A Study on Distance Measure for Effective Anomaly Detection using AutoEncoder

HyunYong Lee, Nac-Woo Kim, Jun-Gi Lee, and Byung-Tak Lee
Honam Research Center (HRC)
Electronics and Telecommunications Research Institute (ETRI)
Gwangju, Korea
{hyunyonglee,nwkim,jungi,bytelee}@etri.re.kr

***Abstract*—Anomaly detection is a popular application in various areas. One challenging issue is to build an anomaly detection model using normal data because collecting potential abnormal data is quite difficult. In this paper, we build an anomaly detection model using just normal data based on adversarial autoencoder for acoustic data. After extracting features using the trained model, we apply a distance-based method for calculating a threshold to be used for anomaly detection. In particular, we propose a method for reflecting differences in dimensions in calculating distance. Through experiments, we show that the proposed dimension-aware distance measure improves anomaly detection accuracy by up to 7% compared to existing distance measure methods.**

***Index Terms*—Anomaly detection, distance, autoencoder, dimension-aware, deep learning.**

I. Introduction

Anomaly detection is to identify abnormal data which raise suspicions by differing noticeably from the majority of the known normal data. Anomaly detection is widely used in various areas including fraud detection, medical anomaly detection, industrial damage detection, and video surveillance [1].

In building a data-centric model for anomaly detection, knowing abnormal data may be greatly helpful. But, the acquisition of abnormal data in advance is very difficult or impossible. For example, abnormal data about rare or catastrophic events are not likely to be allowed in advance. In addition, a few known abnormal data are not likely to represent all possible abnormal data. Therefore, using just known normal data is one popular choice for building a model [2].

Given the normal data, we have two issues. The first issue is to select a proper model to extract features from the normal data. For this, various data-centric models can be used. Recently, deep learning-based models are widely used because of their promising performance. Deep neural network (DNN) [3], convolutional neural network (CNN) [4], long short term memory (LSTM) network [5], and autoencoder (AE) and its variants [6] are popular choices. In the case of DNN, CNN, and LSTM, last or penultimate layer of a trained model can be used to extract features for given normal data. In the case of AE and its variants, latent space can be used to extract features. The second issue is to represent normal data using the extracted features. In other words, we need to derive a threshold for detecting abnormal data. For example, data can be regarded as an anomaly if its value (derived from the features) exceeds a pre-defined threshold. OC-SVM [7] and SVDD [8] are popular choices. The selection of a model and a way for deriving a threshold may heavily depend on the type and distribution of normal data.

In this paper, we are interested in building an anomaly detection model for acoustic data. In particular, with adversarial AE (AAE) model [9], we propose and examine the feasibility of a distance-based threshold. Using AAE, we first extract features. We utilize features in the latent space of AAE. Next, we derive a threshold. For this, in the latent space of AAE, we calculate distance from each latent value to the mean of the latent values and then use the specific percentile of the distance list. In doing this, we are particularly interested in improving distance measure in terms of detection accuracy. For this, we propose a dimension-aware distance measure method that can be applied to existing well-known methods such as Euclidean distance. Our experiment results show that our dimension-aware distance measure improves the detection accuracy by 2 - 7% compared to existing methods. The resulting anomaly detection accuracy is around 90% with the aforementioned methods for the acoustic data, which shows the feasibility of the the proposed approach.

The rest of this paper is organized as follows. Section II describes our model for anomaly detection and examines its feasibility in using normal data only. Section III introduces our dimension-aware distance measure and compares it with existing distance methods. Section IV concludes this paper.

II. Using Normal Data Only

A. Data and Model

For the purpose of experiments, we collect acoustic data in an underground tunnel by generating 10 types of sounds. The ambient sounds include construction (1,573 data), conversation (1,060), drilling (1,160), engine idling (1,005), hammer jack (1,031), rain (1,192), siren (1,379), walk (1,021), water drop (1,015), and water pump (1,324). Each acoustic data is 4 sec. Fig. 1 shows examples of 10 types of acoustic data. We convert raw acoustic data using Mel frequency cepstral coefficient

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20181210301570).

