MONASH UNIVERSITY

Mining high-frequency time-series data and its application for structural health monitoring

by

Dawei Sun

Supervisor: A/Prof. Vincent Cheng-siong Lee and Dr. Ye Lu

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Information Technology, Monash University

June,2018

Declaration of Authorship

I, Dawei Sun, declare that this thesis titled, 'Mining high-frequency data and its application for structural health monitoring' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: 30/06/2018 _____

Abstract

Artificial Intelligence (AI) launches a significant development in recent years. It has aroused dramatic attention from the academy to industry, and many entrepreneurs attempt to transfer their industry or business modes through AI which they call industry version 4.0. Data is one of the most important components in AI to support decision making, and we produce a massive amount of data each day. Extract valuable information from a massive amount of data to help effective decision making is a critical challenge, in particular under a high-frequency data environment.

Recent studies attempt to use AI to provide structural health monitoring (SHM) with an alternative solution. SHM is one of the important areas in civil engineering. It applies sensor technology to collect data from a structure for monitoring its health status and deploys different types of sensors to collect data for further analysis. Any structural damage or failure of sensing or detecting anomaly signals/data could bring catastrophe to both the economy and public safety. Due to the high-frequency feature of sensor data, it produces a massive amount of data each day. How to extract valuable information from a massive amount of sensor data in time is the challenge. Moreover, time-variation would result in dynamic distribution increases the complexity of online analysis. Most of the traditional methodologies use offline analysis, which assumes that the distribution of data is static. They used monotype of sensor data to evaluate the health status of a structure.

This research addresses online anomaly detection for high-frequency SHM data and health status evaluation of a structure using heterogeneous SHM data. I have investigated and designed different approaches for online anomaly detection and heterogeneous data analysis. Extensive empirical evaluations have verified the effectiveness of the proposed approaches with practical SHM data and public UCI dataset.

To handle the online anomaly detection for high-frequency SHM data, this thesis formulates the problem and indicates the challenges of online anomaly detection and proposed three approaches to online analysis. The first approach, Sample entropy gradient (SEG), is highly reliant on a good quality benchmark dataset, I then propose the second approach (ensemble kernel, EK) without relying on a benchmark dataset. The third approach (Multi-dimensional ensemble kernel, MEK) is an extended version of the second approach which can be used for multi-dimensional scenario and heterogeneous data. All proposed approaches have stable and robust performance.

To solve the problem of health status evaluation of a structure for heterogeneous SHM data, this thesis proposes the MEK method to detect anomalies from multi-dimensional SHM data. However, this method is oversensitive for a long-term evaluation. Consequently, this thesis proposes a hybrid intelligence framework to solve the structural health evaluation. The result of the proposed framework produces a health index which indicates the health status of a structure. The effectiveness of the hybrid intelligence framework is verified by practical data and simulation data, and the benchmark arbitrary are defined in our thesis.

The thesis also contributes to the civil engineering field to make efficient decisions. By using machine learning and data mining techniques, abnormal or risky signals can be identified early and efficient maintenance and recovery plans can be developed.

Acknowledgements

Foremost, I would express my deepest sincere gratitude and appreciation to my two supervisors: A/Prof. Vincent C.S. Lee and Dr.Ye Lu for their patient and considerate guidance throughout my PhD over these three and half years. I would never know how to apply machine learning and data mining to engineering problem without them, and never be a skilled and independent researcher. Their constructive comments and creative ideas significantly improved my work and this dissertation. In particular, I appreciate that A/Prof.Vincent and Dr.Lu provided financial support through the Jiangsu Transport Institute. I also appreciate my collaborator Mr.Jiashen Li, who helped me with building the simulation model and producing simulation data.

I also appreciate the help from Ms.Julie Holden who helped me a lot during my PhD period including my dissertation writing and all milestones. Her valuable comments and constructive suggestions guide me through my writing of the dissertation.

Moreover, I would extend my thanks to all academic staff and administrative staff from faculty of information technology. I appreciate Prof.Balasubramaniam Srinivasan, Prof.Wray Buntine, A/Prof. Andrew Paplinski, Dr.Mark Carmen, and Dr.Yuanfang Li who were my panel members and reviewed my work progress, their constructive suggestions inspire my work.

Finally, I would express my gratitude to my loving parents. Their unconditional love supports my overseas study. I still remember that the 14th of February 2010, I

took my first step in Australia to start my overseas study. Without their support, I could not study overseas. In addition, I also need to thank my colleagues and friends, especially their comfort to release my stress.

Contents

Declaration of Authorship i

Abstract ii

Acknowledgements v

List of Figures xi

List of Tables xiii

1 Introduction 1
- 1.1 Motivation . . . 3
- 1.2 Research questions . . . 5
- 1.3 Contributions . . . 6
- 1.4 Structure of thesis . . . 7

2 Literature Review 9
- 2.1 Overall structural health monitoring . . . 10
  - 2.1.1 Sensing technology on SHM . . . 10
  - 2.1.2 Structural health monitoring measurements . . . 13
    - 2.1.2.1 Stress and Strain . . . 14
    - 2.1.2.2 Displacement . . . 16
    - 2.1.2.3 Wind load . . . 16
    - 2.1.2.4 Summary . . . 17
  - 2.1.3 Structural health monitoring analysis methodology . . . 17
    - 2.1.3.1 Traditional method . . . 17
  - 2.1.4 Multi-agent based evaluation system . . . 20
    - 2.1.4.1 Machine learning and data mining method . . . 21
  - 2.1.5 Summary . . . 25

2.2 Traditional Anomaly detection . . . 26
2.2.1 Distance based methodology . . . 28
2.2.1.1 Global outlier method . . . 28
2.2.1.2 Local outlier method . . . 30
2.2.1.3 Summary . . . 31
2.2.2 Model-based anomaly detection . . . 31
2.2.2.1 Classification based anomaly detection . . . 32
2.2.2.2 Probability based anomaly detection . . . 32
2.2.3 Neural network based anomaly detection . . . 33
2.2.3.1 Clustering based anomaly detection . . . 34
2.2.3.2 Summary . . . 35
2.3 Performance measuring . . . 36
2.3.1 Confusion matrix . . . 36

3 An approach to the computation of time-series data similarity using Sample Entropy Gradients 39
3.1 Motivation . . . 40
3.2 Preliminary . . . 42
3.2.1 Approximate entropy . . . 42
3.2.2 Sample entropy . . . 44
3.2.3 Cross sample entropy . . . 45
3.2.4 Sample entropy gradient . . . 47
3.3 Empirical Evaluation . . . 50
3.3.1 Data . . . 51
3.3.1.1 Structural health dataset . . . 51
3.3.1.2 Financial dataset . . . 51
3.3.2 Results . . . 52
3.3.2.1 Case study of between ASX and NASDAQ . . . 55
3.4 Summary . . . 56

4 An ensemble kernel density estimator method for online high-frequency data stream anomaly detection 58
4.1 Motivation . . . 59
4.2 Related Work . . . 60
4.3 Proposed Method . . . 65
4.3.1 Problem formulation . . . 65
4.3.2 Overall description . . . 65
4.3.3 Kernel function . . . 68
4.3.4 Anomaly factor . . . 69
4.3.5 Process of Online EK Anomaly Analysis . . . 72
4.4 Experiments and results . . . 73
4.4.1 EK Evaluation experiment design and result . . . 74
4.4.1.1 Experimental design and procedure . . . 74
4.4.1.2 Experimental result and discussion . . . 77
4.4.2 Comparison experiment and result . . . 77

4.4.2.1 Experimental design and procedure 79
4.4.2.2 Experimental results 79
4.4.3 Large volume data performance evaluation and comparison . 82
4.5 Limitation 84
4.6 Summary 84

5 A new online anomaly detection method: multi-dimensional ensemble kernel 86
5.1 Motivation 87
5.2 Related work 89
5.3 Proposed method : Multi-dimensional Ensemble Kernel 92
5.3.1 Overall 92
5.3.2 Problem Formulation 93
5.3.3 Multi-dimensional Ensemble Kernel 94
5.3.3.1 Windowing 94
5.3.3.2 Densities estimation 95
5.3.3.3 Anomaly hypothesis test 95
5.3.3.4 AF computation 95
5.3.3.5 OAF computation 96
5.3.4 Kernel Density Estimation 96
5.3.5 Anomaly Factor and Overall Anomaly Factor 98
5.3.6 Process of MEK Anomaly Analysis 98
5.4 Experiment and result 99
5.4.1 UCI dataset experiment 102
5.4.1.1 Experimental design and procedure 102
5.4.1.2 Result 104
5.4.2 Practical SHM dataset experiment 105
5.4.2.1 Experimental design and procedure 105
5.4.2.2 Result 111
5.4.3 Long-term performance of MEK 113
5.4.3.1 Experiment design and procedure 113
5.4.3.2 Result 113
5.5 Limitation 114
5.6 Summary 115

6 A hybrid intelligent framework for structural health monitoring 116
6.1 Motivation 117
6.2 Related work 118
6.3 Preliminary 120
6.3.1 Adaptive Resonance Theory(ART) [1] [2] 120
6.4 Hybrid intelligence system for SHM 122
6.4.1 Pre-processing 122
6.4.2 Neural Network 123
6.4.3 Fuzzy Inference System 126

6.5 Empirical Evaluation . . . 128
6.5.1 Practical data . . . 128
6.5.1.1 Data pre-processing . . . 128
6.5.1.2 Neural Network . . . 129
6.5.1.3 Fuzzy Inference system . . . 130
6.5.1.4 Result of practical data . . . 132
6.5.2 Result of simulation data . . . 133
6.5.2.1 FEM model and Data . . . 133
6.5.2.2 Structural health index computation . . . 135
6.5.2.3 Result of simulation data . . . 136
6.6 Summary . . . 136

7 Conclusion and future work **138**
7.1 Contribution . . . 139
7.2 Future work . . . 140

A Appendix A **142**
A.1 Data . . . 143
A.2 Financial analysis . . . 144
A.3 Comparison of SEG with 30-day-trend . . . 151
A.4 SSE and SEG comparison . . . 152
A.5 Summary . . . 155

Bibliography **156**

List of Figures

2.1 Example of a FBG sensor . . . 11
2.2 Example of a PTZ sensor . . . 12
2.3 Example of anomaly . . . 26
2.4 Example of anomaly detection using clustering method. F1,F2 and F3 are recognized as anomalies . . . 34

3.1 Process of SEG . . . 47
3.2 Example process of SEG . . . 50
3.3 Result of 0.2% injection rate . . . 52
3.4 Result of 0.3% injection rate . . . 53
3.5 Result of 0.4% injection rate . . . 54
3.6 Result of 0.6% injection rate . . . 54
3.7 NASDAQ SEG and ASX SEG . . . 56

4.1 Procedures of Ensemble Kernel . . . 66
4.2 Process of Online EK Anomaly Analysis . . . 70
4.3 Wind speed datasets over the time (3-day) . . . 75
4.4 Accuracy . . . 75
4.5 Specificity . . . 76
4.6 Sensitivity . . . 76
4.7 Wind data comparison . . . 79
4.8 Accuracy of UCI datasets . . . 80
4.9 Sensitivity of UCI datasets . . . 81
4.10 Specificity of UCI datasets . . . 81
4.11 Performance Comparison using large volume of wind speed data . . . 82
4.12 Performance Comparison using large volume of temperature data . . . 83

5.1 Procedures of Multi-dimensional Ensemble Kernel . . . 94
5.2 Entire process of MEK anomaly detection . . . 99
5.3 Accuracy . . . 105
5.4 Specificity . . . 106
5.5 Sensitivity . . . 106
5.6 3-day Wind data: blue dots represent the first day data; orange dots represent the second day data; green dots represent the third day data . . . 108

5.7 3-day road surface temperature data : blue dots represent the first day data; orange dots represent the second day data; the green dots represent the third day data . . . 109
5.8 GPS X value : blue dots represent the first day data; orange dots represent the second day data; green dots represent the third day data . . . 109
5.9 GPS Y value : blue dots represent the first day data; orange dots represent the second day data; green dots represent the third day data . . . 110
5.10 GPS Z value : blue dots represent the first day data; orange dots represent the second day data; green dots represent the third day data . . . 110
5.11 Comparison Results using SHM datasets . . . 112
5.12 Long-term performance of MEK . . . 114

6.1 Structure of ART self-organized neural network . . . 121
6.2 Hybrid intelligence Framework for SHM . . . 123
6.3 Process of Self-organized neural network . . . 124
6.4 The network topology of bridge structure health . . . 129
6.5 The network topology of bridge structure health . . . 132
6.6 Side view and top view of real bridge structure . . . 134
6.7 3D side view of bridge model . . . 135
6.8 Health index of simulation data . . . 137

A.1 NASDAQ SEG and AXS SEG . . . 145
A.2 NASDAQ SEG and Nikkei SEG . . . 147
A.3 SSE SEG and NASDAQ SEG . . . 148
A.4 SSE SEG and Nikkei SEG . . . 149
A.5 5-year SEG Euclidean Distance . . . 150
A.6 30-day-trend and SEG . . . 152
A.7 CSE and SEG Euclidean Distance . . . 153
A.8 CSE and Sequence Volatility . . . 154

List of Tables

2.1 Template of confusion matrix 37
2.2 Confusion table summary 38

3.1 Wind speed dataset 51

4.1 Details of Kernels 68
4.2 Detail of UCI datasets 80

5.1 Details of Kernels 96
5.2 UCI dataset description 104

6.1 Result of ART Neural Network 130
6.2 Result of Health Index 132

A.1 International stock indices descriptive statistics 144

Dedicated to my loving parents...

Publications arising from this thesis are list below :

1. The main material in Chapter 3 has been published in IEEE Conference

   D.Sun, V.Lee and Y. Lu, "A Gradient-based Algorithm for trend and outlier prediction in dynamic data streams", proceeding of International Conference on Industrial Electronics and Applications (ICIEA) ERA Rank A Conference, June 2017, Sime Reap, Cambodia. Proceeding of IEEE ICIEA pp.1975-1980.

2. The main material in Chapter 4 has been submitted to Pattern recognition and it under review by Pattern recognition.

3. The main material in Chapter 5 has been submitted to IEEE International conference on data mining (ICDM 2018), and it under review.

4. The main material in Chapter 6 has been published to IEEE Conference

   D.Sun, V.Lee and Y. Lu, "An intelligent data fusion framework for structural health monitoring", proceeding of International Conference on Industrial Electronics and Applications (ICIEA) ERA Rank A Conference, June 2016, Hefei. Proceeding of IEEE ICIEA pp.49-54.

Chapter 1

Introduction

Artificial Intelligence (AI) launches significant developments in recent years. It has aroused dramatic attention from the academy to industry, and most of the entrepreneurs attempt to transfer their industry or business modes through AI which they called industry version 4.0. Data is one of the most important components in AI to support decision making, and we produce a massive amount of data a day. How to extract valuable information from a massive amount of data to help effective decision making is a critical challenge [3], in particular under a high-frequency data environment.

Recent studies attempt to use AI to provide structural health monitoring (SHM) with an alternative solution. SHM is an important area in civil engineering, as it applies sensor technology to collect data from a structure for monitoring its health status [4]. SHM is concerned about damage detection, damage localization, damage assessment, and prediction. It deploys different types of sensors to collect

data for further analysis. Any structural damage or failure of sensing or detecting anomaly signals/data would bring catastrophe to both the economy and public safety. Predicting or detecting potential risky circumstances for a structure is critically important to reduce the occurrence of disasters.

With the tremendous development of sensor technology, various types of sensors have been deployed on a structure to collect relevant data. These sensors collect various types of information through gauges and measurements, such as displacement, stress and strain. All of these data are featured as time-series and high-frequency data because sensor collects data with a certain sampling rate. Traditional SHM analysis methodology uses signal processing technology, such as Fourier Transform (FT) and wavelet-based method. However, FT and wavelet-based methodologies fail to handle online analysis. Moreover, FT and wavelet-based method are used for damage detection or crack detection of a structure or a structural member.

In the recent decades, machine learning and data mining methodology have been introduced to help SHM analysis. Current machine learning and data mining methods mainly used for damage detection and crack detection as well, and most of the current technology cannot handle online analysis either. It is impossible to create a universal algorithm to handle all types of data. A specific algorithm is required to perform a specific task. To analyze high-frequency sensor online, we need to overcome its high-frequency and time-variation which results in a dynamic distribution.

Current machine learning and data mining techniques for SHM problem focus on damage detection and damage localization, and most of these techniques are offline methods. Online analysis and prediction can benefit relevant organizations to help to make the decision to avoid catastrophe. Online analysis provides the real-time information for the relevant organization to avoid risky circumstances.

Due to the high-frequency feature of sensor data, it produces a massive amount of data each day. How to extract valuable information from a massive amount of sensor data in time is the challenge. Most of our research work is based on the SHM background because a massive amount of high-frequency sensor data gives us practical data to test the performance of our proposed method. We proposed different methods for overall structural health evaluation and online anomaly detection on single or multi-dimensional sensor data. Current machine learning and data mining techniques for SHM are offline analysis, so the online analysis provides the real-time information for the relevant organizations to help them to avoid potentially risky circumstances. Moreover, it also can guide relevant organizations to develop maintenance plan and recovery plan efficiently.

1.1 Motivation

Traditional SHM sensor data analysis is based on the signal processing methodology, such as FT and wavelet-based method. Through the FT and wavelet method, a time-series data stream is translated into a spectrum which is a frequency based

analysis methodology. By comparing the frequency between the benchmark frequency with processed data, the damage signal can be identified. However, these methods fail to demonstrate the time information. In other words, we fail to observe changes of a data stream over time. Since the development of machine learning and data mining techniques to SHM problems, most of the current methods focus on damage detection and damage localization. Moreover, the current methodology uses monotype SHM dataset, neither to analysis data synthetically. Therefore, online prediction or evaluation and synthetic analysis of SHM datasets arouse substantial attention from both academics and industry.

To gain a good online analysis result, the online analysis needs to overcome the time-variation problem. Moreover, we also need to consider various types of sensor data (heterogeneous data) that has different features. Most current machine learning and data mining methodologies applied in SHM are based on :

1. Probability-based methodology : uses probability theory, such as Bayesian probability, to predict the degree of damage of a structure.

2. Model-based methodology : uses the neural network to localized or detect damage of a structure.

Most of above methodologies are offline analysis, which analyzes data after the data collection phase. Online anomaly detection is one of the solutions for online analysis of SHM. Current anomaly detection methods are the offline model-based method and distance-based method. To do the online anomaly detection for high-frequency SHM datasets, dynamic distribution is a challenge. Traditional offline

anomaly detection assumes that the distribution of data is static, in fact, the distribution changes over the times. Moreover, current methods are only concerned with monotype sensor data. In fact, engineers evaluate a health status of a structure from different types of sensors. The chanllenge is how to weight various types of datasets when analysing heterogeneous data. To summarise, the challenges of online analysis are listed :

1. High-frequency sensor data: it not only increases the volume of datasets but also introduce the dynamic distribution
2. Online analysis: proposed approaches are offline analysis
3. Heterogeneous data: many proposed approaches are verified with monotype data.

1.2 Research questions

In order to analyze a massive amount of SHM sensor data and discover valuable information for evaluating structural health status, we conducted research on mining SHM sensor data. In my PhD project, I handle the following research questions :

- Research question 1: Can online anomaly detection algorithm be made robust for high-frequency SHM datasets (for single and multi-dimensional SHM dataset) ?
- Research question 2: Can heterogeneous SHM dataset be analyzed synthetically?

1.3 Contributions

In the process of answering the above two research questions, I first investigated and reviewed the challenge of online anomaly detection with existing literature. Then I designed possible approaches to each problem. I verified the performance of my proposed methods extensively with UCI public datasets and practical SHM datasets. Overall, this thesis makes following contributions :

1. A proposed sample entropy gradient method for comparing the similarity between two time-series data. It can be used for anomaly detection and similarity measurement of two time-series datasets (Research Question 1). This method has been published in IEEE ICIEA conference 2017.

2. A proposed ensemble kernel (EK) for online high-frequency data anomaly detection. It detects anomalies for a single dimensional data stream efficiently with a stable performance. (Research Question 1). This method has been submitted to the Journal of Pattern Recognition, and it is under review by pattern recognition.

3. A proposed multi-dimensional ensemble kernel method for online high-frequency data anomaly detection. It detects anomalies for the multi-dimensional data stream with a stable performance. (Research Question 1 and 2). This method has been submitted to the IEEE ICDM 2018, and it is under review.

4. A proposed a structural health monitoring evaluation method using the neural network and fuzzy inference system. It introduces the health index(HI) as

a health status indicator of a structure (Research Question 2). This method has been published in IEEE ICIEA conference 2018.

1.4 Structure of thesis

This thesis organized as follow :

- Chapter 1 gives an overview of the research background, current challenges, proposed approaches and outline of the thesis.
- Chapter 2 firstly introduces the general background of SHM and its relevant challenges. Then, it describes an in-depth review of the relevant theoretical models and design principles of data mining and machine learning. In addition, it also gives a review about current methods used in SHM. This chapter ends with the review of potential applications such as using data mining and machine learning method. The research gaps and novelty of this research project are presented here.
- Chapter 3 introduces a sample entropy gradient method for comparing the similarity between two time-series data, which is derived from cross sample entropy and approximated entropy. In addition, this method can be used for anomaly detection as well. This chapter includes all experimental studies to develop sample entropy gradient.
- Chapter 4 proposes a new method for anomaly detection. In this chapter, I introduce EK method for online high-frequency data anomaly detection.

This study describes the EK method and its performance and includes experimental result compared with other online anomaly detection methods.

- Chapter 5 demonstrates the extended version of EK method which is called multi-dimension ensemble kernel (MEK) method. It includes the process of MEK and relevant experimental results. It enables detecting of anomalies in a multi-dimensional scenario, and give an overall structural evaluation.

- Chapter 6 introduces a structural health monitoring framework for SHM. This framework is based on the hybrid intelligent systems to compute composite structure health index as structure health indicator. It provides a decision-level analysis for structural health monitoring. The outcome of our proposed framework is a structural composite health index which is an aggregated index via optimally weighted variables.

- Chapter 7 gives a conclusion to my PhD work, along with a brief discussion on future perspectives of high-frequency mining and application of data mining and machine learning on SHM problems.

