UNIVERSITY
OF TECHNOLOGY
SYDNEY

Representation Learning for Anomaly Detection: From the Aspects of Data Views and Optimization

by Shaoshen Wang

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Prof. Ling Chen and Prof. Bin Li

University of Technology Sydney
Faculty of Engineering and Information Technology

Dec 2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Shaoshen Wang declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date: 16/12/2021

Abstract

Anomaly detection (a.k.a. outlier detection) is a challenging task in many realistic applications. However, there exist some major issues for anomaly detection: (1) Nowadays, more and more data tend to be collected from multiple sources or views. As a result, multi-view learning emerges as an approach to leverage the information across diverse views to discover intrinsic property of multi-view data. Although the traditional problem setting of anomaly detection focuses on single-view data, multi-view data pose challenges for anomaly detection, since the anomalies now possess more complex patterns and characteristics. Current methods for multi-view anomaly detection identify anomalies by mining the inconsistent features across different views. However, these detectors have their issues correspondingly, for example, they often rely on assumptions on data distribution. In these methods, the data are usually assumed to be categorized into group in each view, which limits the flexibility and application of such multi-view anomaly detectors. There is lack of more efficient and flexible approaches for multi-view outlier detection. (2) In recent years, deep learning has shown remarkable capabilities in learning expressive representations of complex data. Deep neural networks (DNN) have also been broadly used in detecting anomalies and a large number of deep anomaly detector have been developed. For example, AutoEncoder (AE) and its variants are introduced to learn informative representation of data with no or little supervision, which is then used for detecting anomalies. However, DNN has a powerful approximation capability that easily fits both normal and anomalous data simultaneously, which results in an unsatisfactory performance and less reliability during anomaly detection. The mainstream optimization and learning strategies in DNN exacerbate this issue. There is a need for more advanced and reliable optimization framework for DNN-based anomaly detection.

In this thesis, we propose innovative deep representation learning models to tackle anomaly detection problem from aspects of multi-view data and model optimization. We first introduce related work and literature review. The related work includes existing models for general anomaly detection, multi-view anomaly detection and aggregation schemes for DNN optimization. In following chapters, we studied multi-view anomaly detection together with traditional anomaly detection. Firstly, we investigate semi-supervised multi-view anomaly detection via variational generative model, which is applicable to the situation where labelled normal data is available. Then we studied unsupervised multi-view anomaly detection by exploring latent spaces, which is designed for detecting anomalies in data that include both normal and anomalous data. Finally, we

turn to a more general application scenario, which is traditional unsupervised anomaly detection. We investigate more advanced and reliable optimization strategy for DNN-based detectors.

In Chapter 3, we concentrate on the information provided by multiple views in anomaly detection. By means of using the representation learning power of Variational AutoEncoder and controlling the latent spaces in a novel manner, we propose an innovative Bayesian generative latent variable model to classify multi-view abnormal data. The core idea is to model the correlation between multiple views by generating one view from another view. Multi-view anomalies are then detected by higher reconstruction loss comparing to normal instances. The empirical outcome shows that the novel model outperforms the baselines among popular datasets.

In Chapter 4, we further explore the latent spaces by representation learning to provide crucial information for detecting various multi-view anomalies in an unsupervised manner. We develop a novel Cross-aligned and Gumbel-refactored AutoEncoders (CGAEs) architecture, which has the core idea of learning separate latent spaces for different types of anomalies. In CGAEs, we devise a cross-reconstruction module to detect class anomaly by recovering one view from another view. Further, we design a view-alignment module to detect attribute anomaly by the alignment distance among multiple views in latent space. To handle the robustness problem, we put forward a Gumbel-refactored reconstruction loss to replace traditional mean square error in AutoEncoders. Experimental outcomes validate the efficacy of CGAEs model on both benchmark datasets and real-life datasets.

In Chapter 5, we explore the optimization procedure in anomaly detection. We identify issues with widely used deep neural networks and Empirical Risk Minimization optimization strategy on anomaly detection tasks. Existing DNN and Empirical Risk Minimization scheme suffer from overfitting the outliers and generalization issue in unsupervised anomaly detection, resulting in an unsatisfactory and less reliable performance. We propose a novel Diminishing Empirical Risk Minimization (DERM) framework to break the limit. In DERM, the adverse effect of the potential anomalies is suppressed in a dynamic and controllable manner. Analysis and experiments reveal that DERM can directly modify the gradient contribution of each individual loss and perform better than most benchmarks.

