EE357 Midterm Report

Haoxuan Wang, Hao Xia
Course: EE357 - Computer Network

***Abstract*— In this report, we conclude what we have done so far for the course project. Our project's topic is: Application of adversarial learning in financial crisis, which is a combination of new machine learning techniques with traditional financial problems. Our group believes that with the rise of artificial intelligence, dealing with data tasks such as financial risk control is becoming more and more effective and effecient. In the following sections, we state the survey we have done so far and list out assignments for the next three weeks.**

I. Introduction

The topic of financial risk control is quite general and ambiguous, we focus on the perspective of detecting abnormal situations from a large amount of data. This particular problem is called anomaly detection in machine learning and aims to find anomalies in enumerous amount of normal data. In math, it is described as: We have a dataset of $x^1, x^2, ..., x^n$ and we want to determine whether x_{test} is anomalous. Thus, a model $p(x)$ is constructed from the training dataset, and decision is made by:

$$\begin{aligned} p(x_{test}) < \epsilon \rightarrow Anomaly \\ p(x_{test}) \geq \epsilon \rightarrow Normal \end{aligned} \tag{1}$$

where ϵ is a manually set threshold.

Briefly saying, we are trying to find a distribution that can best represent the normal data, and any data that is less likely to follow this distribution is considered to be abnormal. Traditional methods of machine techniques include constructing multivariate Gaussian distributions and reconstructing features. However, with the rise of adversarial learning in modeling data distribution, anomaly detection has achieved a rapid increase in performance. In the following sections, we introduce the current works for anomaly detection as well as what we intend to do in the next few weeks.

II. Related Works

A. Unsupervised Anomaly Detection

Originally, anomaly detection is a kind of unsupervised problem. In this section, we introduce works does not labeled data to do anomaly detection. While supervised methods suffer from the hassle of algorithm picking/parameter tuning, heavy reliance on labels[1], unsupervised methods is able to model the distribution of data samples more properly.

1) Unsupervised Anomaly Detection via Variational Auto-Encoderfor Seasonal KPIs in Web Applications: This paper proposes a method called Donut, which is an algorithm based on Variational Auto-Encoder (a representative deep generative model) with solid theoretical explanation. The researchers from Alibaba faces the problem of dealing with seasonal KPIs (e.g., Page Views, number of online users) to detect anomalies and trigger timely troubleshooting/mitigation[1].

The paper thinks that seasonal KPIs are with local variations and Gaussian noises, thus local variations need to be properly handled. To correctly discriminate between seasonal features and anomaly data samples, variational auto-encoder is introduced to model the distribution of the original data samples. To note here, the modeling include abnormal data as well, which is different from traditionally what we see of anomaly detection. The paper also proved this. To conclude, the paper introduces these new techniques and made these contributions:

- M-ELBO: Allowing the use of abnormal data. And reach the conclusion that it would not be a good practice to train a VAE for anomaly detection using only normal data, although it seems natural for a generative model.
- Missing data injection: A data augmentation method. This method intend to make up for

the missing data points in the original dataset and improve the performance of M-ELBO. However, this does not improve the F-score in practice, which might be due to the fact that injecting missing data points introduces more randomness.
- MCMC imputation: Used in detection, for dealing with abnormal points. Detection is also a vital part in anomaly detection. Using MCMC never hurts the performance, and sometimes obtains significant performance overother settings.

For the studied KPIs, the F-score is able to reach a range from 0.75 to 0.90. Which is of great performance when doing anomaly detection.

2) DOPING: Generative Data Augmentation for Unsupervised Anomaly Detection with GAN: This paper uses GAN to do anomaly detection. GAN is a powerful method of adversarial learning, and many works have been done to do anomaly detection with GAN. We recommend this paper due to its new perspective of oversampling infrequent normal samples instead of modeling the distribution of normal samples or augmenting abnormal samples. The researchers think that infrequent samples are the main cause of high false positive rates and the difficulty in defining a distribution that contains all of the normal data.

Thus, the paper proposes DOPING, which is an AAE (adversarial autoencoder) that can be applied to any dataset or anomaly detection algorithm. DOPING is often used with other anomaly detection algorithms, since it cannot finish the whole job by itself. The AAE uses multivariate Gaussian distribution as the prior distribution, and this can be changed due to different requirements. The datasets used in this experiment are various, including Mammography, Thyroid, Lymphography, Cardiotocography. Large number of datasets are used for testing the robustness of the DOPING algorithm. When using DOPING across different datasets, some achieve significant performance while some achive consistent performance. Thus, adopting this method still depends on the original dataset greatly.

B. Group Anomaly Detection

1) Group Anomaly Detection using DeepGenerative Models: While anomaly detection mainly focuses on settings where we try to find individual abnormal points, there are also problems of finding anomalous collections of data points. In distribution-based group anomalies, points are seemingly regular while their collective behavior is anomalous[3]. Take images of cats as an exmaple: Group anomaly detection discriminates images that follow the expected group behavior and those that do not, such as images of tigers and images of cats that are up side down.

