Automatic Content based Classification of Speech Audio using Multiple Instance Learning

Vivek P, Lajish V L

Abstract*: Audio content understanding is an active research problem in the area of speech analytics. A novel approach for content-based news audio classification using Multiple Instance Learning (MIL) approach is introduced in this paper. Content-based analysis provides useful information for audio classification as well as segmentation. A key step taken in this direction is to propose a classifier that can predict the category of the input audio sample. There are two types of features used for audio content detection, namely, Perceptual Linear Prediction (PLP) coefficients and Mel-Frequency Cepstral Coefficients (MFCC). Two MIL techniques viz. mi-Graph and mi-SVM are used for classification purpose. The results obtained using these methods are evaluated using different performance matrices. From the experimental results, it is marked that the MIL demonstrates excellent audio classification capability.

***Keywords* : Audio classification, Multiple Instance Learning (MIL); Feature extraction; mi-Graph; mi-SVM.**

I. INTRODUCTION

As audio forms a major portion of information disseminated in the world every day many researchers are attempting to classify it based on various criteria [1]. In today's digital world people have access to large amount of news audio and video; on radio, television and the internet. The amount of multimedia data that are available is now so immense that it is infeasible for a human to go through it all and distinguish required files among them. Automatic content based analysis provides useful information for audio classification as well as segmentation. Information about the audio content can be used for classifying the file. Audio content understanding is thus an active research problem in speech analytics. A key step in this direction is a classifier that can predict the category of the input audio. In this paper, a novel audio classification technique is proposed based on Multiple Instance Learning (MIL).

Different supervised and unsupervised learning techniques are available in Machine Learning (ML) approaches. Multiple Instance Learning (MIL) is proposed as a variant of supervised methods for problems with patrial information about labels of training examples. Melih Kandeir et al., [2] conducted a benchmark study over the performance of different MIL methods. In their study, it is reported that mi-Graph and mi-SVM gave considerably better result compared to other MIL methods.

Revised Manuscript Received on February 10, 2020.

Vivek P., Department of computer science, University of Calicut, Kerala, India, Email: vivek_dcs@uoc.ac.in

Lajish V. L., Department of computer science, University of Calicut, Kerala, India, Email: lajish@uoc.ac.in

For the purpose of multimedia classification, features are drawn mainly from the text, audio, and visual modalities. Usually, multimedia approaches are found more often in the literature than text-only approaches. The audio content based approach usually requires fewer computational resources than visual methods [3]. There are different features which provide a compact representation of the given audio signal. Among them, Perceptual Linear Prediction (PLP) coefficients and Mel-Frequency Cepstral Coefficients (MFCC) are commonly used features [4]. Vivek P et al., proposed a news video classification method in which violent incident videos are classified from news video archives using MIL algorithm [5].

In this work, a novel approach for content based speech audio classification using MIL methods is proposed. The experiments are conducted over the own developed news audio database. The results obtained using these methods are evaluated using different performance metrics. The rest of this chapter is organized as follows. Section III describes feature extraction methods used in this study. Section IV describes the proposed news audio classification methodology. Section V discusses the experimental results, and section VI concludes the work

II. REVIEW ON AUDIO CLASSIFICATION USING MULTIPLE INSTANCE LEARNING

The Automatic audio analytics and classification is an emerging research area in multimedia stream. Erling Wold *et al.,* in 1996 introduced an engine for content-based classification, search, and retrieval of audio data [6]. The proposed method lets the sounds to be classified by their audio content. Queries can be based on any one of the acoustical feature or a combination of more than one features. The experiment is conducted either by specifying formerly learned classes based on these features, or by choosing reference sounds and requesting the engine to retrieve sounds that are matching or mismatching to them. A system to retrieve audio documents by acoustic similarity is introduced by Jonathan T. Foote in 1997 [7]. The similarity measure used in this work is based on statistics derived from a supervised vector quantizer.