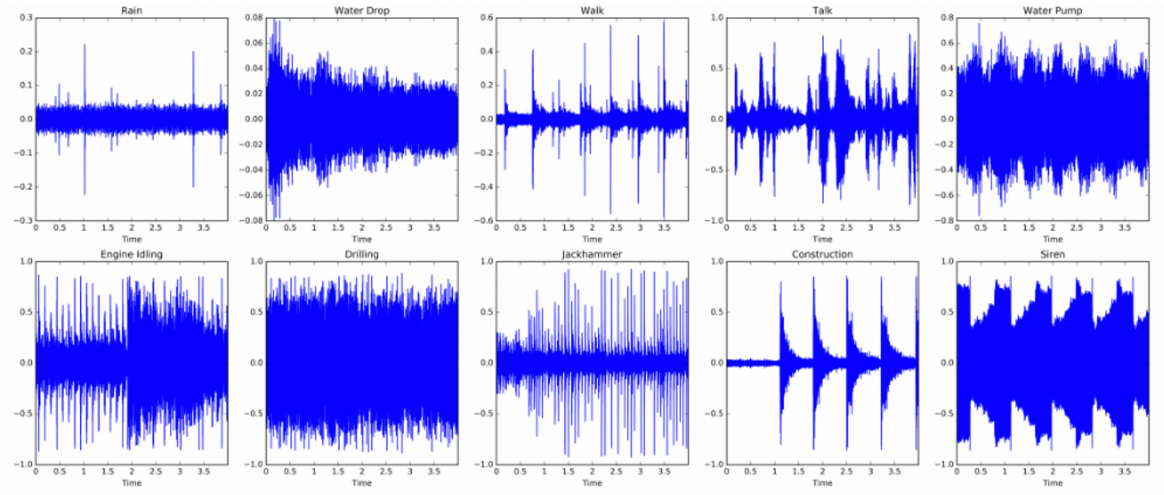

Fig. 1: Examples of 10 types of acoustic data.

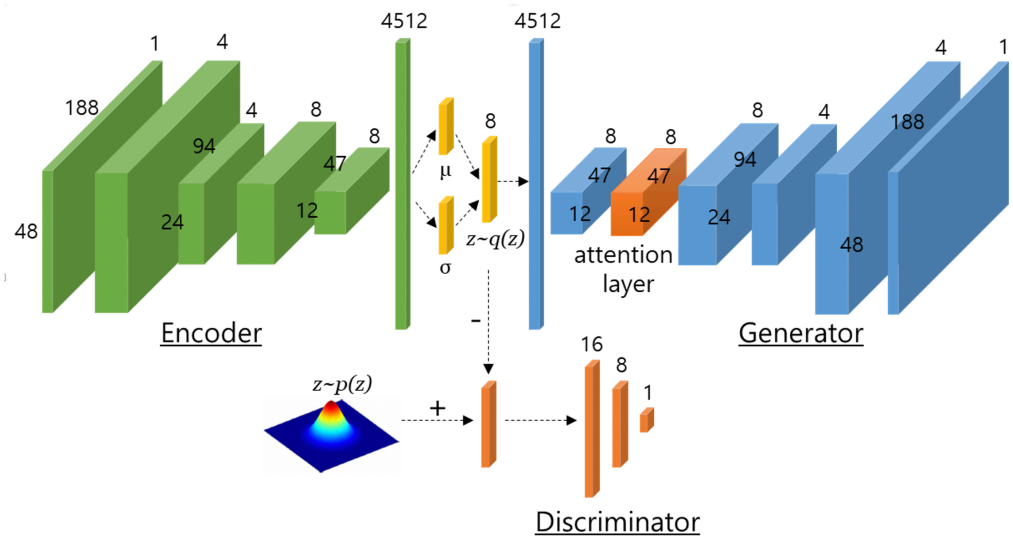

Fig. 2: Examples of 10 types of acoustic data.

(MFCC) [10] and use that as input for an anomaly detection model.

As an anomaly detection model, we use AAE. Fig. 2 shows detailed architecture of AAE. After training AAE model, we use features in latent space (i.e., z in Fig. 2). For anomaly detection scenarios, we use construction and siren as abnormal data. The others are used as normal data. After training AAE model using the most of normal data (i.e., around 90%), we calculate detection accuracy using the rest of normal data and abnormal data.