Chapter 2

Literature Review

Structural health monitoring (SHM) applies various technologies(e.g. sensing technology, guided wave, data analytics) to monitor and analyze the health status of a structure. In current SHM projects, various sensors to collect different types of data. There is a massive volume of data generated each day, how to extract valuable information from a massive volume of data becomes a critical challenge. Machine learning and data mining have a tremendous development in recent decades. It gives a solution to various domains, such as medical, engineering, and finance. There are many applications of using machine learning and data mining techniques on structural health monitoring. Our concerns are anomaly detection on structural health monitoring and overall structural health evaluation. Generally, current machine learning and data mining methods on SHM can be categorized on statistical probability-based method, neural network based method and etc. There are different categories of data type, such as image data, time-series data, and categorical data. Each data type requires a different specific method to process.

In this chapter, I firstly review current sensing technology on SHM. This gives a general view of recent studies of sensing technology applied to SHM.

After demonstrating recent studies of sensing technology, I present current machine learning and data mining techniques applying to SHM. After we discussing current sensing technology, traditional anomaly detection methods are discussed, since anomaly detection is one of the important techniques to help SHM analysis.

Finally, limitations of recent studies are discussed in order to reflect the motivation of our studies.

Apart from brief descriptions of recent studies, I present some popularly used performance measurements in data mining in terms of usability, and measurement we used to evaluate my proposed methods.

2.1 Overall structural health monitoring

2.1.1 Sensing technology on SHM

Current structural health monitoring can be categorised into two classes: one class relates to detection technology, for instance, sensing technology; another class relates to sensor data analytic methodology.

Sensing technology is one of the most important technologies in SHM, it provides the fundamental infrastructure for structural monitoring. Current sensing technology includes optic fibre sensing technology, piezo-electric ceramic sensors,

cement-sensor strain sensors, corrosion sensors, seismometers, and GPS technology and some traditional sensing technology (e.g. accelerometer). Civil experts and engineers can predict and evaluate the health status according to sensor data according to their experience, fieldwork and theoretical model. Although traditional sensing technology and advanced sensing technology have been applied successfully, there are some challenges need to confront in future research and development [5].

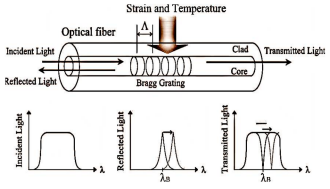


FIGURE 2.1: Example of a FBG sensor

Optic fibre sensor is a new advanced sensor type applying to SHM. It permits a long gauge to measure an overall average strain of a structure. Currently, there are five types of commercial use optic fibre sensor [6]. Some of these optic sensors are based on measuring the light intensity according to change of fibre curvature or reflection of a mirrored surface. The majority of commercial use optic fibre sensors measure the light intensity, which they called fibre Bragg grating(FBG) [7]. They provide a linear response with a wide strain range and temperature compensation (Figure 2.1 ©[8]). The primary advantage of FBG is the immunity to the EMI/RF inference by measuring the light wavelength rather than the signal aptitude. Due to its immunity to the inference, it has been deployed in many complicated electronically environments in practice to avoid the electrical current [9] [10]. However, the durability and reliability of FBG are two most important considerations. The performance of sensors decline with the increasing service

time, which affects the quality of data collection [11]. Optic fibre sensors usually attach to the structure surface. With the changing surface temperature, the material would result in deformation of optic fibre sensors. As a result, the quality of the data becomes unreliable. Other factors, such as oxygen carbonization, acid erosion, material ageing ,have an impact on data collecting. Consequently, the durability and reliability are the first two important considerations.

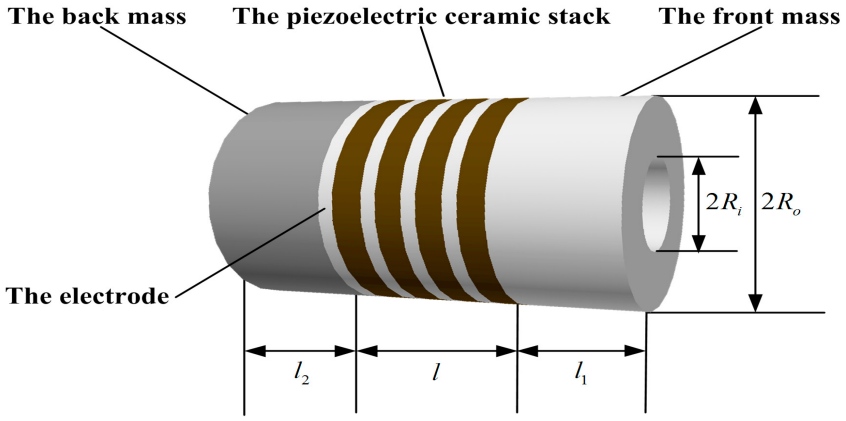


FIGURE 2.2: Example of a PTZ sensor

Piezo-electric ceramic (PTZ) sensor (Figure 2.2 ©[12]) is an acoustic sensor, which has been deployed in many large-scale buildings and structures [13]. It receives the stress wave signal when any damage occurrences within a structure. According to the acoustic emission, a damage index can be calculated. Since the PTZ receives acoustic emission signal, the interference is a critical challenge. Although we have de-noise algorithms and methods, we do not know the effectiveness of de-noise algorithms on the PTZ inference problem.

The cement-based strain sensor is one of most utilized in civil engineering. Cement is capable of solving the incompatibility issue [5] [14]. Moreover, it gives possibility to combine different types of sensors in one cement unit. Such sensors

like conductive nano-sensor, PTZ, and short carbon fibre can be processed using cement. However, cement structures would expand or contract with temperature. Any change(expansion or contraction) of a cement structure would generate some signal which can be sensed by sensors. Consequently, the quality of the sensor data is not good enough to be used for further analysis.

GPS is also an extensively applied sensor in many domains [5]. In SHM domain, GPS sensor is used to measure the displacement of a structure. However, the resolution of commercially used GPS sensors are low, which fails to detect small deformation displacement in a small-scale structure. In addition, GPS sensors are not immutable to the environmental factor. It usually fluctuates to an abnormal level for a period, until the abnormal signal has been tackled.

To sum up, current advanced sensing technology provides rich functions for SHM, but the performance of advanced sensors need to overcome their limitations to ensure the data quality. Except for the physical performance of the sensor, extracting valuable information from a massive amount of sensor data for decision making is also a challenge.

2.1.2 Structural health monitoring measurements

Structural health evaluation is a sub-category of structural health monitoring, it applies different techniques to evaluate a health status of a structure. Fieldwork is a traditional method to evaluate the structure health status. A fieldwork of a structure usually contains :

- Visual observation of crack of a structure member
- Collect signals using special equipment

Based on the result of visual observation and signal analysis, engineers would give a health evaluation on a structure. With the development of sensing technology which we have presented in the previous section, various types of sensors have been installed to help health evaluation. In this section, structural health status measurements and evaluation methods for structural health status are discussed. In this section, we discuss more details in this section in terms of measurements and evaluation methodology.

2.1.2.1 Stress and Strain

Stress and strain are usually measured together and this approach has have been applied in structural health monitoring for many years. Stress and strain reflect forces in material or a structure. These two measurements help to understand the load of the structure and the potential for deformation and cracks. Optic fibre is the latest advanced sensing technology to measure the strain. Optic fibre famous for its FBG is widely used for monitoring strain of a structure. Many researchers have contributed to optic fibre in terms of its functionality and robustness for practice.

Superimposed FBGs proposed by [15] used two superimposed fibre Bragg gratings to allow measuring of dynamic and static strain. By using smoothing filtering

technique and centroid finding technique increase interrogation speed and accuracy with reflection spectrum. This method improved the functionality of FBG sensor and allowing it to measure dynamic and static gauge. [16] used time-division multiplexing method to enrich the functionality of FBG. TDM systems used semiconductor optical amplifier (SOA) which emits a short broadband optical pulse, and FBG sensor is capable of capturing this weak signal. Due to the time efficiency, this TDM based FBG sensing system was able to monitor large-scale structure. [17] combined FBG sensor with a wet etch-erosion procedure to enable an FBG sensor to be a multi-functional sensor which can be not only used for SHM but also environmental monitoring and biochemical sensing.

Researchers also contributed to improving the robustness of FBG sensor used in practice. [18] [19] [20] [21] have proposed and implemented FBG on large-scale structure. These applications demonstrated FBG which can report dynamic strain correctly and accurately. [18] suggested that FBG has the potential to be deployed to a large-scale structure for long-term monitoring. However, the study lacked evidence that FBG can be deployed for long-term monitoring with good quality data. From some reports of some institutions, we found that there is no doubt about the performance of FBG, but installation methods and environmental temperature have an impact on FBG. Current FBG sensors are attached on a structure using some material (e.g. glue), with the change of environmental temperature, the material used for installation has an impact on data gathering of FBG sensor [11].

2.1.2.2 Displacement

Strain-stress can reflect displacement of a structure due to deformation. The global positioning system (GPS) can present the displacement immediately via 3D(x,y and z) coordinates. GPS has been deployed on a structure to monitor its displacement. With its high sampling rate (10Hz), GPS sensor provides the solution to real-time monitoring. [22] [23] [24] [25]reported the applications of GPS sensor on structure health monitoring in practical engineering. However, there are some limitations of using GPS sensors [11]:

- Limited GPS granularity: The precision of current GPS sensor can take to meter unit, which is not good enough for some civil structure monitoring.
- Limited to a specific structure: The application range is limited by its precision limitation, in some civil structure monitoring project, GPS sensor fails to provide enough information about its health status.

2.1.2.3 Wind load

Wind is the primary source of vibration of a structure. Especially, for some large-scale structure like a bridge, the wind could result in vibration for a bridge structure [26]. In some application, the wind is not directly measured instead of measuring the GPS displacement immediately [19], [19] thought it is hard to collect accurate wind data due to the complexity of wind data. [27] [28] reported the applications of using wind sensors in large-scale structures to monitor wind speed.

2.1.2.4 Summary

In this section, I present the details of measurements used in SHM in terms of stress-strain, GPS and wind load. These three measurements are the most important measurements in SHM. Although FBG is famous for its immunity to interference, the environmental factors would have an impact on data quality which reported in practical applications. GPS is specialized in monitoring large-scale structure, but it has some limitations in some civil structural monitoring projects. Wind load is a critical monitoring measurement, in practical application, some argue that it is difficult to measure the wind data due to the complexity of environmental factors, but others used wind sensor to collect wind data directly. Due to the drifted distribution of wind speed, an effective method to analyze wind speed data is required.

2.1.3 Structural health monitoring analysis methodology

2.1.3.1 Traditional method

A massive amount of data are collected from a bridge. After data collection, how do we extract valuable information from a massive amount of sensor data is a critical challenge. Manual observation and analysis are one of the methods to process that massive amount of sensor data. Civil experts and engineers evaluate and predict the health status according to their significant experience. However, manual works could delay their decision making for a massive amount of data.

Machine learning and data mining techniques have been introduced to help to evaluate and predict the health status of a structure.

Traditional manual observation and analysis rely on the spectrum (frequency) analysis and trend analysis of some specific sensors. It helps the experts and engineers to determine whether currently collected data in the defined health range(defined according to model simulation or initial design safe range). The fast Fourier transform and the wavelet transform are two typical signal analysis methods which analyze the data from the frequency domain. The wavelet transform overcome the limitation of FFT. The result of FFT was a summation of a give length signal, which means we fail to indicate the time occurrence of a signal [29]. WT provided the solution to allow signal decomposition with time information. However, there are several wavelet families (each family has multiple wavelet types) for frequency analysis. Only experienced experts and engineers know how to select proper wavelet type.

Nevertheless, researchers still auguring the effectiveness of using Fourier transform or wavelet transform to analysis signal. Researchers argued that Fourier transform fails to reveal the time information, the application of Lamb wave overcomes the problem. Lamb wave method is a new method for damage detection for SHM without a benchmark signal[30]. It has been well-known by it dispersive property and multi-modal where at least two Lamb models existed simultaneously. With the increase of frequency and thickness of an object, more Lamb models can be found. Various of Lamb models provides abundant information in time-domain and frequency domain, but it also increases the complexity to analyze frequency

and time information. [31] proposed a short-space frequency-wave method which is capable of showing the number of wave change during its propagation. This method was a Fourier transform based method, which used 2D Fourier transform to locate the crack within an aluminium plate. [32] is a wavelet-based Lamb wave analysis method, it also used for damage detection. However, environmental factors such as temperature and signal attenuation. This paper only had been verified in some simple samples, and the robustness of this wavelet-based is needed more experimental verification under various circumstances with different samples.

Finite element model (FEM) is another extensively used technique which helps experts and engineers to simulate a structure response situation according to different sensor data [5]. Vibration modelling , for example, uses the FEM technology and theoretical models to simulate a structure. Based on the simulation result, experts and engineers can give an evaluation of a structure. However, this method consumes time to build a structure model to give a reliable evaluation of a structure. Therefore, machine learning and data mining methods started to help experts and engineers to evaluate and predict the health status of a structure.

To sum up, traditional sensor data analysis uses frequency analysis methodology and FEM simulation. Signal processing methodology helps to detect damage or locate cracks within a structure. However, some of these methods are only tested in the laboratory, there is no evidence presented that proposed methods can be applied in practice. Moreover, these methods do not support online analysis and some of them are too time-consuming.

2.1.4 Multi-agent based evaluation system

The multi-agent system is one of artificial intelligence technique where each agent is an autonomous system with a specific task or without. Generally, agents can be categorised into three groups:

- Passive agent: no specific task assigned to an agent
- Active agent: simple tasks assigned to an agent
- Cognitive agent: complex calculation tasks are assigned to agents.

Agents has following features:

- Autonomy: agents are independent to each other (at least palatially), self-organized
- Local view: no agent have a global view
- Decentralization: no agent is designed to control other agents

Some researchers proposed a multi-agent system for SHM problem, especially large-scale SHM monitoring. [33] designed and implemented a multi-agent system for SHM. Generally, the proposed multi-agent system had three layers which are data monitoring layer, data interpretation layer, data diagnostic layer, and information layer respectively.

In the data monitoring layer, agents were assigned with sensing task, which deployed with sensors to reflect the health status of a structure. In their experiment,

they deployed FBG and PZT sensors on a plate structure which divided into several subareas. In the data interpretation layer, significant signals are extracted for further processing, agents in this layer were assigned with data processing task. In the data diagnostic layer, agents used data extract from the previous layer to estimate potential damage. In the information layer, all information are gathered to make a reliable conclusion and reports to users. This multi-agent system had been evaluated via a simulation in the lab [34]. Although this method has been evaluated in the laboratory and demonstrated a good performance, the performance of this multi-agent system in practice project is unknown. [35] and [36] used multi-agent system for wireless sensor monitoring. However, the wireless sensor technology is not robust enough in practice. [37] implemented a simple multi-agent system to monitor a wind turbine system, a number of malfunctions had been detection via the multi-agent system.

2.1.4.1 Machine learning and data mining method

In the recent decade, the tremendous development of artificial development arouses attention from various domains. It offers a potential opportunity to promote multi-disciplinary research. Application of machine learning and data mining methodology in SHM is a new launched in the recent decade. Generally, application of machine learning and data mining can do following tasks[38]:

- Damage detection : gives a qualitative indication of potential damage to a structure
- Damage localization: locates probable damage in a structure

- Damage assessment: assesses the damage severity of a structure
- Prediction: offering information about the safety of a structure, e.g. detect anomalies, estimating a residual life of a structure.

Statistic and probability based are widely used in SHM. [39] is a Bayesian-based approach for SHM. At the initial stage, [39] performed a modal test based on an undamaged structure to build a PDF reference. Then, used these undamaged data to create alarm functions for each substructure. All these probabilities were computed under the naive Bayesian framework. During the monitoring phase, it computes the damage degree and comparing with alarm function. The alarm functions keep updating along with the time and new incoming data. Although the alarm functions kept updating with new incoming data, the PDF reference was not. This kind of sensor data is also considered as a time-series data, we cannot assume that the distribution is static through all the time. [40] is an enhanced Bayesian-based version for SHM. In the initial step, [40] was same with [39] which computed PDF using undamaged structural data as a reference probability. The second step was model updating, which updates the parameters using the expectation-maximum(EM) algorithm. EM is an optimization algorithm which addresses parameters estimation under uncertain circumstances. In this method, some simulation data were used to verify the performance of their proposed method. However, it lacks evidence to support their method that can be applied into practice.

[41] is another statistics-based anomaly detection method for SHM. It applied auto-regressive (AR) and auto-regressive with exogenous (ARX) model to construct a

new reference signal for comparing with a signal. By measuring the distance between the newly constructed signal and a normal signal, anomalies can be found if the distance is large. This method was not an online analysis method, ant it relying on a good-quality of a constructed signal, which we cannot guarantee that the newly constructed signal is qualified to be used as a benchmark. An evaluation method was required to evaluate the result of a newly constructed signal. In addition, regressive functions are easily affected by the density of data, which has an impact on the quality of the newly constructed signal as well as the final result of anomaly detection. [42] is a wavelet-based method which using AR and SVM to detect damage of a structure. It firstly used wavelet to remove the noise and compress the data. Based on the result of the wavelet transform, it used AR to model the data and extract the coefficient. Based on the AR coefficient, SVM classified the coefficient into two group in order to distinguish between damaged structure and undamaged structure. The effectiveness of WAR-SVM has been proven via a case study which detects the damage to a magnetorheological(MR) damper.

Application of neural network is another branch of using machine learning on SHM. Generally, it applied the neural network to detect damage or crack of a structure. [43] proposed to use the artificial neural network (ANN) to detect damage PTZ signal. For engineers, it does not require prior knowledge to analyze the PTZ damage signal. However, training a good quality ANN to detect damage signal requires datasets with various damage features. In this paper, it only applied to a simple structure, which lacks further practical evidence. [44] used similar neural

network method to detect damage PTZ signal, but it applied to different scenarios with :

- on a bolt-jointed aluminium beam
- multi-type and multiple damage detection on a pipe system
- multi-type and multiple damage detection on a real-scale bridge

From a long-term consideration, the impedance signature is easily affected by the external factors such as temperature and degradation of a structure member, so a further investigation and study are required.

Except using the neural network to detect PTZ damage signal, new developed convolutional neural network (CNN) arouses civil engineers attention. Basically, CNN is specialized in pattern recognition in domains like image and voice. In civil engineering, to recognize the crack from a massive amount of image is efficient for engineers to make maintenance decision. In the study of CNN on crack detection [45], they used 332 images for training, and 55 images for testing. The result of recognizing the line cracks can achieve around 97% accuracy. However, in this study, their training samples were simple which is a single crack of a concrete. Under such high accuracy circumstances, the training result was considered over-fitting. In the future study, different types of crack should be considered to add into training samples, and the size of training sample need to increase as well.

To sum up, most of current machine learning applications to SHM are offline methods, and most of these methods concentrate on damage detection and localization.

Some of these methods are only tested in the laboratory, there is a gap between practical and laboratory.

2.1.5 Summary

In this section, a general SHM overview is presented, which includes current sensing technology on SHM, traditional sensor data analysis methodologies and current machine learning and data mining based methodologies. Advanced sensing technologies confront challenges of durability and reliability. Some of the advanced technology, for example, optic fibre can avoid inference in a complicated environment. But optic fibre could be deformed due to environmental factors such as material surface temperature. PTZ sensor and cement-based sensor are widely used in many civil statures to measure the strain, but PTZs are sensitive which is easily interfered by the noise signal.

Frequency analysis methodology is extensively used in SHM problem. Methods like Fourier transform and wavelet transform are two most used frequency analysis technique. Recent, advanced studies attempt to use FT or WT to analyze lamb wave in order to locate the damage location. However, some of these methods have not applied in practice or only tested on some samples. There is a gap between experimental result and practise. To promote these advanced methodologies require more cooperation between academics and industries.

Application of machine learning and data mining methodology arouses dramatic attention. Bayesian-based method and neural network based method helps civil

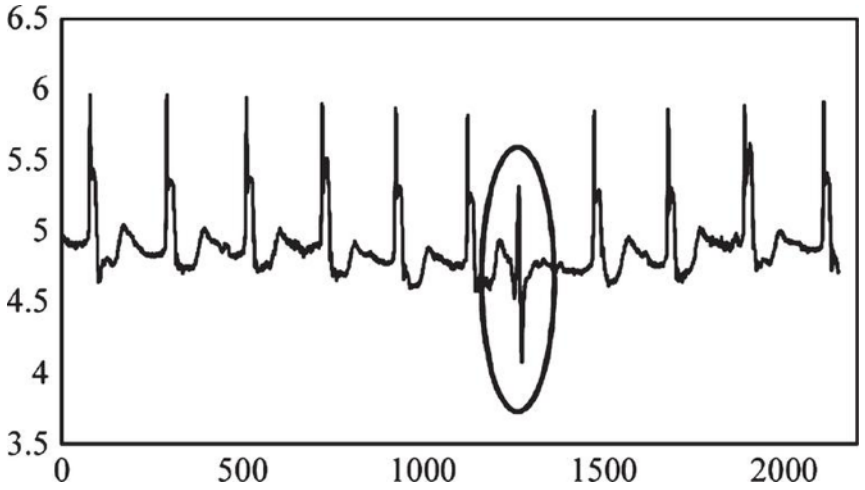


FIGURE 2.3: Example of anomaly

engineers to recognize damage signal of a structure or crack of a concrete. Most of the current application is offline analysis, and some of these methods only tested with some samples. A robust online analysis methodology is required to support decision making on SHM problem.

2.2 Traditional Anomaly detection

Anomaly detection is a branch of data mining technique. It attempts to identify any events, items or observations that not conform expected pattern in datasets (Figure 2.3 [46] is an example of an anomaly). It has been applied in various domains such as finance(i.e. credit card fraud detection), signal processing (i.e. military surveillance) , network security. Anomalies can be categorized into [47]:

- Point anomalies : if a data point is considered as an anomaly according to rest part of data points. Point anomalies are the simplest type of anomaly, and most of the researches focus on this type.

- Contextual anomalies : if data points are considered as anomalous with respect to its context. Most of the contextual anomalies can be found in time-series data.

- Collective anomalies : if a collection of data instances are considered as anomalous with respect to an entire dataset. Most of the collective anomalies can be found from some periodical signal dataset.

Anomaly detection also can be applied into SHM to help to identify abnormal data, which is one of the most critical parts of SHM. Most of SHM sensor data are time-series data , and most of the anomalies in SHM domain are contextual anomalies and collective anomalies.

