ACTIVE LEARNING FOR SOUND EVENT CLASSIFICATION BY CLUSTERING UNLABELED DATA

Zhao Shuyang *Toni Heittola* *Tuomas Virtanen*

Tampere University of Technology, Finland.

ABSTRACT

This paper proposes a novel active learning method to save annotation effort when preparing material to train sound event classifiers. K-medoids clustering is performed on unlabeled sound segments, and medoids of clusters are presented to annotators for labeling. The annotated label for a medoid is used to derive predicted labels for other cluster members. The obtained labels are used to build a classifier using supervised training. The accuracy of the resulted classifier is used to evaluate the performance of the proposed method. The evaluation made on a public environmental sound dataset shows that the proposed method outperforms reference methods (random sampling, certainty-based active learning and semi-supervised learning) with all simulated labeling budgets, the number of available labeling responses. Through all the experiments, the proposed method saves 50%-60% labeling budget to achieve the same accuracy, with respect to the best reference method.

Index Terms: active learning, sound event classification, K-medoids clustering

1. INTRODUCTION

Sound event classification [1] and detection [2] has many applications such as noise monitoring [3, 4, 5], surveillance [6, 7] and home service robots [8]. The development of sound event classification and detection applications requires annotated recordings. Recordings can be made continuously all day around, almost effortlessly. However, reliable annotation takes at least the duration of a recording. As a result, the annotation work is quite often the main cost to build a sound event classifier. To aim at this situation, we attempt a method that optimizes the classification performance with a limited annotation effort, utilizing an abundant amount of audio data that is much more than the amount that can be afforded to annotate.

The maximum number of labels that can be assigned is called a *labeling budget*, which is used to quantify a limited annotation effort. When labeling budget is small, there are two established techniques to utilize the abundant amount of unlabeled data: active learning and semi-supervised learning.

An active learning algorithm actively asks for labeling responses on data selected by the algorithm from a set of unlabeled data. An unlabeled data point is called a sample and the selection of samples to be labeled is called sampling; after labeling, a labeled data point and its label constitutes a training example. Active learning algorithms controls the sampling, in order to avoid redundant examples to optimize the efficiency of labeling effort. Though other types of active learning methods exist, only certainty-based active learning (CRTAL) [9] has been studied in the field of acoustic pattern recognition. It has been proposed to speech recognition in [10]. In certainty-based active learning methods, a small set of samples (selected by the annotator or randomly) are annotated in the beginning. The annotated labels are used to train a classifier and unlabeled samples are classified. A batch of samples with the lowest classification certainties are presented to the annotator for labeling. The classifier is updated after adding new labels to the training material. An experiment on speech recognition has shown that the amount of labels needed to achieve a target word accuracy can be reduced by 60% using CRTAL [11], compared to random sampling.

Semi-supervised learning (SSL) assigns predicted labels to unlabeled data so that unlabeled data is utilized as training examples according to predicted labels. Expectation-maximization based semi-supervised learning has been studied for various acoustic pattern recognition problems such as speaker identification [12] and musical instrument recognition [13]. These methods start by training an initial classifier with labeled data, and they iteratively update predicted labels of either a batch or all unlabeled data. The final classifier is obtained by training with both annotated labels and predicted labels. Gender identification and speaker identification error rates are generally halved using semi-supervised learning with varying proportion of labeled data [12].

All the above-mentioned methods rely on a classifier for uncertainty sampling or label prediction. However, it would require much labeling effort as an overhead to achieve a classifier that produces reasonable classification outputs (predicted class and certainty). As is shown in [11], as long as less than 10% (about 3000) utterances are labeled, performance of CRTAL is behind random sampling. An ideal way to deal with a small labeling budget is to utilize the internal structure of the dataset so that the method starts to outperform random sampling from the very beginning of a labeling process.