Chapter 6 concludes principal content in thesis and discusses potential future research based on this thesis.

Acknowledgments

I would express sincere gratitude to Prof. Ling Chen, who is my principal supervisor, and also my co-supervisor Prof. Bin Li for their support and help during past few years. Prof. Ling Chen supported me in many aspects, from exploring ideas to paper writing. She is always kind to encourage me and provide suggestions to me. Her patience and kindness gave me a lot of support to complete my research work. From Prof. Ling Chen and Prof. Bin Li, I learned essential skills and positive attitude for academic research.

I would utter my sincere thanks to Dr. Yanbin Liu for great help from research idea discussion to academic paper writing. I would thank all my friends and colleagues in my group and FEIT who provided me with support through my research process.

I would be thankful for Prof. Makoto Yamada towards his inspiring thought and help during my intern.

I would utter thanks to UTS, FEIT, School of Computer Science, CAI and AAII for supporting me.

I would also express thanks to my family, who supported my life and research firmly during the past few years.

Shaoshen Wang

Dec 2021

List of Publications

Related to the Thesis :

Conference Papers

C-1. **S. Wang**, L. Chen, F. Hussain and C. Zhang, Semi-supervised Variational Multi-view Anomaly Detection, *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, 2021.

C-2. **S. Wang**, Y. Liu, L. Chen and C. Zhang, Cross-aligned and Gumbel-refactored Autoencoders for Multi-view Anomaly Detection, *The IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2021.

C-3. **S. Wang**, Y. Liu, L. Chen and C. Zhang, Diminishing Empirical Risk Minimization for Unsupervised Anomaly Detection, *IJCNN at 2022 IEEE World Congress on Computational Intelligence*, 2022.

Table of Contents

List of Publications vii

List of Figures xiii

List of Tables xv

1 Introduction 1

1.1 Research Statement . . . 1
1.2 Research Problems . . . 2
1.3 Research Methods . . . 4
1.4 Research Contributions . . . 6
1.5 Thesis Structure . . . 8

2 Literature Review 11

2.1 Anomaly Detection . . . 11
2.1.1 Anomaly Detection types based on labels . . . 12
2.1.2 Multi-view Anomaly Detection . . . 13
2.2 Representation Learning and AutoEncoders . . . 15
2.3 Aggregation Schemes and Re-weighting for Optimization . . . 16
2.4 Comparison methods and Datasets . . . 16
2.4.1 Comparison methods . . . 16
2.4.2 Datasets . . . 18

3 Anomaly Detection under Multiple Views and Semi-supervision via Variational Generative Model 19

3.1 Introduction . . . 19
3.2 Preliminaries . . . 21
3.2.1 Variational AutoEncoder . . . 21
3.2.2 Cross-reconstruction in Generative Models . . . 22

3.3 Methodology . . . 23
3.3.1 Problem Setting . . . 23
3.3.2 Proposed Framework . . . 24
3.3.3 Deduction of objective function . . . 25
3.3.4 Semi-supervised Multi-view Outlier Detection Scoring Designing . 26
3.3.5 Anomaly Detection Algorithm . . . 26
3.4 Experiments . . . 27
3.5 Conclusion . . . 30

4 Unsupervised Anomaly Detection in Multiple Data Views by Exploring Latent Spaces 33
4.1 Introduction . . . 33
4.2 Preliminaries . . . 35
4.2.1 AutoEncoders . . . 35
4.2.2 Autoencoders for Anomaly Detection . . . 35
4.3 Methodology . . . 36
4.3.1 Problem Setting . . . 36
4.3.2 Cross-aligned Autoencoders . . . 37
4.3.3 Gumbel-refactored Reconstruction . . . 39
4.3.4 Extension to Multiple Views . . . 41
4.3.5 Anomaly Scores . . . 41
4.4 Experiments . . . 42
4.4.1 Datasets and Settings . . . 42
4.4.2 Comparison to Baselines . . . 43
4.4.3 Ablation study . . . 47
4.5 Conclusion . . . 50