Current anomaly detection methodology can be categorised by the supervised, semi-supervised and unsupervised method. In supervised anomaly detection method, a number of anomalies need to be labelled, as well as normal data. Based on the labels, models are built according to these labels. However, the anomalies are rare in datasets that the number of training samples is not enough to train models to recognized the anomalies. The imbalanced dataset is the most challenge for supervised anomaly detection. Unlike supervised learning, semi-supervised anomaly detection methods are much more applicable which do not need the labels for anomalies. Based on the label of normal data, the model for normal data are constructed. By comparison with new incoming data with trained models, anomalies can be detected. Unsupervised anomaly detection does not require any training sets and labels, which allows this category applicable widely. However, the challenge of this category increases the false alarm rate. Except for this categorization,

anomaly detection methods can be categorised into a distance-based method and model-based method as well. In the following sections, methods present according to this categorisation.

2.2.1 Distance based methodology

2.2.1.1 Global outlier method

Distance-based anomaly detection methods use distance as the primary arbitrary to identify anomalies. Generally, there are two sub-categories which are the global outlier and the local outlier. Data instance considered as a global outlier if the data instance deviates from a number of other data instances. In this sub-category, researchers are focusing on different purpose:

- Focusing on global outlier score computation Summation of data distance from its neighbours is one of the techniques used in anomaly detection. [48] [49] [50] used this technique to compute the outlier score to detect anomaly detection. In addition, these three methods also concerned about the efficiency of their proposed methods. In [48] study, it used only one dataset (dataset size is 10,000) which showed a good performance. However, one dataset fails to prove that this method is good for other datasets. [49] and [50] concerned with efficiency study, both of them fail to use a standard dataset to prove the performance of their method.

- Focusing on measurements of anomalies Some methods used another way to measure the distance to evaluate an anomaly. Counting the number of

neighbours is alternative way to detect anomalies [51] [52] [53]. This method cannot strictly to be considered as the density-based method since it count number of neighbours around data instance in radius range. The reciprocal of the number of neighbours is the anomaly score to a data instance. Based on the anomaly score, top-k data instances are considered as an anomaly. The hypergraph is another alternative way to detect anomalies. Based on the hypergraph, a strength of connectivity is computed to determine an anomaly in a categorical dataset.

- Focusing on efficiency Efficiency on detection anomaly detection is also critical as well since some anomalies occurred in some domains result in a detrimental impact, such as SHM, credit card fraud detection and network security. To improve the efficiency of detecting anomalies, pruning normal data instance is a key point. The threshold is a simple and common way to prune non-anomalous data. After calculating the distance of data instance, a threshold can screen the weakest score data instances. Partition technique and sampling technique can help to reduce the search space in order to detect anomalies efficiently. [54] used a partitioning technique to pre-process dataset via clustering technique. Based on the result of clustering result, anomaly detection method was performed on the interesting dataset (which contains anomalies). [55] used a sampling technique to enhance efficiency, it computed the distance from small samples of a dataset. Based on the previous result, anomaly detection method was performed to detect anomalies.

2.2.1.2 Local outlier method

The local outlier is another distance-based anomaly detection method. A data instance is defined as a local outlier when a data instance deviates from a number of neighbours around that data instance. Local outlier factor is the first proposed method using local outlier. It computes a local outlier factor for each data instance which equals the ratio of the average local density of a number of nearest neighbours and the local density of the data instance itself. To detect anomalies, the LOF [56] first found the smallest hypersphere from the dataset with k nearest neighbours. Then it computed the local density by dividing k by the size of the founded hypersphere. Any normal data instance within that hypersphere, the local density was similar with its density of neighbours, but abnormal data has a lower density comparing with other normal data in that hypersphere. Consequently, the LOFs of the anomalies are high which can be detected by the algorithm. Following researchers started to focus on different purse base on LOF :

- Efficiency improvement [57] and [58] proposed to improve LOF efficiency by using clustering technique. [57] used clustering technique to find micro-cluster with lower bound and upper bound. Based on the result, it performed LOF calculation to compute LOF. [58] enhanced others version of LOF to enhance the efficiency of LOF. It made some assumption about a problem to prune clusters that do not contain anomalies. LOF score was computed based on the remaining data instance to enhance the efficiency.

- LOF computation Except for the original LOF computation method, many other researchers attempt to the different way to compute LOF score for each data instance. [59] proposed a simpler version of LOF, which called Outlier Detection using In-degree Number (ODIN). It computed ODIN for each data instance, and the inverse of ODIN is the LOF of each data instance.

  [60] proposed a Multi-Granularity Deviation Factor (MDEF), which is a LOF based method. MDEF computed the standard deviation of the local densities of the nearest neighbours for a given data instance. The inverse of MDEF is the LOF of a given data instance. This paper not only used to detect anomalies but also used to detect anomalous clusters, which is called LOCI.

2.2.1.3 Summary

Global outlier based or local outlier based both are the unsupervised and semi-supervised method of anomaly detection. It does not require any assumption of distribution to a dataset. However, the computation of distance to each data instance is expensive and defining the anomalous threshold is a challenge in a complex dataset.

2.2.2 Model-based anomaly detection

Model-based anomaly detection is different from distance-based anomaly detection. It used different methods such as classification, probability, neural network,

and clustering. In the following section, anomaly detection using these methods are presented.

2.2.2.1 Classification based anomaly detection

Classification is considered as a supervised or semi-supervised anomaly detection method since it depends on the assumption that a classifier is capable of recognising the anomalous and normal data instance, which can be learned from during training phase. Generally, the classification-based method can be categorized into two sub-categories which are one-class anomaly detection and multi-class classification. According to existing labels, classes are trained for different purposes. In on-class classification, it assumed only one class label existed, and all normal data instances should be classified into one class. Such SVM method like [61] and [62] proposed methods to compute the class boundary to distinguish the normal data instance and anomalies.

Multi-class anomaly detection method was complicated than one-class anomaly detection method. This method assumed that data instances can be classified into multiple normal classes, any data instance does not belong to any normal classes is considered as an anomaly [63].

2.2.2.2 Probability based anomaly detection

Probability-based anomaly detection also classifies to supervised or semi-supervised anomaly detection method. Bayesian network is an extensive used and studied

probability-based method. It computes posterior probability from labelled normal classes and labelled anomalous classes. In general, this method aggregated all posterior probabilities from each test instance and based on the aggregated probabilities to determine the label of test instances. This method has been applied into intrusion detection with outstanding contribution [64][65].

2.2.3 Neural network based anomaly detection

There are many types of the neural network which can be used for anomaly detection. It can be considered as supervised and unsupervised anomaly detection, which depends on its type of neural network. In a tradition neural network, normal data are used to train for normal classes, and the test instances are used to test the training result. If the neural network accepted the test instance, which indicates this instance is normal, while any rejected instances are considered as anomalies[63]. Adaptive resonance theory [66] based neural network is an unsupervised neural network. It classified the input variables into different neurons by computing the forward weight and feedback weight. When a new neuron is set up when it fails in vigilance test and no more neurons to test [67]. If an input is greater than the pre-defined threshold, the weights are updated according to the input. This method like clustering method, which normal data and anomalies classify into separated clusters. Multi-layer perceptron (MLP) neural network also can be used for anomaly detection [47]. In the MLP neural network, there are a number of hidden layers, an original data instance is reconstructed through hidden layers, which uses the output of the previous layer as input for next layer. This

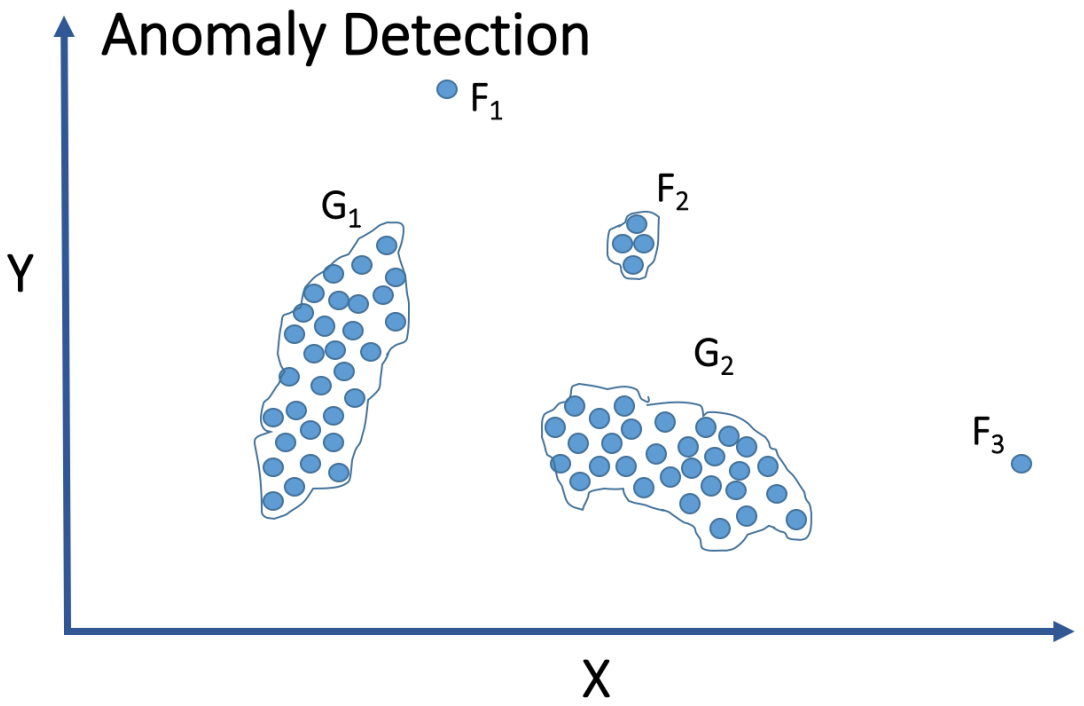


FIGURE 2.4: Example of anomaly detection using clustering method. F1,F2 and F3 are recognized as anomalies

method is a one-class anomaly detection method, the final result of data instances are similar, while anomalies are deviates from normal instances [68].

2.2.3.1 Clustering based anomaly detection

Clustering based anomaly detection method is usually considered as unsupervised anomaly detection, but some researchers also proposed some semi-supervised methods (Figure 2.4 [69] is an example of anomaly detection using clustering method). It can be categorized into three sub-categories: centroid based method, density based method and clusters based method [47].

The centroid is introduced with the development of clustering, which means a centre of a cluster. The centroid-based method assumes that normal data within a cluster should lies closely, while anomalies deviate from centroids. In centroid

method, data instances are clustered into clusters first. An anomaly factor is computed according to the distance to a centroid within a cluster. K-means [70], Self-organized map (SOM) [71], and Expectation maximization(EM) are three typical centroid based method. Unlike K-means and EM, SOM is a semi-supervised method which has been applied in intrusion detection, fault detection and fraud detection.

The density-based method assumes normal data should be classified into a large and dense cluster, but anomalies should be classified into a small and sparse cluster. Evaluation of size and density of a cluster is critically important for the density-based method. [72] proposed cluster-based local outlier factor(CBLOF) score to evaluate of data instance including the size of the cluster, and the distance of a data instance to its centroid. Other methods like [73] and [74] used threshold to screen normal dataset.

The cluster-based method assumes that anomalies fail to classify into any cluster, while normal data instances should belong to a cluster. [75] proposed FindOut algorithm to detect anomalies. However, the disadvantage of this cluster-based method is they are classifying data instances to a cluster rather than focusing on detecting anomalies.

2.2.3.2 Summary

Most of the model-based anomaly detection can be deployed by supervised and semi-supervised method. For these methods, they make some assumptions about the datasets followed by some models or distributions. During the training phase,

we have to provide enough labelled data instances to train models. Clustering can be deployed for unsupervised learning without assumptions. Comparing the distance-based method, the distance-based method is much more computationally expensive than model-based method. But there are many methods proposed to improve the efficiency of them.

2.3 Performance measuring

Once new methods are proposed, how to evaluate the performance of a method is important. There are many measurements to evaluate a method from different aspects. The most extensively used measurements to evaluate a method in data mining and machine learning are accuracy, specificity, sensitivity and precision. All of these measurements are derived from the confusion matrix. In this section, commonly used measurements in machine learning and data mining field are introduced.

2.3.1 Confusion matrix

The confusion matrix is used to evaluate the performance of a model in machine learning. This table contains two rows and two column reporting true positive(TP), false positive(FP), true negative (TN) and false negative(FN). Table 2.1 is the template of the confusion table. All measurements are derived from this confusion matrix.

	Condition positive	Condition negative
Predicted positive	TP	FP (Type 1 error)
Predicted Negative	FN (Type 2 error)	TN

TABLE 2.1: Template of confusion matrix

Accuracy(ACC) measures the quantity of that quantity's true value, equation 2.1 defined the accuracy.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.1}$$

Specificity (SPC) measures the proportion of negatives that have been correctly identified, equation 2.2 defined the specificity.

$$SPC = \frac{TN}{FP + TN} \tag{2.2}$$

Sensitivity (SEN) also known as recall measures the proportion of all positives that have been correctly identified, equation 2.3 defined the sensitivity.

$$SEN = \frac{TP}{TP + FN} \tag{2.3}$$

Precision(PRE) measures the proportion of correctly identified positives and all predicted positives, equation defines the precision.

$$PRE = \frac{TP}{TP + FP} \tag{2.4}$$

	Condition positive	Condition negative	
Predicted positive	TP	FP (Type 1 error)	$PRE = \frac{TP}{TP+FP}$
Predicted Negative	FN (Type 2 error)	TN	$NPV = \frac{TN}{TN+FN}$
	$SEN = \frac{TP}{TP+FN}$	$SPC = \frac{TN}{FP+TN}$	$F = 2 * \frac{SEN*PRE}{SEN+PRE}$

TABLE 2.2: Confusion table summary

Negative predictive value (NPV) measures the proportion of correctly identified negatives and all predicted negatives, equation 2.5defines the NPV.

$$NPV = \frac{TN}{TN + FN} \tag{2.5}$$

F score, also known as the F1 score, which is a harmonic average of sensitivity(recall) and precision. It gives equal weight to both precision and sensitivity. Equation 2.6 defines the F1 score :

$$F = 2 * \frac{SEN * PRE}{SEN + PRE} \tag{2.6}$$

Many researchers argued that F1 score is not reliable under some special circumstances due to its bias. In the natural language processing (NLP) field, the F1 score is widely used for evaluating NLP algorithm.

To sum up, table 2.2 shows the summary of above equations.

Chapter 3

An approach to the computation of time-series data similarity using Sample Entropy Gradients

Traditional anomaly detection method like clustering based methodologies compare to local density; or distance to centroids; or a number of neighbour clusters. The distance-based methodologies attempt to compute a global outlier score or local outlier score to determine anomaly with respect defined thresholds. In a time-series dataset, time is a unique characteristic of time-series data. The challenge of time-series data is the dynamic distribution that the distribution is change over the time. Consequently, we developed a gradient-based algorithm using data sample entropy (SEG) for trend and outlier prediction in high-frequency time series data streams. To answer the first question, we investigated cross sample entropy which is applied widely in medical time-series data analysis and propose our SEG

method. We conduct practical data experiments on SEG algorithm to two application areas: dynamic wind speed data stream; and financial time series data. Our experiments demonstrate that SEG algorithm could be feasibly used in online implementation to derive predictive early warning signals to a domain-specific decision maker.

3.1 Motivation

Cross-sample entropy (CSE) has been applied in many domains, including medical, engineering, and financial analyses [76–78, 78–82], where it has provided good results. It specializes in analyzing time-series and comparing the similarity of two time-series data. However, it is still necessary to specify certain parameters (i.e., r and m) to execute the computation, especially r, that indicates a tolerance of similarity. In fact, however, it is difficult to control tolerance in order to compare similarity. Except for similarity, CSE fails to indicate the similarity over the time stamp between two datasets/signals, so it is incapable to be applied into anomaly detection domain. Since the Cross-sample entropy only gives value to indicate the similarity between two time-series datasets. Therefore, we proposed our method called 'sample entropy gradient' (SEG), which detects the correlations of two time-series datasets/signals without using a tolerance parameter and able to apply to anomaly detection. The SEG was developed according to the CSE method, which is specialized for measuring the similarity and correlations of time-series datasets. It also can be used to anomaly detection problems, since the nature of anomaly detection is trying to find the particular dissimilarity.

Cross-sample entropy is devised based on sample entropy (SE), which is an enhanced version of approximate entropy (AE). These methods are concerned with nonlinear dynamic analysis for biological data. AE method observes the similarity of a given dataset. AE [83] and SE observe time-series datasets for similarity. Informally, given a dataset, AE and SE measure the similarity of two sequences that are subsets of the given dataset. A lower value of AE or SE indicates the higher regularity and lower complexity of a given dataset. Importantly, AE counts a sequence as matching itself, whereas SE does not. In fact, SE is an enhanced version of AE that avoids the limitations of AE in measuring the complexity and regularity of a given dataset.

Cross-sample entropy extends the theory of AE to compare the similarity of two datasets. Other than comparing sequences of a given dataset, CSE has also been used to compare sequences from two datasets. A higher value of CSE reflects a lower degree of synchrony or dissimilarity of two datasets/signals and vice versa.

The computation of AE, SE or CSE [84] requires specification of two parameters, m and r, i.e., the length of a sequence (subset) and the tolerance of similarity, respectively. Consequently, tolerance r is the critical parameter that has an impact on the final result. Theoretically, people usually use standard deviation times 0.2 as the similarity tolerance when computing AE and SE. For CSE, 0.2 is also common selection. Furthermore, CSE can only identify the similarity and synchronicity of two time-series datasets/signals. It is unable to demonstrate the relationship between two time-series datasets/signals or to quantify the correlation between two time-series datasets/signals.

Sample Entropy gradient (SEG) is a new method for measuring the similarity between two time-series datasets. It is able to reveal the relationships between time-series without requiring a tolerance r parameter. Instead of measuring the difference between sequences, SEG measures the gradient of sequences (subsets), avoiding the use of a tolerance parameter to measure similarity. Moreover, SEG can be extended to anomaly detection problems by comparing with a benchmark dataset.

In this study, we computed and compared SEGs of two time-series datasets. As an illustrative example, we conducted experiments by injecting a number of outliers into a normal dataset to see whether our SEG method is able to detect these outliers. In our experiment, we introduced practical sensor data from a cable bridge.

3.2 Preliminary

3.2.1 Approximate entropy

Approximate entropy was developed by Grassberger and Procaccia and by Eckmann [85] and Ruelle [83, 86]. AE requires two important parameters, m and r, where m is the length of the sequence (a subset of a dataset) and r is the tolerance. AE compute the natural logarithm of the conditional probability of two similar sequences (with length m). It reflects the complexity and repeatability (regularity) of a dataset. A higher value of AE indicates the higher complexity and lower

regularity, and vice versa. The formula is defined in 3.1 :

$$AE(S_n, m, r) = ln(\frac{C_m(r)}{C_{m+1}(r)}) \tag{3.1}$$

where $C_m(r)$ is defined in 3.2

$$C_m(r) = \frac{\sum_k^m C_{im}(r)}{n - m + 1} \tag{3.2}$$

where $C_m(r)$ is defined in 3.3

$$C_{im} = \frac{N_{im}(r)}{n - m + 1} \tag{3.3}$$

where the $N_{im}(r)$ is the number of similar sequences (subsets). The similarity is computed by 3.4:

$$|S_{Ni} - S_{Nj}| < r, x < y < N \tag{3.4}$$

where S_{Ni} and S_{Nj} are the data in the sequences (subsets). Importantly, AE counts similar sequences including itself. In other words, the number of similar sequences includes a self-match count. As AE includes the self-matching count to avoid the $ln(0)$ circumstance, it introduces bias to the result [87].

Let A_1 equals to probability of similar sequences (with length m), and B_1 equals to probability of similar sequences (with length m+1). A_1 and B_1 are computed according to the equation 3.2 and $N_{im}(r) \geq 1$, therefore $\frac{A_1}{B_1} > 0$.

If we use the same sequence without self-count, the A_1 and B_1 are denoted as A_1' and B_1' respectively. Since without the self-count, the $C_m(r)$ of A_1' and B_1' are ≥ 0,

which $N_{im}(r) \geq 0$. Therefore, $\frac{A'_1}{B'_1} \geq 0$.Therefore, $\frac{A'_1}{B'_1} < \frac{A_1}{B_1}$ which indicates AE is biased with self-count. Sample entropy is introduced for reducing bias.

3.2.2 Sample entropy

As AE introduces bias to the computation, the most straightforward way to remove that bias is to ignore the self-matching count. Sample entropy is an enhanced version of AE and is a bias-free method. SE is same as AE in that it requires two important parameters for computation, namely, m and r (length of the sequence (subsets)) and tolerance. Equation 3.5 defines SE.

$$SE(S_n, m, r) = ln(\frac{A}{B}) \tag{3.5}$$

where A is defined in equation 3.6, B is defined in equation 3.7 respectively.

$$A = A^m(r) = \frac{\sum_{i=1}^{N-m} A_i^m(r)}{n-m} \tag{3.6}$$

where $A_i^m(r)$ is the probability of matched sequences with m+1 length.

$$B = B^m(r) = \frac{\sum_{i=1}^{N-m} A_i^m(r)}{n-m} \tag{3.7}$$

where $A_i^m(r)$ is the probability of matched sequences with m length. If A or B is equal to 0, it indicates that there is no complexity or regularity of a given dataset.

If we want to measure the similarity of two time-series datasets, SE fails to meet the requirements since it only applies to a single dataset. As a result, CSE is introduced for measuring the similarity of two datasets.

