Initial training samples selection using clustering algorithm for audio recognition

Rongyan Wang, Yan Leng

1 School of Information Management, DeZhou University, Daxue Road 566, 253000, Dezhou, China
2 School of Physics and Electronics, Shandong Normal University, East Wenhua Road 88, 250014, Jinan, China

Abstract. The quantity and quality of training samples are one of the most critical factors for the performance of the classifier, whatever traditional classification method or deep learning algorithm. To improve the quality of the training samples with less manual tagging effort, this paper proposed a representative samples selection method based on clustering algorithm to choose more representative samples supplied to human to annotate. These representative ones can describe class distribution better than samples annotated at random. On the one hand, these selected samples can train a more generalized classifier. On the other hand, this samples selection method can save expensive annotation work which is thought a lot by people now. The results of experiments in melodrama Friends corpus show better classification performance compared to traditional annotation methods.

1. Introduction

Audio recognition has been proven essential to fields such as human perception[1][2], with recent applications in surveillance[3][4][5] and multimedia retrieval[6][7][8]. There have been many studies on automatic audio recognition referring to many classical methods in pattern recognition such as supervised learning[6][7], semi-supervised learning[8] and unsupervised learning[9]. During the recent years, deep learning technology has also been applied to the field of audio classification and recognition. In 2009, Dr Andrew Ng of Stanford University and Dr. Honglak Lee of the University of Michigan introduced the restricted Boltzmann machine into the audio classification task in document[10] and good performance was achieved in several classification tasks. After that, other depth learning algorithms such as convolutional neural networks, was also introduced into the field of audio classification[11][12]. However, all of these methods need labeled or unlabeled samples as training set to learn recognition model. Therefore, the quantity and quality of training samples naturally become one of the most important factors that affect the performance of classifier. Traditionally, we get training samples by annotating audio stream in sequence. In fact, the samples annotated in sequence may not be the most useful one and samples annotation is often difficult, expensive, and time consuming. Consequently, the selection of initial instances which should be annotated is a very important question.

For this reason, active learning has been the research hotspot in pattern recognition and machine learning. Generally speaking, active learning needs some annotated samples which used to select more
useful unlabeled samples to annotate. Then all these samples are used to train better recognizer. However, the quantity of initial samples infects the performance of the useful samples learning. In the same way, the supervised learning performance is infected by the quantity of the training instances annotated at the beginning.

Consequently, the selection of initial instances which should be annotated is an important question. However, very little work has been done in this problem. Traditionally, we annotate the instances of audio stream in sequence but these samples may not be the most useful ones.

Inspired by the idea of active learning, if we select the more representative samples feed back to people for annotating at the beginning, the model trained by these samples may have more recognize ability than traditional annotation method. As a result, we propose a method to select the most representative samples. The idea refers clustering algorithm which cluster all unlabeled samples to different clusters in this paper. Then samples are selected from these clusters according to some rules described in Sec.3. These samples selected are the representative ones which can describe the rough distribution of samples. The experiments results in Sec.4 show encouraging results of audio events recognizer regarding the selection method of samples which should be annotated at first.

2. Audio recognition

Regarding audio recognition, our primary objective is to identify individual sound events in audio stream. In this paper, events are often defined as structurally meaningful units such as speech event or laugh event. A “speech event” would then consist of the continuous speech between pauses or transitions where the speaker changes—additional processing would be necessary to identify phonemes, words, or sentences. In our example, we just think about speech events as female speech and male speech.

The first step in audio recognition problem is to extract features from wav files which are used for classification. The choice of feature set play key role in classification performance. Our own past work[13] has monitored discrimination of different events in six features (short-time energy, zero crossing rate, 8 dimensional Mel Frequency Cepstral Coefficients(MFCC) and its first time derivatives, sub-band spectral flux, sub-band energy ratio, brightness, bandwidth, harmonicity prominence, high zero crossing ratio, low energy ratio and spectrum flux) that are specifically adapted to the sounds in Friends.