B. Approach

One common challenging issue in building a deep learning-based system is the lack of data. Please note that deep learning is a data-driven approach, which typically requires a large amount of good data. For example, a simple approach to build a deep learning-based anomaly detection system is to collect normal and abnormal data and to train a system to detect the pre-defined normal or abnormal cases. If we collect data about all possible cases, we may be able to build a nice deep learning-based anomaly detection system. However,

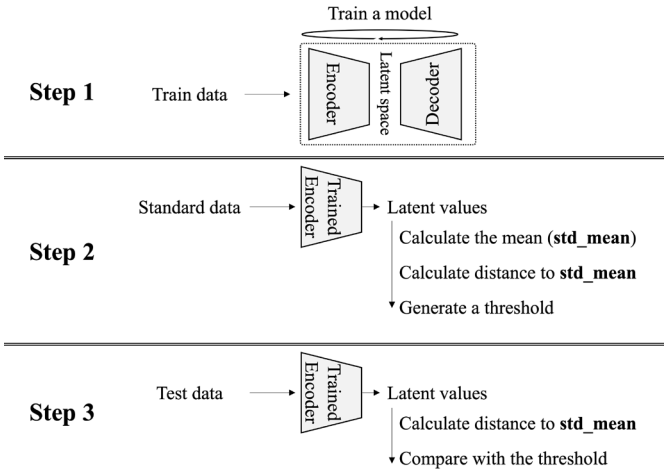

Fig. 3: An overview of distance-based anomaly detection using normal data only.

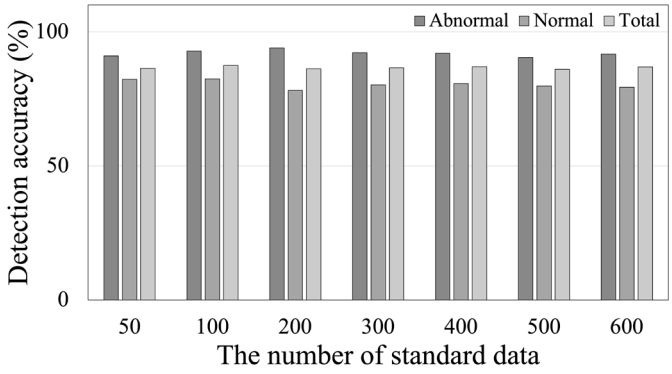

Fig. 4: An effect of the number of standard data on detection accuracy.

we may have two problems in practical cases. One problem is that collecting enough data is not easy in many practical cases, particularly if the data is related to very rare events or catastrophic events. The other problem is that the system is not likely to detect cases not included in the train data.

In this paper, we try to solve the problems above by using normal data only because it is quite easy to collect normal data. Fig. 3 shows our approach for distance-based anomaly detection using normal data only. A basic concept of our approach is to generate a detection threshold using normal data. For this, we divide normal train data into train data (to be used for training a model) and a standard data (to be used for generating a detection threshold). At the first step, with the train data, we first train AAE model. At the second step, using the trained encoder of AAE model, we get latent values of standard data. We calculate the mean of the latent values (i.e., **std_mean**). Then, we calculate the distance from each latent value of the standard data to **std_mean** and generate a list of distance. From the list of distance, we get a specific percentile of the list as the detection threshold. At the third step, latent values of test data are calculated using the trained encoder of AAE model. After calculating the distance from the latent values of test data to **std_mean**, the distance is compared with the detection threshold. If the distance is larger than the detection threshold, that test data case is regarded as an anomaly.

C. Feasibility Study

To examine the feasibility of our distance-based anomaly detection using normal data only, we conduct experiments. For the sake of simplicity, we focus on Euclidean distance in this subsection, but other methods of measuring a distance will be discussed in Section III.

We first examine the effect of the number of standard data on detection accuracy because the standard data is used for generating a detection threshold. Fixing the train data (not to affect the training of the model), we take standard data from the remaining normal data (i.e., 1,364 data). The remaining normal data excluding the standard data is used as normal test data. The number of abnormal test data is 1,204. We change the number of standard data from 50 to 600 and we examine the detection accuracy of the normal test data, abnormal test data, and total test data. Interestingly, we do not observe any noticeable change with the increasing number of standard data. To study this further, we examine **std_mean** with a different number of standard data (Table I). As shown in Table I, **std_mean** of six dimensions does not noticeably change with the increasing number of standard data. Please note that the latent space is in six dimensions. This means that the standard data has a similar distribution in the latent space. It also means that a small number of standard data (e.g., around 100) may be enough to generate a detection threshold as long as normal data has a similar distribution in the latent space. From now on, we use 100 normal test data s the standard data.

TABLE I: The normalized mean value with different number of standard data.