In the early days, classification experiments were conducted on simple cases such as speech-music classification, speech-silent classification *etc*. Pfeiffer *et al.,* proposed a theoretic framework and various applications of automatic audio content analysis using certain perceptual features [8]. In a work, D. Kimber *et al.,* classified audio recordings into speech, silence, laughter, and non-speech sounds, to segment discussion recordings in meetings from other sounds [9].

Zhang and Kuo introduced a method to classify audio recordings into various categories such as songs and speeches over music, based on a heuristic-based model [10].

Audio-based approaches are also found more in video classification literature rather than the texts and video based approach. Advantages of audio approaches includes the usage of less computational resources than visual approaches and more reliable than text. In the earlier works time domain features were used widely and later, researches started using combined features from both time and frequency domains for better recognition accuracy [11, 12]. Among these features MFCC is identified as the most used and trustable one [13].

Liu *et al.*, in 1998 considered the problem of discriminating five types of news into commercial, basketball game, football game, news report, and weather forecast [14]. They have designed an ergodic HMM using the clip based features as observation vectors. A filter predictor for audio event classification and extraction is introduced in 2017 by Visser *et al.* [15]. Deep Neural Networks (DNN) is widely used for extracting required target audio files from unlabeled training data.

The research on MIL problems initially started for the task of digit recognition, here a neural network was trained with the information on the presence of a given digit without specifying its position [16]. Another early application of MIL was to discover the drug in which the bags were molecules of the drug and the instances were conformations of those molecules [17]. MIL has also been applied for detecting objects in images [18], video classification to match names and faces [19], and to text classification [20], in which documents are considered as bags and sentences as instances. Many approaches have been introduced for MIL, including mi-Graph, Gaussian Process Multiple Instance Learning (GPMIL), MILBoost, mi-SVM and Bag key instance SVM (B-KI-SVM) [21]. The use of weakly supervised machine learning technique can reduce the computational cost in a large manner. Till now no significant work has been proposed on the use of Multiple Instance Learning (MIL) approach on speech audio classification.

III. FEATURE EXTRACTION FROM NEWS AUDIOS FOR CLASSIFICATION

In this section audio feature extraction for MIL classifier is discussed in detail. To classify the news audio input, as a pre-processing the audio part is extracted and is split into overlapping segments. The signals are divided into constant-time segment of 25ms blocks [22]. Speech signal analysis usually considers speech as a non-stationary and exhibit quasi-stationary behavior in short durations [23]. It is generally performed over short-time frames with a fixed frame length (FFL) and a fixed frame rate (FFR). This method is simple to implementation due to the ease of comparing blocks of the same length.

Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) coefficients are extracted as features and further used for classification purpose. The algorithms used for MFCC and PLP based feature extraction techniques are described below.

A. Mel Frequency Cepstral Coefficient (MFCC) Feature

Mel Frequency Cepstral Coefficients called MFCC is one of the most popular spectral based feature used in speech recognition problems. MFCC considers human perception sensitivity with respect to frequencies for better audio recognition [24]. The procedure to determine MFCC is described in the following algorithm.

Algorithm

Step 1: Segmentation of voiced speech signal into 25 ms-length frames.

Step 2: Calculate the periodogram estimate of the power spectrum for each frame.

Step 3: Apply the mel filterbank to the power spectra and take the sum of energy in each filter.

Step 4: Calculate the logarithm value for all filterbank energies.

Step 5: Find the DCT of the filterbank energies (log).

Step 6: Find the MFCC as the amplitudes of the subsequent spectrum.

The mel-scale frequency mapping is formulated as:

$$m(f) = 1125\ (1 + \frac{f}{700}) \quad (1)$$

B. Perceptual Linear Prediction (PLP) Feature

The PLP model, proposed by Hermansky [25], is based on the concept of psychophysics of human hearing. PLP throw-outs irrelevant information in the speech and makes significant improvement in speech recognition rate. The procedure to determine PLP coefficients are described as follows:

Algorithm

Step 1: The N- point DFT is applied on the segmented input signal *x(n).*

Step 2: The power spectrum is convolved with the piece-wise approximation of the critical-band curve to calculate the critical-band power spectrum.