Classification model is needed for audio event recognition after features are extracted. There are two classical models for classification in pattern recognition. One is Generative probability models such as Hidden Markov model (HMM) and Gaussian Mixture Models (GMM). This kind of models provide a principled way of treating missing information and make strong assumptions about the data. The other one is discriminative methods such as Support Vector Machines (SVM) which enable us to construct flexible decision boundaries. During the last years, SVM has become extremely successful discriminative approaches to pattern classification.

SVM is the model that offers a discriminative solution to classification problems with strong bounds on error minimization. In SVM training, kernel tricks are used to map non-linearly feature vectors space into a high-dimension space and a hyperplane is then searched in the new feature space to separate the data points of the classes with a maximum margin. In this paper, we use SVM as multi-class classification.

3. Criterion for Representative samples selection

In this section, firstly, we give the formal definition of representative audio samples in the condition we have no labeled examples at the beginning. The representative audio samples are selected for annotation. They are used as the training set for supervised learning. Secondly, we present the process of representative audio samples selection which consists of two sub-phases: feature space clustering and samples selection from the clusters.
3.1. Representative audio samples
When the training set is not big enough, the training performance is very sensitive to the effectiveness
of each training example. That is, the statistical characteristics of the labeled audio stream archives
will highly affect the performance of the audio recognition system.

On one hand, if these audio samples are too similar to each other, there will be too much
redundancy which decreases the information capacity; on the other hand, if there are little consistency
among them, the learning algorithm will encounter great difficulty in training a reasonable classifier.
In order to make the training more smoothly, we should provide some representative samples for user
labeling, which should have the following three properties.

• The audio samples should have consistency. Here, the consistency means that these samples
should have similar behavior in training the classifier.
• The audio samples should not contain too much redundancy.
• The audio samples should refer every class to recognition.

To guarantee the first property, we should select representative samples from a sub audio set with
consistent characteristics instead of the whole database. That is, these samples contain a subset of
audios, which are most likely relevant to each other.

To guarantee the second property, we detect audio elements that are most representative for the
semantic content, that is, the key audio elements are selected from the same audio set, and they are not
very similar to each other.

To guarantee the third property, we detect audio elements from different sets which include every
class.

3.2. Representative samples selection method based on clustering
Before human annotate audio samples, there are no supervised information to be used. For training
samples should cover the three properties referred in Sec.3.1, we utilize the advantage of clustering
learning methods to save human labor. That is to say, temporal signal segments with similar low-level
features are grouped into natural clusters. We call these signal segments as audio elements. Audio
elements in the same cluster are similar with each other. We choose audio elements far away from
each other to eliminate redundancy in some extent in the same cluster. These samples are
representative ones. Because the boundary between the clusters lies in low-density region which may
be the boundary of classes, we get representative samples covering different classes by selecting
samples from different clusters. These representative samples are provided to people to annotate.

Regarding clustering method, we use K-Means clustering algorithm which is one of the important
cluster analysis methods of data mining. K-means clustering algorithm is one of the simplest
unsupervised learning algorithms that solve the well known clustering problem. The procedure follows
a simple and easy way to classify a given data set through a certain number of clusters (assume k
clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. The next step is to
take each point belonging to a given data set and associate it to the nearest centroid. When no point is
pending, the first circulation is completed. At this point we need to re-calculate k new centroids as
centers of the clusters resulting from the previous step. After we have these k new centroids, a new
binding has to be done between the same data set points and the nearest centroid. Finally, this
algorithm aims at minimizing an objective junction, in this case a squared error function. The
objective function:

\[ J = \sum_{j=1}^{k} \sum_{i=1}^{n} || x_i^{(j)} - c_j ||^2 \]  

(1)

Where \( || x_i - c_j ||^2 \) is a chosen distance measure between a data point \( x_i(j) \) and the cluster centre \( c_j \) is
an indicator of the distance of the n data points from their representative cluster centers.
After samples are clustered, representative ones are selected based on different strategies as following. That means select samples from different region in sample distribution.

