Technical Note

Low-cost Unsupervised Outlier Detection by Autoencoders with Robust Estimation

Yoshinao Ishii[1,a)] Masaki Takanashi[1,b)]

Received: October 18, 2018, Accepted: December 4, 2018

Abstract: Recently, an unsupervised outlier detection method based on the reconstruction errors of an autoencoder (AE), which achieves high detection accuracy, was proposed. This method, however, requires a high calculation cost because of its ensemble scheme. Therefore, in this paper, we propose a novel AE-based unsupervised method that can achieve high detection performance at a low calculation cost. Our method introduces the concept of robust estimation to appropriately restrict reconstruction capability and ensure robustness. Experimental results on several public benchmark datasets show that our method outperforms well-known outlier detection methods and at a low calculation cost.

Keywords: outlier detection, unsupervised learning, autoencoder, robust estimation

1. Introduction

In real datasets, it often happens that some samples have different values or features from that of the majority (*inliers*). Such samples are called outliers, and detecting outliers from the given data is generally called outlier detection. Since outlier detection plays an important role in detecting anomalies in target systems, in creating normal models from real datasets, and so on, outlier detection methods with high accuracy are required. In particular, unsupervised outlier detection methods are important because real datasets are not often labeled. Some common unsupervised methods include distance-based methods [1], [2], density-based methods [3], and linear methods [4]. These methods basically calculate *outlier scores* that indicate the outlierness of each sample, and detect the outliers using their outlier scores.

Distance-based methods and density-based methods derive outlier scores using the distance and density ratio between samples, respectively. Although both methods can obtain outlier scores by considering the nonlinear features in the data, it is difficult to achieve highly accurate detection in the case of high-dimensional data. In linear methods, we regress the high-dimensional data to a low-dimensional linear model, and calculate the outlier scores from the residuals of the samples. These methods utilize the property that the squared residuals of outliers tend to be large because the obtained regression model fits well to the inliers (majority) rather than the outliers. We could detect outliers with high accuracy from high-dimensional data. However, the nonlinear data detection accuracy might be low in some cases because these methods cannot extract the nonlinear characteristics in the data.

[1] Toyota Central R&D Laboratories, Inc., Nagakute, Aichi 480–1118, Japan
[a)] y-ishii@mosk.tytlabs.co.jp
[b)] m-takanashi@mosk.tytlabs.co.jp

As a generalization of linear methods, an unsupervised outlier detection method based on the reconstruction errors of an autoencoder (AE) [5], which is a special type of multi-layer neural network has been proposed. We refer to this unsupervised method as the *AE-based method*. In the AE-based method, an AE is trained by constraining its reconstruction capability in order to prevent identity mapping, and the reconstruction errors of the samples are taken as their outlier scores. An encoder is a nonlinear mapping (regression) from an original data space to a feature space, a decoder is a nonlinear mapping from the feature space to the original data space, and the reconstruction errors correspond to the residuals in the linear methods. Therefore, the AE-based method is regarded as a generalization of linear methods. The AE-based method is capable of achieving high detection accuracy even for data having high dimensionality and nonlinear features. However, the reconstruction errors of the outliers as well as the inliers are small if its reconstruction capability is properly restricted. This leads to a low detection accuracy or namely overfitting. Due to this drawback, most of the AE's reconstruction error-based methods have been explored with regard to semi-supervised outlier detection requiring normal labels, and almost no unsupervised method has been proposed [6].

Recently, *RandNet* [6], an AE-based method that overcomes the aforementioned drawback and achieves high detection accuracy was proposed. In that study, an ensemble scheme was introduced and various randomly connected AEs (100 AEs were used in the experiment) with different structures and connection densities were prepared. In RandNet, each AE is trained independently, and then the median of the reconstruction errors of all the trained AEs are taken as outlier scores of the samples. This ensemble scheme constrains the reconstruction capability and improves robustness; therefore, RandNet achieves a high detection accuracy. However, its calculation cost is huge because it is necessary to train a large number of AEs and pre-train each

AE layer-wise. In particular, when parallel computational environment is not available, the high computational cost needed to independently train a large number of AEs becomes a significant problem.