Step 3: Equal loudness pre-emphasis is applied on the down-sampled *θ(B)* and then intensity-loudness compression is performed.

Step 4: Inverse DFT is performed for getting the equivalent autocorrelation function.

Step 5: Autoregressive modelling followed by conversion of the autoregressive coefficients into cepstral coefficients to compute PLP coefficients

IV. CONTENT BASED AUDIO CLASSIFICATION USING MIL

This work aims for the automatic classification of news audios by categorizing it based on the content. This categorization will reduce the search cost of analytic applications. Given a list of news audios of interest, the proposed method will produce a discriminative model to distinguish them. In the following sections, implementation details of the MIL approach have been discussed along with a description on the mi-Graph and mi-SVM methods which have been used for classification are explained in detail.

A. Preparation of Malayalam News Audio Corpus (MNAC)

Initially, news text corpus is generated from the online news portals of the popular Malayalam dailies.

These news text sentences are then classified into five categories which includes state news, national news, international news, sports news and news related to cultural importance. The first three categories viz. state news, national news and international news represent news related to the current affairs excluding sports and culture categories. A total of 5250, news text Sentences of various such categories are collected for dataset creation.

An own developed Malayalam News Audio Corpus (MNAC) consists of news audio samples spoken by 35 speakers (both male and female) from different age group are then created. Each speaker uttered 150 sentences taken randomly from any of the five news categories of the text corpus (MNSTC) created for this purpose. All these 5,250 spoken sentences are labelled with the News category id, sample id, speaker id and gender/age of the speaker that they belongs to. The average length of the news audio samples present in the dataset is 5.35 seconds.

B. MIL for News Audio Classification

Initially, the input news audio files taken from the MNAC audio corpus are split into 25 ms length overlapping segments for feature extraction. Instances are created from each audio segments by extracting features from it. Group of features (also called as instances) extracted from the same news files are categorized into a single bag. Bags and instances are labelled properly. The bags are labelled in such a way that, the bag label is the maximum of the instance labels inside the bag. Then, these bags along with their corresponding labels are provide for the MIL classifier. If there is at least one positively labelled instance inside a bag then it is labelled as positive and for a negatively labelled bag, all instances are identified as negative labels. Thus, as shown in figure 1, interested newsgroups are represented by the positive bag and other news sets by the negative bag. The schematic diagram of the proposed MIL based audio classification methodology is shown in figure 2.

The variation of MIL with supervised learning is that it usually deals with problems with partial knowledge about the labels of training examples. MIL is a binary problem as, a bag is labelled negative if all instances in the bag is negative and as positive if any one instance is positive. That is the MIL training set consists of bags $\{X_1, X_2, \ldots X_n\}$ and bag labels $\{y_1, y_2, \ldots y_n\}$, where $X_i = \{x_{i1}, x_{i2} \ldots, x_{im}\}$, $x_{ij} \in$ X and $y_i \in$ {-1, 1}. The goal of MIL is to either train an instance classifier h(X): X → Y or a bag classifier H(X): X^m → Y.