	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6
50	0.62	0.56	0.61	0.52	0.61	0.50
100	0.60	0.55	0.58	0.51	0.59	0.50
200	0.60	0.56	0.59	0.52	0.60	0.51
300	0.60	0.56	0.60	0.52	0.61	0.51
400	0.58	0.53	0.57	0.50	0.58	0.49
500	0.60	0.56	0.59	0.52	0.61	0.52
600	0.60	0.56	0.59	0.53	0.61	0.52

We also examine the effect of the percentile to be used to generate a detection threshold. As we described before, the detection threshold is a specific percentile in the list of distance. We change the percentile value from 50 to 95. As the percentile value increases, the detection accuracy of the abnormal cases decreases because the detection threshold gets loose for them. On the other hand, as the percentile value increases, the detection accuracy of the normal cases increases,

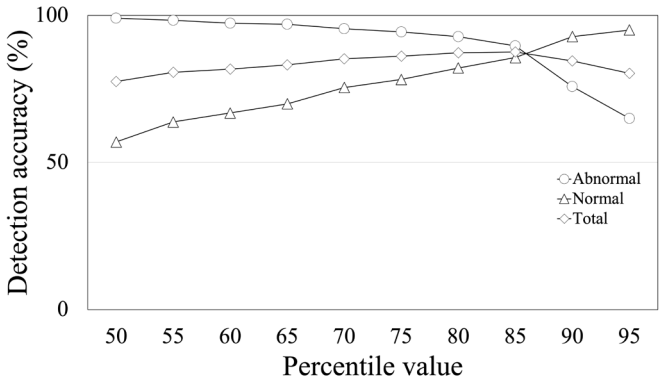

Fig. 5: An effect of the number of standard data on detection accuracy.

which shows a trade-off. However, interestingly, total detection accuracy does not change noticeably, which confirms the trade-off. The detection accuracy for all test data shows its best with the 80th percentile. From now on, we use 80th percentile in calculating the detection threshold.

III. Methods of Measuring Distance

A. Existing Methods

We detect abnormal cases using distance in the latent space of AAE. Therefore, methods of measuring distance may affect the detection accuracy. In this subsection, we study four methods of measuring distance: Euclidean distance, City-block distance, Minkowski distance, and Chebyshev distance [11].[1]

We first briefly describe the four methods of measuring distance. Euclidean distance computes the square root of the sum of the squares of the differences between the coordinates of pairs of data. The equation for Euclidean distance between data x and y can be written as:

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

City-block distance (or, Manhattan distance) computes the absolute differences between coordinates of pair of data. The equation for City-block distance between data x and y can be written as:

$$\sum_{i=1}^{n}|x_i - y_i| \tag{2}$$

Minkowski distance is a generalization of both Euclidean distance (i.e., with p=2 in Equation (3)) and City-block distance (i.e., with p=1 in Equation(3)). The equation for Minkowski distance between data x and y can be written as:

$$(\sum_{i=1}^{n}|x_i - y_i|^{1/p})^p \tag{3}$$

Throughout this paper, we calculate Minkowski distance with p=3. Chebyshev distance computes the absolute magnitude of the differences. The equation for Chebyshev distance between data x and y can be written as:

$$\max_i |x_i - y_i| \tag{4}$$

We compare the four methods of measuring distance in terms of detection accuracy. For this, we use 100 standard data and 80th percentile in calculating the detection threshold. Table II shows the detection accuracy for abnormal case, normal case, and total case (i.e., including both abnormal and normal cases). Euclidean distance shows the best accuracy in detecting abnormal cases. In detecting normal cases, Minkowski distance shows the best accuracy. In total, Euclidean distance shows the best detection accuracy. Chebyshev distance shows the worst detection accuracy.

TABLE II: The detection accuracy of methods of measuring distance.

	Euclidean	City-block	Minkowski	Chebyshev
Abnormal	**92.77**	90.37	89.45	78.82
Normal	82.12	81.57	**84.02**	81.17
Total	**87.32**	85.86	86.67	80.02

B. Dimension-aware method

During our experiments, we found that existing methods of measuring distance is not good at reflecting differences of different coordinates. For example, data D1 = (5, 4, 3, 2, 1) and data D2 = (3, 2, 7, 6, 9) show the same Euclidean distance to data D3 = (1, 2, 3, 4, 5) even though D1 and D2 are different. Therefore, in this paper, we propose one approach for overcoming this limitation. We call it a dimension-aware distance. Algorithm 1 shows the pseudo-code for the dimension-aware distance. Given the two input data (i.e., D1 and D2), we shift elements of D2 and calculate a distance between D1 and the shifted D2. We repeat this as the number of elements of D2 minus 1. During the repeats, we cumulate the distance and use the final cumulated distance as the dimension-aware distance. For calculating the distance between D1 and the shifted D2, any existing methods of measuring distance can be used, which means our dimension-aware distance can be applied to them for improving their ability to reflect the differences of different dimensions. For example, if we use the dimension-aware Euclidean distance, the distance between D1 and D3 is 21.06 while the distance between D1 and D2 is 41.97.