Robust Deep Autoencoders (RDAs) [7] have also been proposed in recent years as an AE-based method with a similar purpose. Since RDA learns not to reconstruct samples considered as outliers it is more robust than a normal AE. RDA decomposes a data matrix X nonlinearly into a matrix L_D which is easy to reconstruct and a sparse error matrix S which is difficult to reconstruct by an AE, subject to $X = L_D + S$. Then, S can be identified with outliers by outlier detection. In order to perform this decomposition, RDA optimizes parameters of AE and S in a alternating manner. However, there is a drawback that the optimization phase of S does not always optimize the whole objective function. Therefore, it could take long time to converge or in other words impose a high computational cost. Additionally, the experimental results in Ref. [7] show that RDA highly depends on its parameter choice. This is another drawback. Owing to these drawbacks, RDA is hard to use as an unsupervised anomaly detection method for real datasets.

Therefore, in this paper, we propose a novel AE-based method that can achieve high detection performance at a low calculation cost. Our method introduces the concept of robust estimation [8] to restrict the reconstruction capability and ensure robustness. An outline of the AEs and robust estimation is provided in Section 2. Section 3 discusses our proposed method, and Section 4 discusses the experimental results using real datasets.

2. Autoencoders and Robust Estimation

An AE is a special type of multi-layer neural network in which the number of nodes in the output layer is the same as that in the input layer. Generally, the model parameters are trained to minimize the reconstruction error (loss function) L, which is defined by the following equation.

$$L = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i\|_2^2. \tag{1}$$

In this Eq. (1), $\{\boldsymbol{x}_i\}(i = 1, \ldots, N)$, $\{\bar{\boldsymbol{x}}_i\}(i = 1, \ldots, N)$, and N respectively denote the training data, the outputs from the AE corresponding to each training data, and the number of samples.

Robust estimation is a technique for regressing inliers to the model, avoiding the strong adverse effect of outliers present in the data. Robust estimation has been studied for a long time, and it is capable of overcoming the low robustness against outliers of the least squares method, which is the most commonly used in regression. Some of the most common estimators include M estimator [9], least median of squares estimator (LMS) [10], least trimmed squares estimator (LTS) [10], and so on. In this paper, we focus on the LTS estimator. The LTS estimator is defined as

$$\hat{\beta}_{LTS} = \arg\min_{\beta} \sum_{i=1}^{m} r_i^2(\boldsymbol{\beta}), \tag{2}$$

which minimizes the sum up to the m-th order statistic. $\boldsymbol{\beta}$ and $r_1^2(\beta) \le r_2^2(\beta) \le \ldots r_N^2(\beta)$ denote a regression parameter vector and the ordered squared residuals with β, respectively. That is, the LTS estimator avoids the adverse effects of outliers by not using samples with higher squared residuals in the regression for parameter estimation.

3. Proposed Method

Our proposed method uses a loss function that incorporates concepts of the LTS estimator in order to restrict the reconstruction capability and ensure robustness of AEs. Specifically, our method utilizes a mini-batch learning approach to minimize the loss function L_{prop} defined by the following equation.

$$L_{prop} = \frac{1}{B}\sum_{i=1}^{B} w_i \cdot e_i, \tag{3}$$

where B denotes the mini-batch size, $e_i = \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i\|_2^2$ holds, and w_i satisfies the following equation.

$$w_i = \begin{cases} 1 & (e_i \le c) \\ 0 & (e_i > c). \end{cases} \tag{4}$$

Here, c denotes the α_p-th percentile value of $\boldsymbol{e} = \{e_1, \ldots, e_B\}$, α_p is a parameter, and w_i is updated in every mini-batch learning.

Namely, in our proposed method, samples with higher reconstruction error are not used for updating the parameters during batch learning since their losses are forced to 0. The final outlier scores of the samples in our method are e_i obtained from the trained model. The additional calculation cost of our proposed method compared to normal AE, due to this update process, is only a derivation of the α_p percentile value.

The use of the proposed loss function results in the following effects. First, the reconstruction errors of the inliers with $w_i = 0$ tend to be small since the inliers with $w_i = 1$ are trained to be reconstructed. This is because the inliers exist close to each other and make up a majority regardless of the values of $w_i = 0$ and $w_i = 1$. Second, even if several outliers with $w_i = 1$ are obtained in a training step, such outliers are less likely to be reconstructed than the inliers. This is because, in general, there are few similar samples around the outliers. The w_i of outliers is set to 0 in the successive steps. As a result, the outliers are not reconstructed as the training progresses, and finally only the reconstruction errors (i.e. outlier scores) of the outliers will be large.