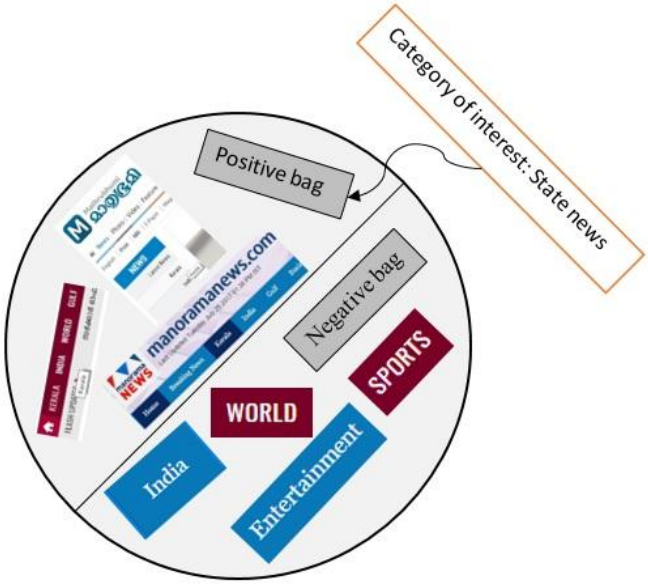


Fig. 1.MIL approach for news audio classification considering state news as the area of interest

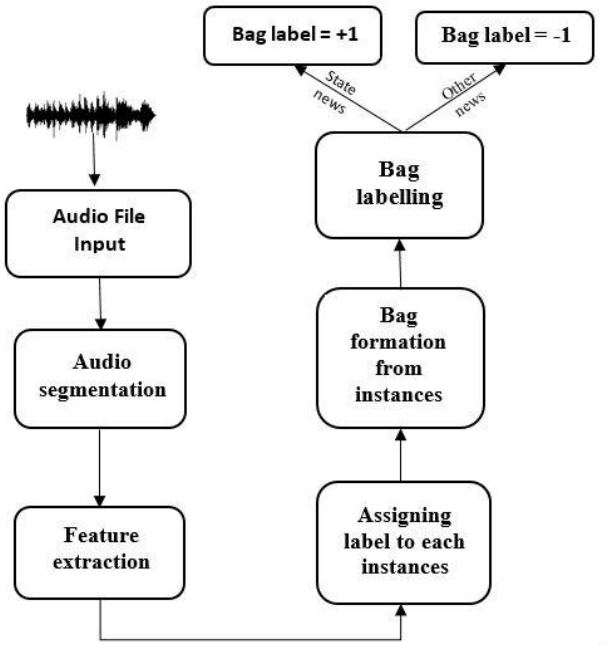


Fig. 2.Schematic diagram of proposed news audio classification methodology

The brief description of two MIL based classification methods viz. mi-Graph and mi-SVM, used in this study are given in the following subsections.

1) *mi-Graph:* mi-Graph is an efficient and simple MIL method which represents each bag by a similarity graph [26]. In this method cross-similarities of bag instances are computed by an instance-level kernel function $K_{inst}(x_i, x_j)$. A graph structure is then created accordingly.

Each instances are represented as nodes and node pairs are connected only if their exist a similarity between them over a threshold δ. Let W_b be the affinity matrix of bag b, whose entry is w^b_{nm}= 1, if there is an edge between the nodes of instances n and m, and w^b_{nm}= 0 otherwise. Accordingly, similarity among two bags b and c are calculated by the following kernel function:

$$K_{bag}(X_b, X_c) = \frac{\sum_{n=1}^{N_b}\sum_{m=1}^{N_c} v_{bn} v_{cm} k_{inst}(X_{bn}, X_{cm})}{\sum_{n=1}^{N_b}\sum_{m=1}^{N_c} v_{cm}} \quad (2)$$

Where, $v_{bn} = 1/\sum_{u=1}^{N_b} W^b_{nu}$, $v_{bn} = 1/\sum_{u=1}^{N_c} W^c_{mu}$ are the sum of the weights of the edges incident to node n of bag b and m of bags c. Training of an arbitrary kernel learner is then performed for computing bag-level Gram matrix. The importance of these kernel value is that bags having larger number of similar instances has smaller value and instances differ has larger value. Thus the influence of odd instances within bags are increased, and others are reduced.