C. Feasibility Study

To examine the feasibility of the dimension-aware distance, we apply the proposed approach to the four methods of measuring distance. Table 3 compares the results of existing

[1]We also examined other methods of measuring distances such as Bray-curtis distance, Canberra distance, and so on. But, we do not have an acceptable detection accuracy with those methods.

Algorithm 1 Dimension-aware distance

procedure DIMDISTANCE($D1, D2$)
  $dim\_dist = 0$
  $N = length(D1) - 1$
  for $k \leftarrow 1$ to N **do**
    $D\_shifted = element\_shift(D2, k)$ ▷ Shifting elements
    $dim\_dist+ = calc\_distance(D1, D\_shifted)$ ▷ Calculating distance
  end for
  return dim_dist
end procedure

TABLE III: Comparison between naive methods and proposed methods.

	Abnormal		Normal		Total	
	Naive	Prop.	Naive	Prop.	Naive	Prop.
Euclidean	92.77	**95.27**	82.12	**83.31**	87.32	**89.14**
City-block	90.37	**91.45**	81.57	**84.34**	85.86	**87.8**
Minkowski	89.45	**94.85**	**84.02**	83.07	86.67	**88.82**
Chebyshev	78.82	**92.28**	81.17	**83.62**	80.02	**87.84**

naive methods and existing methods with our approach. In detecting abnormal cases, our dimension-aware distance improves the detection accuracy of all existing methods. With our approach, Euclidean distance and Chebyshev distance increases the detection accuracy by 2.5% and 13.45%, respectively. In detecting normal cases, our dimension-aware distance improves the detection accuracy of existing methods excluding Minkowski distance. In total, our dimension-aware distance improves the total detection accuracy by up to 7.82%. This result shows that the dimension-aware distance is better than existing naive methods at conducting anomaly detection.

IV. CONCLUSION

A selection of a proper model and threshold for detecting abnormal cases is important in building a good anomaly detection model. In this paper, we study the feasibility of using AAE as the feature extraction model and distance-based threshold for acoustic data. In particular, we propose the dimension-aware distance measure method for improving anomaly detection accuracy. As future work, it would be interesting to study a way for calculating the dimension-aware distance for high-dimensional data so as to reduce the computational overhead.

REFERENCES

[1] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *CoRR*, vol. abs/1901.03407, 2019.

[2] S. Maya, K. Ueno, and T. Nishikawa, "dLST: A new approach for anomaly detection using deep learning with delayed prediction," *International Journal of Data Science and Analytics*, vol.8, pp.137-164, May 2019.

[3] F. K. Lodhi, S. R. Hasan, O. Hasan, and F. Awwadl, "Power profiling of microcontroller?s instruction set for runtime hardware trojans detection without golden circuit models," in *Proc. of the Conference on Design, Automation & Test in Europe*, 2017.

[4] Z. Zhang, X. Zhou, X. Zhang, L. Wang, and P. Wang, "A model based on convolutional neural network for online transaction fraud detection," *Security and Communication Networks*, vol.2018, Aug. 2018.

[5] X. Lp, W. Yu, T. Luwang, J. Zheng, X. Qiu, J. Zhao, L. Xia, and Y. Li, "Transaction fraud detection using gru-centered sandwich-structured model," in *Proc. of IEEE CSCWD*, 2018.

[6] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, "Training adversarial discriminators for cross-channel abnormal event detection in crowds," *arXiv preprint arXiv:1706.07680*, 2017.

[7] D. Dotti, M. Popa, and S. Asteriadis, "Unsupervised discovery of normal and abnormal activity patterns in indoor and outdoor environments," in *Proc. of International Conference on Computer Vision Theory and Applications*, 2017.

[8] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and largescale anomaly detection using a linear one-class svm with deep learning," *Pattern Recognition*, vol.58, pp.121-134, 2016.

[9] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *CoRR*, vol.abs/1511.05644, 2015.

[10] I. A. Sulistijono, R. C. Urrosyda, and Z. Darojah, "Me-frequency cepstral coefficient MFCC for music feature extraction for the dancing robot movement decision," in *Proc. of ICIRA*, 2016.

[11] V. Moghtadaiee and A. G. Dempster, "Determining the best vector distance measure for use in location fingerprinting," *Elsevier Pervasive and Mobile Computing*, vol.23, pp.59-79, Oct. 2015.