3.2.3 Cross sample entropy

CSE[84, 88] is introduced to deal with the comparison of two different time-series datasets. It is able to describe the similarity or synchrony of two time-series datasets/signals. The definition of CSE is similar to that of SE but, instead of comparing the similarity of sequences from a single dataset, it compares a pair of sequences (subsets) from two datasets individually. Let v and v be the two different n length datasets, each of which can be divided into m-length sequences (subsets), namely, $x_m = (u(i), u(i+1), \ldots, u(i+m-1))$ and $y_m(j) = (u(j), u(j+1), \ldots, u(j+m-1))$, where $1 \leq i, j \leq N - m$. The CSE is defined in equation 3.8.

$$CSE(m, r) = -ln(\frac{A^m(r)(v||u)}{B^m(r)(v||u)}) \tag{3.8}$$

where $A^m(r)(v||u)$ is defined in equation 3.9

$$A^m(r)(v||u) = \frac{\sum_{i=1}^{n-m} A_i^m(r)(v||u)}{n-m} \tag{3.9}$$

where $A^m(r)(v||u)$ is defined in equation 3.10

$$A_i^m(r)(v||u) = \frac{N(s[x_{m+1}(i), y_{m+1}(j)])}{n-m} \tag{3.10}$$

where $s[x_{m+1}(i), y_{m+1}(j)]$ is defined in equation 3.11

$$s[x_{m+1}(i), y_{m+1}(j)] = max(|u(i+k)| - |v(j+k)|); \tag{3.11}$$

The CSE measures the similarity of each pair of data points and uses the maximum value as the similarity of those two sequences. The definitions of $B^m(r)(v||u)$ and $B_i^m(r)(v||u)$ are similarly shown in 3.12 and 3.13:

$$B^m(r)(v||u) = \frac{\sum_{i=1}^{n-m} B_i^m(r)(v||u)}{n-m} \tag{3.12}$$

where

$$B_i^m(r)(v||u) = \frac{N(s[x_m(i), y_m(j)])}{n-m} \tag{3.13}$$

CSE needs parameters, like SE and AE, to complete the computation. It compares two datasets rather than a single dataset. Thus, it is crucially important to select a proper tolerance. In general, the selection of tolerance r ranges from 0.1 to 0.25. In cases that require high precision in the measurement of similarity, the tolerance selection range from 0.1 to 0.25 [77, 81, 89] is no longer useful. For instance, we assume that we have two sequences, [0.11, 0.12, 0.13 . . .] and [0.12, 0.12, 0.11. . .], respectively. If we still adopt the tolerance range from 0.1 to 0.25, these two sequences are recognized as similar, whereas by observation we would expect the result to be dissimilar. If an inappropriate tolerance is used, the entire result is affected detrimentally. In other words, subsequent analysis and interpretation of the data are misinterpreted.

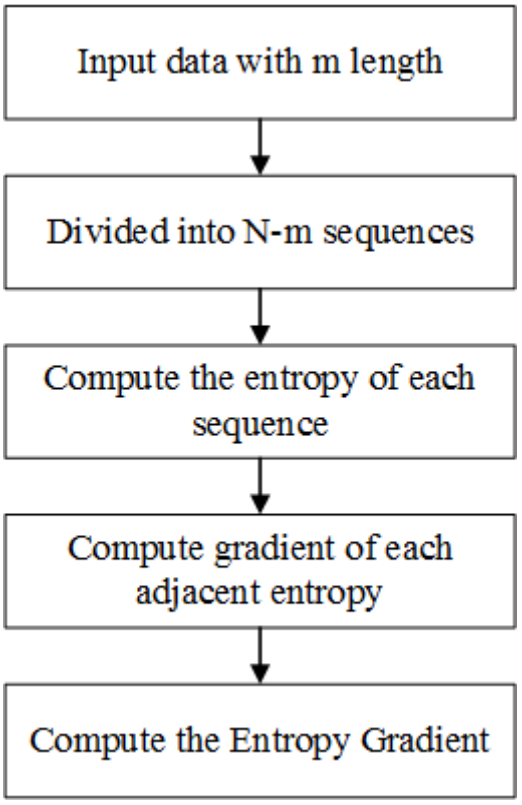


FIGURE 3.1: Process of SEG

Moreover, CSE cannot demonstrate inter- and intra-period correlation. It can only provide an overall value to indicate the degree of the similarity and synchrony between two datasets. As well, it fails to quantify the similarity of different periods.

3.2.4 Sample entropy gradient

By exploiting time-series data, we propose an approach to compute the similarity (Euclidean distance) of two time-series using SEGs derived from real data. The SEG is a novel measure that has not previously been described. The process of our approach for similarity computation is in Fig.3.1:

The SEG is a new method for comparing two time-series datasets/signals. It enables us to investigate the correlations between two time-series datasets/signals

with different periods. In addition, it is also able to find the outliers by comparing with a standard benchmark dataset. The SEG overcomes the limitations of CSE in that there is no need to select the tolerance in order to avoid its impact. Furthermore, the SEG provides a mechanism to observe the outliers. Unlike CSE, we use the SEG to measure both correlation and similarity, avoiding the need for a tolerance parameter. In addition, the SEG implies the outliers of a dataset that can be used for real-time anomaly detection.

Algorithm 1: Main CSE algorithm

Data: dataset u and v
Result: Distance of x_m and y_m
$x_m(i) \leftarrow(\mathrm{u}_i, u_(i+1), u_{i+2}, ..., u_{i+k})$ // x_m `contains all` $x_m(i)$
$y_n(i) \leftarrow(\mathrm{v}_i, v_(i+1), v_{i+2}, ..., v_{i+k})$ // y_m `contains all` $y_m(i)$
$Entropy(x_m) \leftarrow \phi$
$Entropy(y_m) \leftarrow \phi$
$D(x_m, y_m) \leftarrow \phi$
while $i \leq n\text{-}m$ **do**
  $Entropy(x_m) \leftarrow Entropy(x_m(i))$
  $Entropy(x_m(i)) \leftarrow Entropy(x_m(i))$
end
while $i \leq sizeofEntropy(x_m$ *and* $i \leq sizeofEntropy(y_m)$ **do**
  $D(x_m, y_m) \leftarrow Distance(Entropy(x_m(i), y_m(i))$
end
return $D(x_m, y_m)$

Generally, the principle of the SEG is similar to that of CSE. A complete set is divided into several sequences (subsets) according to the length m, which indicates the time period as a unit. For each sequence, entropy is computed based on the frequency distribution of each sequence/subset. For two adjacent entropies, we compute the slope them. Fig. 1 shows the entire process of the SEG. For two time-series datasets $u = \{u_i, u_{i+2}, u_{i+3}, ..., u_{i+k}\}$ and $v = \{v_i, v_{i+2}, v_{i+3}, ..., v_{i+k}\}$, the sequences can be formed :

$$x_m(i) = (u_i, u_(i+1), u_{i+2}, ..., u_{i+k})$$

$$y_m(i) = (v_i, v_(i+1), v_{i+2}, ..., v_{i+k})$$

The SEG can be defined by equation 3.14 and 3.15

$$SEG(x_m) = L(Entropy(x_m(i), x_m(i+k))) \tag{3.14}$$

$$SEG(y_m) = L(Entropy(y_m(i), y_m(i+k))) \tag{3.15}$$

where $entropy(.)$ is computed by 3.16 and L is a linear regression function:

$$Entropy = -\sum_{i=1}^{n} -ln(p(s_i))p(s_i), s_i \in x_m or y_m \tag{3.16}$$

where $p(.)$ is the probability of s_i.

Fig. 3.2 shows an example of SEG. For any two time-series data stream, both time-series data streams are equally divided into identical number of sequences; we computed entropy value for each sequence of both time-series data stream; in case Fig 3.2, we linearize each two adjacent entropy to compute its SEG (In general case, the number of linearization process can be customised according to the specific requirements). For each pair of the sequence in both time-series data stream, we compute the Euclidean distance (L2) to determine the similarity in that time-stamp.

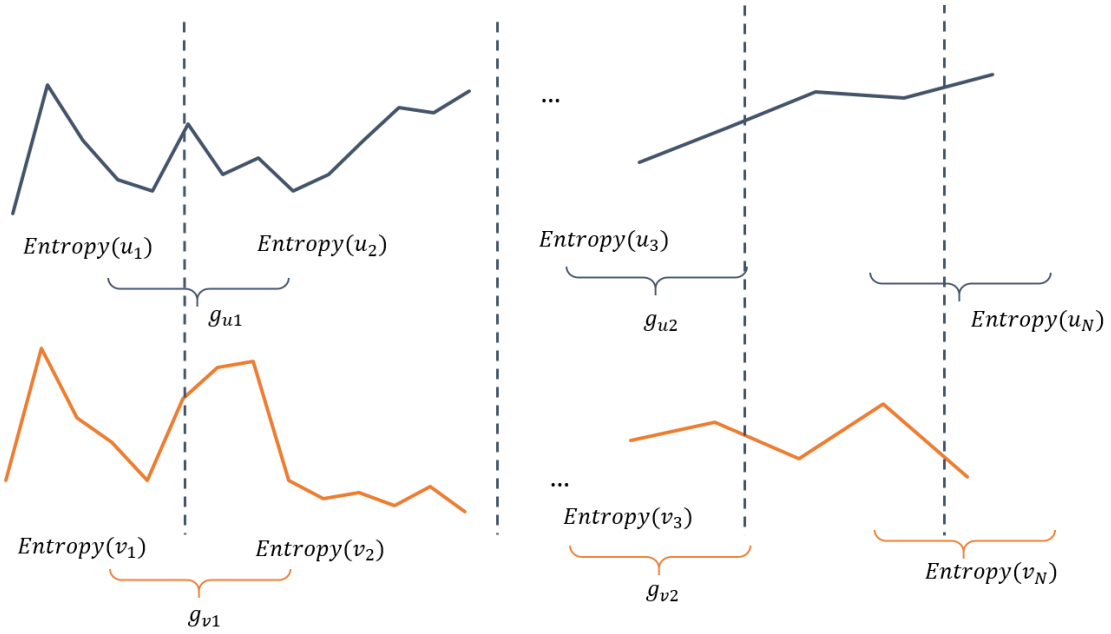


FIGURE 3.2: Example process of SEG

3.3 Empirical Evaluation

To evaluate our proposed approach, we used two different domains of time-series data, which are wind velocity data collected from a cable bridge and financial data. We investigated if our method is able to detect outliers we injected in the dataset. In anomaly detection experiment, we injected a different amount of outliers into dataset to see whether our method is affected by the density of outliers. If the place of detected outliers are consistent with our arranged injecting place, which can indicate our method is able to detect the outliers.

From the financial study, we investigated the correlation between Australian stock market (NASDQ) and American stock market (ASX). Our SEG method gives a quantitative analysis between NASDQ and ASX, and we also survey relevant financial report and papers to prove our quantitative analysis (Further detail of financial study is presented in Appendix A).

Data length	Outlier rate
86400	0.2%
86400	0.4 %
86400	0.6 %
86400	0.8%

TABLE 3.1: Wind speed dataset

3.3.1 Data

3.3.1.1 Structural health dataset

We introduced the sensor data collected from a cable bridge. Specifically, we used the wind load data as our experiment data. The length of the wind load data is 86,400, which collects data every second (1Hz). The difference within a minute (60 seconds) is slight, so we aggregate the data into 1440 points by computing the average of every minute. In this experiments, we have four different altered datasets which inject with a different number of outliers. Table 3.1 showed the outliers status.

3.3.1.2 Financial dataset

All finance data were downloaded from Yahoo Finance [89, 90] from March 1, 2000 to September 1, 2015. The share market indices were the ASX and NASDAQ In this study, we used only the closing price over the given period.

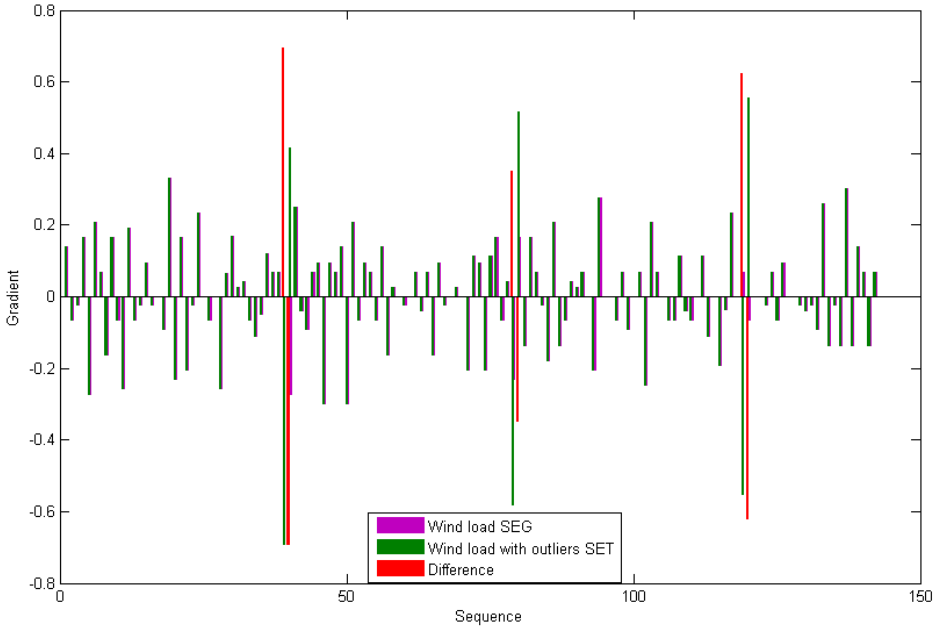


FIGURE 3.3: Result of 0.2% injection rate

3.3.2 Results

Fig.3.6 shows the result of 0.6% outlier rate. We injected an outlier at every 150th of the original data. We computed the entropy of every 10 points, consequently, at sequence 15, 30, 45 and so forth. Totally, there are nine outliers are detected by our method. Any outlier detected by our SEG method showed by the red bar, otherwise, the SEG of altered data and original data should be identical. For each outlier, it would result in two adjacent SEG change. The result reveals that the injecting place is consistent with where we detected outliers.

Fig.3.5 shows the result of 0.4% outlier rate, which we injected an outlier at every 200th of the original data. We computed the entropy of every 10 points, consequently, at sequence 20, 40, 60 and so forth. Totally, there are 7 outliers are detected by our method. Any outlier detected by our SEG method is shown by the

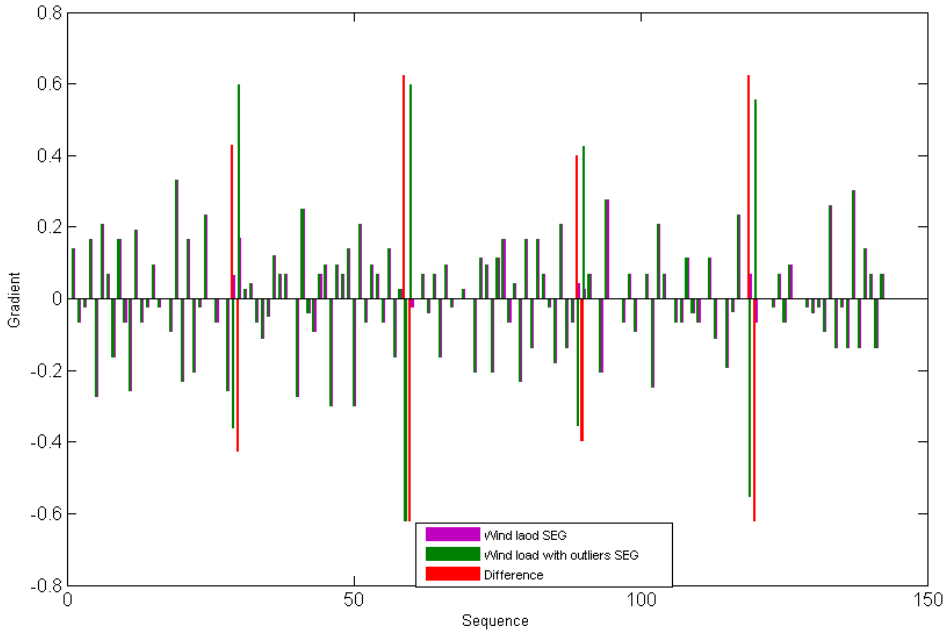


FIGURE 3.4: Result of 0.3% injection rate

red bar, otherwise the SEG of altered data and original data should be identical. The result shows that the injecting place is consistent with where we detected outliers.

Fig.3.4 shows the result of 0.3% outlier rate, which we injected an outlier at every 300th of the original data. We computed the entropy of every 10 points, consequently, at sequence 30, 60, 90 and 120. Totally, there are 4 outliers are detected by our method. Any outlier detected by our SEG method is shown by the red bar, otherwise the SEG of altered data and original data should be identical. The result shows that the injecting place is consistent with where we detected outliers.

Fig.3.3 shows the result of 0.2% outlier rate, which we injected an outlier at every 400th of the original data. We computed the entropy of every 10 points, consequently, at sequence 40, 80 and 120. Totally, there are 3 outliers are detected

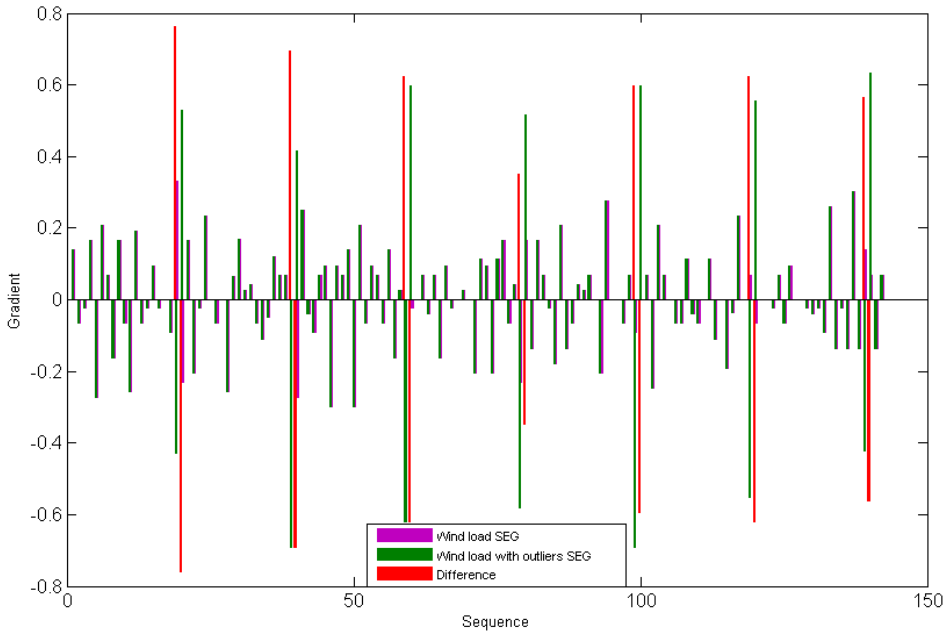


FIGURE 3.5: Result of 0.4% injection rate

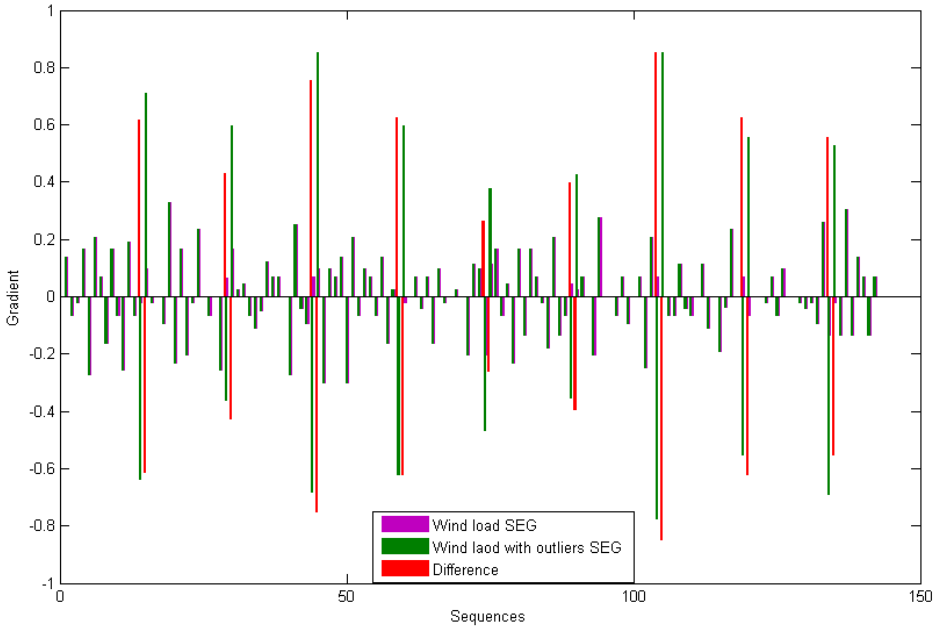


FIGURE 3.6: Result of 0.6% injection rate

by our method. Any outlier detected by our SEG method is shown by the red bar, otherwise the SEG of altered data and original data should be identical. The result shows that the injecting place is consistent with where we detected outliers.

In conclusion, our method is able to detect the outliers and consistent with our arranged place. However, benchmark dataset is the most important component in our experiment. Every altered dataset needs to be compared with a benchmark dataset for detecting the difference between them.

3.3.2.1 Case study of between ASX and NASDAQ

Fig.3.7 shows the entropy gradients of the NASDAQ and the ASX indices. It is evident that the ASX index is correlated with the NASDAQ index, especially over the first six sequences (the first 180 trading days); sequence 21 to sequence 34 and sequence 74 to 85 are also highly correlated. From the political perspective, Australia is an ally of the U.S., sharing information and interacting closely [91]. Also, branches of many international corporations are located in Australia.

Reference [92] provides a detailed report of corporations financed by the U.S. From this report, JP Morgan Chase and Citibank were two major investors in some of Australia's largest corporations including the banking, mining, and retailing sectors and so forth. These two corporations accounted for at least 11% of the shares of those companies (details can be found in the report). In banking industry especially, JPMorgan Chase and Citibank held at least 30% of the shares of four of the most important Australian banks (Commonwealth, National Australia Bank, Westpac, and ANZ). Thus the U.S. has a tremendous impact on the Australian

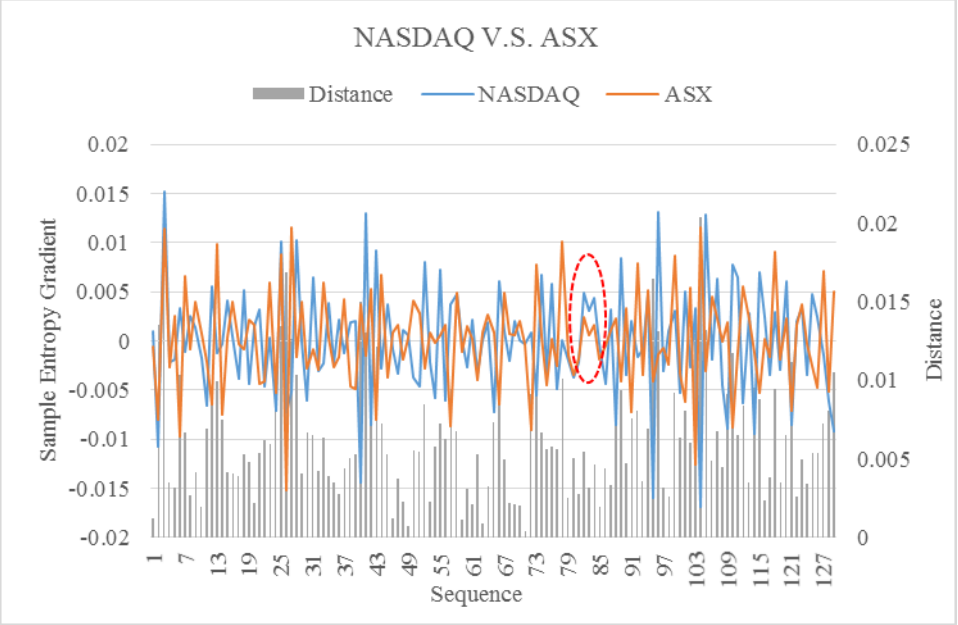


FIGURE 3.7: NASDAQ SEG and ASX SEG

financial market. Consequently, the ASX index correlates with the NASDAQ index over many periods. Fig. 3.7 demonstrates the Euclidean distance between the ASX index entropy gradient and the NASDAQ index entropy gradient as well by the grey bar chart. The gradient distance reflects that the difference between the two entropy gradients is small, ranging from 0.0011 to 0.02. Most of the distances between these two entropy gradients are around 0.01. Thus, we can conclude that the ASX is correlated with the NASDAQ over this period.