We propose a method to optimize the sound event classification performance when labeling budget is limited and only a small portion of data can be annotated. The proposed method is called medoid-based active learning (MAL). K-medoids clustering is performed on sound segments, and the centroids of clusters (medoids) are selected for labeling. The label assigned to a medoid is used to derive predicted labels for other cluster members. An advantage of MAL over traditional SSL and CRTAL is that it does not depend on a model that would require many labels as an overhead to achieve reliable performance on uncerntainty sampling and label prediction. In the evaluation, labels are produced to a training dataset through the proposed method or reference methods, simulating a limited number of labeling responses. A classifier is trained according to the produced labels and its classification accuracy on a test dataset is used to evaluate the performance of the whole process. Selecting cluster representatives for labeling has been originally proposed for text classification in [14], but it does not use representatives to predict labels. Similar studies have not been found in the field of acoustic pattern recognition.

The proposed method is described in Section 2. The evaluation of the proposed system and the discussion about the results is given in Section 3. The conclusion is drawn in Section 4.

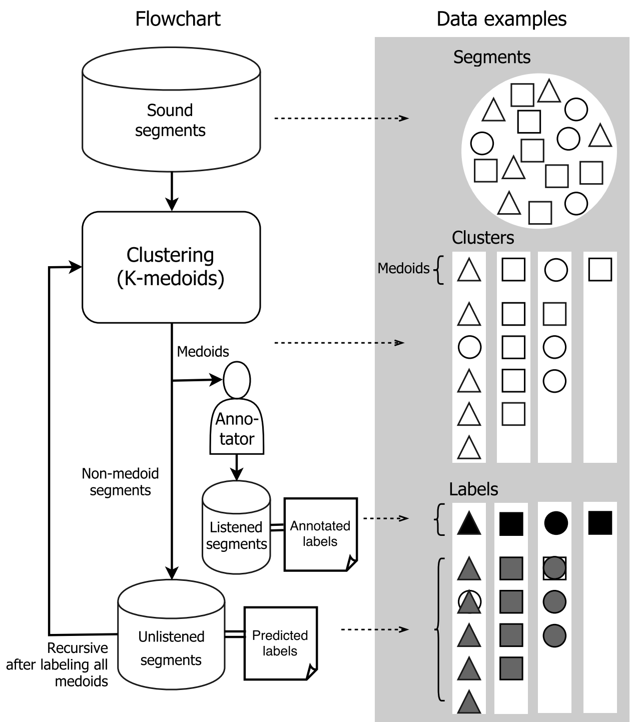

Fig. 1. Overview of the proposed method. The shape of a geometric drawing in data examples represents the ground truth class or label of a segment. Ground truth classes, annotated labels and predicted labels are represented by unfilled drawings, black filled drawings and gray filled drawings, respectively.

2. THE PROPOSED METHOD

The procedure of the proposed method is shown in Figure 1. The proposed method takes sound segments as input and labels of segments are produced as the output. Sound segments are typically sliced from audio recordings. The production of labels requires an annotator who listens to presented segments and assign labels for them. The labels are chosen from a closed set of pre-defined classes.

Segments in the dataset are originally unlabeled and marked as unlistened. Each segment in the dataset is represented by a multivariate Gaussian distribution and the dissimilarity between a pair of segments is measured by Kullback-Leibler (KL) divergence. Segments are clustered using K-medoids algorithm based on the dissimilarity to each other. The medoid of each cluster is presented to annotators for labeling. Medoids are the representatives of local distributions so that they have two useful properties. Firstly, medoids are assured to span different local distributions, thus redundant examples densely distributed within a small area can be avoided. Secondly, a cluster consists of segments around the medoid, thus predicted labels for other cluster members can be derived from the medoid. In case the labeling budget is more than the number of clusters, the annotation proceeds with another round of clustering on unlistened segments. The details of the processing is described in more detail in the following subsections.

2.1. Sound segment representation

Mel-frequency cepstral coefficients (MFCCs), its first-order and second-order derivatives are used as acoustic features. A sound segment is represented by a multi-variate Gaussian distribution, based on the mean and the covariance of the corresponding features. In a preliminary study, using a diagonal covariance matrices gave better performance than using full covariance matrices, thus diagonal covariance matrices are used in this study.

