Unsupervised Anomaly Detection Using Variational Auto-Encoder based Feature Extraction

Rong Yao
Department of Automation
Tsinghua University
Beijing, China
yaor17@mails.tsinghua.edu.cn

Chongdang Liu
Department of Automation
Tsinghua University
Beijing, China
liucd16@mails.tsinghua.edu.cn

Linxuan Zhang
Department of Automation
Tsinghua University
Beijing, China
lxzhang@mails.tsinghua.edu.cn

Peng Peng
Department of Automation
Tsinghua University
Beijing, China
pengp17@mails.tsinghua.edu.cn

Abstract—**Anomaly detection is a key task in Prognostics and Health Management (PHM) system. Specially, in most practical applications, the lack of labels often exists which makes the unsupervised anomaly detection very meaningful. Furthermore, unsupervised anomaly detection is also considered as a challenging task due to the diversity and information-lack of data. Variational Auto-Encoder (VAE) is a stochastic generative model which is designed to reconstruct input data as close as possible. In this paper, VAE is applied to extract valuable features for the unsupervised anomaly detection tasks. Comparison experiments are conducted on KDD CUP 99 dataset and MNIST dataset. Results show that features obtained by VAE can make unsupervised anomaly detection approaches perform better. Auto-Encoder (AE) and Kernel Principle Component Analysis (KPCA) were applied as comparisons. The result demonstrates that VAE gets best performance among them.**

Keywords—unsupervised anomaly detection, VAE, feature extraction

I. Introduction

In data science area, the task of searching "abnormal" samples from a large amount of "normal" samples is known as anomaly detection, outlier detection or novelty detection. With the rapid development of Internet of Things (IoT) and big data technologies, anomaly detection has played an increasingly important role in industries, where defective products or failures can be regarded as abnormal samples. Anomaly detection techniques have also attracted a lot of attention in Condition Based Monitoring (CBM) of Prognostics and Health Management (PHM) system, as they can detect failures more rapidly and accurately. Basically, anomaly detection can be categorized into three classes depending on the labels' availability in the dataset [1]:

Supervised anomaly detection uses both fully labeled training data and test data, and tries to get a model to classify the normal and abnormal instances in training data. Then, the model will be applied to the test data. Such problem is almost equivalent to an unbalanced classification problem.

Semi-supervised anomaly detection trains a model with only normal instances and get a standard "normal" model so that the characteristics of normal instances should have can be pretty learned. Then data deviated greatly from the "normal" model, which can be calculated by high scores, are classified as anomalies.

Unsupervised anomaly detection deals with unlabeled data mainly and doesn't have a distinction between training data and test data. Specially, it has a strong assumption that the detect data consists of few anomalies and a majority of normal instances. The models can learn the intrinsic information of data. And the instances that deviate from the major instances will be reported as anomalies.

Within above three classes, unsupervised anomaly detection is considered as the most difficult task due to the diversity and information-lack of data. Furthermore, in practical applications, it is worth emphasizing that labeling is an expensive and manpower-required work [2]. Therefore, the lack of labels often exists in most certain cases which makes the unsupervised anomaly detection techniques very meaningful and applicable. This paper will focus on the unsupervised anomaly detection problems.

Essentially speaking, whether labels available or not, anomaly detection problem can be considered as a special unbalanced classification problem. It is well known that the data quality has large effects to the results of data-driven approaches. Better features usually mean better results. But in most cases, data quality can't be guaranteed because of the existence of irrelevant and redundant information in the data. The useless information will not only diminishes the detection accuracy but also increases the processing time [3]. Under this condition, dimension reduction is thought to have the ability to improve anomaly detection accuracy if it abandons useless information and remains the influential information as much as possible.

The rest of this paper is organized as follows: Section II introduces the related work of unsupervised anomaly detection. Section III describes the theory and feature representation ability of Variational Auto-Encoder (VAE), as well as the proposed approach. Section VI displays the experiment results and related discussion. Section V presents overall conclusions and future works.

II. Related Works

In recent years, a variety of methods have been applied to unsupervised anomaly detection problem. Typically, the related works include but not limit to the following categories: (1) Nearest-neighbor based, K-nearest-neighbor (KNN) [4, 5] focuses on distance to neighbors, and Local Outlier Factor (LOF) [6] cares about the local outliers. (2) Clustering based,

978-1-5386-8357-6/19/$31.00 ©2019 IEEE

978-1-5386-8357-6/19/$31.00 ©2019 IEEE

Cluster-Based Local Outlier Factor (CBLOF) [7] performs clustering first and then estimates the local density of data, and CD-trees [8] declares the sparse clusters as anomalies directly. (3) Statistical based, several statistics tests such as student's t test [9] and χ^2 test [10] have been applied for anomaly detection. (4) Subspace based, Principle Component Analysis (PCA) projects data into subspace where anomalous instances are more obvious. (5) Classifier based, One-Class Support Vector Machine (OCSVM) [11] trains a classifier to distinguish abnormal data from normal data. (6) Hybrid methods, mix the above approaches [12].