4. Evaluations

4.1 Experimental Settings

In this paper, we utilize 15 types of datasets published in Outlier Detection DataSets (ODDS) [11], which is usually used as the benchmark for outlier detection methods. The summary of the datasets is shown in **Table 1**. We normalize the values of all the datasets into the range from -1 to +1 for each dimension. We use the following common network structure and parameters of our proposed method over all the datasets in order to consider unsupervised learning. We apply a fully connected network for the network structure, in which the number of hidden layers is three and the numbers of the neurons are $[D, \sqrt[2]{D}, \sqrt[4]{D}, \sqrt[2]{D}, D]$ from the input to the output, respectively. The decimal points of the neurons are rounded up. We apply activation functions {relu, none, relu, none} from the input to the output, and utilize a Glorot normal distribution [13] as an initial parameter for the network. The

Table 1 Summary of the datasets.

Dataset	Dims.	Samples	Outlier ratio [%]
Arrhythmia	274	452	14.60
Cardio	21	1,831	9.61
Cover	10	286,048	0.96
KDDCUP-Rev [12]	118	121,597	20.00
Mnist	100	7,603	9.21
Musk	166	3,062	3.17
Optdigits	64	5,216	2.88
Satelite	36	6,435	31.64
Satimage-2	36	5,803	1.22
Seismic	28	2,584	6.58
Shuttle	9	49,097	7.15
Smtp	3	95,156	0.03
Speech	400	3,686	1.65
Thyroid	6	3,772	2.47
Vowels	12	1,456	3.43

batch size is $N/50$. We apply *adam* [14] ($\alpha = 0.001$) as an optimization function to train the network, and complete the training when the total number of epochs reaches 400.

We utilize Chainer [15] (ver. 1.21.0) for the network implementation. The value of α_p is 70 for the proposed method. We utilize the AUC (area under the ROC curve) computed using outlier scores as the evaluation indices. Because AUC depends on the initial parameters of the networks, we generate 20 types of initial parameters in each dataset and derive the averaged AUC as the evaluation index.

In this paper, we also apply the following five methods for comparison.

Normal autoencoder (N-AE)

This is a normal AE which utilizes (1) as a loss function. We derive the outlier scores from the reconstruction error e_i, which is obtained from the model after training. We apply network parameters that are equivalent to our proposed method. Incidentally, the parameters of our method in the above are empirically derived so that N-AE achieves a high AUC on average.

RandNet

The number of ensembles is 100, and the other parameters are set to the ones equivalent to the parameters stated in Ref. [6], for example, the structure parameter is 0.5. We evaluate the performance with the averaged AUC over 20 trials.

One-class support vector machine (OC-SVM) [16]

We utilize the OC-SVM implemented in scikit-learn, which is a machine learning library for Python, and use the default values for the parameters.

Local outlier factor (LOF) [3]

We utilize the LOF implemented in scikit-learn, set k to 20 for the k-nearest neighbors and use the default values for the rest of the parameters.

Isolation Forest (IForest) [17]

We utilize the IForest implemented for outlier detection libraries in Python *pyod* [18], and apply the parameters recommended in Ref. [17] for this evaluation. The performance is evaluated with AUC averaged over 20 trials.

4.2 Experimental Results

We show the experimental results in **Table 2**. Here, "Prop." denotes the proposed method, and "Avg. rank" denotes the ranking averaged over all the datasets. In each dataset, each method is given a rank from one to five. In this paper, we refrain from discussing the results of RandNet due to the computational complexity. We utilize the computational environment [*1] for this evaluation. Under this computational environment, the entire evaluation process of RandNet is estimated to takes 40 days [*2]. This process takes much longer than the proposed method and N-AE, which take 415 minutes and 324 minutes, respectively. We can readily say that our proposed method outperforms RandNet in terms of the computational complexity.

Table 2 AUC from our proposed method and the well-known methods.

