the first stage, each segment gets a label, either annotated or predicted. In the second stage, a batch of B samples are selected for annotation in each iteration. The selection is based on the prediction mismatch between two classifiers: nearest-neighbor prediction based on $\mathcal{A}$ and a model-based classifier trained with $\mathcal{A}$. The segments are further ranked by the distance to the nearest annotated segment. In the second stage, the clusters are updated, using $\mathcal{A}$ as cluster centroids and assigning each unexamined segment to its nearest annotated segment.

3.1. Distance matrix

The proposed method relies on a distance metric relevant to the target classification problem. The distances between segments under the same class should be generally smaller, compared to segments under different classes. We compute a distance matrix consisting of pairwise distances between all the sample.

Mel-frequency cepstral coefficients (MFCCs), its first-order and second-order derivatives are used as acoustic features. The MFCCs within a sound segment is represented by a multi-variate Gaussian distribution, based on the mean and the variance. Symmetric Kullback–Leibler (KL) divergence is used to measure the dissimilarity between a segment pair. The measured dissimilarity between two segments x and y is called distance for simplicity, and denoted as $d(x, y)$, though KL divergence is not distance. The distance from a segment to itself is zero and the distance matrix $D^{N\times N}$ is symmetric with diagonal values being zero.

The MFCCs-Gaussian-KL as a similarity measurement has been widely used in acoustic information retrieval [15, 16]. Besides MFCC-Gaussian-KL as a static programmed similarity measurement, there are studies on machine-learned metrics, which outperformed static programmed similarity metrics in problems such as content-based music recommendation [17] and sound event query by voice-imitated examples [18]. However, this does not suit the targeted situation, since a learned metric itself requires labeled data to train.

3.2. Stage one: Clusters with representatives

K-medoids clustering is performed based on the distance matrix. The clustering algorithm finds a set of K medoids $\mathcal{M} = \{m_1, m_2, ..m_k\}$, that minimizes the total distance from each segment to its nearest medoid, as $\sum_{x\in\mathcal{S}} min\{d(x, y)|y \in \mathcal{M}\}$. This can be interpreted that $\mathcal{M}$ is an optimized set of segments to make nearest-neighbor prediction to the rest. Thus, medoids are presented to the annotator for labeling. The annotated label assigned to a medoid propagates to the whole cluster as predicted labels. The label propagation is equivalent to nearest neighbor prediction based on $\mathcal{M}$.

3.2.1. K-medoids clustering

K-medoids [19, 20] is a partitioning-based clustering algorithm, similar to more widely-used K-means. The main difference is that K-medoids uses real data point as centroids, whereas in K-means, a cluster centers at an arbitrary data point.

The initialization of medoids is based on farthest-first traversal [21]: a traversed set starts as a singleton of a random segment and the farthest segment to the current traversed set (the distance from a

point x to a set $\mathcal{S}$ is defined as $d(x, \mathcal{S}) = min\{d(x, y)|y \in \mathcal{S}\})$ is iteratively added to the traversed set. Farthest-first traversal has been proved to give an efficient approximation of k-center problem [22].

3.2.2. Choosing the number of clusters

We analyse the number of clusters K inversely, using a factor $KI = \frac{N}{K}$, where KI can be interpreted as the average cluster size. KI controls the trade-off between quantity and accuracy of generated predicted labels. In the previous MAL study [12], KI has been fixed to four, based on a preliminary experiment on a small scale dataset. However, the best choice of KI varies along with each dataset.

We propose a median neighborhood test method to determine KI, estimating the largest cluster size that an annotated label can reliably propagate to. The test needs to manually annotate a small number of segments. Firstly, we choose a pivotal segment p, the segment that has the median distance to its nearest neighbor among $\mathcal{S}$, targeting on a segment with average neighborhood density. A counter i is initially set to one. The algorithm queries the label for the ith nearest neighbor of p. The counter increments if the label of the ith nearest neighbor matches with p. Otherwise, we settle with $KI = i$ and runs K-medoids clustering with it. In case KI ends up to be one, the method will be equivalent to random sampling. This happens when the distance metric is highly irrelevant to the target classification problem.

3.3. Stage two: Mismatch-first farthest-search

The sample selection in the second stage is iterative. In each iteration, a batch of B samples are selected for annotation, denoted as $\mathcal{B}$. The selection is based on mismatch-first farthest-search, targeting on segments with wrong predicted labels.