Moreover, with the rapid development of big data, some stochastic process and deep learning methods have been more and more popular in anomaly detection: Pang et.al [13] employ Guassian Process (GP) to anomaly detection of data stream. Nanduri et.al [14] use Recurrent Neural Network (RNN) to provide anomaly detection of time-series data and Feng et.al [15] outline a Long Short-Term Memory (LSTM) anomaly detection method for industrial control systems. Yao et.al [16] propose a Convolutional Neural Network (CNN) structure to the time-delayed attacks detection. In addition to these, the generative models such as Generative Adversarial Network (GAN) [17] and VAE [18] are also employed for unsupervised anomaly detection, since the generative models mainly learn from major data so the losses of GAN and VAE can be used as the anomaly scores.

As mentioned in Section I, in most cases, origin data of actual system, especially the high-dimension data, often contains some irrelevant and redundant information. During above anomaly detection methods, the existence of these useless information will affect the detection accuracy. Therefore, many dimension reduction approaches have been applied to get better detection results in anomaly detection, including feature extraction and feature selection. Gan et.al [19] use Partial Least Square (PLS) for feature extraction and improve the anomaly intrusion detection in large-scale data. Zhao et.al [20] combines kernel Principal Component Analysis and kernel Independent Component Analysis (ICA) as feature extraction method for anomaly detection in hyperspectral imagery. Besides, some feature selection approaches [21, 22] have verified to be helpful to anomaly detection.

Specially, VAE has been demonstrated to be a new non-linear feature extraction method, which can represent the structure of data effectively and stably [23]. Meanwhile, in the traditional unsupervised anomaly detection, the feature quality is a very important factor. In this paper, we present an framework that apply VAE to extract valuable features for several traditional unsupervised anomaly detection tasks. Results show that most of the unsupervised anomaly detection approaches perform better with features obtained by VAE.

III. Methodologies

A. Variational Auto-Encoder

As we know, VAE is developed from Auto-Encoder (AE) which is trained to reconstruct the input data. Auto-Encoder consists of an encoder and a decoder, the encoder map the input data to a latent variable while the decoder reconstruct the input data with the latent variable. In comparison with VAE, AE use a simple network layer to denote the latent variable. Thus, Auto-Encoder is more considered as a deterministic model which doesn't have the ability of generating new samples. To overcome the weakness of Auto-Encoder, VAE adds a variational constrain that the latent variable z is subject to a normal distribution and the decoder starts with sampling from the distribution. Accordingly, VAE can map the training data to a normal distribution and generate new samples from the distribution. The structure of VAE is shown in Fig. 1.

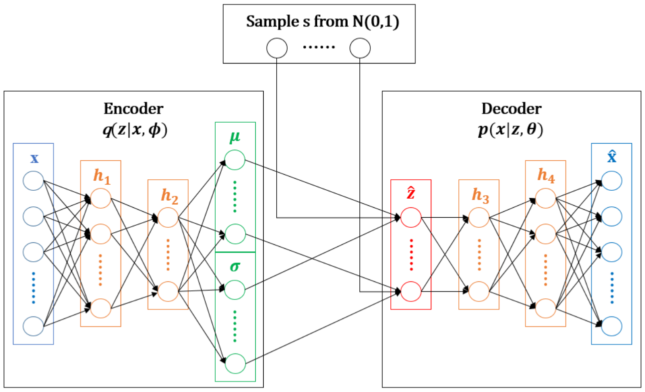

Fig. 1. Structure of Variational Auto-Encoder

In Fig. 1, $x/\hat{x}$ denote the input/reconstruct data, μ/σ^2 denote guassian distribution's mean/variance of latent variable z, $\hat{z}$ represents the sample from $N(\mu,\sigma^2)$, and h represents the hidden layer in the network. The main purpose of VAE is also to train a network which reconstructs its own input data x with $\hat{x}$ as close as possible by minimizing:

$$L(x,\hat{x}) = \| x - \hat{x} \| \tag{1}$$

As shown in Fig. 1, VAE is a directed probabilistic graphical model (DPGM) which also consists of an encoder and a decoder. Different from Auto-Encoder, VAE is a stochastic generative model in which encoder and decoder are given by probabilistic function $q(z|x,\phi)$ and $p(x|z,\theta)$. Specifically speaking, $q(z|x,\phi)$ is the approximate posterior called adversarial model or encoder, while $p(x|z,\theta)$ is the likelihood of x given z called generative model or decoder. After fully training, the encoder is thought to approximate the posterior distribution very well, and be able to map the input data x to the latent space. Different from Auto-Encoder, VAE doesn't represent the latent space with simple value, but maps input data to a stochastic variable z. In this paper, z is subject to guassian distribution which is determined by its mean μ and variance σ^2. After encoding, $\hat{z}$ can be sampled from $N(\mu,\sigma^2)$, and the decoder is trained to output the reconstruct result with $\hat{z}$. The marginal likelihood of the data points is given by:

$$\log p_\theta(x) = \sum_{i=1}^{n} \log p_\theta\left(x^{(i)}\right) \tag{2}$$

where n denotes the number of data, and θ is the parameter of decoder network. VAE is trained to get network parameters which maximizing the marginal likelihood. The equation (2) can be rewritten for each individual data point $x^{(i)}$ [17]:

$$\log p_\theta\left(x^{(i)}\right) = KL\left(q\left(z\middle|x^{(i)},\phi\right), p\left(z\middle|x^{(i)},\theta\right)\right) + L\left(\theta,\phi;x^{(i)}\right) \tag{3}$$

where $p(z|x,\phi)$ represents the true posterior distribution, $KL(,)$ is KL divergence function defined to calculate the distance of two distributions, and:

$$L\left(\theta,\phi;x^{(i)}\right) = -KL(q\left(z\middle|x^{(i)},\phi\right), p(z,\theta)) + E_{q\left(z\middle|x^{(i)},\phi\right)}\left[\log p\left(x^{(i)}\middle|z,\theta\right)\right] \tag{4}$$

Due to the non-negativity of KL function, equation (3) can be rewritten as:

$$\log p_\theta\left(x^{(i)}\right) \geq -KL(q(z|x^{(i)},\phi), p(z,\theta)) + E_{q\left(z\middle|x^{(i)},\phi\right)}\left[\log p\left(x^{(i)}\middle|z,\theta\right)\right] \tag{5}$$

The first term on the right side of inequality (5) is the KL divergence between the approximate posterior and prior of latent variable z. The second term on the right side of inequality (5) can be considered as the expected reconstruction error. The KL divergence can be obtained straightly if the prior $p(z,\theta)$ is subject to a normal distribution $N(\mu,\sigma^2)$. However, the second term requires Monte Carlo estimate of the expectation. Before decoding, we need to sample $\hat{z}$ from unknown distribution $q\left(z\middle|x^{(i)},\phi\right)$. And it's hard to take the derivative of the sampling process, which makes the network unable to use gradient-based training. To solve it, Kingma et.al in [24] introduced a reparameterization of z, called reparameterization trick:

$$z = \mu + \sigma s \tag{6}$$

where, $s \in N(0,I)$ represents the sample from the standard normal distribution, μ and σ^2 denote the mean and variance of latent variable distribution obtained by the encoder of VAE. Through reparameterization trick, as shown in Fig.1, VAE set the sampling process independent of the network, so the network can be trained through gradient-based methods directly. The final objective function of VAE is expressed as:

$$L_{VAE} = \sum\nolimits_{i=1}^{n} (\frac{1}{L}\sum\nolimits_{j=1}^{J} \log p\left(x^{(i)}|z^{(i,j)},\theta\right) - KL(q\left(z\middle|x^{(i)},\phi\right), N(0,I)) \tag{7}$$

where, $z^{(i,j)}$ is calculated by equation (6) with the jth sample $s^{(j)}$ from the standard normal distribution for the ith data point. Specially, by changing the train object from the likelihood function to equation (7), parameter ϕ of encoder is optimized along with parameter θ of decoder. In summary, VAE is a special generative model trying to reconstruct its input data with latent variable. In its training process, VAE needs only data which makes it a good unsupervised method to fulfill our requirement.

B. *Feature extraction via Variational Auto-Encoder*

In VAE, input data is firstly projected to a stochastic distribution of the latent variable through encoder, then the latent variable is sampled from the distribution, and the decoder will reconstruct the input data based on the latent variable. In other words, a VAE network has trained to reconstruct training data accurately also means that the latent variable has included enough information of input data. Furthermore, in most cases, the dimension of the latent variable is smaller than input data, so a low-dimension representation of input data can be obtained by the encoder. In another word, VAE can be applied to some data processing work such as feature extraction. In the previous work [23], it has been demonstrated that VAE has good feature representation ability in classification tasks. As a new non-linear feature representation method, because of its reconstruction training purpose, features obtained by VAE can not only remain key information of data, but also remove irrelevant and redundant information. Since VAE has the probabilistic structure instead of determined structure, compared with AE, features of VAE is thought to contain more information.

In this paper, we use the feature representation ability of VAE in anomaly detection tasks and test it with several traditional approaches.