2.2. Segment-to-segment dissimilarity measurement

Dissimilarity measurement between segments is needed to perform clustering. Symmetric KL divergence is a dissimilarity measurement between multi-variate Gaussian pairs, which has been used in various applications such as in music information retrieval [15] and audio texture creation [16]. Symmetric KL divergence is also used in this study to measure the dissimilarity between a pair of sound segments. The KL divergence between two multi-variate Gaussian distributions $\mathcal{P}_0$ and $\mathcal{P}_1$ is calculated as

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{P}_0 \| \mathcal{P}_1) = &\frac{1}{2}(\mathrm{tr}(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_0) \\ &+ (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \mathbf{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ &+ \ln(\frac{\det \mathbf{\Sigma}_1}{\det \mathbf{\Sigma}_0}) - k), \end{aligned} \tag{1}$$

where $\boldsymbol{\mu}_0$ and $\mathbf{\Sigma}_0$ are mean and covariance of distribution $\mathcal{P}_0$, respectively. The mean and covariance of distribution $\mathcal{P}_1$ are denoted as $\boldsymbol{\mu}_1$ and $\mathbf{\Sigma}_1$.

KL divergence is not a commutative operation so that $D_{\mathrm{KL}}(\mathcal{P}_0\|\mathcal{P}_1)$ is different from $D_{\mathrm{KL}}(\mathcal{P}_1\|\mathcal{P}_0)$. In order to obtain a symmetric dissimilarity matrix, the average of both way KL divergence is used to measure the dissimilarity between two segments as

$$D(\mathcal{P}_0\|\mathcal{P}_1) = D(\mathcal{P}_1\|\mathcal{P}_0) = \frac{D_{\mathrm{KL}}(\mathcal{P}_0\|\mathcal{P}_1) + D_{\mathrm{KL}}(\mathcal{P}_1\|\mathcal{P}_0)}{2}. \tag{2}$$

The segment-to-segment dissimilarity matrix $\mathbf{D}$ is formed as $\mathbf{D}_{ij} = D(\mathcal{P}_i\|\mathcal{P}_j)$. Since distance matrix is a more common term used for clustering algorithms, the term distance matrix is used instead of dissimilarity matrix in next section. However, it has to be mentioned that KL divergence is not a distance measure.

2.3. K-medoids clustering

K-medoids clustering algorithm [17, 18] is performed based on the segment-to-segment distance matrix. K-medoids is a partitioning-based clustering algorithm, similar to K-means. K-medoids uses a data point in the dataset as a centroid whereas K-means uses an arbitrary point in the coordinate space as a centroid. K-medoids typically outperforms K-means, in terms of accuracy, and the advantage increases with the size of the dataset [19]. Furthermore, a medoid, as the centroid of a cluster, is intuitively the best sample to estimate the most frequent class in a cluster if only one sample can be taken.

In a bit more detail, K-medoids is performed by assigning each segment to the nearest medoid among all k medoids. The medoids are initialized and iteratively updated to minimize the total distance of all segments to the nearest medoids until no medoid can be swapped to reduce the total distance.

The initialization of medoids is based on farthest-first traversal [20]. Farthest-first traversal has been proved to give an efficient approximation of k-center problem [21]. A traversed set starts as a singleton of a random segment. The farthest segment to the current traversed set (the distance from a point x to a set $\mathcal{S}$ is defined as $d(x, \mathcal{S}) = min_{y\in\mathcal{S}} d(x, y)$) is added to the traversed set until the traversed set reaches the size of K. The traversed set is then used as the initial medoids.

The choice of the number of clusters k gives a trade-off between bigger cluster size (more predicted labels can be derived from a single label assignment) and better accuracy of predicted labels. Let us denote the number of unlistened segments as n. We choose $k = n/4$, which can be interpreted that the average size of clusters is four.

2.4. Assigning labels

The medoids of clusters are presented to an annotator in a sequence sorted by cluster size in descending order. Only one medoid is played at a time and the annotator assign label to the medoid by selecting a class from a list of pre-defined classes. Assigning a label consumes labeling budget by one. The label assigned to a medoid is seen as an *annotated* label. The label of the medoid is derived as *predicted* labels for the rest of cluster members. Largest clusters are labeled first so that high number of predicted labels can be derived with low listening budget.

2.5. Recursive process

Initially, all the segments are flagged as *unlistened*. Once a medoid segment is annotated, the segment is flagged as *listened*. The target situation is small labeling budget so that we do not aim on an optimal performance when the budget is more than the number of clusters. In case all medoids are annotated, we simply perform another round of clustering on unlistened segments and the annotation process continues with medoids in the latest round of clustering. Annotated labels overrule predicted labels received in previous rounds.

