AI in Anomaly Detection and Image Compression

A Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Kaushal Kishore
(111601008)

under the guidance of

Dr. Chandra Shekar

INDIAN INSTITUTE
OF TECHNOLOGY
PALAKKAD

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**AI in Anomaly Detection and Image Compression**" is a bonafide work of **Kaushal Kishore** (**Roll No. 111601008**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Palakkad under my supervision and that it has not been submitted elsewhere for a degree.*

Dr. Chandra Shekar

Assistant Professor

Department of Computer Science & Engineering

Indian Institute of Technology Palakkad

Acknowledgements

Apart from the efforts of myself, the success of any project depends largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project. I would like to show my greatest appreciation to Dr. Chandra Shekar. I can't say thank you enough for his tremendous support and help. I feel motivated and encouraged every time I attend his meeting. Without his encouragement and guidance, this project would not have materialized.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter discusses anomaly detection, its use cases and some major challenges.

1.1 Anomaly Detection

Anomaly detection (also outlier detection) is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data. Typically, the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies also known as outliers, novelties, noise, deviations and exceptions.

Three broad categories of anomaly detection techniques exist:

1. Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set.

2. Supervised anomaly detection techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier (the key difference to

many other statistical classification problems is the inherently unbalanced nature of outlier detection).

3. Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given normal training data set, and then test the likelihood of a test instance to be generated by the learnt model.

We will restrict ourselves to unsupervised anomaly detection and semi-supervised anomaly detection problem.

1.2 Use Cases

The ability to detect anomalies has significant relevance, and anomalies often provides critical and actionable information in various application domains.

Identification of potential outliers is important for the following reasons: [1]

1. An outlier may indicate bad data. For example, the data may have been coded incorrectly, or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).

2. In some cases, it may not be possible to determine if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting. In any event, we typically do not want to simply delete the outlying observation.

For example, anomalies in credit card transactions could signify fraudulent use of credit cards. An anomalous spot in an astronomy image could indicate the discovery of a new star. An unusual computer network traffic pattern could stand for unauthorised access. These applications demand anomaly detection algorithms with high detection accuracy and fast execution.

1.3 Masking and Swamping

Masking and swamping is the biggest problem affecting any anomaly detection algorithm.

Masking is the existence of too many anomalies concealing their own presence. It happens when anomaly clusters become large and dense. For example, if we are testing for a single outlier when there are in fact more outliers, these additional outliers may influence the value of the test statistic enough so that no points are declared as outliers.

On the other hand, swamping refers to situations where normal instances are wrongly identifying as anomalies. It happens when the number of normal instances increases, or they become more scattered. For example, if we are testing for two or more outliers when there is in fact only a single outlier, both points may be declared outliers.

Masking is one reason that trying to apply a single outlier test sequentially can fail. For example, if there are multiple outliers, masking may cause the outlier test for the first outlier to return a conclusion of no outliers. So the testing is not performed for any additional outliers.

1.4 Concept drift

In the case of streaming data, the anomaly context can change over time. For example, consider a user's behaviour change from one system to another. The anomaly detection algorithm should adapt to this change in the behaviour of the external agent. This deviation of the normal behaviour time to time is called concept drift. Any online anomaly detection algorithm must have a way to deal with this.

1.5 Organization of The Report

This chapter [1] provides a background for the topics covered in this report. We provided a description of anomaly detection problem and discussed some use cases. Then we discussed some challenges to anomaly detection problem: masking, swamping and concept drift. In

the next chapter2 we will discuss a very efficient ensemble method Isolation Forest for anomaly detection. In chapter3 we will discuss another ensemble method PIDForest which has been recently developed. The major drawback of the above mentioned algorithms is that they are used in offline setting without dealing with concept drift. Most of the anomaly detection algorithm is offline and fail to address the problem of concept drift. In chapter4 we will present some methods to address these issues. In chapter5 we will review our work did till mid-sem. And finally in chapter6, we conclude with some future works.

Chapter 2

Isolation Forest

Survey comes hear

2.1 Section name

write ....

2.2 Conclusion

This chapter provided details of the some of the existing distributed algorithms for constructing a CDS in wireless ad-hoc networks. The results of these evaluations are summarized in table **??**. In next chapter, we discuss our distributed Algorithm I, for constructing a small backbone in ad-hoc wireless network.

Chapter 3

PIDForest

give details of your algorithm

3.1 Conclusion

In this chapter, we proposed a distributed algorithm for construction of xyz. The complexity of this algorithm is $O(n \log n)$. Next chapter presents another distributed algorithm which has linear time complexity based on xyz.

Chapter 4

Contributions

The algorithm presented in previous chapter has $O(n)$ time complexity. We further propose another distributed algorithm in this chapter based on xyz which has linear time complexity.

4.1 Construction

Write ...

4.2 Improved Method

Write...

4.3 Conclusion

In this chapter, we proposed another distributed algorithm for XYZ. This algorithm has both time complexity of $O(n)$ where n is the total number of nodes. In next chapter, we conclude and discuss some of the future aspects.

Chapter 5

AI in Image Compression

This chapter contains information about work done before COVID-19 pandemic lockdown.

5.1 Construction

Write ...

5.2 Improved Method

Write...

5.3 Conclusion

In this chapter, we proposed another distributed algorithm for XYZ. This algorithm has both time complexity of $O(n)$ where n is the total number of nodes. In next chapter, we conclude and discuss some of the future aspects.

Chapter 6

Conclusion and Future Work

write results of your thesis and future work.

References

[1] Pham H, *Handbook of Engineering Satistics*. Springer, 2006.