2) mi-SVM: This method is a variant of semi-supervised learning problem considering the positive bag instances as latent variables [27]. These latent variables are served to the optimization problem and inferred from data.

$$\min_{y} \min_{w,b,\xi} \frac{1}{2}\|W^2\| + C\sum_{i=1}^{N}\xi_b, \quad (3)$$
$$\text{s.t } y_i\left(W^T\phi(X_i)\right) \geq 1 - \xi_i, \forall i,$$
$$\xi_i \geq 0, \forall i,$$
$$\max(y_b) = Y_b, \forall b.$$

where w is the vector of model parameters, C is the regularization constant, ξ_b are slack variables, and $\phi(.)$ is a function that maps an instance from the original feature space to a Reproducing Kernel Hilbert Space (RKHS) [33]. During every iteration, the estimated solution is found as follows: trains an instance-level standard SVM based on the current assignments of the latent variables, then update these variables by making predictions with the learned SVM.

V. SIMULATION EXPERIMENTS AND RESULTS

The evaluation of the proposed MIL based audio classification is performed on MNAC news audio archive. The MFCC and PLP features and two MIL techniques viz. mi-Graph and mi-SVM have been used for the experiments. The experiment is conducted over the resultant audio samples obtained after the conduct of keyword spotting experiments. The evaluation of the proposed method is conducted by considering the news audio samples present in the dataset as two classes *viz*. state news audio and non-state news audio. Similarly, non-state news can be further categorized into different binary classes like national and non-national, sports and non-sports as well as news with cultural and non-cultural importance. The block diagram of the evaluation model for the preposed MIL based news audio classification is shown in figure 3. The keyword spotted audio files are given as an input to the MIL classifier. The MIL classifier classifies the audio file into either positive bag or negative bag. The file is considered as state news if it is labelled as positive. The details of the performance evaluations are discussed in this section.

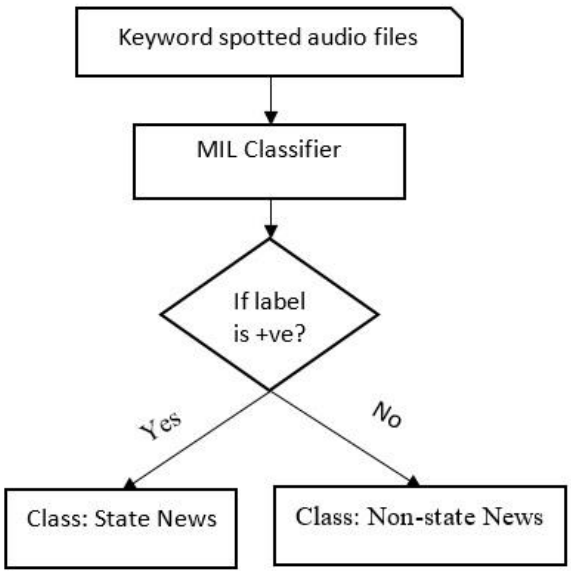


Fig. 3.Evaluation model for the MIL based news audio classifier

As the first stage the audio signals are segmented into 25 ms frames. Frames are considered as the instances of the audio signal. MFCC and PLP features have been extracted from each frame. Following four performance metrics are used for audio classification evaluation of the proposed MIL classifier.

Accuracy: measurement (%) of how close a result comes to the true value.

F1 score: Function of precision and recall.

AUC-ROC: Area under Receiver Operating Characteristics (ROC) curve.

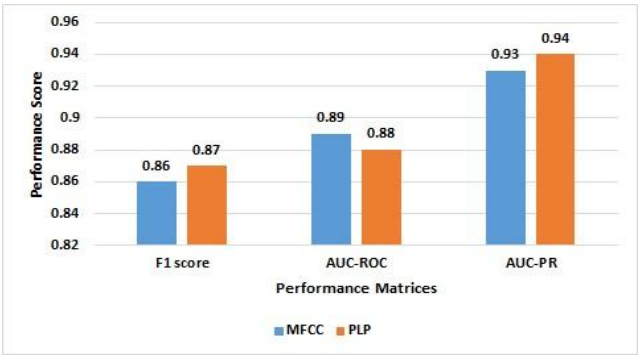
AUC-PR: Area under precision–recall curve.