C. *Proposed Unsupervised Anomaly Detection Approach*

As is discussed in Section 3.2, VAE has been demonstrated to have the ability of feature representation. Based on this, a novel unsupervised anomaly detection approach is proposed, which is given in Fig. 2 :

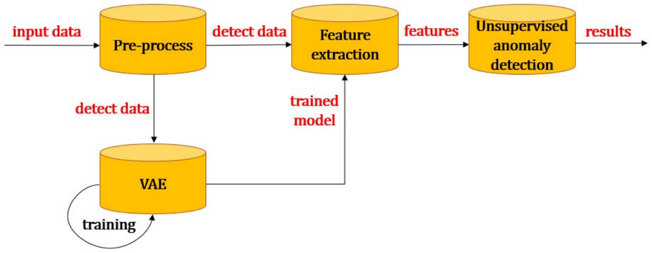

Fig. 2. schematic of proposed unsupervised anomaly detection method

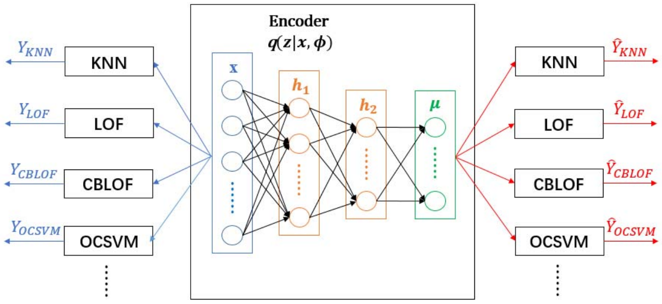

Fig. 3. Feature extraction structure of VAE in unsupervised anomaly detection

To begin with, data preprocess is conducted to clean data and get the detect data. Then, VAE is trained to learn the property of data, and the encoder of trained model can be used as the feature extraction model. To extract the best form of latent variable, through experiments, we employ μ, the expectation of latent variable's guassian distribution $N(\mu,\sigma^2)$, as extracted features.

With the encoder model, the detect data can be represented with lower-dimension features efficiently and stably. To verify the effectiveness of the extracted features, four traditional unsupervised anomaly detection methods are performed to detect the features represented by our model, and the results can be compared with the same results obtained by the original detect data. The structure of the model is shown in Fig. 3.

In VAE's objective function equation (7), we have known that the KL divergence part represents the distance between the approximate posterior distribution $q(z|x,\phi)$ and the real posterior distribution $p(z|x,\phi)$. While the other part considers more about the reconstruct error. Taking into account our particular structure, the ability of the encoder to map data to the latent variable should get more attention. So, a factor to adjust the weight of two parts of equation (7) is added :

$$L_{VAE} = \alpha \sum_{i=1}^{n} (\frac{1}{L}\sum_{j=1}^{J} \log p\left(x^{(i)}|z^{(i,j)},\theta\right) \\ -KL\left(q\left(z|x^{(i)},\phi\right), N(0,I)\right) \qquad (8)$$

where, $\alpha \in (0,1]$ is a weight factor. It means that during training, the model pay more attention to keeping the form of the approximate posterior $q(z|x,\phi)$ than the reconstruct error.

D. Performance Evaluation

In order to indicate whether an instance is an anomaly or not, there are mainly two types of output in anomaly detection approaches: *label* and *score*. Label type usually uses 1 to denote anomalies while uses 0 to denote normal instances, and the output will be either 1 or 0. In general, supervised anomaly detection approaches often use labels as outputs because they can be taken as two-class classification problems. Score type is a confidence value between 0 and 1 which reflects the possibility that an instance is abnormal. In semi-supervised or unsupervised anomaly detection approaches, scores are used more commonly because they contain more information than simple labels. In practical applications, anomaly score often indicates the abnormal possibility of instance, which make it more applicable. In this paper, all four approaches are trained to output anomaly scores so the anomalies can be detected through the scores. Generally, several instances with top scores are usually reported as anomalies. Typically, there are four report results in anomaly detection problem as shown in Table I.

TABLE I. FOUR REPROT RESULTS IN ANOMALY DETECTION

	normal	***abnormal***
report normal	True Negative (TN)	False Negative (FN)
report abnormal	False Positive (FP)	True Positive (TP)

As shown in Table I, in an unsupervised anomaly detection problem, TP/FP are the quantities of correct/wrong anomaly judgement instances while TN/FN are corresponsive to the correct/wrong normal judgement instances. Based on these results, two factors to evaluate the results are proposed:

$$FPR = \frac{FP}{FP + TN} \qquad (9)$$

$$TPR = \frac{TP}{TP + FN} \qquad (10)$$

Typically, True Positive Rate (TPR) denotes the ratio of correct anomaly judgement instances in anomaly judgement class to all abnormal instances, similarly, False Positive Rate (FPR) is the proportion of correct normal judgement instances in all normal instances. Obviously, to get the biggest TPR value, we only need to report all instances as anomalies, but FPR will also become 1 in this case. Ideally, a high-TPR and low-FPR model is expected. As a comprehensive standard, Receiver Operator Characteristic (ROC) curves are finally used as an evaluation method in common. In ROC curves, x-axis means FPR and y-axis means TPR. Basically, the area under the curve (AUC) is often adopted as a detection performance measurement. With the number of reported anomalies increases, both TPR and FPR value will increase to 1, but we still look forward to finding a model needs only few report-anomalies to get a high TPR.

IV. EXPERRIMENTS

In order to verify the feature representation ability of VAE in unsupervised anomaly detection, we test our approach in KDD CUP 99 dataset and MNIST dataset. With the feature extraction of VAE, four unsupervised anomaly detection approaches are used to show the effect of our approach. Moreover, to explain the performance of VAE, two other feature extraction approaches are applied as comparisons: Auto-Encoder and KPCA.