Our first sample selection criteria is based on committee-based sample selection [14]. The principle is to select samples with mismatched prediction results from different types of classifiers, trained with the same material, as a decision committee. It is based on two assumptions. The first one is that a classifier is more likely to be wrong when another type of classifier makes a mismatched prediction, compared to the case that the whole committee agrees on the prediction. The other assumption is that a classifier benefits more from a counter example, where the classifier makes mistakes, than an example where the classifier succeeds. Every selected sample is a counter example to at least one classifier in the committee, thus the committee as a whole efficiently improves with the selected samples.

The proposed method intrinsically involves two types of classifiers: the nearest-neighbor classifier for label prediction and the model-based classifier trained after the labeling budget runs out. The model-based classifier is trained with $\mathcal{L}_\mathcal{A}$ and the prediction results on $\mathcal{U}$ are compared with the $\mathcal{L}_\mathcal{U}$. The prediction mismatch between the two classifiers is the first criteria in the sample selection.

There are typically multiple unexamined segments with mismatched predictions. The second criteria is the distance of label propagation, assuming that the label propagating the largest distance is most likely to be wrong. Thus, the segments with mismatched predictions are further ranked by the distance to its nearest annotated segment. Practically, the segments with mismatched prediction are added to $\mathcal{B}$ based on farthest-first traversal, as is defined in Section 3.2.1, adding the sample that has the farthest distance to $\mathcal{A} \cap \mathcal{B}$ to $\mathcal{B}$ until $\mathcal{B}$ reaches the size of B. In case that less than B segments have mismatched predictions, farthest-first traversal continues to segments of matched predictions.

After annotating a batch of segments, the predicted label of each unexamined segment is updated based on its nearest annotated segment. This is equivalent to replacing $\mathcal{M}$ by $\mathcal{A}$ as medoids and updating the partition in K-medoids clustering. Since $\mathcal{M} \subseteq \mathcal{A}$ in the second stage, the sizes of updated clusters are equal or smaller compared to the first stage.

4. EVALUATION

In order to evaluate an active learning algorithm, we use the obtained labels to train a supervised classifier, with which the classification accuracy on a test dataset is used for evaluation. The labels obtained with different active learning algorithms vary in terms of quantity and accuracy, thus the resulted classifier is used for evaluation.

4.1. Dataset

Previous study on MAL used UrbanSound8K [24] for evaluation. We use the same dataset in this study for consistency. UrbanSound8K is a public environmental sound dataset, based on real field-recordings. The dataset includes 8 732 manually annotated sound segments with maximum duration of 4 seconds, totalling 8.75 hours. The dataset includes 10 sound event classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren and street music. The dataset provides a 10-folds division for cross validation.

4.2. Experimental setup

The experimental setup also follows the previous MAL study. An active learning algorithm output training material that requires a supervised classifiers to evaluate. Since the purpose of the evaluation is not to find the best model, we simply use a support vector machine (SVM) classifier, the baseline classifier of the UrbanSound8K dataset with radial basis function as kernel.

The acoustic feature extraction in the supervised learning also follows the baseline in UrbanSound8K, using the following summary statistics of MFCCs in each segment: minimum, maximum, median, mean, variance, skewness, kurtosis and the median and variance of the first and second derivatives. MFCCs used in the similarity measurement and supervised learning are the same. The audio signal is divided into frames with 24 ms length and 50% frame overlap. We compute 1st to 25th MFCCs from 40 Mel bands between 25 Hz and 22 050 Hz.

In each round of evaluation, nine folds are used for training and one fold is used for testing. The labels provided by the dataset are used as ground truth. In a training set, the ground truth labels are initially all hidden. Annotating a sound segment consumes the labeling budget by one. The annotated labels are always simulated with the ground truth.

Unweighted accuracy is used to evaluate the performance. It weighs different classes the same, regardless to the number of instances. The classification accuracy is reported averaging the accuracy across all 10 folds. Due to the random elements, medoid initialization and random sampling, in the experiments, all the experiments are repeated three times and the averaged results are reported.

4.3. Reference methods

Random sampling is commonly used as a baseline in active learning studies [11, 12]. It presents the data to the annotator in a random permutation.

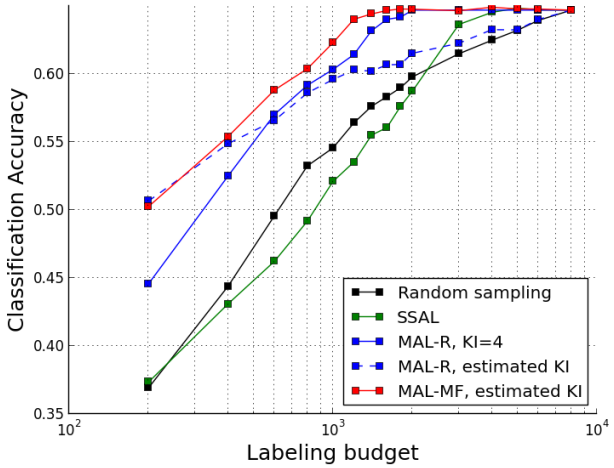

Fig. 2. Classification accuracy as a function of labeling budget. The proposed method, MAL-MF, is evaluated with SSAL [11], MAL-R [12] and random sampling as reference methods.