Dataset	Prop.	N-AE	OC-SVM	LOF	IForest
Arrhythmia	77.60	75.78	78.74	75.86	**81.17**
Cardio	**94.94**	84.64	92.98	63.72	93.26
Cover	83.74	86.03	**91.81**	52.62	87.73
KDD-Rev	**95.87**	13.56	81.39	35.34	77.52
Mnist	**84.33**	82.38	81.99	71.53	81.01
Musk	**100.00**	60.91	93.11	42.71	99.89
Optdigits	**79.58**	73.70	53.39	58.69	75.21
Satelite	**76.81**	63.50	59.94	54.36	69.34
Satimage-2	**99.92**	78.25	98.01	55.14	99.21
Seismic	67.21	**69.99**	59.30	59.05	67.11
Shuttle	98.74	75.42	98.26	52.39	**99.72**
Smtp	88.06	79.58	76.91	90.23	**90.54**
Speech	46.96	46.84	46.39	**50.68**	47.14
Thyroid	92.59	88.37	85.01	80.74	**98.08**
Vowels	90.46	86.03	57.32	**94.42**	75.58
Avg. AUC	**85.15**	71.00	76.97	62.50	82.83
Avg. rank	**1.87**	3.47	3.47	4.07	2.13

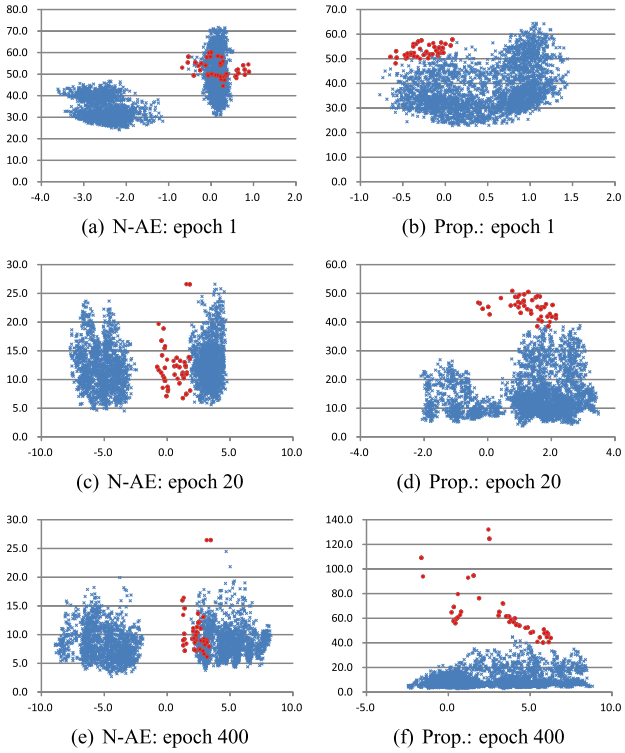
From Table 2, we can see that our proposed method outperforms N-AE in terms of AUC over almost all the datasets. Here, we discuss the reason. First, we focus on one of the datasets, namely "musk", in which our proposed method outperforms N-AE considerably, and show a comparison of their training transitions (epoch = $[1, 20, 400]$) in **Fig. 1**. The vertical axis and horizontal axis in each figure denote a dimensional value in a bottleneck layer and an unweighted reconstruction error e_i, respectively. The blue cross and the red circle denote an inlier sample and an outliner sample, respectively. The figures on the left-hand side show the transitions of N-AE, and the ones on the right-hand side show the transitions of our proposed method.

Next, we focus on the transitions of the reconstruction errors over all the samples on N-AE, where overfitting occurs in the middle of the training process. Although the performance is expected to improve by making the network more robust, it is difficult to identify the structure in the case of unsupervised outlier detection. On the other hand, we can see that the reconstruction errors, which are expressed by e_i of the outlier samples, steadily increase over all the samples as the training step progresses in the proposed method. We can say that our proposed method avoids overfitting by not utilizing highly ranked samples on the reconstruction errors and estimates the outliers with high accuracy even though we utilize the same network as that for N-AE. We can also say that our proposed method avoids overfitting and improves the detection performance for almost all other datasets.

Finally, we focus on the results of "cover" and "seismic" datasets where our method is less accurate than N-AE. These

[*1] OS: ubuntu14.04 64 bit, Memory: 31.3 GB, CPU: 3.50 GHz*12, GPU: GeForce GTX TITAN X, no parallelization.

[*2] = 2 minutes (trianing time for each AE training process) * 100 (the number of ensenbles) * 20 (the number of trials) * 15 (the number of datasets)



(a) N-AE: epoch 1 (b) Prop.: epoch 1

(c) N-AE: epoch 20 (d) Prop.: epoch 20

(e) N-AE: epoch 400 (f) Prop.: epoch 400

Fig. 1 Training transitions of N-AE and our method.

results are thought to be due to a drawback of our method described below. Since our method considers those samples different from the majority as outliers, if some inliers form a small cluster and differ from the majority, our method seemingly identifies them as minority and learns them as outliers which causes false positives. These two datasets seemingly have such a data characteristic. However, the accuracy deterioration is not significant compared to the other methods, and we can say that our proposed method is generally superior in practical uses.