A. KDD CUP 99 dataset

1) Dataset description

KDD CUP 99 dataset is a basic benchmark of network intrusion field. The dataset is popular in intrusion detection classification. What is worth mentioning, in original KDD99 dataset, many attacks (anomalies) are defined as a set of instances so that the attacks are hard to be detect without labels. To use this most popular dataset in unsupervised situation, Goldstein et.al in [25] adopted some processing methods including using HTTP traffic data only and limiting DoS traffic from the dataset. Through these processing methods, a dataset with 620089 instances and 0.17% anomalies is available for our problem, and each instance includes 29 variables. In our experiments, we take 900 normal instances and 100 abnormal instances from this dataset randomly, perform multiple experiments and compare their average performance.

2) Performance of Features of VAE in Unsupervised Anomaly Detection

In KDD99, different attributes usually have different ranges. As a preprocessing, a Min-Max normalization method is applied to our dataset. After preprocessing, the unsupervised anomaly detection based-on feature extraction via VAE can be employed to the detect dataset. In the training of VAE, the feature-dimension of KDD99 dataset is set to 15, nodes number of all 4 hidden layers (each 2 in encoder and decoder) are set to 80, and the activation functions in two hidden layers and latent variable are set to elu, tanh and sigmoid correspondingly. Specially, the weight factor α is set to 0.8 so that the model will still pay a lot of attention to the reconstruction error. Once VAE has been trained, the network is able to map the input data to a lower-dimension space and reconstruct the input data with a lower-dimension variable. Then encoder of trained VAE is taken as a feature extraction model of detect dataset. After getting the features, the last step is to employ the features generated by the encoder to several tradition unsupervised anomaly detection approaches. To show the universality of our method, we test original data at KNN, LOF, CBLOF, and OCSVM, and then test features of VAE at the same approaches. As an example, the ROC curves of KDD99 are shown in Fig. 4.

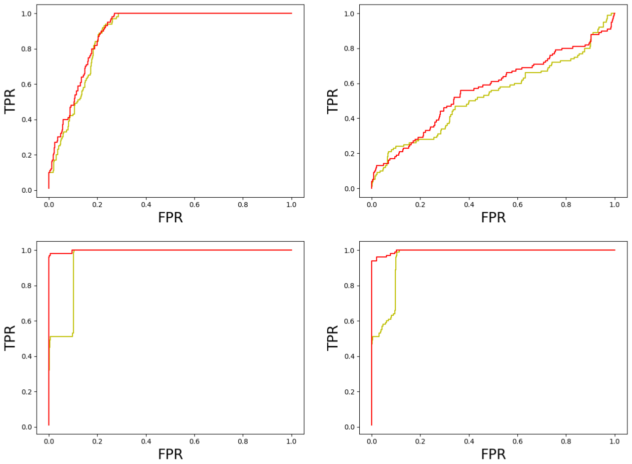

Fig. 4. ROC curves of KDD99 using original data (yellow) and features of VAE (red), from left to right, top to bottom: KNN, LOF, CBLOF, OCSVM

As shown in Fig. 4, obviously, the curves illustrate that our approach with features of VAE makes all four approaches perform better.

3) *Comparison*

In order to show the superiority of our approach, in this section, we employ other two feature representation methods as comparisons of VAE: KPCA and Auto-Encoder. In this paper, the KPCA with kernel function Radial Basis Function (RBF) is employed, while the Auto-Encoder is the basis of VAE, and both KPCA, AE and VAE are used as nonlinear feature extraction methods. In order to maintain the consistency of the experiment, the feature dimensions obtained by above three methods are all set to the same number. Additionally, to show the different improvements of different feature representation approaches, for convenience of visualization, three methods' results mapping KDD99 data to 2-dimension features are shown in Fig. 5.

In Fig. 5, red points denote anomalies while green points denote normal data. In results of KPCA and AE, There are some areas where the anomaly point is not significantly deviated from the normal point. But for same results in VAE, the difference between the abnormal point and the normal point is more obvious. In other words, our model can map detect data into feature space where anomalies are more obvious,

As a common detection performance measure of anomaly detection research, AUC values of different methods are calculated. The results of KDD99 are shown in Table II and Fig. 6 after 100 times running. In experiment of KDD99 dataset, results of four basic anomaly detection approaches and their improve methods with three feature extraction methods are given in Table II. In addition to the best and average AUC performance, the average AUC increments under different features are also calculated. Meanwhile, Fig. 6 is the AUC histogram of different approaches in KDD99. As we can see, features of VAE help three approaches perform better, among VAE, AE and KPCA, VAE is the best-performing feature extraction method in our experiments.