SSAL [11] is used as the second reference method. In the first stage, 200 samples are randomly selected. In the second stage, the sample selection is iterative. In each iteration, the annotated labels are used to train a classifier. In each iteration, the least confident 50 samples to the classifier are selected for annotation. When the labeling process ends, unexamined segments get predicted labels from the classifier, and all the obtained labels are used to train a final classifier. Originally in the SSAL study, it has a maximum confidence threshold for sample selection and samples are randomly selected under the threshold. In addition, it has a minimum confidence threshold for label prediction. Our reference method does not use these two thresholds, since there is not an established rule to set them.

Previous MAL [12], named here MAL-recursive (MAL-R), has the similar procedure as the first stage of the proposed method, with fixed $KI = 4$. It runs a recursive process, repeating the first stage process on unexamined segments, after all the medoids are annotated. We firstly evaluate MAL-R with $KI = 4$, as it has been originally proposed. Additionally, we evaluate MAL-R with the KI estimated using the proposed median neighborhood test.

The proposed method, medoid-based active learning with mismatch-first farthest-search (MAL-MF) uses median neighborhood test to determine KI. The batch size in the second stage is set to 50, the same as the experimental setup on SSAL.

4.4. Results

Figure 2 illustrates the performance of the proposed method (MAL-MF), compared to MAL-R, SSAL and random sampling. All segments in the training set get annotated labels when the labeling budget is 8 000. When all the segments are labeled as ground truth, the obtained classifier achieves an accuracy about 64.7%, which is the ceiling performance of all compared methods. Experimentally in some cases, a few errors in predicted labels result in a classifier with higher accuracy. As a result, some results in the illustration may be slightly higher than the ceiling performance. We call a result to approximate the ceiling performance when the difference in accuracy is lower than 0.5%.

The result shows that the proposed method outperforms all the reference methods through the experiments. The proposed method requires only 20% of the training data to be manually annotated to approximate the ceiling performance. In comparison, SSAL outperforms baseline only when the labeling budget is more than 25% of the training data. The main reason is that the labels predicted with SVM are much less accurate than the labels propagated from the local representatives, when the labeling budget is low.

The proposed method and MAL-R shares the same process in the first stage. The proposed method uses KI estimated separately for each fold. Based on the proposed median neighborhood test, the choice of KI ranges in $[4, 16]$ across the ten folds, with the median of 12. When MAL-R uses fixed $KI = 4$ as previously proposed, the cluster size is relatively small,thus the purity of the clusters is more than 97%. It approximates the ceiling performance by annotating all the medoids, using 25% of unlabeled data as labeling budget. The proposed method, considering the median case $KI = 12$, produces labels three times fast as $KI = 4$, with the purity of clusters dropping to 85%. The higher number of obtained labels allows better performance on small labeling budget. The second stage process allows the proposed method to effectively correct the errors in predicted labels. As a result, the proposed method approximates the ceiling performance using only 20% of unlabeled data as labeling budget. When MAL-R uses the same KI estimated with the proposed median neighborhood test, it has the same performance to MAL-MF with low labeling budget, however the accuracy of MAL-R increases slowly as labeling budget grows, due to its non-optimal second stage.

In order to analyse the sample selection performance in the second stage, we observed the label prediction error rate in unexamined segments, unexamined segments with mismatched predictions and selected segments. From the beginning of the second stage to where the performance approximates the ceiling, the prediction error rate of segments with mismatched predictions is typically 1.5 times to the error rate in all unexamined segments. The selected segments, the segments with mismatched prediction and ranking top 50 by the distance to the nearest annotated segment, has 3-10 times label prediction error rate, compared to error rate in all unexamined segments. Typically the ratio grows from three to ten along with the labeling process.

5. CONCLUSIONS

This study proposes an active learning algorithm to control the labeling process on sound event data, to save the annotation effort to prepare training material. The proposed method has two stages. In the first stage, K-medoids clustering is performed on an unlabeled dataset and the medoids are selected for annotation. The annotated label on a medoid propagates to its cluster. In the second stage, the selection is based on mismatch-first farthest-search, an extension and committee-based sample selection. The predicted labels are updated using nearest-neighbor prediction, based on the annotated data.