If the listening budget is sufficient so that multiple rounds of clustering have been performed, there would be multiple, possibly different predicted labels given to an unlistened segment. In supervised learning, all the different predicted labels for an unlistened segment are used, by taking the segment as an training example of each labeled class.

3. EVALUATION

The performance of the proposed method is evaluated as the classification accuracy using labels produced with the proposed method.

3.1. Dataset

The goal of the proposed method is to save annotation effort. In order to approximate the target situation, the used dataset has to be large enough so that reducing annotation effort is worthy attempting. In addition, a public dataset designed for sound event classification is preferred.

We use UrbanSound8K dataset [22], a public environmental sound dataset, consisting of 10 classes of sound events: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren and street music. All the sounds in the dataset are real field-recordings from urban environments. The dataset includes 8 732 labeled sound segments with maximum duration of 4 seconds, totaling 8.75 hours. A 10-folds division is provided by the dataset for cross validation. The division is made using a random allocation process that keeps segments originating from the same recordings allocated to the same fold, meanwhile trying to balance the number of segments per fold for each sound class.

3.2. Experimental setup

MFCCs are used as frame-wise features. 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. To calculate the segment-to-segment distances, the mean and covariance of MFCCs are used as is discussed in Section 2.2. In supervised learning, the following summary statistics of MFCCs are used as segment-wise features: minimum, maximum, median, mean, variance, skewness, kurtosis and the median and variance of the first and second derivatives.

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. A labeling budget m allows a learning algorithm to query labeling responses for up to m segments. The labels obtained directly through labeling responses are called annotated labels, whereas other labels generated using the proposed method or SSL are called predicted labels.

Two annotators are simulated: an oracle annotator that always answers the ground truth and an artificial weak annotator [23] that produces noisy labels. The labeling accuracy of our artificial weak annotator is set to 75%, which is the lowest reported human sound event recognition rate in found studies [5, 8, 24, 25]. The probabilities that the artificial annotator mislabels a class to any other classes are even.

Obtained labels are used to perform supervised learning. Support vector machine (SVM) with radial basis function as kernel is used as classification model. Since this study does not aim on optimal parametrization, we use default settings of Python Scikit-learn [26]. A training example consists of a segment-wise feature vector and a target class according to the label.

Since the distribution of classes in the dataset is not even, we use unweighted accuracy to weigh different classes the same regardless to the number of instances. The classification accuracy is reported averaging the accuracy across all 10 folds. There are random elements (medoid initialization, random sampling and labeling errors from the weak annotator) in the experiments that affect on the performance. Therefore, all the experiments are repeated five times and the averaged results are reported.

3.3. Reference methods

Random sampling is used as a baseline, where a random subset with the size of labeling budget in the training dataset is annotated. The purpose of random sampling is to simulate the performance of passive learning as a benchmark.

CRTAL [10] is used as the second reference method. Half of the labeling budget is used for the initial samples that are randomly selected. The other half of the labeling budget is used for uncertainty selection. A batch size five is used so that the least confident five samples to the current model, in each iteration, are selected for labeling and the model is updated after adding new labels to training material.

SSL [12] is coupled with random sampling and CRTAL, respectively, as the third and the fourth reference method. The annotated labels are obtained though either random sampling or CRTAL. An initial classifier is trained with annotated labels and all unlabeled segments get predicted labels based on the classification output using the initial classifier. The predicted labels and the classifier are updated with five iterations. This way of combining of CRTAL and SSL is called a serial combined learner [27].

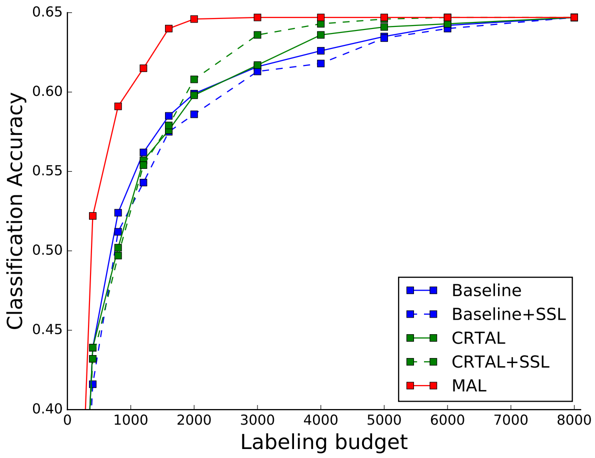

Fig. 2. Classification accuracy as a function of labeling budget, simulated using an oracle annotator.