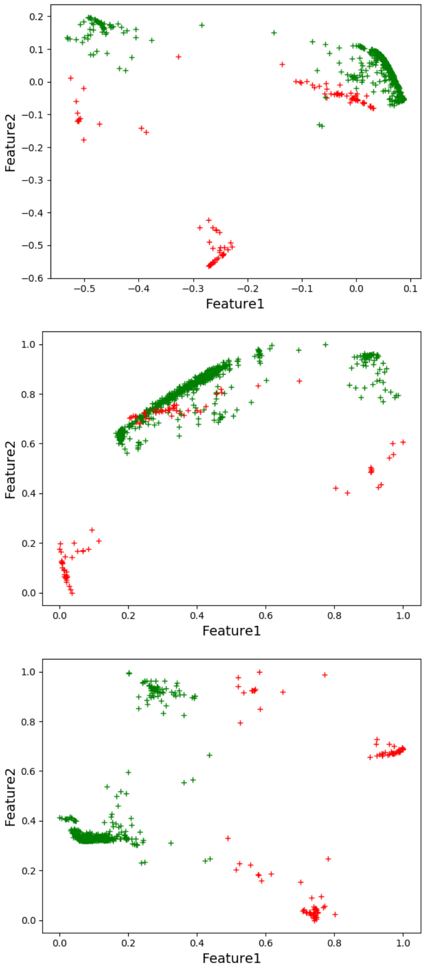

Fig. 5. three methods' results mapping KDD99 data to 2-dimension features (normal: green; abnormal: red), from top to bottom: KPCA, AE, VAE

TABLE II. PERFORMANCE FOR DIFFERENT APPROACHES IN KDD99

Approaches		*Base*	*VAE*	*AE*	*KPCA*
KNN	Best	0.935	0.952	0.902	0.926
	Average	0.885	0.904	0.846	0.867
	Increment	0.000	0.021	-0.039	-0.018
LOF	Best	0.652	0.611	0.664	0.639
	Average	0.578	0.554	0.574	0.567
	Increment	0.000	-0.024	-0.004	-0.012
CBLOF	Best	0.998	**0.999**	0.996	0.998
	Average	0.962	0.979	0.963	0.958
	Increment	0.000	0.017	0.001	-0.004
OCSVM	Best	0.970	0.998	0.995	0.970
	Average	0.959	**0.990**	0.967	0.958
	Increment	0.000	**0.031**	0.008	-0.001

[a] Base: performance of basic algorithms with origin detect data

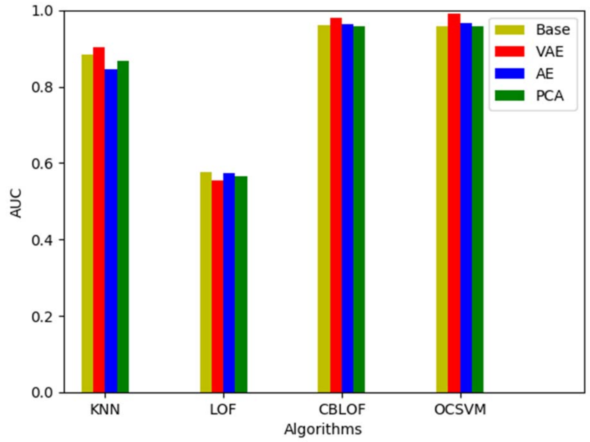

Fig. 6. AUC histogram of different approaches in KDD99

B. MNIST dataset

MNIST dataset of handwritten digits was created from NIST database in [26]. Both train set and test set are composed of 28×28 gray level images, and each image corresponds to a handwritten number 0~9. In our study, we only take the test set which consists of 10000 images. Specifically, the written images of '7' are taken as normal set, and other numbers' images are considered as anomalies. However, as mentioned above, unsupervised anomaly detection has a strong assumption that the detect data should consist of few anomalies and lots of normal instances, so the models can identify anomalies deviated from the majority. To make this assumption satisfied, a pre-processing work is adopted as shown in Fig. 7.

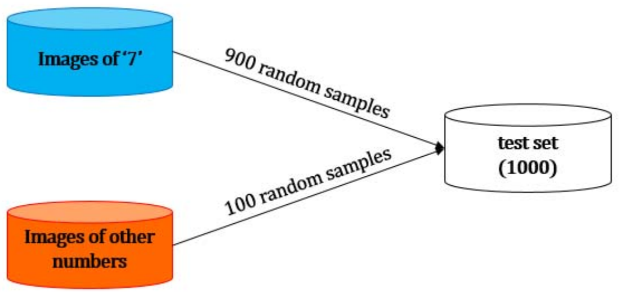

Fig. 7. Pre-process work in MNIST dataset

To satisfy the unbalanced requirement, 900 normal instances and 100 anomalies are sampled as final MNIST test set after extending to 784-dimensional vectors. The weight factor α in MNIST is set to 0.01 which means the model pay more attention to keeping the form of the approximate posterior, and the latent variable and 4 hidden layers' dimension set to 400 and 3000. Similar AUC results of MNIST are shown in Table III and Fig. 8. As we can see, features of VAE not only help all four traditional anomaly detection approaches get better AUC results, but also perform better than features of KPCA and AE.