Besides the above, we can also show that our method outperforms the well-known conventional non AE-based methods. This is because we can retain the extraction performance of high-dimensional features, which AE inherently possesses, by applying the concepts of robust estimation. As already mentioned, there is a small difference in the computational duration between N-AE and the proposed method. The proposed method could detect outliers at a lower calculation cost than RandNet, which is also an AE-based method.

5. Summary

In this paper, we propose a novel unsupervised outlier detection method in which we combine an AE-based unsupervised outlier detection method with the concepts of robust estimation. We show that our method requires lower computational cost compared to a conventional AE-based method and achieves a more accurate detection performance than some of the well-known non AE-based outlier detection methods. In the future we will introduce estimation values other than the LTS estimator, and conduct a theoretical analysis of the LTS estimator and a comparison with other AE-based methods.

References

[1] Knox, E.M. and Ng, R.T.: Algorithms for mining distancebased outliers in large datasets, *Proc. International Conference on Very Large Data Bases*, pp.392–403 (1998).

[2] Angiulli, F. and Pizzuti, C.: Fast outlier detection in high dimensional spaces, *European Conference on Principles of Data Mining and Knowledge Discovery*, pp.15–27 (2002).

[3] Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J.: LOF: Identifying density-based local outliers, *ACM Sigmod Record*, Vol.29, No.2, pp.93–104 (2000).

[4] Shyu, M.L., Chen, S.C., Sarinnapakorn, K. and Chang, L.: A novel anomaly detection scheme based on principal component classifier, *MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING* (2003).

[5] Hinton, G.E. and Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks, *science*, Vol.313, No.5786, pp.504–507 (2006).

[6] Chen, J., Sathe, S., Aggarwal, C. and Turaga, D.: Outlier detection with autoencoder ensembles, *Proc. 2017 SIAM International Conference on Data Mining*, pp.90–98 (2017)

[7] Zhou, C. and Paffenroth, R.C.: Anomaly detection with robust deep autoencoders, *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.665–674 (2017)

[8] Rousseeuw, P.J. and Hubert, M.: Robust statistics for outlier detection, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol.1, No.1, pp.73–79 (2011).

[9] Huber, P.J.: Robust estimation of a location parameter, *The annals of mathematical statistics*, Vol.35, No.1, pp.73–101 (1964).

[10] Rousseeuw, P.J.: Least median of squares regression, *Journal of the American statistical association*, Vol.79, No.388, pp.871–880 (1984).

[11] Rayana, S.: ODDS Library (online), available from ⟨http://odds.cs.stonybrook.edu⟩ (accessed 2018-10-12).

[12] Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D. and Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection, *Proc. International Conference on Learning Representations* (2018).

[13] Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, *Proc. 13th International Conference on Artificial Intelligence and Statistics*, pp.249–256 (2010).

[14] Kingma, D. and Ba, J.: Adam: A Method for Stochastic Optimization, arXiv preprint arXiv:1412.6980 (2014).

[15] Chainer: A flexible framework for neural networks (online), available from ⟨https://chainer.org/⟩ (accessed 2018-10-12).

[16] Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J. and Williamson, R.C.: Estimating the support of a high-dimensional distribution, *Neural computation*, Vol.13, No.7, pp.1443–1471 (2001).

[17] Liu, F.T., Ting, K.M. and Zhou, Z.H.: Isolation forest, *Proc. IEEE International Conference on Data Mining*, pp.413–422 (2008).

[18] Zhao, Y.: pyod (online), available from ⟨http://pyod.readthedocs.io/en/latest/⟩ (accessed 2018-10-12).

Yoshinao Ishii received his B.E. and M.E. degrees from Keio University, Yokohama, Japan, in 2009 and 2011 respectively. In 2011, he joined Toyota Central Research & Development Laboratories, Incorporated, Aichi, Japan. His research interests include optimization techniques, automated software testing and machine learning based anomaly detection.

Masaki Takanashi received his B.E., M.E. and Ph.D. degrees from Hokkaido University, Sapporo, Japan, in 2001, 2003 and 2007, respectively. In 2007, he joined Toyota Central Research & Development Laboratories, Incorporated, Aichi, Japan. His research interests include super-resolution techniques, ultrawide-band systems, signal processing for wireless communications and machine learning. He is a member of the IEICE and IEEE.