5 Advanced Optimization Strategy for Unsupervised Anomaly Detection 51
5.1 Introduction . . . 51
5.2 Preliminaries . . . 53
5.3 Methodology . . . 54
5.3.1 Diminishing Empirical Risk Minimization (DERM) . . . 54
5.3.2 Comparison with Existing Methods . . . 58
5.3.3 Collaborative Autoencoders (cAE) . . . 60
5.4 Experimental Results . . . 61
5.4.1 Comparing to other competitors . . . 61

5.4.2 Study of Ablation and Parameter Analysis 63
5.5 Conclusion 64

6 Conclusion and Future work **67**
6.1 Conclusion 67
6.2 Future Work 68

A Appendix **71**

Bibliography **73**

List of Figures

Figure **Page**

1.1 Architecture of Variational Auto-Encoder. 5

3.1 Three genres of outliers defined in multiple data-view setting. 20

3.2 VAE structure. E represents encoder and D represents decoder. 22

3.3 Pipeline of proposed Multi-VAE. This architecture is to learn cross-view correlation and relationship via reconstruction in crossing manner. 23

3.4 AUCs with respect to outlier proportion on Pima data. 30

4.1 The proposed CGAEs model. Data from two views are input to different autoencoders to generate the cross-reconstruction loss, the view-alignment loss and regular autoencoder loss. 37

4.2 Sorted anomaly scores for *Vowel* dataset. The instances are sorted from left to right with the vertical bar corresponding to their anomaly scores. Here, red bars represents class anomaly, grey bars represents normal data, and blue bars represents attribute anomaly. The baseline with neither L_{A} nor L_{C} obtains AUC= 0.516. 47

4.3 Changes of AUC as the values of anomaly rate increases on Ionosphere and Vowel dataset. 49

4.4 AUCs with different dimensionalities of the latent code. 49

4.5 AUCs with various combinations of α and β on *Zoo* dataset. 49

5.1 Illustration of DERM framework in the training phase. A mini-batch of training data is randomly sampled as the input for k ($k = 2$ in this figure) collaborative autoencoders. The reconstruction loss for each instance is computed. Then, the total loss for an instance from the two autoencoders is summed to obtain a batch of loss. DERM aggregation scheme is applied on the loss and back propagation is performed to update parameters in neural networks of both autoencoders. 55
5.2 Gradient comparison between DERM and TERM [42] on real and synthetic datasets. The height of grey bar represents the gradient weights of the normal instances. The height of red bar represents that of anomalous instances. The gradient weight is obtained by $\omega_i / \frac{1}{N}$ for clear visualization. Blue dash line separates the normal and anomalous data. In (4), a base-10 logarithm scale is applied on the y-axis for a better visualization. 59
5.3 Change of average weight of gradients for normal (blue) and anomalous (red) instances in $vowels$ and $pendigits$ dataset w.r.t. training iterations in DERM framework. 64
5.4 Mean AUCs among 18 datasets with various t via DERM and TERM respectively. 65

List of Tables

Table **Page**

3.1 Information of experiment datasets. 28

3.2 AUC scores (mean and standard deviation) of semi-supervised multi-view AD on 7 datasets with outlier rate = 0.05. (A: Attribute outlier; C: Class outlier; C-A: Class-Attribute outlier) 29

3.3 AUC values of all baselines at WebKB data with four views. 30

4.1 AUCs on five sets of data. In first column, the name means "Dataset-NumberOfViews-ClassAnomalyRatio-AttributeAnomalyRatio". For example, "NewsM-3-2-8" denotes NewsM dataset with 3 views, 2% class anomaly, and 8% attribute anomaly. The results of all the comparison methods are reported by [65]. . . . 45

4.2 AUCs on five datasets of Setting 2. "L2" and "Gauss" denote the L2-norm and Gaussian kernel utilized as similarity metrics in HOAD [22] and AP [49]. We follow [30] to generate three genres of anomalies: attribute anomaly, class-attribute anomaly, and class anomaly. Experimental outcomes of all comparison methods are reported by [30]. 46

4.3 The effect of Gumbel-refactored reconstruction. 48

5.1 AUC values (mean±std) on 18 datasets across diverse domains. 61

5.2 Ablation study of our proposed method (DERM + cAE). cAE represents collaborative AEs with MSE loss and MSE anomaly score. In all cAE, number of AEs k is set to 2. 63