The evaluation is based on the classification accuracy on a test dataset, using a support vector machine classifier, trained based on labels obtained in the active learning process. The results show that only 20% of the data needs to be manually annotated with the proposed method, to achieve the performance with all the data annotated. Furthermore, it clearly outperforms all the reference method, SSAL and MAL-R, through all the experiments.

In the future, the proposed method can be tried to save labeling budget to classify other media type, if there is a exists a similarity metric that gives decent retrieval performance.

6. REFERENCES

[1] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[2] Karol J. Piczak, "ESC: dataset for environmental sound classification," in *23rd Annual ACM Conference on Multimedia Conference*, 2015, pp. 1015–1018.

[3] Panu Maijala, Zhao Shuyang, Toni Heittola, and Tuomas Virtanen, "Environmental noise monitoring using source classification in sensors," *Applied Acoustics*, vol. 129, no. 6, pp. 258267, January 2018.

[4] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Transaction on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 279–288, 2016.

[5] Ha Manh Do, Weihua Sheng, and Meiqin Liu, "Human-assisted sound event recognition for home service robots," *Robotics and Biomimetics*, vol. 3, no. 1, pp. 7, Jun 2016.

[6] Emre Çakir, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[7] Juncheng Li, Wei Dai, Florian Metze, Shuhui Qu, and Samarjit Das, "Comparison of deep learning methods for environmental sound detection," in *2017 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '17*, 2017, pp. 126–130.

[8] Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani, "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," in *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, 2003, pp. 58–65.

[9] Neil Rubens, Mehdi Elahi, Masashi Sugiyama, and Dain Kaplan, "Active learning in recommender systems," pp. 809–846, 2015.

[10] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016.

[11] Wenjing Han, Eduardo Coutinho, Huabin Ruan, Haifeng Li, Bjrn Schuller, Xiaojie Yu, and Xuan Zhu, "Semi-supervised active learning for sound classification in hybrid learning environments," *PLOS ONE*, vol. 11, no. 9, pp. 1–23, 09 2016.

[12] Shuyang Zhao, Toni Heittola, and Tuomas Virtanen, "Active learning for sound event classification by clustering unlabeled data," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, 2017, pp. 751–755.

[13] Zixing Zhang and Björn W. Schuller, "Semi-supervised learning helps in sound event classification," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*, 2012, pp. 333–336.

[14] Hyunjune S. Seung, Mike Opper, and Haim Sompolinsky, "Query by committee," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, New York, NY, USA, 1992, pp. 287–294.

[15] Marko Leonard Helén and Tuomas Virtanen, "Audio query by example using similarity measures between probability density functions of features," *EURASIP J. Audio, Speech and Music Processing*, vol. 2010, 2010.

[16] Dmitry Bogdanov, Martín Haro, Ferdinand Fuhrmann, Anna Xambó, Emilia Gómez, and Perfecto Herrera, "Semantic audio content-based music recommendation and visualization based on user preference examples," *Inf. Process. Manage.*, vol. 49, no. 1, pp. 13–33, 2013.

[17] Rui Lu, Kailun Wu, Zhiyao Duan, and Changshui Zhang, "Deep ranking: Triplet matchnet for music metric learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, 2017, pp. 121–125.

[18] Yichi Zhang and Zhiyao Duan, "IMINET: convolutional semi-siamese networks for sound search by vocal imitation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2017, New Paltz, NY, USA, October 15-18, 2017*, 2017, pp. 304–308.

[19] Leonard Kaufman and Peter J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley, 1990.

[20] Hae-Sang Park and Chi-Hyuck Jun, "A simple and fast algorithm for k-medoids clustering," *Expert System with Application*, vol. 36, no. 2, pp. 3336–3341, 2009.

[21] Sugato Basu, Arindam Banerjee, and Raymond J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *4th SIAM International Conference on Data Mining*, 2004, pp. 333–344.

[22] Dorit S. Hochbaum and David B. Shmoys, "A best possible heuristic for the *k*-center problem," *Mathematics of Operations Research*, vol. 10, no. 2, pp. 180–184, 1985.

[23] Giuseppe Riccardi and Dilek Hakkani-Tür, "Active learning: theory and applications to automatic speech recognition," *IEEE Transaction on Speech and Audio Processing*, vol. 13, no. 4, pp. 504–511, 2005.

[24] Justin Salamon, Christopher Jacoby, and Juan P. Bello, "A dataset and taxonomy for urban sound research," in *2014 ACM International Conference on Multimedia*, 2014, pp. 1041–1044.