3.4 Summary

In this chapter, we proposed a SEG method for comparing two time-series datasets/signals for similarity and anomaly detection. The SEG method avoids the need to use a tolerance parameter, allowing the method to be adapted to various cases. Moreover, it allows comparison of time-series datasets/signals for specified time

segments with quantified differences, enabling us to observe similarity or abnormalities over the different time segments. Our experiment demonstrated that the method could identify the outliers by comparing with a benchmark dataset. Our financial case study demonstrates there is a correlation between ASX and NASDAQ in different timestamps, and surveys and reports can support our analysis and the further financial study are presented in Appendix A. For anomaly detection, the limitation of our method is dependent on a well-constructed benchmark dataset. The main material of this chapter has been published in IEEE International Conference on Industrial Electronics and Applications (ICIEA 2017).

Publication arising in this chapter:

- D.Sun, V.Lee and Y. Lu, "A Gradient-based Algorithm for trend and outlier prediction in dynamic data streams", proceeding of International Conference on Industrial Electronics and Applications (ICIEA) ERA Rank A Conference, June 2017, Sime Reap, Cambodia. Proceeding of IEEE ICIEA pp.1975-1980.

Chapter 4

An ensemble kernel density estimator method for online high-frequency data stream anomaly detection

In Chapter 3 we proposed the SEG method to detect anomalies and analyze two time-series datasets. However, there are some limitations of our SEG method:

- Highly relying on the benchmark datasets for anomaly detection
- High-quality of benchmark datasets is expensive to produce
- Hard to determine the euclidean distance for detecting anomalies under poor benchmark datasets

Therefore, a benchmark dataset free method is required. To answer the first research question, in this chapter I proposed an ensemble kernel density method for online anomaly detection. It applied ensemble analysis, kernel density estimator and sliding window to detect anomalies for a data stream. To evaluate the performance of our EK method in terms of accuracy, specificity and sensitivity, we compared it with other method using UCI and structural health monitoring (SHM) data in which accuracy and sensitivity, in particular, are significant measurements. Any type-2 error (false negative) can have a fatally detrimental impact on both structural and public safety. The results show that our EK method is capable of capturing anomalies with a rapid reaction time compared to other online anomaly detection methods.

4.1 Motivation

Online anomaly detection is a branch of anomaly detection that helps monitor the status of incoming data in real time. Traditional anomaly detection suites focus on assumptions of static data distribution and environments. Researchers and scientists have worked extensively on this problem across a wide range of domains, with various results, as reviewed in [93]. Online detection of anomalies in data streams has been undertaken in recent years, where the challenge lies in the dynamically changing data environments. Application of current online anomaly detection methods to data stream anomaly detection is inappropriate for rapidly changing data environments. Time variation, in which the data stream possesses uncertainties and non-stationary process characteristics, is a unique characteristic

which differentiates dynamic from static data environments. For instance, a data point may be recognized as non-anomalous based on an offline method, but it may actually be an anomaly in a specific time-segment. To overcome this challenge, we proposed an ensemble kernel density estimator (EK) method for online anomaly detection in high-frequency data streams. Our method adopted a sliding window principle [94] and used an ensemble analysis principle for recent temporal data densities to evaluate abnormalities in incoming data[95]. We conducted experiments on UCI datasets and SHM data collected from a highway steel cable suspension bridge in China. SHM deploys various types of sensors on a structure to detect any abnormal status. Wind speed is one critical data type of concern in SHM. With increasing wind speed, a structure (especially a cable bridge) can bear dramatical variations in wind load and structural stress. We introduced high-frequency wind speed data to test the performance of the EK method and compare it with other methods. Our experimental findings suggest that the proposed EK method has stable and significantly better performance in terms of accuracy, sensitivity, specificity and efficiency in of online anomaly detection.

4.2 Related Work

Anomaly detection problems have been extensively studied for years. Previous anomaly detection work has been based on the assumption of static and stationary data distribution. In real climate data environments, streaming data are non-stationary, yet, coupled with the need for timeliness of decision making, online anomaly detection is essential. For instance, civil structural engineers' concerns

about the real-time safety of a structure (e.g. a bridge, a building) belong to a branch of SHM. Offline anomaly detection would clearly be inappropriate in such context, a situation that leads us to online anomaly detection, the most important challenge resulting from dynamic changes within an incoming data stream.

Two general categories of anomaly detection methods exist, the distance-based method and the model-based method. Distance-based methods use the distance of a data point away from an arbitrary threshold as the criterion for judging the anomaly of an actual data point. Global outliers and the local outliers are two sub-categories of distance-based anomaly detection methods. A data point is defined as a global outlier if it deviates more than the distance R (a domain specific threshold) from k data points. Ref. [96] first proposed the use of global outlier approach to detect outliers. However, because R is a constant threshold for measuring global distance, that approach fails to produce a good performance (in terms of accuracy) for heterogeneous datasets. A local outlier differs from global outliers, in that any data point is defined as a local outlier if it deviates more than distance R from the k nearest neighbours. [97] and [98] each presented enhanced versions of LOF that could be applied to data stream anomaly detection. However, these local outlier approaches assumed that distribution was static, without any impact on performance. Furthermore, it is difficult to select the k nearest neighbours in a dynamically changing environment, especially in high-frequency data stream environments. Adaptive online outlier detection for data stream (AODDS) [99] is another distance-based method for data stream. AODDS adopts a global deviation factor (GDF) and a local deviation factor (LDF) to determine anomalies

in data streams. However, AODDS is not efficient in detecting anomalies especially in a high frequency environment, because it consumes a large amount of time for computing both local distance and global distance, as demonstrated in our experiment in Section 5.4.2.2.

Model-based anomaly detection is an alternative method category for anomaly detection, which attempts to capture anomalies via modelling data. [100] and [101] are two examples of model-based methods for offline anomaly detection. [100] adopted a clustering method based on a graph-based method to extract clustering and outliers at the same time. [101] proposed a support vector domain description to establish a tight boundary for describing a dataset in order to distinguish anomalies by measuring the distance to the boundary. However, both methods fail to be applied to a high-frequency data environment. The relevance-weighted ensemble model [95] and the method of outlier detection in stream data by K-means clustering [102] are two model-based methods. [95] attempted to model normal data in a previous period by clustering according to previous normal clusters. That clustering formed a relevant normal model for detecting anomalies in the current period. However, the method was designed for switching data streams rather than high-frequency data streams. [102] adopted a sliding window principle; that applied an incremental K-means clustering method to capture outliers. However, an increase in data volume had a negative impact on accuracy and detection time of the method.

Ensemble analysis enhances the performance of anomaly detection methods and

[93] has been adopted widely in anomaly detection. An analyst's primitive understanding of a dataset is subjective, and the selection of a model(s) or function(s) contains elements of the analyst's preference. Thus, the result produced by the selected functions or models would also be biased toward the analyst's choices. Consequently, ensemble analysis has been introduced in recent years to alleviate such subjectivity and dependence on model or function. Ensemble analysis can reflect whether the categorization is determined by component dependence or categorized by constituent components. Categorization by component dependence classifies the method in terms of dependence; specifically, whether a model or function is dependent on another model(s) or function(s). Sequential ensemble and independent ensemble are two main sub-types of this categorization. Sequential ensemble analysis is the output of an algorithm or a set of algorithms which have an impact on the next algorithm(s). [103] and [104] proposed two typical sequential ensemble approaches. In an independent ensemble, the outputs of different or the same algorithm(s) are independent of each other, but the final result is aggregated from all outputs. Many approaches have adopted independent ensemble analysis, such as the approaches of [105] and [96].

Categorization by constituent components classifies approaches as either model-centred or data-centred. Model-centred approaches aggregate different results given by different models. The approaches of [96] and [105] are two examples of model-centred ensemble analysis. Data-centred ensemble analysis aggregates the results from different parts or samples of an entire dataset. [106] used a data-centred ensemble analysis. This data-centred method is also known as a feature

bagging or subspace ensemble method [107]. In another example of a data-centred ensemble method, In [108] employed a genetic approach to detect anomalies via evaluating data behaviours in sub-spaces.

Our proposed EK method is model-based (distance-based or model-based categorization) and data-centralized (ensemble categorization). It models the density of a number of short-term period data for evaluating the abnormality of new incoming stream data in the current period, in which the period is defined according to the width of a sliding-window.

Our proposed EK method makes the following contribution:

1. It can detect anomalies under a high-frequency data stream environment
2. It overcomes the dynamically changing environment challenge of online data stream anomaly detection
3. It is capable of capturing anomalies efficiently with around 92% to 94% accuracy
4. It is capable of remaining at least 90% accurate under a massive volume of data.

4.3 Proposed Method

4.3.1 Problem formulation

A given finite data stream $X = \{x_1, x_2, x_3, ..., x_n\}$ is divided in to n windows with length m, denoteed as $W^m = \{w_1^m, w_2^m, w_3^m, ..., w_N^m\}$, where $w_i^m = \{x_i, x_{i+1}, x_{i+n}...x_{i+n}\}(i+n = m)$.The number of windows and the width of a window are constant, but the data within windows are change dynamically over time, which is also known as a sliding window. The principle of a sliding window has been applied in many domains [94], and we adopte this principle in our EK method. The densities (distributions) of $w_i^m \in W^m$ denote as $\theta(W^m) = \{\theta_1^m, \theta_2^m, \theta_3^m, ..., \theta_n^m\}$. Any incoming data point x_i is substituted into $\theta(W^m)$. The result of $\theta(W^m)$ is a vote vector (where $\theta_i^m \in \theta(W^m)$) is denoted as $V = \{v_1, v_2, v_3, ..., v_n\}$. For each vote vector the EK method computes an anomaly factor(AF),described in Section 4.3.4, which is denoted as $a_i \in A$ where $A = \{a_1, a_2, a_3, ..., a_n\}$. Any $a_i \geq K | 0 \leq K \leq 1$ is considered as an anomaly, where K is the threshold of AF depending on the problem domain.

4.3.2 Overall description

The EK method uses an ensemble analysis based on a kernel estimator for online anomaly detection in high-frequency data streams. It can be categorized as an independent ensemble analysis (in terms of the component independence category)

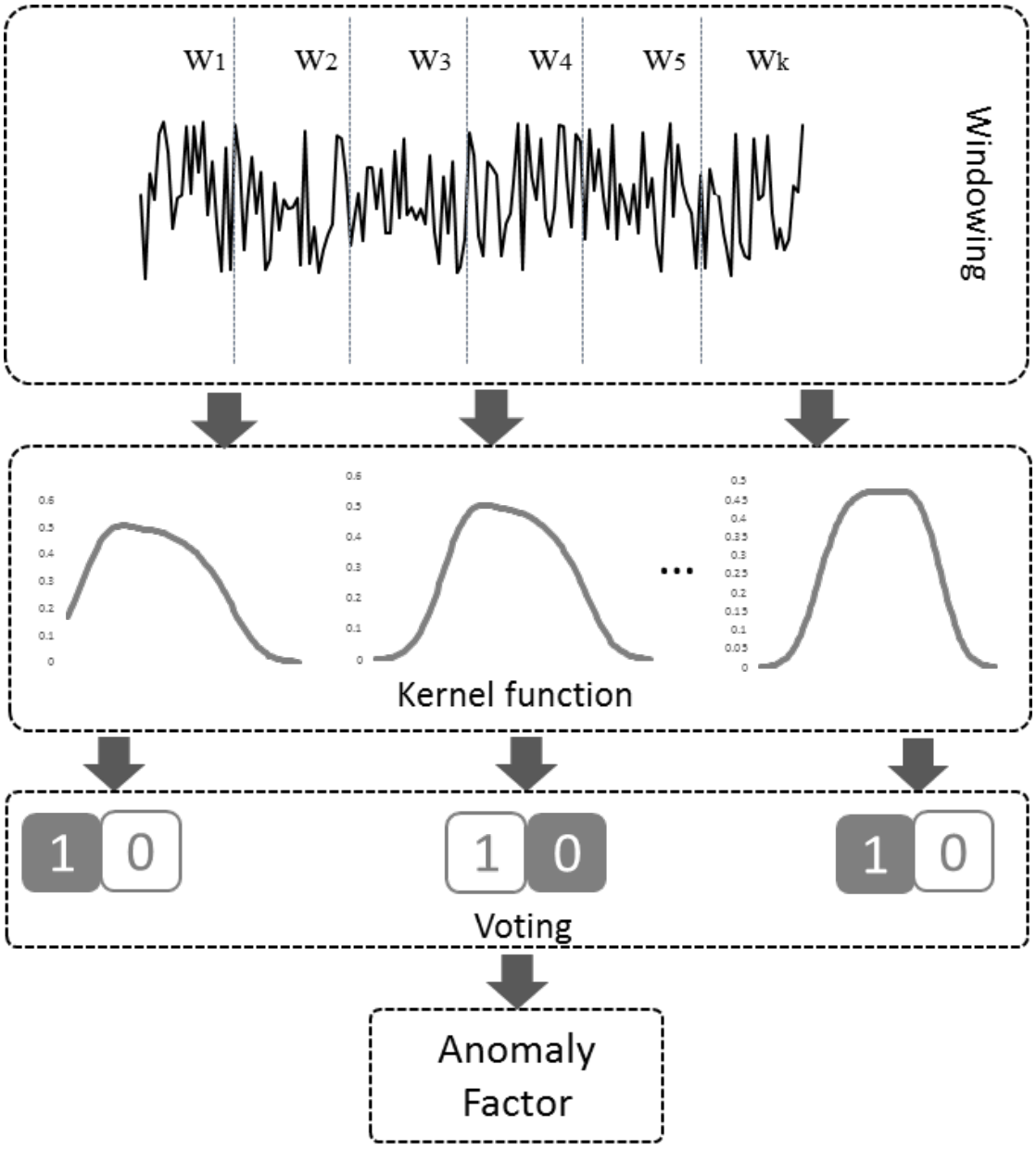


FIGURE 4.1: Procedures of Ensemble Kernel

or a data-centred ensemble analysis (in terms of the constituent component category), which uses a kernel estimator method (model-based anomaly detection). Our proposed EK method can detect anomalies 1) under dynamic distribution environments and high-frequency environments, 2) in real-time (online detection), and 3) with a rapid detection time. Generally, the EK method contains following steps (shown in Fig. 4.1):

1. Windowing

   In a finite data stream $X(X = \{x_1, x_2, x_3, ..., x_n\})$, we have n sliding windows of size m. The n and m are two parameters to define the sliding windows, which are denoted as $W^m = \{w_1^m, w_2^m, w_3^m, ..., w_n^m\}$. The optimized m and n are 1200 and 20. Discussion of this result is presented in Section 4.4.1.

2. Kernel function

   For each window w_i^m, we use the kernel density estimator to compute the density individually, denoted as $\theta^m(W^m) = \{\theta_1^m, \theta_2^m, \theta_3^m, ..., \theta_n^m\}$. For new incoming data x_i, x_i is tested via a hypothesis test supported by the density θ_i^m. An incoming data point x is substituted into the density of a window θ_i^m. To compute the density value in a window, the density at x is defined in Equation (5.1) :

   $$\theta_i^m(w_i^m) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{4.1}$$

   where $K(.)$ is the kernel function and $x_i \in w_i^m$ and h is a smoothing parameter which controls the size of the neighbourhood around x_i). Detail of the Kernel function is provided in Section 4.3.3.

3. Voting

   For each window, we have a hypothesis that in the present data x is normal ($H_0 : x = 0$(indicating that our null hypothesis for data x is normal). If the density of x is less than a threshold value (p-value), this hypothesis (that the present data x is normal) can be rejected, or vice versa. Then, this window votes 1(abnormal data) or 0(normal data).

uniform	$K(u) = \frac{1}{2}$
triangle	$K(u) = (1 - \|u\|)$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$
Epanechincov	$K(u) = \frac{3}{4}(1 - u^2)$
Quartic	$K(u) = \frac{15}{16}(1 - u^2)^2$
Triweight	$K(u) = \frac{35}{32}(1 - u^2)^3$
Cosine	$K(u) = \frac{\pi}{4} \cos(\frac{\pi}{2}u)$

TABLE 4.1: Details of Kernels

4. Aggregating

   In the final stage, by aggregating all votes from all windows, this factor can be used as our primary reference to detect anomalies (details are provided in Section 4.3.4).

4.3.3 Kernel function

The kernel density estimator (KDE) [109, 110] is our core method for estimating the density function for each window, without having to select the parameter(s). The definition of KDE can be found in Equation (5.1). Several kernel functions [111] are available, namely uniform, triangle, Gaussian, Epanechnikov, quartic, triweight and Cosine(details of kernel functions are shown in Table 5.1). Although many kernel functions are available, their impact on the final result is slight [112]. However, the bandwidth selection has an influential impact on the result. The mean integrated squared error (MISE) [113] is the criterion used in the optimization process to find an optimized bandwidth, and Equation (5.3) defines the MISE which enables us to find the proper bandwidth [93, 104].

$$MISE(h) = E[\int (f'_h(x) - f(x))^2 dx] \tag{4.2}$$

where $E[.]$ is the expectation value $f_h^{'(x)}$ is the unknown density and $f(x)$ is the density estimation based on the given sample. If we use a Gaussian kernel in a practical dataset, the bandwidth h is defined in Equation (5.4).

$$h = 1.06\delta N^{-\frac{1}{5}} \tag{4.3}$$

where δ is the mean of a given sample and N is the number of training examples. In our experiment, we use Gaussian distribution as our kernel function.

Compared with the most frequently used density function, a histogram, KDE has two main advantages:

1. Smoothness

   The result of a histogram is not smooth, as it is represented by squared bars or lines.

2. Options of kernel functions

   A number of kernel functions are available for cases with different circumstances and distributions assumptions.

3. Dependence on the width of the bin

   The result of a histogram is profoundly affected by the width of the bin.

4.3.4 Anomaly factor

The AF is the primary measurement of the EK. It expresses the degree of abnormality of a data point within a certain period. When the voting process

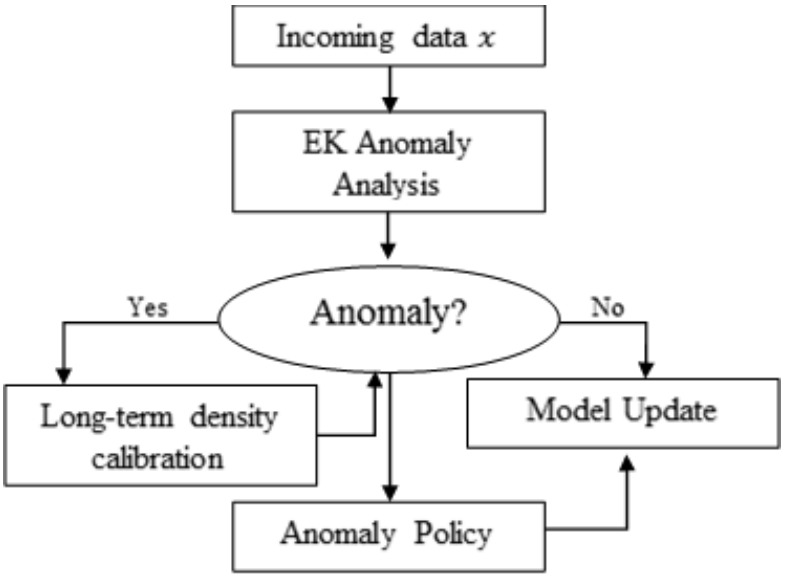


FIGURE 4.2: Process of Online EK Anomaly Analysis

has been completed, our method computes the AF with the vote vector $V = \{v_1, v_2, v_3, ..., v_n\}$, defined in equation (5.2):

$$AF = \frac{\sum_{i=1}^{N} v_i}{N} \tag{4.4}$$

The higher the AF, the more likely the data point is to be abnormal, and vice versa. The threshold of the AF depends on the problem domain. In cases which require high sensitivity, the threshold is adjusted to a low level. If $AF > K$, the new incoming data point is an anomaly. The threshold K is determined by the user, where $0 < K < 1$. In our experiment, K is 0.8 indicating 80% likelihood as an anomaly.