TABLE III. PERFORMANCE FOR DIFFERENT APPROACHES IN MNIST

Approaches		***Base***	***VAE***	***AE***	***KPCA***
KNN	Best	0.965	**0.973**	0.965	0.966
	Average	0.936	**0.952**	0.937	0.936
	Increment	0.000	0.016	0.001	0.000
LOF	Best	0.905	0.959	0.922	0.907
	Average	0.860	0.921	0.877	0.864
	Increment	0.000	0.061	0.017	0.004
CBLOF	Best	0.936	0.970	0.944	0.939
	Average	0.859	0.945	0.884	0.867
	Increment	0.000	**0.086**	0.025	0.008
OCSVM	Best	0.923	0.971	0.944	0.924
	Average	0.897	0.939	0.916	0.897
	Increment	0.000	0.042	0.019	0.000

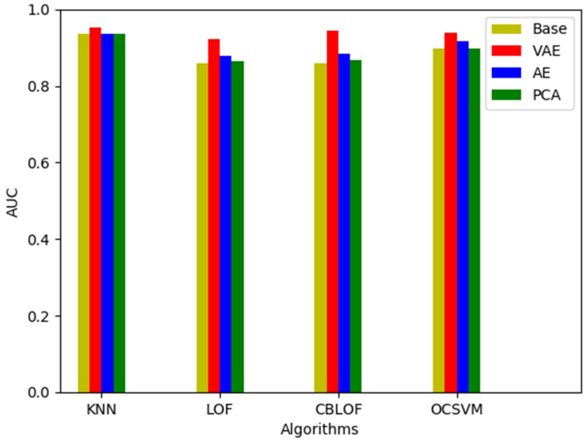

Fig. 8. AUC histogram of different approaches in MNIST

C. Discussion

Similar results are obtained in both KDD99 and MNIST dataset. In general, among VAE, AE and KPCA, VAE is the best-performing feature extraction method in our experiments. Specifically, VAE gets the both largest average and best AUC values. As two non-linear network-based feature extraction methods, both VAE and AE make some improvements to all four unsupervised anomaly detection methods. But in our experiments, features of KPCA even play an opposite role and make the result even worse. That's because KPCA is a feature reduction model which aims to discard the unimportant components and leave the principle components. Even kernel functions are applied in our experiment, it's still unavoidable that KPCA has its own information loss which affects the result.

Among three feature extraction methods, features of VAE have biggest influence in unsupervised anomaly detection problem. Even some methods (LOF and CBLOF in MNIST) with poor performance have performed very well through our approach. As previously analyzed, different from AE, VAE is a stochastic generative model so the encoder of VAE is able to map the input data to a stochastic distribution. Compared with the determined value obtained by AE, VAE can learn the expression of data better, correspondingly, better result can be obtained.

V. Conclusion and future work

In this paper, we employ the feature representation ability of VAE in the unsupervised anomaly detection. Our approach takes only encoder part of VAE as a feature extraction model which maps the input data to low-dimension latent space used to anomaly detection. In the experiment, our proposed approach is applied to four traditional unsupervised anomaly detection approaches. By comparing the detection results with VAE's features and origin detect data, it has been demonstrated that VAE makes the detection results performs better. As comparisons, we also adopt Auto-Encoder and KPCA to obtain their features to the same dimension. The final results show that VAE is a more efficient method in our problem. Our future work will mainly focus on following:

(1) Make use of the decoder: In this paper, we only take the trained encoder part of VAE for the anomaly detection. In future, we can employ the decoder part to reconstruct more normal or anomaly data and improve our results.

(2) More deep learning methods such as generative adversarial network and probabilistic graphical model can be also applied in unsupervised anomaly detection problems.

References

[1] M. Goldstein and S. Uchida, "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data," Plos One, vol. 11, no. 4, Apr 19 2016, Art no. e0152173.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," (in English), Acm Computing Surveys, Review vol. 41, no. 3, p. 58, 2009, Art no. 15.

[3] N. Y. Almusallam, Z. Tari, P. Bertok, and A. Y. Zomaya, "Dimensionality Reduction for Intrusion Detection Systems in Multi-data Streams-A Review and Proposal of Unsupervised Feature Selection Scheme," Emergent Computation: a Festschrift for Selim G. Akl, vol. 24, pp. 467-487, 2017 2017.

[4] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in Principles of Data Mining and Knowledge Discovery. 6th European Conference, PKDD 2002. Proceedings, T. Elomaa, H. Mannila, and H. Toivonen, Eds.: Springer-Verlag, 2002, pp. 15-26.

[5] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," (in English), Sigmod Record, Article; Proceedings Paper vol. 29, no. 2, pp. 427-438, Jun 2000.

[6] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," (in English), Sigmod Record, Article; Proceedings Paper vol. 29, no. 2, pp. 93-104, Jun 2000.

[7] Z. Y. He, X. F. Xu, and S. C. Deng, "Discovering cluster-based local outliers," (in English), Pattern Recognition Letters, Article vol. 24, no. 9-10, pp. 1641-1650, Jun 2003.

