FEATURE BASED FUSION SYSTEM FOR ANOMALOUS SOUNDS MONITORING

Technical Report

Jisheng Bai

LianFeng Acoustic Technologies Co., Ltd.
Xi'an, China
baijs@mail.nwpu.edu.cn

Chen Chen, Jianfeng Chen

Northwestern Polytechnical University
Xi'an, China
cc_chen524@mail.nwpu.edu.cn
cjf@nwpu.edu.cn

ABSTRACT

Anomaly detection has a wide range of application scenarios in industry such as finding fraud cases in financial industry or finding network intrusion in network security. And finding anomaly condition of machines in factories can prevent causing damage. In this paper, we introduce our system for Task2 of Dcase 2020 challenges (Unsupervised Detection of Anomalous Sounds for Machine Condition Monitoring). We focus on finding relationship between different kinds of features and different types of anomaly sounds. MFCC, log-mel, log-linear and HPSS are fed into a deep autoencoder (DAE). We use the baseline DAE as our primary network, meanwhile, compared to Isolation Forest (IF) and One-class SVM (OC-SVM). Experiment results shows that for different machine type, different features may improve the detection results respectively, and DAE is more likely to preform much better than machine learning techniques.

Index Terms— Anomaly detection, autoencoder, feature fusion

1. INTRODUCTION

Anomalies are also referred to as abnormalities, deviants, or outliers in the data mining and statistics literature [1]. The purpose of anomaly detection algorithm is to find a boundary between normal data and anomalous data. The challenge of anomaly detection is the lack of anomaly data and the uncertain types of anomaly data. To extract general feature of normal data, machine learning techniques such as PCA (Principal Components Analysis), IF and OC-SVM are widely applied in anomaly detection. But traditional anomaly detection algorithms can not handle high dimensional data and is weak in feature extracting. Deep anomaly detection (DAD) can learn hierarchical discriminative features from data , and advocates to solve the problems and is developed rapidly in recent years. Autoencoder (AE) is one of the common DAD algorithms, GAN (Generative Adversarial Network) , VAE (Variational autoencoder) and OC-NN (one class neural network) are generally applied in various scenes [2]. AE can compress the input data into a lower dimension in a unsupervised way and decode the data to initial input data. By minimizing the distance between decoded data and initial input data (reconstruction error), the encoded data can greatly represent the input data. The AE, trained with normal data can hardly reconstruct the anomaly data, so a significant reconstruction error will occur when the anomaly data are fed into the AE.

Using MFCC (Mel Frequency Cepstral Coefficents), log-mel and HPSS (The harmonic percussive source separation) spectorgrams as sound representations can explore different source of urban sounds [3]. Considering 6 various types of machinery anomaly sound, to apply different features in anomaly detection may improve the performance of each machine type in task2.

This paper is organized as follows: the data splits and setup of task2 will be introduced in Section 2. Section 3 gives a brief introduction of our anomaly detection algorithms. In Section 4, the details of features extraction will be showed. The evaluation results and discussion are presented in Section 5.

2. TASK DATASET AND SETUP

ToyADMOS and MIMII Dataset are used as the primary dataset in the task [4] [5]. ToyADMOS contains two types of sound, ToyCar and ToyConveyor. MIMII contains four types of sound, fan, pump, slider and valve. Each recording is a 10-second audio, for each machine type split, there are three or four different machine IDs.

Development dataset: For each machine ID, there are around 1,000 samples of normal sounds for training and 100-200 samples each of normal and anomalous sounds for the test.

Additional training dataset: This dataset includes around 1,000 normal samples for each Machine Type and Machine ID used in the evaluation dataset.

Evaluation dataset: It consists of the same Machine Types' test samples as the development dataset. The number of test samples for each Machine ID is around 400, none of which have a condition label (i.e., normal or anomaly).

This task is evaluated with the area under the receiver operating characteristic (ROC) curve (AUC) and the partial-AUC (pAUC). The pAUC is an AUC calculated from a portion of the ROC curve over the pre-specified range of interest.

3. ANOMALY DETECTION ALGORITHMS

Traditional machine learning algorithms try to find the boundary of normal data or the hyper plane between normal and anomalous data. Isolation forest is an unsupervised anomaly detection method for continuous data [7].