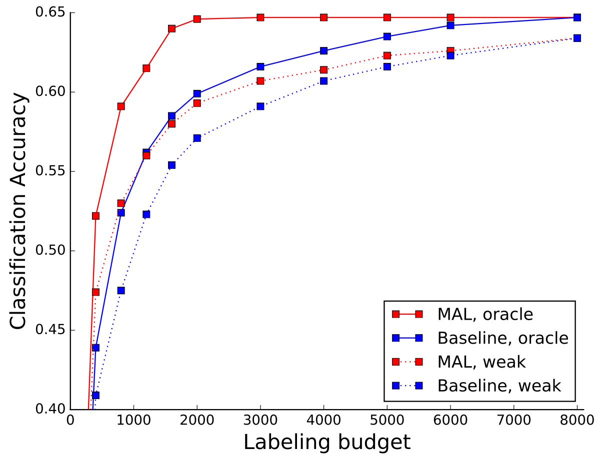

Fig. 3. Classification accuracy as a function of labeling budget, simulated using an oracle annotator (oracle) and an artificial weak annotator (weak).

3.4. Results

Figure 2 illustrates the performance of the proposed method compared with reference methods, simulating oracle labeling responses. 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 65%, which is the ceiling performance of all compared methods.

The proposed method (MAL) performs the best with all simulated labeling budget until all methods converge to the ceiling performance. Reference methods need 2-4 times of labeling budget, compared to the proposed method, to achieve the same accuracy . An interesting benchmark is listening budget 2 000, where each segment has received a label, either annotated or predicted using the proposed method. We have observed that the accuracy of predicted labels is about 97%. The high labeling accuracy makes the resulted classifier approximates the ceiling performance.

CRTAL does not outperform the baseline until labeling budget of 3 000. An active learning study on speech recognition [11] shows a similar trend. When labeling budget is small, the most uncertain segments selected within a batch are often similar to each other, which makes the selected training material more redundant than when using baseline.

The effect of SSL goes divergent along with baseline and CRTAL. The performance is improved when SSL is used together with CRTAL, but similar improvement is not observed with the baseline. Uncertain segments are labeled out with CRTAL, and there remains confident segments to predict. As a result, the label prediction accuracy is much higher when CRTAL is used compared to baseline.

Figure 3 illustrates the difference in performance between the resulted classifiers using the oracle annotator and the artificial weak annotator. The results show that the proposed method also outperforms the baseline when the weak annotator is used. However, the advantage of the proposed method is smaller compared to using the oracle annotator: the baseline needs less than double sized labeling budget to achieve the same accuracy. Intuitively, this phenomenon is due to the predicted label derivation mechanism of the proposed method. Mislabeling the medoid makes a whole cluster of segments wrongly labeled to another class, which may lead to a strong confusion between the two classes. In comparison, when the same amount of wrong labels are evenly distributed to all classes, the performance of the resulted classifier seems to be affected much less. As a summary, the proposed method might be less effective when using weak annotators.

4. CONCLUSION

We propose a novel method, medoid-based active learning (MAL), to improve sound event classification performance when labeling budget is small, compared to the number of unlabeled data.

In the evaluation using an oracle annotator, when the labeling budget was less than 10% of unlabeled data, the resulted classifier using the proposed method gave about 8% better accuracy than using the best reference method. Furthermore, as the listening budget grew, the proposed method kept to outperform reference methods. Through all the experiments, the proposed method used generally 50%-60% less labeling budget to achieve the same classification accuracy with respect to the best reference method.

In this study, the number of clusters k was set to a rather big number (only four segments per cluster in average). However, the performance of the proposed method could be potentially further improved by tuning k according to the listening budget, e.g. a smaller k for a tight budget. In preliminary experiments, the classification accuracy with a tight listening budget (400) was further improved by 5% when k was halved.