The news audio classification experiments are conducted using MFCC and PLP features separately based on two different MIL techniques *viz*. mi-Graph and mi-SVM. The news audio classification results are performance matrices obtained by taking the state news as positive bags is given in table I.

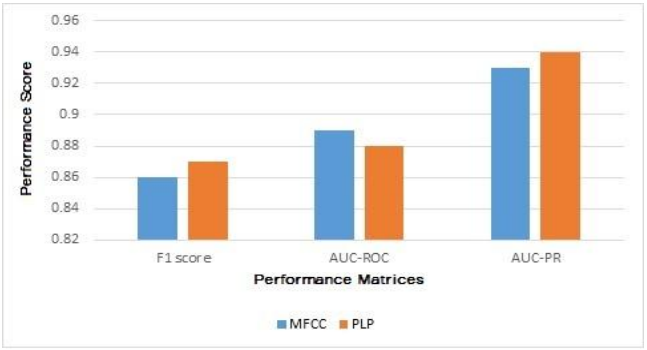
Table- I: MIL based news audio classification results and performance matrices

MIL method	Feature	Accuracy (%)	F1 score	AUC-ROC	AUC-PR
mi-Graph	MFCC	95.8	0.96	0.98	0.99
	PLP	93.2	0.93	0.97	0.97
mi-svm	MFCC	85.0	0.86	0.94	0.96
	PLP	80.3	0.86	0.89	0.93

From the experimental result it is evident that the MIL classification method works effectively in speech audio classification. It is also evident that mi-Graph with MFCC feature give better result compared to other methods. Figure 4 shows the graphical representation of the performance score obtained for mi-graph and mi-SVM based audio classification.



a. mi-Graph



b. *mi-SVM*

Fig. 4.Performance scores for (a) mi-Graph (b) mi-SVM based news audio classification

VI. CONCLUSION

In this study, a novel method for content based audio classification using MIL approach is presented. The news audio files taken from the indigenous MNAC audio dataset are classified using mi-Graph and mi-SVM techniques. mi-Graph models a direct relationship between bag and instances and mi-SVM is semi-supervised in its nature. The news audio classification experiments are conducted using MFCC and PLP features. Performance evaluation of the proposed mi-Graph and mi-SVM methods using MFCC and PLP parameters are also carried out. mi-Graph using MFCC features appears as the best-performing method with 95.8% audio classification accuracy and 0.96 F1 score which is comparative with the other audio classification results reported earlier. Many audio, multimedia and speech analytics applications would certainly benefit from the ability of the proposed MIL based audio classifier to classify and retrieve audio samples into different categories based on its content.