Isolation forest detects outliers by isolating sample points. Specifically, the algorithm uses a binary tree to isolate samples. Because the number of outliers is scarce and they are deviated from most samples, outliers will be isolated earlier, that means, outliers

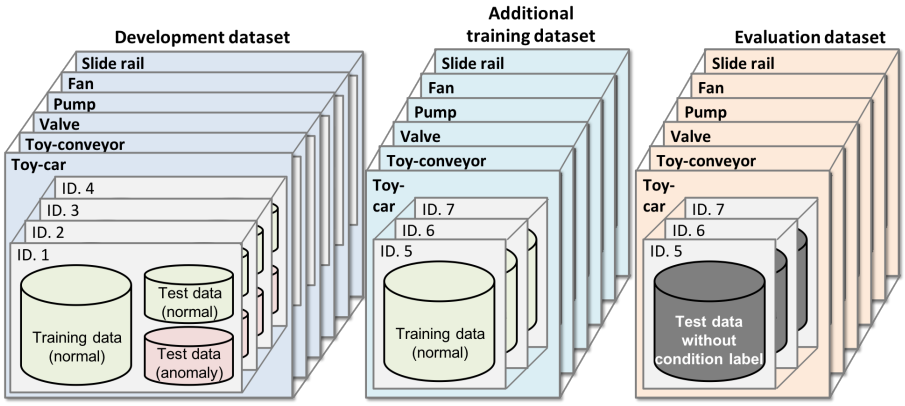

Figure 1: Task2 unsupervised detection of anomalous sounds for machine condition monitoring [6]

Input	Input dim
Dense*4	128
Dense	5
Dense*4	128
Dense	Input dim

Table 1: AE architecture

will be closer to the root node, while normal values will be farther away from the root node.

OC-SVM or SVDD (Support vector domain description) uses a hypersphere instead of a hyperplane to partition. The algorithm obtains the spherical boundary around the normal data in the feature space, and expects to minimize the volume of the hypersphere. So if a data is not surrounded by the hypersphere, it is considered to be a outlier or anomaly.

Autoencoder is a kind of artificial neural networks used in semi supervised learning and unsupervised learning. It can learn the efficient representation of input data, so it is widely applied in dimensionality reduction and anomaly detection. An autoencoder takes an input $\mathbf{x} \in \mathcal{R}^d$ and first maps it to the latent representation $\mathbf{h} \in \mathcal{R}^{d'}$ using a deterministic function of the type $\mathbf{h} = f_\theta = \sigma(Wx + b)$ with parameters $\theta = \{W, b\}$. This "code" is then used to reconstruct the input by a reverse mapping of $f : \mathbf{y} = f_{\theta'}(h) = \sigma\left(W'h + b'\right)$ with $\theta' = \{W', b'\}$. Each training pattern x_i is then mapped onto its code h_i and its reconstruction y_i. The parameters are optimized, minimizing an appropriate cost function over the training set $\mathcal{D}_n = \{(x_0, t_0), \ldots, (x_n, t_n)\}$ [8].

In our experiments, an AE is used to extract feature embedding to train IF and OC-SVM, and to detect anomalous data as well. Scikit-learn provides IF and OC-SVM functions, the estimators of IF are 256 and the ν of OC-SVM is 1e-4. The AE architecture of the baseline is described in Table 1 and it is our primary network.

4. FEATURES EXTRACTION

Recordings are loaded with default sample rate and applied short time Fourier transform (STFT) with a Hanning window size of 1024 and hop length of 512 samples. Mel and linear filters with bands of 128 are used to transformed STFT spectrogram to mel and linear spectrogram. Then the mel spectrograms are used to generate MFCCs of 128 bands.

The harmonic percussive source separation (HPSS) [9] can split a signal $w(t)$ into harmonic part $h(t)$ and percussive part $p(t)$ and there are several approaches to separate. We can simplify the separation procedure as follows [10]

$$w(t) \overset{\text{HPSS}(l)}{\longrightarrow} h(t), p(t) \tag{1}$$

The harmonic and percussive spectrograms are generated from STFT spectrogram with librosa decompose function. All the spectrograms are calculated by the log algorithm to get log spectrograms.