Algorithm 2: EK Algorithm

Data: $X = x_1, x_2, x_3, ..., x_n, windowwidth = m, \#windows = n, densities = \theta'$
Result: Anomaly factor $A = a_1, a_2, a_3, ..., a_n$
$\theta^m \leftarrow \theta'$;
`// assign short-term density vector`
$K \leftarrow k$;
`// constant threshold, threshold set based on problem domain`
while $x_i \neq \phi$ **do**
  $w^m_{n+1} \leftarrow x_i$; `//` **if** $w^m_{n+1} = n, update\theta$
  $a_i \leftarrow AFComputation(x_i, \theta, K)$
  `//` **if** $a_i \geq K$ **then**
    $AnomalyPolicy(x_i)$
  else
    if $size(w^m_{n+1} \neq m)$ **then**
      $w^m_{n+1} \leftarrow x_i$
      `//`
    else
      $UpdateModel(\theta, w^m_{n+1})$
      `//` $w^m_{n+1} \leftarrow \phi$
      `//`
    end
  end
end

Algorithm 3: AFComputation Algorithm

Data: x_i, θ, K
Result: a
$p_t \leftarrow 0.0001$;
`// p-value threshold for hypothesis test`
for $\theta^m_i \in \theta^m$ **do**
  $p \leftarrow \theta^m_i(x_i)$;
  `// use kernel density function in equation (5.1)`
  if $p \geq p_t$ **then**
    $v_i \leftarrow 0$;
  else
    $v_i \leftarrow 1$;
  end
end
$a = \frac{\sum_{i=1}^{n} v_i}{n}$;
if $a \geq K$ **then**
  $a = LongtermCalibration(V, x_i)$
else
end

Algorithm 4: Long-term calibration Algorithm

Data: V,x
Result: a
$C \leftarrow 0.7$;
// Calibration factor, impact of long-term calibration
$p_t \leftarrow 0.05$;
$p \leftarrow \theta_l(x)$;
// compute the p value of x in distribution Θ
if $p \geq p_t$ **then**
  for $v_i \in V$ **do**
    if $v_i == 1$ **then**
      $v_i \leftarrow v_i(1 - c)$
    else
      // do nothing
    end
  end
else
  // do nothing
end
Return $a \leftarrow \frac{\sum_{i=1}^{n}}{n}$;

Algorithm 5: UpdateModel Algorithm

Data: w_{n+1}^m, θ^m
Result: θ'^m
$\theta^m \leftarrow \theta^m(w_{n+1}^m)$;
for $\theta_i^m \in \theta^m$ **do**
  if $\theta_i^m \neq n$ **then**
    $\theta_i'^m \leftarrow \theta_{(i+1)}^m$
  else
    $\theta_i'^m \leftarrow \theta_n^m$
  end
end

4.3.5 Process of Online EK Anomaly Analysis

Fig. 4.2 shows the entire process of online EK anomaly detection for a data stream. For each incoming data x, we use the EK to test its abnormality. If the present x is an anomaly, long-term calibration is triggered to test the presenting data x again. If the result of calibration is the same as the hypothesis test of the previous

step, anomaly policies are triggered; otherwise the result is corrected by long-term density calibration; if the present data x is not an anomaly, a model update process is triggered to renew the density of the windows. Algorithms 2, 3, 4 and 5 show the algorithms of the EK method.

4.4 Experiments and results

To test the performance of the EK, we conducted three groups of experiments using UCI datasets and SHM data provided by a provincial Transport Research Institute from China (Because of confidentiality agreements and national policy and security issues, we do not to publish the name of the institute or the dataset). SHM uses different technologies to monitor the health status of a structure. Online anomaly detection in a data stream is one of the crucial components in SHM, as a public safety measure to prevent fatalities via reporting abnormal/hazardous situations. All datasets were collected from a highway steel cable suspension bridge. In the first group of experiments, we used the wind speed dataset to evaluate the performance of EK method in terms of accuracy, sensitivity and specificity. In the second group of experiments, we compared the EK with sliding-window K-means anomaly detection for data stream, LOF, and AODDS with UCI datasets and SHM wind speed data. In the third group of experiments, we introduced large-volume wind speed and road surface data to test the performance of the EK method compared with the K-means method. All datasets used were collected from the same bridge structure.

4.4.1 EK Evaluation experiment design and result

4.4.1.1 Experimental design and procedure

In this experiment, we evaluated our EK method in terms of accuracy, specificity and sensitivity, which are defined in Equation (2.1), (2.3) and (2.2).

The size of the dataset used in this experiment was 259,200, which was collected during December 2016 (1 Hz sampling rate). Structural experts for the Transport Research Institute had labelled some anomalies from datasets. Because the rate of anomalies was really low and our method could detect all these anomalies, we introduced extra anomalies into the data to test the performance of the datasets. We randomly injected outliers into the original dataset. The number of outliers was controlled by the injection rate (10% in this experiment). For each data point, we generated a random number between 0 and 1. If this random number was greater than (1- rate), this data point was added as a random number between 5 and 10 as an outlier. During the injection process, we also labelled the altered data points as anomalies. To prove the optimized result, we set different window widths, namely 5-minute window(300 data points if the frequency is 1Hz), 10-minute window, 20-minute window, 30-minute window and 60-minute window. We also set different numbers of windows, namely 5, 10 and 20. If we had increased the number of windows above 20, it would no longer be a short-term duration evaluation. In our experiment, we set the AF threshold at 0.8. The p-value for the short-term hypothesis test was 0.0001 (to reduce type-2 error under short-term density support) and 0.05 was used for long-term calibration.

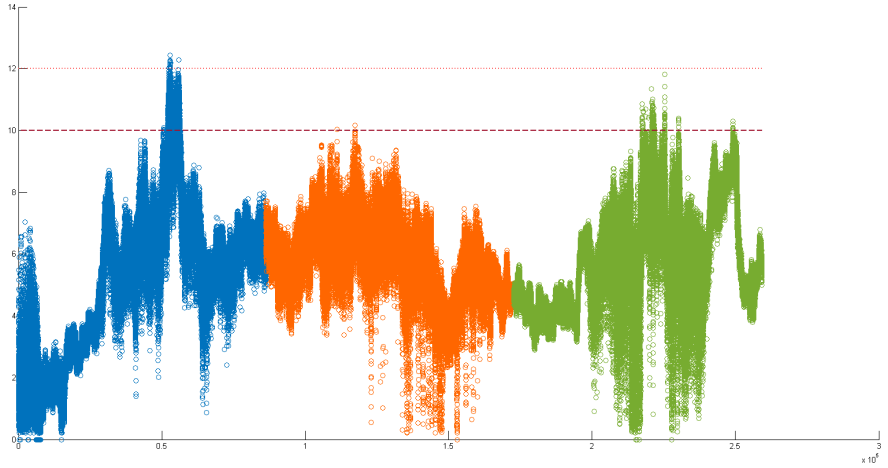


FIGURE 4.3: Wind speed datasets over the time (3-day)

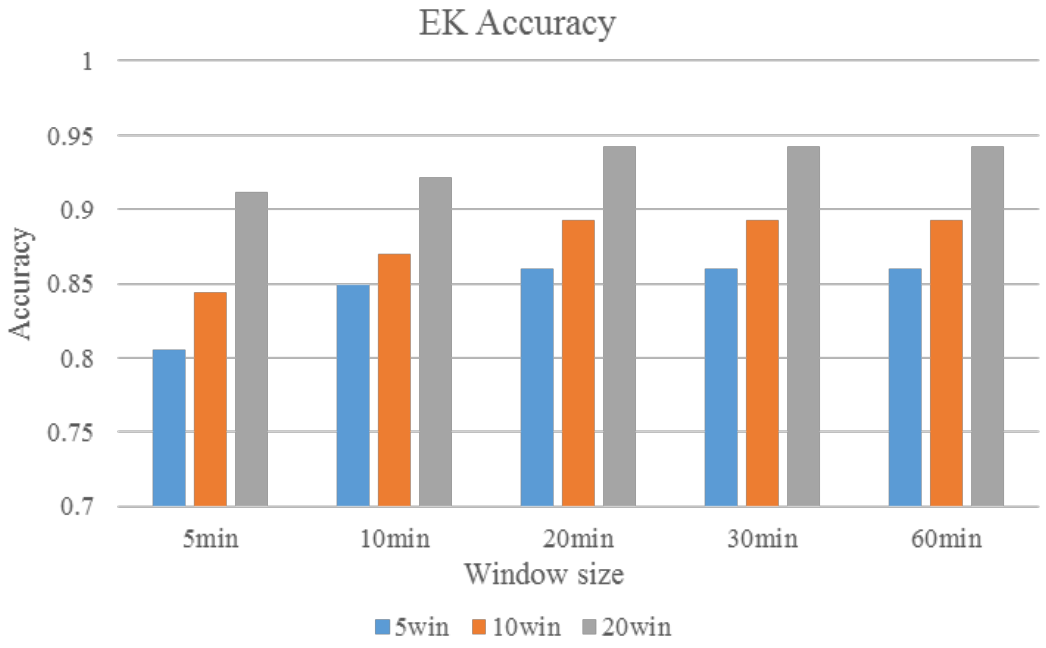


FIGURE 4.4: Accuracy

Fig.4.3 shows the scatter of our 72-hour(3-day data) wind speed dataset. We used three different colours to represent data points over three days. From observation, the distribution changed over time.

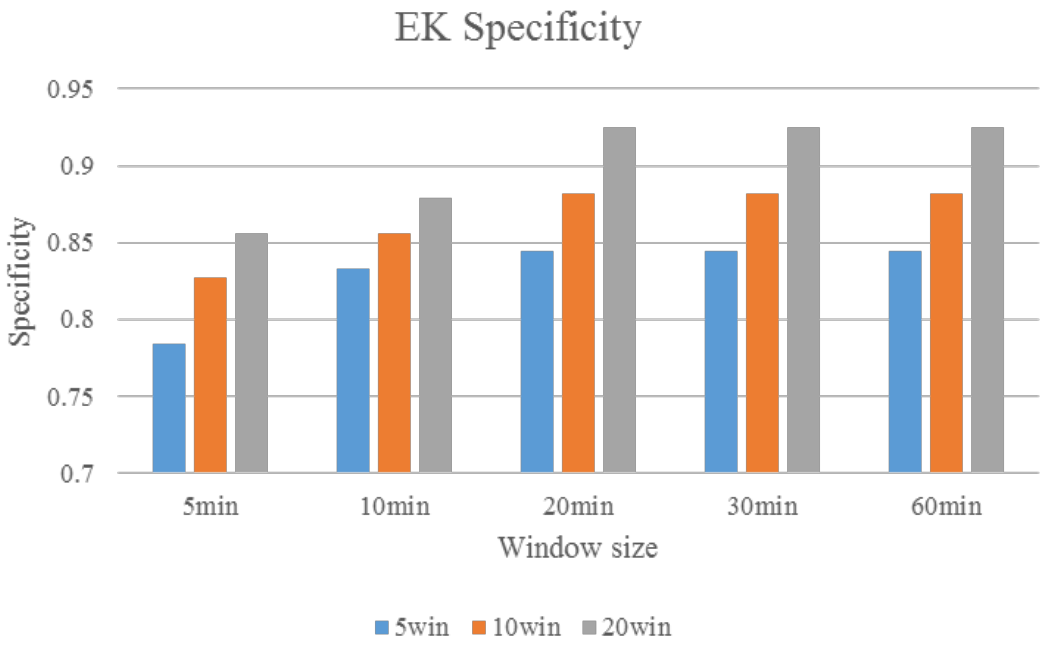


FIGURE 4.5: Specificity

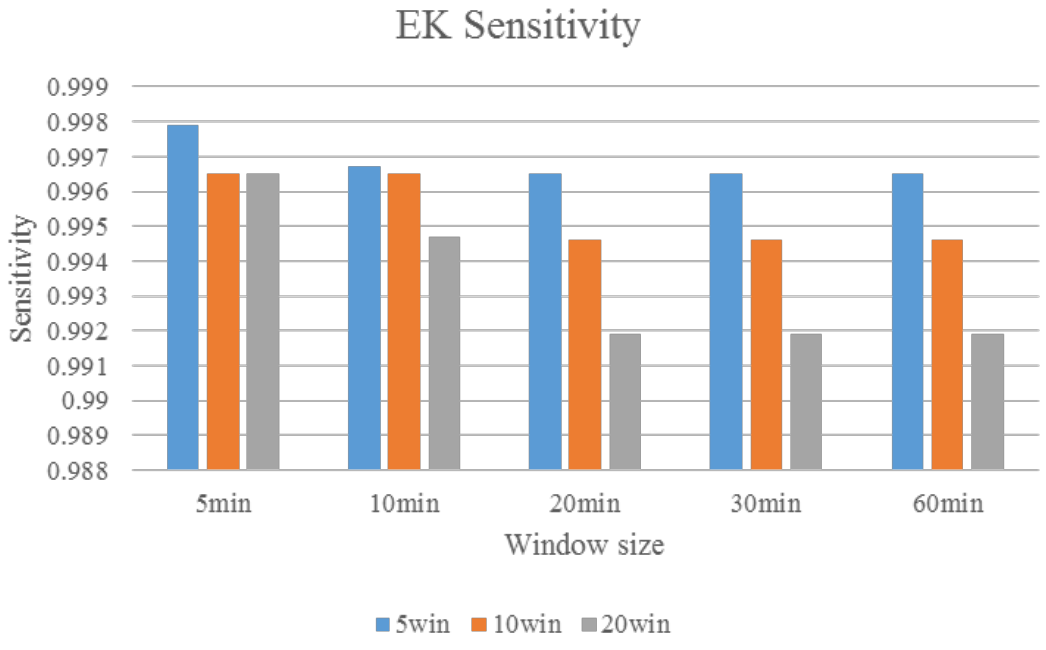


FIGURE 4.6: Sensitivity

4.4.1.2 Experimental result and discussion

Fig. 4.4, 4.5 and 4.6 show the accuracy, specificity and sensitivity of the EK method respectively. From Fig. 4.4 and 4.5 we found that with the increase in window width and the number of windows, accuracy and specificity increased significantly. In 4.6, we observed that with the increase in window width and number of windows, sensitivity decreases slightly. However, if we observed the value of sensitivity, the differences between groups are slight, because all results are above 0.99. The performance of the EK method with regard to accuracy, specificity and sensitivity increase to a certain point, and then remains at the same level. In the present case, the performance remains steady when the window width is a 20-minute window (1200 data points) with 20 windows. Consequently, we adopted a 20-minute window width1200 data points) and 20 windows as our optimized parameter for subsequent experiments.

To conclude, our EK method produced a superior result for SHM in terms of accuracy, specificity and sensitivity. Sensitivity is the most important measurement to consider in SHM, as any type-2 error (false negative) can have a fatally detrimental impact on both structure and public safety.

4.4.2 Comparison experiment and result

In these experiments, we compared our EK method with the method of K-means for data streams [102], LOF [97], and AODDS [99] using UCI datasets and wind speed datasets. We evaluated these methods in terms of accuracy and detection

time, which was the time taken to test an incoming data point. K-means for data stream employs a K-means clustering method (a model-based method) to detect outliers in a data stream. The LOF, an ensemble distance-based method, measures the density of local (nearest) neighbours to indicate the degree of abnormality. The equations and relevant proofs can be found in [96]. The AODDS, a distance-based method, is another data stream detection method, which computes the GDF and local deviation factor LDF to determine an outlier. Any point which is greater than three times the standard deviation of GDF or LDF is recognized as an outlier. Details of the GDF and LDF can be found in [99]. The EK method, online K-means for data streams and AODDS adopt the sliding-window principle to perform online anomaly analysis. Consequently, we used different window widths for these methods. Although LOF has no sliding window principle, we still implemented a sliding window principle in this experiment for comparison. That addition did not have a negative impact on the result; in fact, it enhanced the performance of LOF which because it performed an anomaly analysis for a period of data in a window rather than for the entire large dataset.

By comparisons with a distance-based method(AODDS), an ensemble distance-based method (LOF) and a model-based method (sliding window K-means), we observed the performance of different method categories on high-frequency data streams.

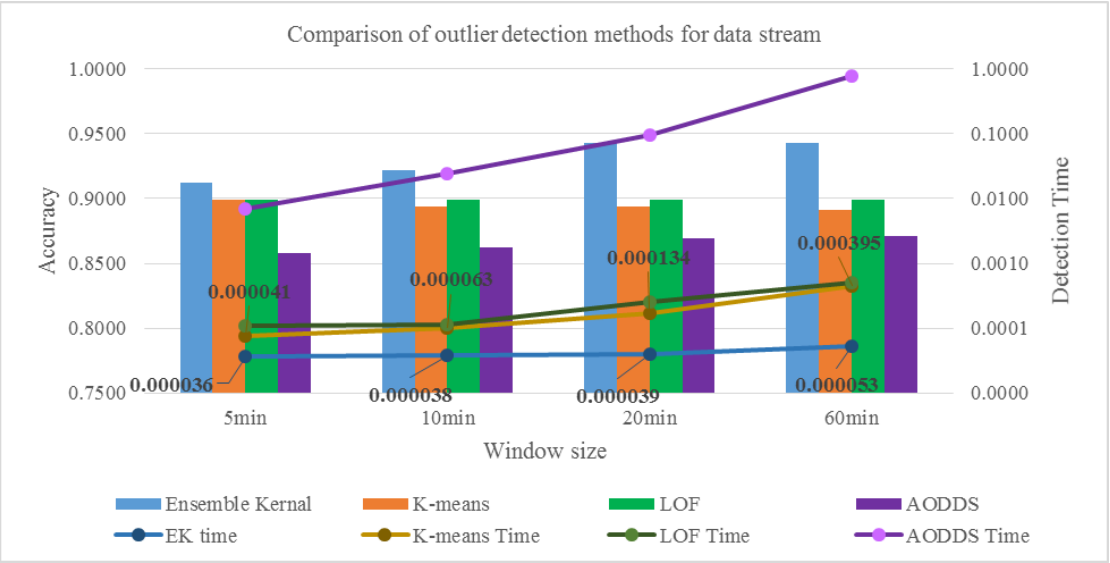


FIGURE 4.7: Wind data comparison

4.4.2.1 Experimental design and procedure

The wind speed datasets were the same as in Section 4.4.1. We chose Forest Cover, HTTP, SMTP, Mammography, Shuttle, and Mulcross from UCI. In these experiments, the AF threshold was set at 0.8; the p-value for the short-term hypothesis test was 0.0001 (to reduce type-2 error under short-term density support) and 0.05 was used for long-term calibration.

4.4.2.2 Experimental results

In the wind speed dataset (Fig. 4.7), our EK method produced the best performance of all the methods, reaching approximately 94% accuracy. From the efficiency perspective, our EK method had the fastest reaction time, which remained at almost at the same level except for 60-minute window width. The performance of K-means and LOF methods had similar performance at around 90% accuracy. For confidentiality reasons we do not publish our SHM dataset.

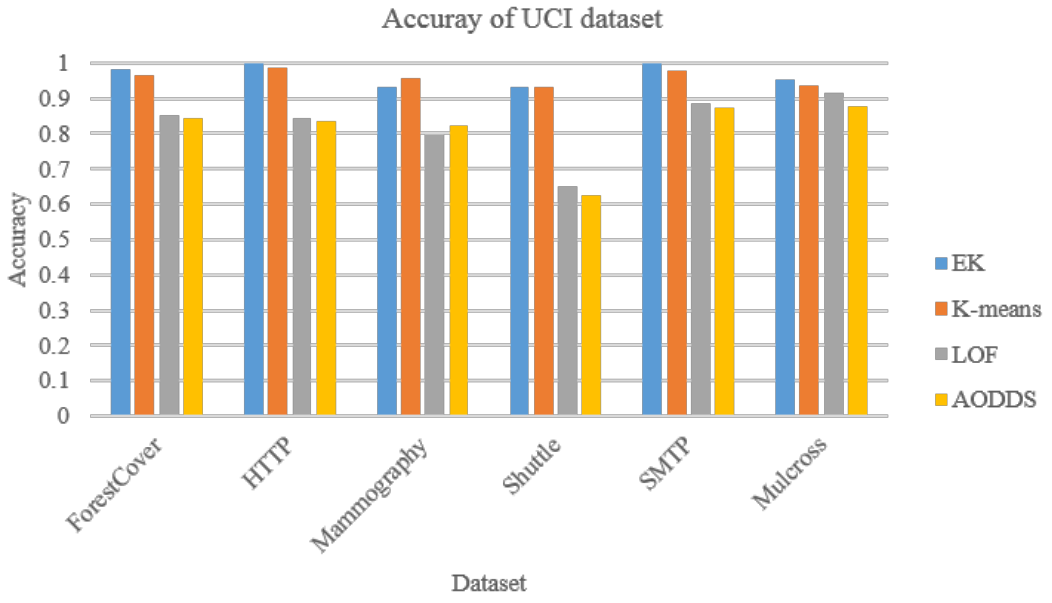


FIGURE 4.8: Accuracy of UCI datasets

We also compare our method using UCI datasets. Table 4.2 shows information about the UCI datasets.

Dataset Name	Size	Outlier Rate
HTTP	567479	0.4%
SMTP	95156	0.03%
ForestCover	286048	0.9%
Mulcross	262144	10%
Mammography	11183	2.32%
Shuttle	49097	7%

TABLE 4.2: Detail of UCI datasets

Fig. 4.8 4.9 and 4.10 show the comparative results using UCI datasets. The results show that the performance of our EK method is significantly superior to that of the other methods, with different outlier rates in terms of accuracy, sensitivity and specificity. Only the Mammography and Shuttle have a slightly lower accuracy compared with the K-means method with lower accuracy. The reason for this problem is that the Mammography and Shuttle datasets were smaller than those of the other datasets.