- Random selection: annotate samples in sequence without selection.
- Clustering based selection: select samples which are far away from each other in the same cluster. We realize this though compute the distance between samples and center. Representative samples are those which have some certain distances from center. The other clusters are accounted for the same way.

Different representative sample sets are selected based on the above two strategies. In section 4 experiments are done based on the different selection strategy.

4. Experiments
The following experiments aim to evaluate whether representative samples selected by clustering can make recognizer more generalized than traditional method.

4.1. database
In our experiments, typical audio events are from melodrama Friends. We select audio stream of ten episodes from season one to construct the database and annotate them with semantic concept including laugh, music, female sounds and male sounds. In fact, the audio includes many overlapped parts. For example, there are laugh sound and speech in the same audio. In our experiments, we do not consider these overlapped parts. Besides that, lots of speech sounds are filtered because most of the audio are speech in these melodramas and too much speech vitiates the effect of clustering.

The format of audio is wav, 16 kHz in pulse code modulation (PCM) format with a single audio channel. As mentioned in section 3.1, before classification, the audio features are extracted first. Audio signal is divided into frames. Given the sampling frequency of 16 kHz, the frames are of 400 samples (25 ms) each, with 50% (200 samples or 12.5 ms) overlapped in each of the two adjacent frames. The segment features contain short-time energy, zero crossing rate, 8 dimensional Mel Frequency Cepstral Coefficients and its first time derivatives, sub-band spectral flux, sub-band energy ratio, brightness, bandwidth, harmonicity prominence, high zero crossing ratio, low energy ratio and spectrum flux. Then audio segment features are extracted from each non-silent frame. The segment length is 1s which is always long enough to imply semantic meaning. The segment features are gotten through computing the means and standard deviations of the above original frame features over every audio segment. The adjacent segments are half overlapping. Thus all features are formed to 81-dimensional feature vector. In order to avoid data imbalance, we choose about 1400s audio samples as training set for each kind of audio events.

To evaluate the classification performance, accuracy is used. They are defined as,

$$\text{Accuracy} = \frac{\text{number of correct output AEs of system}}{\text{number of all output AEs of system}}$$  \hspace{1cm} (2)

AEs is audio events.

4.2. Experiments

4.2.1. Clustering analysis
This experiment aims to evaluate the purity of cluster by average precision and average recall of the largest class which covers the most samples in the cluster. They are defined as followings:

$$\text{precision} = \frac{\text{number of samples account for the largest part of the a cluster}}{\text{all samples in the clusters}}$$  \hspace{1cm} (3)

$$\text{recall} = \frac{\text{number of samples account for the largest part of the a cluster}}{\text{number of corresponding samples in all clusters}}$$  \hspace{1cm} (4)

Average precision and average recall is the mean value of all k clusters.
All the samples of 3200 samples of four classes are clustered into k clusters. We take different values of k to evaluate its influence to purity of clusters. The results are shown in Figure 1.

As shown in Figure 1, the average recall is decreasing when the clusters increasing just as we thought of. That is not difficult to understand. The samples included in one cluster become fewer when cluster number increase. That indicates fewer samples of some class are included in one cluster while samples of corresponding class outside this cluster become more. So the average recall becomes lower with clusters becoming more. In the same time, the change of average precision is not obviously when the samples are clustered more than 3 clusters. That implicates the purities of cluster are not high even though the cluster become smaller when clusters become more. That is mainly because the boundary surface between clusters can not respect the boundary between classes which lead to a cluster including many samples from each class.

4.2.2. Representative samples selection criterion

This experiment aims to determine whether the representative samples selection method is helpful for the improving of classification performance.