As it is introduced in the baseline system, a spectrogram of the input $X = \{X_t\}_{t=1}^{T}$ where $X_t \in R^F$, and F and T are the number of mel-filters and time-frames, respectively. Then, the acoustic feature at t is obtained by concatenating before/after several frames of outputs as $\psi_t = (X_{t-P}, \ldots, X_{t+P}) \in R^D$, where $D = F \times (2P + 1)$ and P is the context window size. Because the size of frequency axis is 640 for log-mel, log-linear and MFCC but 513 for hpss-h and hpss-p, it is necessary to decrease the P dimension of hpss-h and hpss-p. So The P is set to 2 and 1 respectively.

5. RESULTS AND DISCUSSION

We extract the encoded embeddings of dense layer with size of 8 and train them with IF or OC-SVM. Then different features are fed into a AE.

Table 2 shows all the AUC/pAUC scores of algorithms and features for each machine type. The results of Encoder+IF and Encoder+OC-SVM are much worse than AE. This may due to that the highly compressed nonlinear embeddings can not be classified by these algorithm.

Log-mel gets the best scores of ToyCar, ToyConveyor and fan. As for the other 4 features, MFCC also preforms well in ToyCar, ToyConveyor and fan. Log-linear gets good results in ToyCar and slider. For the HPSS, harmonic spectrogram gets highest score of 0.801/0.628 in pump and performs well in ToyConveyor and fan. The hpss-p can explore slider and valve sounds which may contains percussive components, it obtains 0.917/0.782 and 0.845/0.661 on these two machine types, about 7.9% and 27.1% improvement from the baseline.

Algorithm	Feature	ToyCar	ToyConveyor	Fan	Pump	Slider	Valve
AE	Log-mel	0.801/0.672	0.727/0.607	0.652/0.526	0.726/0.600	0.850/0.669	0.665/0.506
Encoder+IF	Log-mel	0.430/0.485	0.505/0.509	0.530/0.516	0.460/0.524	0.501/0.525	0.536/0.512
Encoder+OC-SVM	Log-mel	0.455/0.502	0.500/0.507	0.536/0.531	0.509/0.490	0.626/0.527	0.526/0.503
AE	MFCC	0.791/0.668	0.705/0.580	0.647/524	0.739/0.602	0.844/0.660	0.670/0.508
AE	Log-linear	0.750/0.618	0.672/0.564	0.597/0.508	0.686/0.594	0.911/0.761	0.741/0.542
AE	Hpss-h	0.638/0.561	0.593/0.529	0.533/0.522	0.612/0.579	0.917/0.782	0.845/0.661
AE	Hpss-p	0.662/0.538	0.724/0.606	0.642/0.525	0.801/0.628	0.810/0.595	0.567/0.504

Table 2: Best AUC/pAUC scores of algorithms and features for each machine type

Final output is fused with the results of different features, based on the best performance for each machine type on development dataset.

6. CONCLUSION

In this paper, we present a feature based system for Anomalous Sounds Monitoring. In our approach, five different features are generated as inputs of the networks. Then, an AE is applied for detecting anomalous data. Finally, we fused different results of the AE according to the AUC scores. It can be concluded that different features can improve the performance for some machine types respectively. The fusion method can make a great improvement than the baseline system. For further work, encoder plus machine learning algorithm for anomaly detection will be studied, and advantages of detecting different machinery sounds with features will be researched as well.

7. REFERENCES

[1] C. C. Aggarwal, "An introduction to outlier analysis," in *Outlier analysis*. Springer, 2017, pp. 1–34.

[2] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019.

[3] J. Bai, C. Chen, and J. Chen, "A multi-feature fusion based method for urban sound tagging," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1313–1317.

[4] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, November 2019, pp. 308–312. [Online]. Available: https://ieeexplore.ieee.org/document/8937164

[5] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, November 2019, pp. 209–213. [Online]. Available: http://dcase.community/documents/workshop2019/proceedings/DCASE2019Workshop_Purohit_21.pdf

[6] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *arXiv e-prints: 2006.05822*, June 2020, pp. 1–4. [Online]. Available: https://arxiv.org/abs/2006.05822

[7] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.

[8] J. Masci, U. Meier, C. Dan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Artificial Neural Networks Machine Learning-icann - international Conference on Artificial Neural Networks*, 2011.

[9] D. Fitzgerald, "Harmonic/percussive separation using median filtering," 2010.

[10] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 228–237, 2013.