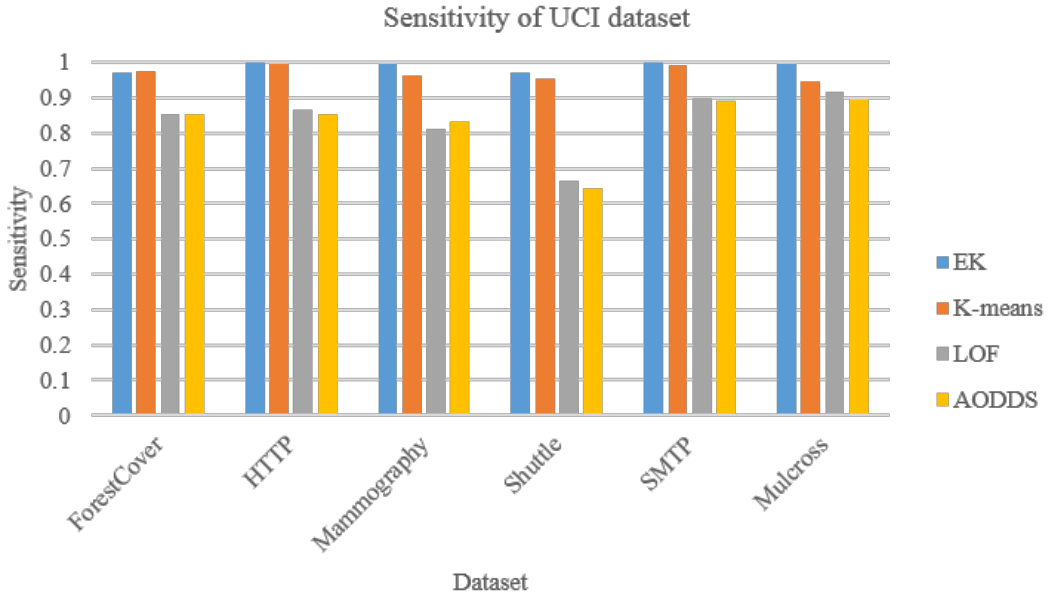


FIGURE 4.9: Sensitivity of UCI datasets

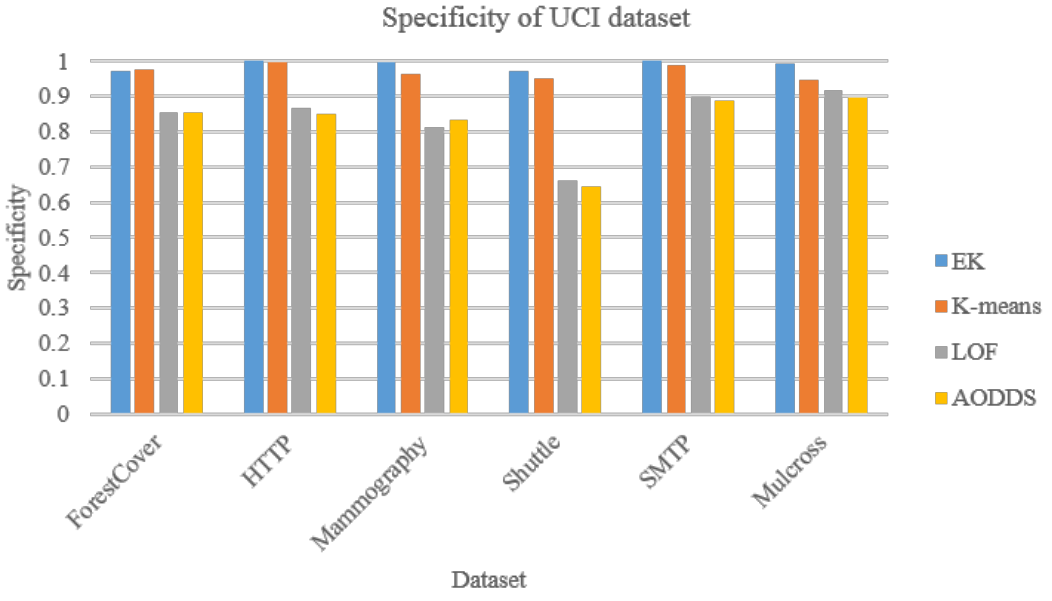


FIGURE 4.10: Specificity of UCI datasets

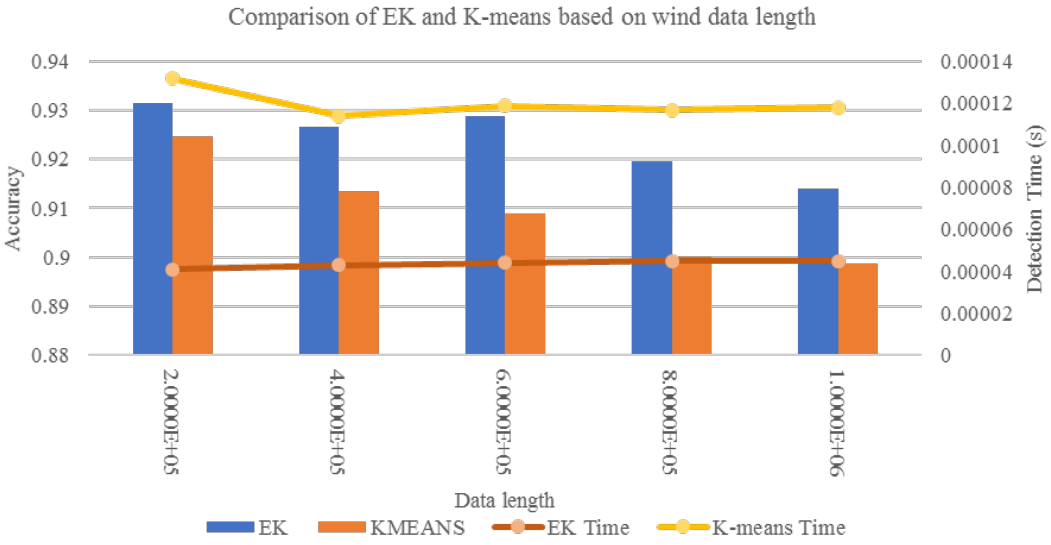


FIGURE 4.11: Performance Comparison using large volume of wind speed data

To summarize, our EK method demonstrated stable performance with respect to accuracy and detection time compared to the other methods in our experiment, using the wind speed dataset. In addition, in the UCI dataset experiment, our EK method showed significantly better performance than other methods. However, through the UCI experiments, we found that the performance of our EK method is affected by the size of the dataset.

4.4.3 Large volume data performance evaluation and comparison

Throughout the above two experiments, our EK method demonstrated stable performance. To test the performance of our EK method under large-volume datasets, we conducted experiments with massive wind speed datasets and temperature datasets in order to observe the performance of the EK method in terms of accuracy and detection time. In addition, we compared our EK method with the

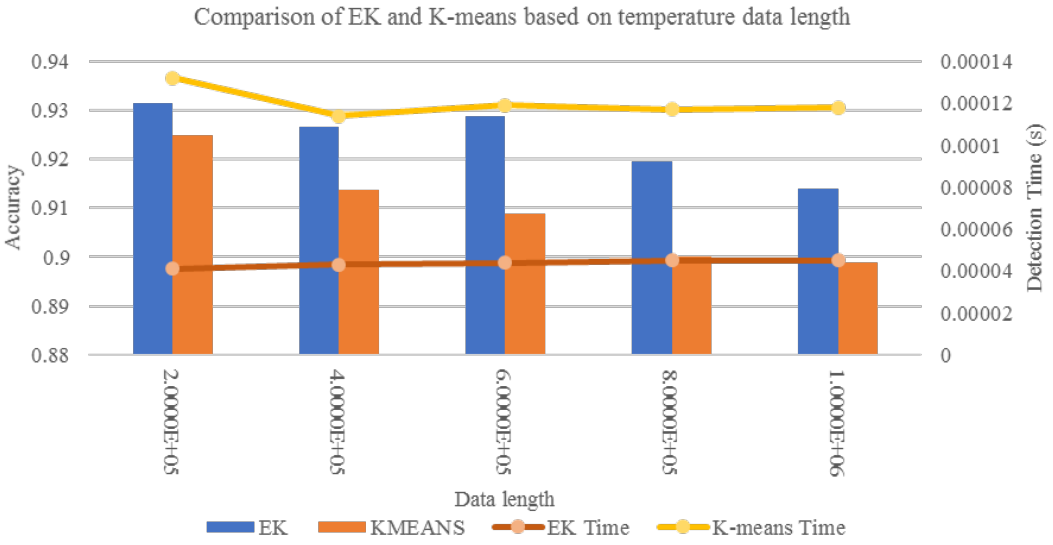


FIGURE 4.12: Performance Comparison using large volume of temperature data

K-means method described in Section **??**. Because the K-means method showed good performance in comparisons with remaining methods, we compared the EK method with the K-means method in the present experiments. Fig. 4.11 and 4.12 show the result based on the volume of data. Generally, the performance of both methods declines with the increasing volume of the dataset, but the results for our EK method are better than those for K-means. The average of our method remained approximately 93% accuracy, and the average detection time was around 0.00003 seconds, which was significantly less than that of the K-means method. Interestingly, the detection time of the K-means method decreased to around 0.00012 seconds and stabilized at around 0.00012 seconds after 40,000 data points for both wind speed and temperature datasets.

4.5 Limitation

Our EK method was capable of detecting anomalies with around 93% accuracy within a reasonable detection time. Moreover, under massive volumes of data, the EK method remained 90% accuracy and responded to anomalies within a rapid processing interval. However, it had some limitations. Our EK method was influenced by the size of the dataset. Specifically, our EK method showed poorer performance on small dataset.

In future work, we intend to focus on an adaptive AF threshold definition and on enhancing the performance of the EK method under low frequency and low volatility environments.

4.6 Summary

Dynamically changing environments present a challenge in the online detection of data stream anomalies. Our proposed EK method adopts a sliding window principle and ensemble analysis to capture anomalies in data streams. We tested our method using practical SHM data and the results showed the method to be capable of detecting anomalies in data streams efficiently, especially in high-frequency and high-volatility environments. Furthermore, the EK method retained above 90% accuracy and rapid detection time even when applied to a massive data volume. Thus, it would enable structural engineers to monitor structural health status

in real-time. When applied to public UCI datasets, the performance of our EK method was significantly superior to that of other methods.

The main material of this chapter has been submitted to the journal of pattern recognition, and it has been under review by pattern recognition.

Chapter 5

A new online anomaly detection method: multi-dimensional ensemble kernel

In Chapter 4 we proposed EK method for online anomaly detection. It can detect anomalies for a data stream in real-time efficiently. To answer the first research, in this chapter, we extended our EK method to multi-dimensional ensemble kernel method to detect anomalies in n dimensions. It also can answer part of the second research problem. Our proposed method, the MEK method is capable of detecting anomalies in real time for multi-dimensional data, using ensemble analysis and kernel density estimation. To evaluate the performance of our MEK method in terms of accuracy, specificity, and sensitivity, we used UCI standard datasets and practical structural health monitoring (SHM) data, in which accuracy and sensitivity, in particular, are significant measurements. In SHM, any type-2 error (false

negative) can have a fatally detrimental impact on both structural and public safety. The results show that the MEK method is capable of capturing anomalies with high sensitivity and accuracy compared to other online anomaly detection methods in both practical datasets and UCI datasets. Under a long-term and massive amount of data scenario, MEK can maintain good accuracy and sensitivity. Our MEK method is capable of detecting anomalies from heterogeneous data such as various SHM data. It also capable to handle data with different characteristics, such as periodical or contextual data streams. Moreover, our proposed method is computationally efficient with stable performance for a long-term period or a massive amount of data scenario.

5.1 Motivation

Anomaly detection aims to capture or detect any unexpected behaviours or data in a dataset. This technique has achieved a profound result with various applications in domains such as engineering[114], finance [115], and security[116]. Many researchers and scientists have been contributed to anomaly detection with remarkable results and achievements. With the increasing number of demands for real-time information, online anomaly detection has aroused interest in research communities. Online anomaly detection attempts to capture uncomfortable behaviours or data points in real time. Most traditional anomaly detection work is based on the assumption of a static data environment or the assumption of a static single/multi-dimensional data environment. Online anomaly detection is unlike traditional anomaly detection: data distribution and environment change

dynamically [117], especially for high-frequency and high-volatility data. Time variation is a further challenge in online anomaly detection, bringing additional uncertainties to anomaly detection. For instance, whereas a data point might be captured as an anomaly in some traditional offline methods, it may, in fact, be an anomaly in a specific time-segment captured by an online anomaly detection method. Consequently, we proposed a multi-dimensional ensemble kernel (MEK) method for detecting anomalies in various multidimensional online data. Our MEK uses ensemble analysis and a kernel density estimator to detect anomalies in different time periods. We apply and test our method using structural health monitoring (SHM) data.

With tremendous developments in civil construction such as short-span or long-span bridges, high or low buildings, railways, or mega-structural public buildings, managing and monitoring the health status of a structure is critically important, since any potential hazards could result in catastrophe for both public safety and economy[118, 119]. Sensor technology provides a solution to SHM, enabling civil engineers or related organization to monitor the health status of a building in real time. In a SHM project, they would deploy displacement sensor (via accelerometer, GPS, fibre optic sensor), strain-stress sensors(optic fibre sensors, accelerometer, and piezo-electric ceramic), and load sensors (i.e. traffic load sensor, wind load sensors). Although sensors help civil engineers to collect data for analysis, analyzing a massive amount of structural data collected from sensors is a challenge for civil engineers. In particular, reporting abnormal data in real-time is vital for reducing potential risks and avoid catastrophe. Online anomaly detection for SHM

helps civil engineers to develop a plan for the effective structural maintenance plan or dealing with emergencies.

Our proposed MEK can be applied in both single or multi-dimensional anomaly detection. In this chapter, we introduce our the MEK method and a series of experiments evaluating the accuracy, sensitivity and specificity. Our MEK method provides decision support for the bridge monitoring and management of anomaly detection, enabling engineers to troubleshoot efficiently and devise maintenance and emergency plans help relevant organizations to controls expenditure. Our contribution to anomaly detection includes:

- MEK has a greater performance compared with other methods [120–122] in terms of accuracy, sensitivity and specificity
- MEK is more scalable to be applied to data with different characteristics
- MEK method is more computationally efficient for heterogeneous data

5.2 Related work

Anomaly detection has been extensively studied in recent decades, with many remarkable results. Traditional anomaly detection methods have generally developed based on the assumption of a static distribution or static data environment and can be categorized into distance-based methods or model-based methods. Distance-based methods have two sub-categories, global outlier and local outlier.

A data point is detected as a global outlier if it deviates more than the threshold distance R from k points. [96] is the first proposed method that employed the global outlier to detect anomalies. [48], [49] and [50] used the summation of data distance from neighbours to determine the anomaly. In [48], it used only one dataset (size 10,000) showing a good performance. However, one dataset failed to prove that this method was good for other datasets. [49] and [50] were concerned with efficiency , both of them fail to use a standard dataset to prove the performance of their method.

A local outlier is similar to a global outlier. A data point is recognised as a local outlier if it deviates more than a threshold distance R from k the nearest neighbour data points. [97] introduced the first local outlier method, the local outlier factor(LOF). Based on the LOF, [98] developed an enhanced version for online anomaly detection. Distance-based methods are based on the distance between a data point and a threshold distance R to determine the abnormality. However, the threshold R is a constant threshold that not adaptive to change with a dynamic data environment. Under the high-frequency and dynamic data environment, LOF-based method failed to select nearest neighbours in an efficient manner. Under this scenario, continuously selecting nearest neighbours is computationally expensive. References [123], and [124] employed other distance-based methods, that use Euclidean distance as the key measurement to recognize anomaly. However, these two methods are for off-line anomaly detection. References [57] and [58] improved LOF efficiency by using a clustering technique. [57] used a clustering technique to find micro-clusters with lower and upper bounds. Based on

the result, it computed LOF. AODDS [99] is a distance-based method for online anomaly detection using both global-distance and local-distance. It computed a global deviation factor(GDF) and a local deviation factor (LDF) for each incoming data points. Any GDF or LDF greater than three times of the standard deviation of LDF or GDF is recognised as an anomaly. Although AODDS has good performance in online anomaly detection, the computation time for each detection is inefficient under a high-frequency data environment.

Model-based detection is another branch of anomaly detection. It detects anomalies via building model(s) from data samples. [100] is an example of an offline model-based method that employs a clustering technique to find outliers. [95] and [102] are two model-based online anomaly detection methods. Reference [95] attempted to use a clustering technique to train previous data in different period into clusters. Based on the previous clusters, new clusters were formed according to normal data for comparison with current cluster. Anomalies were recognised based on the similarity(distance) between the normal cluster and the current period cluster. [102] also applied a clustering technique and adopted a sliding-window principle to capture anomalies. These methods only considered single-dimensional online anomaly detection.

Multi-dimensional anomaly detection detects anomalies according to various dimensions. Most traditional multi-dimensional anomaly detection uses offline analysis. Reference [122] proposed an unsupervised method for anomaly detection, and it has been applied to data captured by wearable equipment. It trained previous data to extract features of normal and abnormal activity as primitive

discovery, then uses clustering to establish an activity structure, then used the activity structure to develop multi-dimensional time series data to detect abnormal activity. Although this method is unsupervised, its accuracy is not robust enough to be applied in SHM. From the reported results, it had 85% accuracy in detecting abnormalities. In [**?**], a similar method uses statistical and smoothed trajectory(SST) extracted from historical data and non-linear prediction to evaluate the abnormality of a data point. In [120], it used the subspace and linear regression method to detect anomalies in multi-dimensional time series data. In our experiments we made comparisons with references[120–122].

5.3 Proposed method : Multi-dimensional Ensemble Kernel

5.3.1 Overall

Our MEK adopts a sliding-window principle and ensemble analysis for online multi-dimensional time series anomaly detection. For a given set of time-series data, we use a kernel function to estimate the density of a period where the period is defined by the size of a sliding window. In each window, an anomaly hypothesis test is performed based on support from an individual window. By computing the anomaly factor (AF) from different windows in one dimension, any data with a high AF is considered an anomaly in a dimension. The multi-dimensional AF can

be computed via averaging AFs from different dimensions. Any anomaly would have an impact on the overall anomaly factor (OAF).

5.3.2 Problem Formulation

For multi-dimensional data $D = \{D_1, D_2, D_3, ..., D_n\}$, where n is the number of dimensions, in dimension D_i, all dimensions share a synchronized time-stamp denoted as $T = \{t_1, t_2, t_3, ...t_n\}$ and their data are denoted as $D_i = \{d_{t1}, d_{t2}, d_{t3}, ..., d_{tn}\}$ where t_i is the time-stamp. In a dimension D_i, a number of windows are set for further computation, denoted as $W_{d_i}^{(t_i,t_{i+k})} = \{w_1, w_2, w_3, ..., w_n\}$ where $n \in \mathbb{R}$ is the number of windows and (t_i, t_{i+k}) is the time period which from t_i to t_{i+k}. The density of $W_{d_i}^{(t_i,t_{i+k})}$ is denoted as $\Theta(W_{d_i}^{(t_i,t_{i+k})}) = \{\theta_{w_1}^1, \theta_{w_2}^2, \theta_{w_3}^3, ..., \theta_{w_n}^n\}$. In each density, we have a hypothesis test for new incoming data point in each dimension. The results of all densities are denoted as $V_\Theta(W_{d_i}^{(t_i,t_{i+k})}) = \{v_{\theta_1}, v_{\theta_2}, v_{\theta_3}, ..., v_{\theta_n}\}$, where $v_{\theta_1} \in (1||0)$. The anomaly factor (AF) of a new incoming data in D_i is denoted as $a_{d_i}(x_{di})$, where a_{d_i} is $0 < a_{d_i}(x) < 1, a_{di} \in a$. A higher a_{d_i} indicates that new incoming data in a dimension is an anomaly, and vice versa. The OAF indicates the overall abnormality from different dimensions, denoted as O_i. A lower of O_i indicates a higher degree of abnormality of new incoming data, where $O_i \in [0, 1]$. For the incoming data vector $x = \{x_{d1}, x_{d2}, x_{d3}, ..., x_{dn}\}$,our proposed method attempts to compute the OAF in an efficient manner to capture the anomaly in various dimensions.

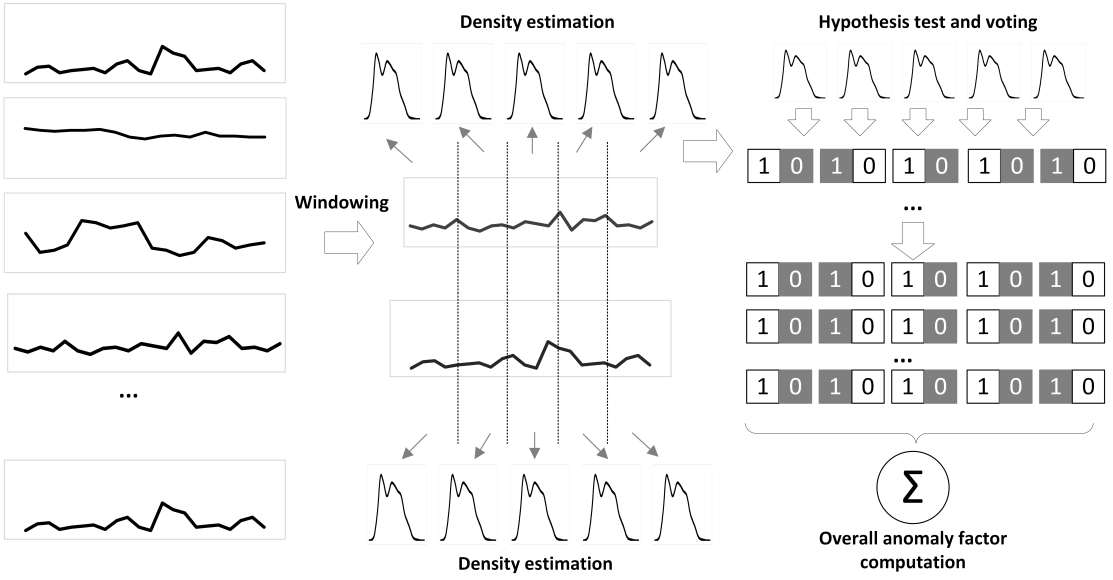


FIGURE 5.1: Procedures of Multi-dimensional Ensemble Kernel

5.3.3 Multi-dimensional Ensemble Kernel

Our proposed method MEK method adopts ensemble analysis based on the kernel estimator for online multi-dimensional high-frequency data. It can be categorized as an independent ensemble analysis (in terms of the component independence category) or a data-centred ensemble analysis (in terms of the constituent component category). Fig.5.1 shows the procedures of our MEK. This method has following steps:

5.3.3.1 Windowing

In this phase, we define the effective time-span of each window and the number of windows. Through our experiment, when the time-span of a window $t = 1200(secs)$,the number of windows (n = 20) can achieve the optimized result. Recent historical data are divided into n number of windows, and with the incoming data, all windows would be re-divided into n windows.

5.3.3.2 Densities estimation

For each window, we use a kernel density estimator to compute the density, which is defined in equation (5.1).

$$\theta^{i}_{w_i} = \frac{1}{n}\sum_{i=1}^{n} K(\frac{x_{di} - x_i}{h}) \tag{5.1}$$

where $K(.)$ is the kernel function and $x_i \in w_i^m$. Detail of the Kernel function is provided in Section 5.3.4.

5.3.3.3 Anomaly hypothesis test

For an incoming data point x_{di} in a dimension, a hypothesis test is computed based on support from the densities of each individual window. The H_0 is that the incoming data point x_{di} is normal data. If the p-value is less than the threshold ($p < \alpha$), then this hypothesis is rejected vice verse. In other words, this incoming data point x is recognised as an anomaly.

5.3.3.4 AF computation

Based on the result of each individual hypothesis, if a null hypothesis has been rejected, then the vote of a window is 1, vice versa. In the equation (5.2), the anomaly factor of an incoming data point in dimension d_i is computed.

$$AF = \frac{\sum_{i=1}^{n} V_{\Theta}(W_{d_i}^{(t_i, t_{i+k})})}{n} \tag{5.2}$$

5.3.3.5 OAF computation

After computation of the completed AFs in each dimension, the OAF is computed based on the result according to the Equation (5.3)

$$OAF = 1 - \frac{\sum_{i=1}^{n} AF_{d_i}}{n} \tag{5.3}$$

The higher the OAF score, the less likely there is an anomaly.