REFERENCES

1. Christel, Michael, Scott Stevens, and Howard Wactlar. "Informedia digital video library." In Proceedings of the second ACM international conference on Multimedia, pp. 480-481. ACM, 1994.
2. Kandemir, Melih, and Fred A. Hamprecht. "Computer-aided diagnosis from weak supervision: A benchmarking study." Computerized Medical Imaging and Graphics 42 (2015): 44-50.
3. Kandemir, Melih, and Fred A. Hamprecht. "Computer-aided diagnosis from weak supervision: A benchmarking study." Computerized Medical Imaging and Graphics 42 (2015): 44-50.
4. Mporas, Iosif, Todor Ganchev, Mihalis Siafarikas, and Nikos Fakotakis. "Comparison of speech features on the speech recognition task." Journal of Computer Science 3, no. 8 (2007): 608-616.
5. Vivek P, Kumar Rajamani, Lajish V L, "Effective News Video Classification Based On Audio Content: A Multiple Instance Learning Approach", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 7 (6) , page: 2556-2560, ISSN: 0975-9646, 2016
6. Wold, Erling, Thom Blum, Douglas Keislar, and James Wheaten. "Content-based classification, search, and retrieval of audio." IEEE multimedia 3, no. 3 (1996): 27-36.
7. Foote, Jonathan T. "Content-based retrieval of music and audio." In Multimedia Storage and Archiving Systems II, vol. 3229, pp. 138-148. International Society for Optics and Photonics, 1997.
8. Pfeiffer, Silvia, Stephan Fischer, and Wolfgang Effelsberg. "Automatic audio content analysis." In Proceedings of the fourth ACM international conference on Multimedia, pp. 21-30. ACM, 1997.
9. Kimber, Don, and Lynn Wilcox. "Acoustic segmentation for audio browsers." Computing Science and Statistics (1997): 295-304.
10. Zhang, Tong, and C-C. Jay Kuo. "Video content parsing based on combined audio and visual information." In Multimedia Storage and Archiving Systems IV, vol. 3846, pp. 78-90. International Society for Optics and Photonics, 1999.
11. Dinh, Phung Quoc, Chitra Dorai, and Svetha Venkatesh. "Video genre categorization using audio wavelet coefficients." ACCV 2002 (2002).
12. Jasinschi, Radu S., and Jennifer Louie. "Automatic tv program genre classification based on audio patterns." In Euromicro Conference, 2001. Proceedings. 27th, pp. 370-375. IEEE, 2001.
13. Thomas F Quatieri. Discrete-time speech signal processing: principles and practice. Pearson Education India, 2006.
14. Liu, Zhu, Jincheng Huang, and Yao Wang. "Classification TV programs based on audio information using hidden Markov model." In Multimedia Signal Processing, 1998 IEEE Second Workshop on, pp. 27-32. IEEE, 1998.
15. Visser, Erik, Yinyi Guo, Lae-Hoon Kim, Raghuveer Peri, and Shuhua Zhang. "Deep neural net based filter prediction for audio event classification and extraction." U.S. Patent 9,666,183, issued May 30, 2017.
16. James D Keeler, David E Rumelhart, and Wee Kheng Leow. Integrated segmentation and recognition of hand-printed numerals. In Advances in neural information processing systems, pages 557–563, 1991.
17. Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence, 89(1):31–71, 1997.
18. Yixin Chen and James Z Wang. Image categorization by learning and reasoning with regions. Journal of Machine Learning Research, 5(Aug):913–939, 2004.
19. Jun Yang, Rong Yan, and Alexander G Hauptmann. Multiple instance learning for labeling faces in broadcasting news video. In Proceedings of the 13th annual ACM international conference on Multimedia, pages 31–40. ACM, 2005.
20. Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In Advances in neural information processing systems, pages 577–584, 2003.
21. Melih Kandemir and Fred A Hamprecht. Computer-aided diagnosis from weak supervision: a benchmarking study. Computerized Medical Imaging and Graphics, 42:44–50, 2015.
22. Young, Steve. "A review of large-vocabulary continuous-speech." IEEE signal processing magazine 13, no. 5 (1996): 45.
23. Tan, Zheng-Hua, and Ivan Kraljevski. "Joint variable frame rate and length analysis for speech recognition under adverse conditions." Computers & Electrical Engineering 40, no. 7 (2014): 2139-2149.
24. Vergin, Rivarol, Douglas O'shaughnessy, and Azarshid Farhat. "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition." IEEE Transactions on Speech and Audio Processing 7, no. 5 (1999): 525-532.
25. Hermansky, Hynek. "Perceptual linear predictive (PLP) analysis of speech." the Journal of the Acoustical Society of America 87, no. 4 (1990): 1738-1752.
26. Zhou, Zhi-Hua, Yu-Yin Sun, and Yu-Feng Li. "Multi-instance learning by treating instances as non-iid samples." In Proceedings of the 26th annual international conference on machine learning, pp. 1249-1256. ACM, 2009.
27. Andrews, Stuart, Ioannis Tsochantaridis, and Thomas Hofmann. "Support vector machines for multiple-instance learning." In Advances in neural information processing systems, pp. 577-584. 2003.