When comparing random annotating method and clustering based representative samples selection method on the same unlabeled set, we take samples of each class annotated in sequence as many instances sampled by clustering based method. The number of each class is the same. In fact, if we want to get the same number samples of each class in sequence as clustering based sampling method, more annotation work should be done, because the speech event samples usually appear more than
laugh event and music event samples. In the Figure 2, we take the same number samples of both methods, and samples number of each class is equal using random method. We can see from figure 2 that the precision of classifier based on clustering method is higher than sequential annotated method. That is to say the clustering based method can make the classifier have better generalization capability. In the same time, this method can save human annotation work.

It is worth noting that the precision of classifier varied steady when the number of clusters increases. That suggests the performance of classifier trained by the instances sampling based on clustering method is not sensitive to clustering number. We can use this method conveniently and need not consider how many clusters should be clustered to.

5. Conclusion

In this paper, we exploit a representative samples selection method using method clustering to audio events recognition. These samples are selected and given to human for annotation when there are no annotated samples at the beginning. The clustering method can make the audio segments with similar low-level features group into same cluster. In the same time, dissimilar segments, which maybe belong to different classes, are grouped into different sets. The samples are chosen from different clusters according to some rules. That is, they come from the same cluster to guaranty the similarity. Besides that, they have some distance from each other to eliminate the redundancy. And we choose from every cluster to cover all the classes. So, samples selected through this method usually more representative to the class distribution than samples lie in sequence. Then these selected samples are returned to people for annotation while samples are annotated in sequence of traditional method. And these representative samples annotated can make recognizer more efficient than ones trained using traditional annotation method.

Acknowledgments

This work is supported by Shandong Province Natural Science Foundation ,China(No.ZR2016FB25).

References

[1] A. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: MIT Press, 1990.
[2] C Micheyl, B Tian, RP Carlyon, AJP Rauschecker, Perceptual Organization of Tone Sequences in the Auditory Cortex of Awake Macaques, Neuron, 2005, 48 (1):139
[3] Tuomas Virtanen,Computational analysis of acoustic events in everyday environments. Journal of the Acoustical Society of America. 141(5):3451-3451 · May 2017
[4] H Lozano, I Hernáez, A Picón, J Camarena, E Navas, Audio Classification Techniques in Home Environments for Elderly/Dependant People, Springer Berlin Heidelberg, 2010, 6179:320-323
[5] RM Alsina-Pagès, J Navarro, F Alías, M Hervás, homeSound: Real-Time Audio Event Detection Based on High Performance Computing for Behaviour and Surveillance Remote Monitoring:Sensors, 2017, 17 (4)
[6] E. Wold, T. Blum, D. Keislar, and J. Wheaton, “Content-based classification, search and retrieval of audio,” IEEE Multimedia Mag., vol. 3, pp. 27–36, July 1996.
[7] G Guo, SZ Li,Content-based audio classification and retrieval by support vector machines, IEEE Transactions on Neural Networks, 2003 , 14 (1):209
[8] Lie Lu, Member, and Alan Hanjalic, Member, Audio Keywords Discovery for Text-Like Audio Content Analysis and Retrieval, IEEE Transactions on Multimedia, VOL. 10, NO. 1, JANUARY 2008
[9] Honglak Lee Yan Largman Peter Pham Andrew Y. Ng, Unsupervised feature learning for audio classification using convolutional deep belief networks , Advances in Neural Information Processing Systems 22 (NIPS 2009), pp:1096-1104.
[10] M. Ravanelli ; B. Elizalde ; K. Ni ; G. Friedland, Audio concept classification with Hierarchical Deep Neural Networks, Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European, Page(s):606 – 610.

[11] K. J. Piczak, Environmental sound classification with convolutional neural networks, Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on, Page(s):1 – 6.

[12] Lars Hertel, Huy Phan, Alfred Mertins, Comparing Time and Frequency Domain for Audio Event Recognition Using Deep Learning, Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), July 2016.

[13] Rongyan Wang; Gang Liu; Jun Guo; Zhenxin Ma; Semi-supervised learning for automatic audio events annotation using TSVM , 10.1109/ICCASM.2010.5620310, Page(s): V4-530 - V4-534,2010