[8] H. L. Sun, Y. B. Bao, F. X. Zhao, G. Yu, and D. L. Wang, "CD-Trees: An efficient index structure for outlier detection," (in English), Advances in Web-Age Information Management: Proceedings, Article; Proceedings Paper vol. 3129, pp. 600-609, 2004.

[9] C. Surace and K. Worden, "A novelty detection method to diagnose damage in structures: An application to an offshore platform," in 8th International Offshore and Polar Engineering Conference (ISOPE-98), Montreal, Canada, 1998, CUPERTINO: International Society Offshore& Polar Engineers, 1998, pp. 64-70.

[10] N. Ye and Q. Chen, "An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems," (in English), Quality and Reliability Engineering International, Article vol. 17, no. 2, pp. 105-112, Mar-Apr 2001.

[11] S. Saxena, K. Myeongsu, X. Yinjiao, and M. Pecht, "Anomaly Detection During Lithium-ion Battery Qualification Testing," in 2018 IEEE International Conference on Prognostics and Health Management: Ieee, 2018, pp. 6 pp.-6 pp.

[12] J. Camacho, A. Perez-Villegas, P. Garcia-Teodoro, and G. Macia-Fernandez, "PCA-based multivariate statistical network monitoring for anomaly detection," (in English), Computers & Security, Article vol. 59, pp. 118-137, Jun 2016.

[13] P. Jingyue, L. Datong, L. Haitao, P. Yu, and P. Xiyuan, "Anomaly detection based on data stream monitoring and prediction with improved Gaussian process regression algorithm," 2014 IEEE International Conference on Prognostics and Health Management (PHM), pp. 7 pp.-7 pp., 2014 2014.

[14] A. Nanduri, L. Sherry, and Ieee, "ANOMALY DETECTION IN AIRCRAFT DATA USING RECURRENT NEURAL NETWORKS (RNN)," in Integrated Communications Navigation and Surveillance Conference (ICNS), Herndon, VA, 2016, NEW YORK: Ieee, 2016.

[15] C. Feng, T. Li, D. Chana, and Ieee, "Multi-level Anomaly Detection in Industrial Control Systems via Package Signatures and LSTM networks," in 47th IEEE/IFIP Annual International Conference on Dependable Systems and Networks (DSN), Denver, CO, 2017, 2017, pp. 261-272.

[16] Y. Yao, Y. Wei, F. X. Gao, and Y. Ge, "Anomaly intrusion detection approach using hybrid MLP/CNN neural network," (in English), Isda 2006: Sixth International Conference on Intelligent Systems Design and Applications, Vol 2, Proceedings Paper pp. 1095-1100, 2006.

[17] T. Schlegl, P. Seebock, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," in Information Processing in Medical Imaging. 25th International Conference, IPMI 2017. Proceedings: LNCS 10265, M. Niethammer et al., Eds.: Springer International Publishing, 2017, pp. 146-157.

[18] D. Kim et al., "Squeezed Convolutional Variational AutoEncoder for Unsupervised Anomaly Detection in Edge Device Industrial Internet of Things," (in English), Conference Proceedings of 2018 International Conference on Information and Computer Technologies (Icict), Proceedings Paper pp. 67-71, 2018.

[19] X. S. Gan, J. S. Duanmu, J. F. Wang, and W. Cong, "Anomaly intrusion detection based on PLS feature extraction and core vector machine," (in English), Knowledge-Based Systems, Article vol. 40, pp. 1-6, Mar 2013.

[20] C. H. Zhao, Y. L. Wang, and F. Mei, "Kernel ICA Feature Extraction for Anomaly Detection in Hyperspectral Imagery," (in English), Chinese Journal of Electronics, Article vol. 21, no. 2, pp. 265-269, Apr 2012.

[21] C. Pascoal et al., "Robust Feature Selection and Robust PCA for Internet Traffic Anomaly Detection," 2012 Proceedings Ieee Infocom, pp. 1755-1763, 2012 2012.

[22] U. Ravale, N. Marathe, and P. Padiya, "Feature Selection Based Hybrid Anomaly Intrusion Detection System Using K Means and RBF Kernel Function," International Conference on Advanced Computing Technologies and Applications (Icacta), vol. 45, pp. 428-435, 2015 2015.

[23] D. Chenxi, X. Tengfei, and W. Cong, "The Feature Representation Ability of Variational Autoencoder," in 2018 IEEE Third International Conference on Data Science in Cyberspace: IEEE Computer Society, 2018, pp. 680-684.

[24] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," ArXiv e-prints, vol. 1312, Accessed on: December 1, 2013[Online]. Available: http://adsabs.harvard.edu/abs/2013arXiv1312.6114K

[25] M. Goldstein, "Unsupervised Anomaly Detection Benchmark," DRAFT VERSION ed: Harvard Dataverse, 2015.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," (in English), Proceedings of the Ieee, Review vol. 86, no. 11, pp. 2278-2324, Nov 1998.