The models used for detection are AAE and VAE. And the experiment results do get a better performance over state-of-art methods. While the authors only discusses how this group detection method can be adopted to image datasets, we further discuss our thoughts about how this method can be used for our anomaly detection settings.

C. Anomaly Detection With Large Datasets

1) Expected Similarity Estimation for Large-Scale Batch and Streaming Anomaly Detection: Different from the previous work, this paper try to use a simple kernel method to efficiently compute the similarity between new data points and the distribution of regular data. The estimator is formulated as an inner product with a reproducing kernel Hilbert space embedding and makes no assumption about the type or shape of the underlying data distribution. Since it does not use any neural network, it has a very good computational performance. To process the large scale batch, this paper propose a method that can do parallel and distributed processing.

D. Semi-Supervised Anomaly Detection

1) GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training: This is another paper that uses GAN to do anomaly detection. This paper mainly focuses on the application of anomaly detection in computer vision. Although the title contains the word "semi-supervised", it basically uses the normal samples for training and few abnormal samples for testing.

The network proposed by this paper is called "GANomaly". The main contributions of this paper are: (1) It used a encoder-decoder-encoder sub-network instead of a traditional encoder-decoder generator; (2) It both learns the distance between the original graph and the reconstructed graph and the distance between the latent vector and the reconstructed latent vector.

To test the actual performance, some categories from MNIST and CIFAR10 are selected as normal samples, and some others as abnormal samples. The results are based on the area under the curve (AUC) of the Receiver Operating Characteristic, true positive rate (TPR) as a function of false positive rate (FPR) for different points, each of which is a TPR-FPR value for different thresholds. It overperforms the AnoGAN and EGBAD and it also has a good computational performance.

Despite the difference between CV and data mining, it can inspire our ideas for the unsupervised condition in the anomaly detection.

III. Our Thinking

The survey we have done include, but is not limited to the papers stated above. From what we have known about the topic of anomaly detection, we come up with the following thinkings.

- Unsupervised methods are most frequently used and they achieve great performance. AAE, VAE and GAN are the most popular structures. Autoencoders have a relatively simple structure and are easy for training while GAN is more complicated. The most frequently encountered datasets are of images, but we tend to deal with data that has time series, which is consistent with the topic of financial risk control.
- From the papers, we have also obtained these ideas: (1) When trying to model the normal data distribution, training with anomalies might improve the performance depending on dataset features; (2) Despite the anomalies, dealing with infrequent normal data samples is also important since we evaluate performance by F-score; (3) Detection methods of anomalies include setting thresholds, and can also be dealt with using probabilistic models such as MCMC.
- The choice for datasets is essential. Different datasets possess different properties which determine what problems we would face and what methods we are required to use.
- Thus, our choice for dataset should be careful, and we should go into the dataset before we decide which method to use.

IV. Our Work

We formulate the project type into an anomaly detection problem with raw data. Thus, our choice of data should be careful and finding the properties of the dataset is important. Our expriment design is as follows.

A. Dataset

Financial data such as KPIs of large companies are often private and dirty. Getting clean and high quality data is of vital importance. We consider one very popular dataset $KDD-99$, a dataset from Yahoo $Yahoo-S5$. They are all data with time series, and pretty clean and easy to deal with. These datasets have labels as well, which makes it easier for us to deal with.

B. Training Details

There are three main models for us to choose from: AAE, VAE and GAN. All of which has rather simple layers, but the inner deatiled implementations are quite different.

- First, we would train with only normal data and test its performance. Then, we would include abnormal data in the training dataset to see whether they can make the model more robust.
- The datasets we have mentioned does not have missing data points. Thus missing data injection is not needed in our process.
- If the dataset has much infrequent normal samples, we would adapt DOPING to augment the dataset. Comparison can be made between the performance not using augmented data and using augmented data.
- For detection, we primarily intend to use threshold for determining anomalies. And use F-score as the evaluation metric.

V. Conclusion

In this report, we point out our perspective of the project topic, which is doing anomaly detection on financial data (such as KPIs). We list part of the works we have went through and comment on their application situations, mainly faced problems, novel ideas and performance. At last, we state what we plan to do in the following few weeks in detail based on the survey we have done and the discussions we have made. We have some confidence that our method will perform well in practice.

REFERENCES

[1] Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu et al. Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications. DOI: 10.1145/3178876.3185996. arXiv:1802.03903.

[2] Lim S K , Loo Y , Tran N T , et al. DOPING: Generative Data Augmentation for Unsupervised Anomaly Detection with GAN[J]. 2018.

[3] Chalapathy R , Toth E , Chawla S . Group Anomaly Detection using Deep Generative Models[J]. 2018.

[4] Markus Schneider, Wolfgang Ertel, Fabio Ramos. Expected Similarity Estimation for Large-Scale Batch and Streaming Anomaly Detection[J]. 2016

[5] Samet Akcay, Amir Atapour-Abarghouei, Toby P. Breckon. GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. 2018.