The experiment using an artificial weak annotator shows that the proposed method is less effective if the annotator gives too many wrong labels. This suggests a future study about using weak annotators. In case of very weak annotator, clustering may be used to improve the labeling accuracy (listening to all segments in a cluster and label the whole cluster using majority vote) instead of active learning, which leads to another study.

As a conclusion, the proposed method can effectively improve the sound event classification performance when the labeling budget is small. In future, datasets with different number of segments and possible classes can be studied. Furthermore, it would be helpful to evaluate the performance using realistic human annotators. At last, it would be useful to study alternative acoustic models, e.g. neural network, to compare how they work along with less accurate labels.

5. REFERENCES

[1] Karol J. Piczak, "Environmental sound classification with convolutional neural networks," in *25th IEEE International Workshop on Machine Learning for Signal Processing, (MLSP)*, 2015, pp. 1–6.

[2] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6440–6444.

[3] Antonio J. Torija and Diego P. Ruiz, "Automated classification of urban locations for environmental noise impact assessment on the basis of road-traffic content," *Expert System with Applications*, vol. 53, pp. 1–13, 2016.

[4] Buket Barkana and Burak Uzkent, "Environmental Noise Classifier Using a New Set of Feature Parameters Based on Pitch Range," *Applied Acoustics*, vol. 72, no. 11, pp. 841–848, Nov. 2011.

[5] Paul Gaunard, Corine Ginette Mubikangiey, Christophe Couvreur, and Vincent Fontaine, "Automatic classification of environmental noise events by hidden markov models," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1998, pp. 3609–3612.

[6] 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.

[7] Sébastien Lecomte, Régis Lengellé, Cédric Richard, François Capman, and Bertrand Ravera, "Abnormal events detection using unsupervised one-class svm - application to audio surveillance and evaluation.," in *International Conference on Advanced Video and Signal-Based Surveillance*, 2011, pp. 124–129.

[8] 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. 1–12, 2016.

[9] David Cohn, Les Atlas, and Richard Ladner, "Improving generalization with active learning," *Machince Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[10] Dilek Z. Hakkani-Tür, Giuseppe Riccardi, and Allen L. Gorin, "Active learning for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. 3904–3907.

[11] 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.

[12] Pedro J. Moreno and Shivani Agarwal, "An experimental study of em-based algorithms for semi-supervised learning in audio classification," in *2003 International Conference on Machine Learning (ICML) Workshop on the Continuum from Labeled to Unlabeled Data*, 2003.

[13] Aleksandr Diment, Toni Heittola, and Tuomas Virtanen, "Semi-supervised learning for musical instrument recognition," in *21st European Signal Processing Conference (EUSIPCO)*, 2013, pp. 1–5.

[14] Zhao Xu, Kai Yu, Volker Tresp, Xiaowei Xu, and Jizhi Wang, "Representative sampling for text classification using support vector machines," in *European Conference on Information Retrieval*, 2003, pp. 393–407.

[15] Michael I. Mandel and Dan Ellis, "Song-level features and support vector machines for music classification," in *6th International Conference on Music Information Retrieval*, 2005, pp. 594–599.

[16] Toni Heittola, Annamaria Mesaros, Dani Korpi, Antti J. Eronen, and Tuomas Virtanen, "Method for creating location-specific audio textures," *EURASIP Journal of Audio, Speech and Music Processing*, vol. 2014, no. 9, 2014.

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

[18] 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.

[19] Tagaram S. Madhulatha, "Comparison between k-means and k-medoids clustering algorithms," in *1st International Conference in Advances in Computing and Information Technology (ACITY)*, 2011, pp. 472–481.

[20] 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.

[21] 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.

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

[23] Shankar Vembu and Sandra Zilles, "Interactive learning from multiple noisy labels," in *European Conference in Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, 2016, pp. 493–508.

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

[25] Selina Chu, Shrikanth S. Narayanan, and C.-C. Jay Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Transaction on Audio, Speech & Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] Malte Darnstädt, Hendrik Meutzner, and Dorothea Kolossa, "Reducing the cost of breaking audio captchas by active and semi-supervised learning," in *13th International Conference on Machine Learning and Applications*, 2014, pp. 67–73.