5.3.4 Kernel Density Estimation

The kernel density estimator (KDE) is the key function for estimating the density The KDE [109, 110] is our core method for estimating the density function for each window, without having to select parameter(s). The definition of KDE can be found in Equation (5.1). Several kernel functions [111] are available, namely uniform, triangle, Gaussian, Epanechnikov, Quartic, Triweight and Cosine(details of kernel functions shown in Table 5.1). Although many kernel functions are

uniform	$K(u) = \frac{1}{2}$
triangle	$K(u) = (1 - \lvert u \rvert)$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$
Epanechnikov	$K(u) = \frac{3}{4}(1 - u^2)$
Quartic	$K(u) = \frac{15}{16}(1 - u^2)^2$
Triweight	$K(u) = \frac{35}{32}(1 - u^2)^3$
Cosine	$K(u) = \frac{\pi}{4} \cos(\frac{\pi}{2}u)$

TABLE 5.1: Details of Kernels

available, their impact on the final result is slight [112]. However, the bandwidth selection has an influential impact on the result. The mean integrated squared

error (MISE) [113] is the criterion used in the optimization process to find an optimized bandwidth, and Equation (5.4) defines the MISE, which enables us to find the proper bandwidth[93, 104].

$$MISE(h) = argmin(E[\int (f'_h(x_{di}) - f(x_{di}))^2 dx]) \tag{5.4}$$

where $E[.]$ is the expectation value, $f_h^{'(x_{di})}$ is the unknown density and $f(x_{di})$ is the density estimation based on the given sample. If we assume that the density is close to the Gaussian distribution, the bandwidth h is defined in Equation (5.5).

$$h = 1.06\delta N^{-\frac{1}{5}} \tag{5.5}$$

where δ is the mean of a given sample and N is the number of training examples. In our experiment, we use the Gaussian kernel as our kernel function.

Compared with the most frequently used density function, which is the histogram, KDE has two main advantages:

1. Smoothness

   The result of a histogram is not smooth, as it is represented by squared bars or lines.

2. Options of kernel functions

   Many kernel functions are available for cases with different circumstances and distributions assumptions.

3. Dependence on the width of the bin

The result of a histogram is profoundly affected by the width of the bin.

5.3.5 Anomaly Factor and Overall Anomaly Factor

The AF is a primary measurement of the MEK method. It states the degree of abnormality of a data point within a certain period of a dimension. When the voting process has been completed, our method computes the AF with the vote vector, which is defined in equation (5.2). The higher the AF, the more likely the data point is to be abnormal in a dimension D_i, and vice versa. The threshold of the AF depends on the problem domain. In cases which require high sensitivity, the threshold is adjusted to a low level.

The OAF is an indicator which demonstrates the overall degree of abnormality from n dimensions. When all AF computations from dimensions have been completed, the OAF is computed according to equation (5.3). A higher OAF indicates that incoming data points x (x is a vector includes all dimensional data) is less likely to be abnormal, and vice versa. In SHM problems, a higher OAF suggests that health status of the overall structure is normal or safe at a given period.

5.3.6 Process of MEK Anomaly Analysis

Fig.5.2 shows the entire process for multi-dimensional data streams. For each incoming data points x in all dimensions, we compute its abnormality. If the presenting data points x are abnormal, long-term calibration is triggered to test the presenting data x again. If the result of calibration is the same as the hypothesis

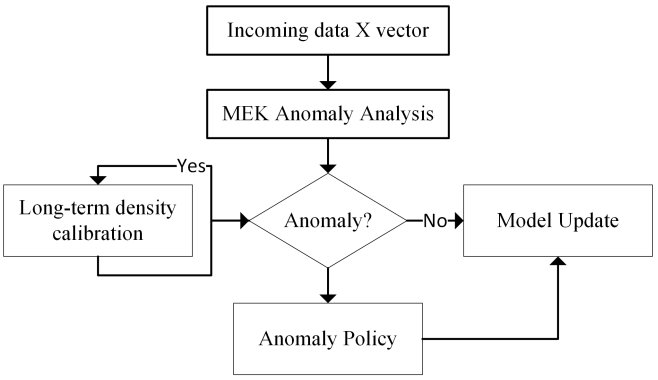


FIGURE 5.2: Entire process of MEK anomaly detection

test of the previous step, anomaly policies are triggered; otherwise, the result is corrected by long-term density calibration; if the present data points x are not an anomaly, a model update process is triggered to renew the density of the windows respectively. Algorithms 6, 7, 8 and 9 show the algorithms of the MEK method for single dimension evaluation. Algorithm 6 shows the entire main process of MEK method; Algorithm 7 is the process of testing abnormality of new incoming data points, any anomaly revokes the long-term calibration for reducing the false positive rate;Algorithm 8 shows the long-term calibration algorithm for OFA; Algorithm 9 shows the process of updating models, when a number of new incoming data points can form a new window in each dimension, the model updating is revoked to update all the models.

5.4 Experiment and result

To test the performance of the MEK, we conducted three groups of experiments using UCI dataset and SHM data provided by a provincial Transport Research Institute from China (Because of confidentiality agreements and national policy and

Algorithm 6: MEK Algorithm

Data: $X = x_{d1}, x_{d2}, x_{d3}, ..., x_{dn}, windowwidth = m, \#windows = n, densities = \theta'$

Result: Overall Anomaly factor $OFA = O_1, O_2, O_3, ..., O_n$

$\theta^m \leftarrow \theta'$;

// assign short-term density vector

$K \leftarrow k$;

// constant threshold, threshold set based on problem domain

while $x_i \neq \phi$ **do**

  $w^m_{n+1} \leftarrow x_i$; // if $w^m_{n+1} = n, update\theta$

  $a_{di} \leftarrow AFComputation(x_{di}, \theta, K)$

  // **if** $a_i \geq K$ **then**

    $AnomalyPolicy(x_i)$

  else

    if $size(w^m_{n+1} \neq m)$ **then**

      $w^m_{n+1} \leftarrow x_i$

      //

    else

      $UpdateModel(\theta, w^m_{n+1})$

      // $w^m_{n+1} \leftarrow \phi$

      //

    end

  end

end

if $OFA(a) \leq K$ **then**

  $LongtermCalibration(V,x)$

else

end

// Long-term calibration to review the abnormal candidates

security issues, we do not publish the full name of the institute and the dataset). SHM uses different types of technologies to monitor the health status of a structure. Online anomaly detection in a data stream is one of the crucial components in SHM, which as a public safety measure to prevent fatalities via reporting abnormal/hazardous situations. All datasets were collected from a highway steel cable suspended bridge. In the first group of experiments, we use UCI public dataset to compare with other methods. In the second group of experiments, we used

Algorithm 7: AFComputation Algorithm

Data: x_i, θ, K
Result: a
$p_t \leftarrow 0.0001$;
// p-value threshold for hypothesis test
for $\theta_i^m \in \theta^m$ **do**
  $p \leftarrow \theta_i^m(x_i)$;
  // use kernel density function in equation (5.1)
  if $p \geq p_t$ **then**
    $v_i \leftarrow 0$;
  else
    $v_i \leftarrow 1$;
  end
end
$a = \frac{\sum_{i=1}^{n} v_i}{n}$;
Return a

Algorithm 8: Long-term calibration Algorithm

Data: V,x
Result: OFA
$C \leftarrow 0.8$;
// Calibration factor, impact of long-term calibration
$p_t \leftarrow 0.05$;
$p \leftarrow \theta_l(x)$;
// compute the p value of in each dimension Θ
if $p \geq p_t$ **then**
  for $v_i \in V$ **do**
    if $v_i == 1$ **then**
      $v_i \leftarrow v_i(1 - c)$
    else
      // do nothing
    end
  end
else
  // do nothing
end
Return $OFA \leftarrow \frac{\sum_{i=1}^{n} a}{n}$;

Algorithm 9: Update Model Algorithm

Data: w_{n+1}^m, θ^m
Result: θ'^m
$\theta^m \leftarrow \theta^m(w_{n+1}^m)$;
for $\theta_i^m \in \theta^m$ **do**
  if $\theta_i^m \neq n$ **then**
    $\theta_i'^m \leftarrow \theta_{(i+1)}^m$
  else
    $\theta_i'^m \leftarrow \theta_n^m$
  end
end

SHM datasets including wind speed, road surface temperature, and GPS to evaluate the MEK method's performance compared with other methods. In these two experiments, we compared with the multi-dimensional data cube (MDC), multi-dimensional intrusion detection (MID) and Unsupervised clustering (UC) for with both road surface temperature data and wind speed data. In addition, we conducted the third experiment using wind speed data, road surface data, and GPS data with different sizes from 1 million to 10 million. In this experiment, we aim to test the performance of our MEK under the long-term scenario.

5.4.1 UCI dataset experiment

5.4.1.1 Experimental design and procedure

To test the performance of our method, we use public datasets to demonstrate its performance compared with other methods. We choose Forest Cover, HTTP, SMTP, Mammography, Shuttle, and Mulcross from UCI dataset collections [125]. Details of the datasets are shown in Table 5.2. These datasets had various outlier rates from 0.03% to 10% with different sizes. ForestCover and Shuttle have

more dimensions than HTTP, SMTP and Mulcross datasets, which are 10 and 9 respectively. The rest of the dimension information is presented in Table 5.2.

- HTTP: The 1999 KDD cup dataset has 4 attributes(service, duration, source bytes, destination bytes). The Original dataset has 41 attributes (34 continuous, 7 categorical), but they are reduced to 4 attributes which regarding HTTP data. HTTP, SMTP, FTP, FTP_data, and other subsets are integrated into the original dataset, and only HTTP service used in our experiment. The size of original dataset contains 3,925,651 records, and it has been condensed into 567,479 records for the 1999 cup dataset.

- SMTP: The 1999 KDD cup dataset, has 4 attributes(service, duration, source bytes, destination bytes). 1999 KDD cup dataset, has 4 attributes(service, duration, source bytes, destination bytes). The Original dataset has 41 attributes (34 continuous, 7 categorical), but they are reduced to 4 attributes which regarding HTTP data. HTTP, SMTP, FTP, FTP_data, and others subsets are integrated into the original dataset, and only SMTP service used in our experiment. The original dataset contains 3,925,651 records and has been condensed into 95,156 records for the 1999 cup dataset.

- ForestCover : used for predicting forest cover type from cartographic variables. All data are collected from a study of the Roosevelt National Forest of northern Colorado.This dataset has 54 attributes (10 quantitative variables, 4 binary wilderness areas and 40 binary soil type variables). For anomaly detection, only 10 quantitative attributes can be used. The outlier rate is 0.9%

- Mulcross : generated from a synthetic data generator with 4 dimensions.
- Shuttle: is a multi-class classification dataset with 9 dimensions.

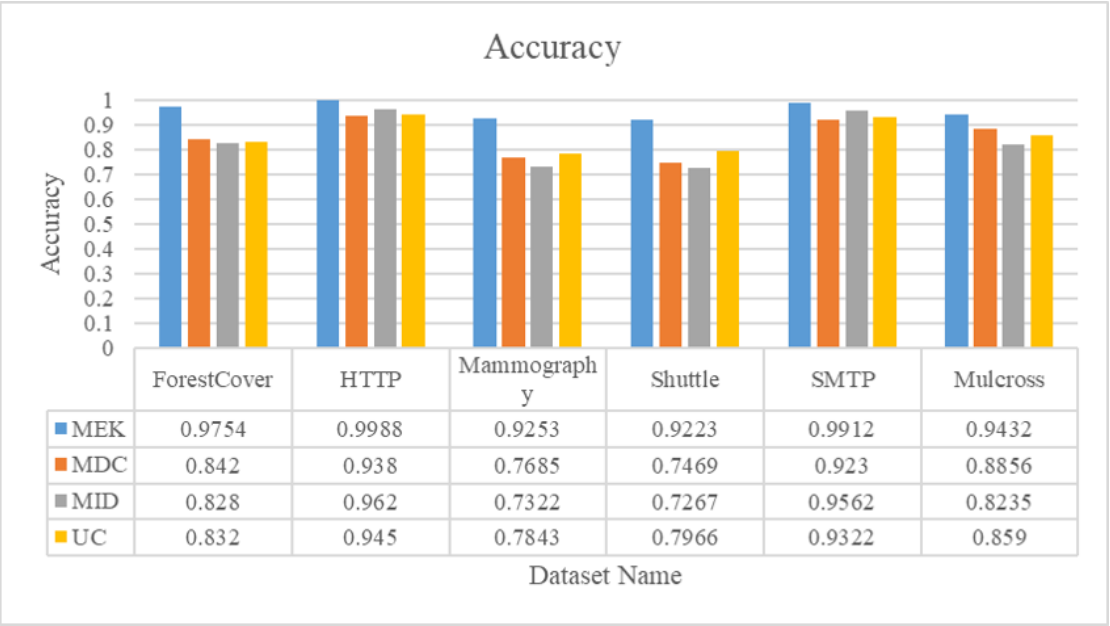
By using different datasets with various outlier rate, we can observe the performance of our MEK under different scenarios. In this experiment, the AF threshold was set to 0.8; the p-value for the short-term hypothesis test was 0.0001 (to reduce type-2 error under short-term density support), and 0.05 was used for long-term calibration.

Dataset Name	Size	Outlier Rate	Dimensions
HTTP	567479	0.4%	4
SMTP	95156	0.03%	4
ForestCover	286048	0.9%	10
Mulcross	262144	10%	4
Shuttle	49097	7%	9

TABLE 5.2: UCI dataset description

5.4.1.2 Result

Fig.5.3 5.4 5.5 are the results of MEK method comparing with other methods. Our MEK method is significantly better than other methods in terms of accuracy, sensitivity and specificity. For those low outlier rate (i.e. HTTP and SMTP), our method can detect anomalies easily which almost achieved 100% due to its low outlier rate. Mammography and Shuttle have worse performance in terms of accuracy, sensitivity and specificity compared with other datasets. These two datasets are small-size datasets with higher outlier rate. Base on this observation, we found that the size of a dataset has an impact on the performance. In fact, our MEK method is affected by the size of a dataset since its need enough short-term



	ForestCover	HTTP	Mammography	Shuttle	SMTP	Mulcross
MEK	0.9754	0.9988	0.9253	0.9223	0.9912	0.9432
MDC	0.842	0.938	0.7685	0.7469	0.923	0.8856
MID	0.828	0.962	0.7322	0.7267	0.9562	0.8235
UC	0.832	0.945	0.7843	0.7966	0.9322	0.859

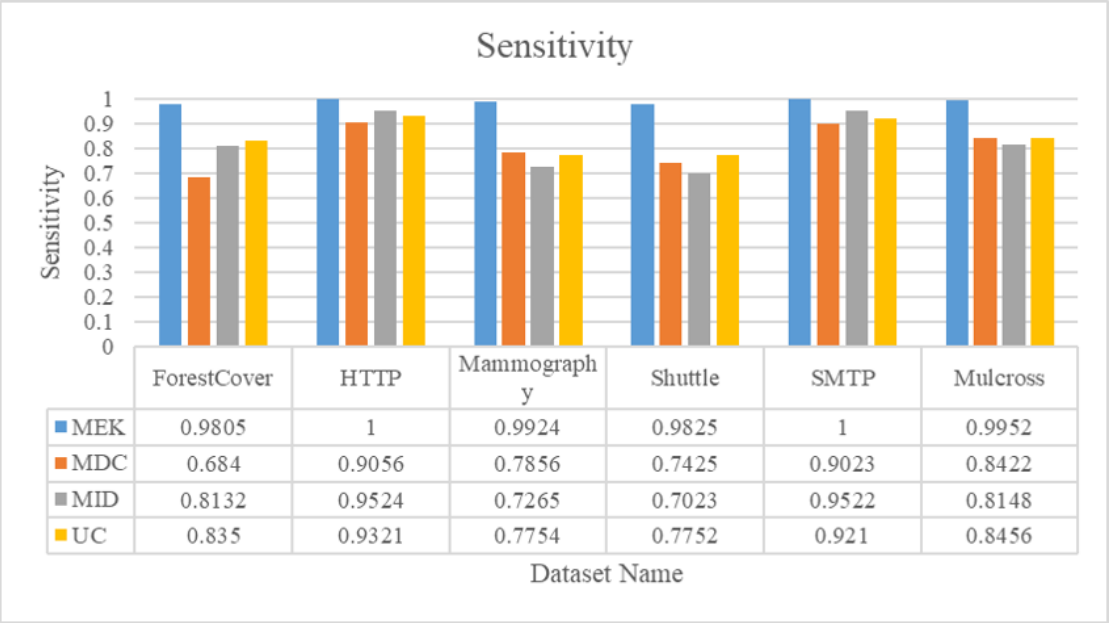
FIGURE 5.3: Accuracy

historical data to support to perform a hypothesis test. MID method is good at detecting HTTP and SMTP anomalies since it specializes for network anomaly detection. Other methods remain around 85% to 79% accuracy to all datasets. However, all of these methods have a good performance on HTTP and SMTP dataset, since its low outlier rate. Through current result, we fail to conclude that the dimension has an impact on performance since ForestCover has a better performance comparing with Shuttle, where the dimensions are 10 and 9 respectively.

5.4.2 Practical SHM dataset experiment

5.4.2.1 Experimental design and procedure

In this experiment, we evaluated our MEK method in terms of accuracy, specificity and sensitivity using surface temperature data, wind speed data, and GPS data



	ForestCover	HTTP	Mammography	Shuttle	SMTP	Mulcross
MEK	0.9805	1	0.9924	0.9825	1	0.9952
MDC	0.684	0.9056	0.7856	0.7425	0.9023	0.8422
MID	0.8132	0.9524	0.7265	0.7023	0.9522	0.8148
UC	0.835	0.9321	0.7754	0.7752	0.921	0.8456

FIGURE 5.4: Specificity

	ForestCover	HTTP	Mammography	Shuttle	SMTP	Mulcross
MEK	0.9916	0.9988	0.9864	0.9511	0.9966	0.9904
MDC	0.8542	0.9456	0.8423	0.8123	0.9386	0.8956
MID	0.8266	0.9656	0.7422	0.7311	0.965	0.8245
UC	0.821	0.9487	0.7885	0.8033	0.936	0.8621

FIGURE 5.5: Sensitivity

in terms of x, y and z. The sizes of all datasets were different because sensors had different sampling rates. We used the lowest sampling rate to synchronize all sensors data which was 1Hz; in total the size of each dimension was 259,200. All outliers were labelled by the Transport Institute, but there were few outliers with the outlier rate around 0.002%. Consequently, to test the performance of our method, we introduced anomalies into the data to increase the outlier rate. We randomly injected outliers into the original dataset. The number of outliers was controlled by the injection rate (10% in this experiment).For each data point, we generated a random number between 0 and 1. If this random number was greater than the (1-rate), this data point was added as a random number between 5 and 10 as an outlier. During the injection process, we also labelled the altered data points as anomalies. The AF threshold in our experiment was set at 0.8. The p-value for the short-term hypothesis test was 0.0001 (to reduce type-2 error under short-term density support) and 0.05 was used for long-term calibration.

Before we present our result, we first look into our datasets. Fig.5.6 is the scatter of the 3-day wind speed dataset which contains 259,200 data points, and each colour represents data of 1-day. The sampling rate of the wind sensor is 1Hz. Over the 3-day data, the distribution of the wind data changes over the time if we set the duration unit to one day. Especially in the third data, the value varies from the lowest to the highest speed.

Fig.5.7 shows the 3-day data of road surface temperature containing 5,184,000 data instances collected by temperature sensors with 25 Hz. Overall, we observe

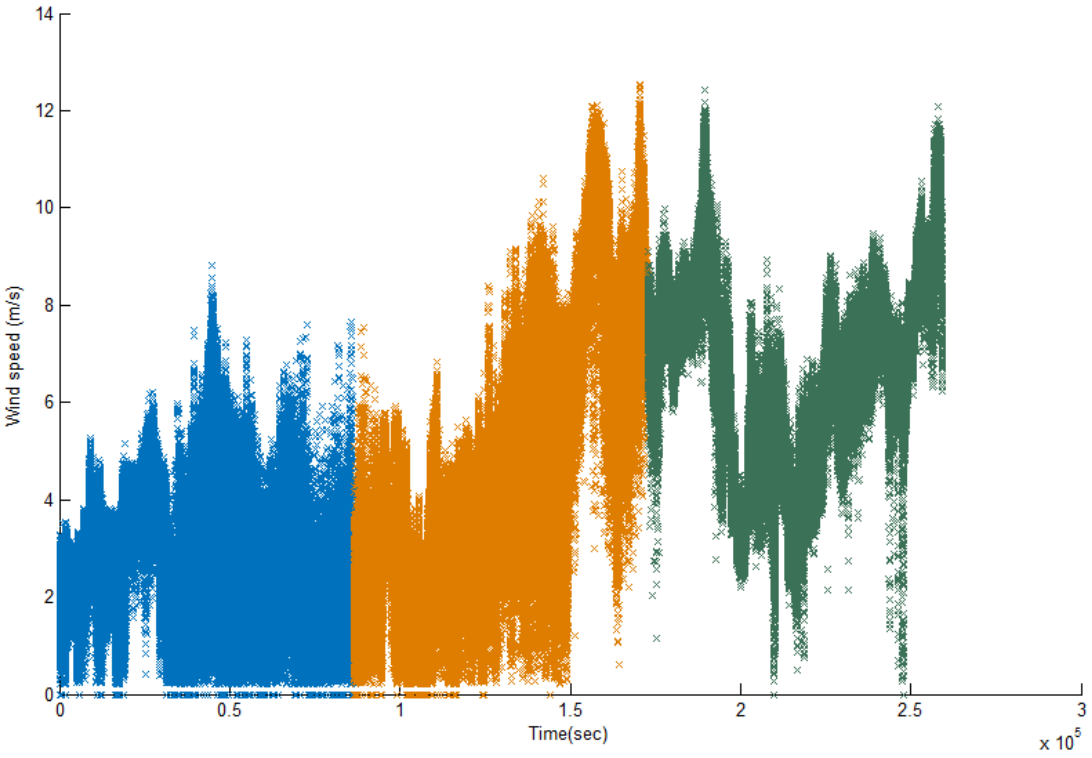


FIGURE 5.6: 3-day Wind data: blue dots represent the first day data; orange dots represent the second day data; green dots represent the third day data

that the change of temperature is periodical within a range, and only a few outliers can be observed.

Fig.5.6,5.7,5.8,5.9 and 5.10 present the 3-day GPS data in terms of X,Y, and Z dimensions. The GPS data measures the displacement of the bridge in terms of horizontal displacement(X), perpendicular displacement (Y) and vertical displacement(Z). The GPS sensor rate is 10 Hz and in total there are 2,592,000 data instances. From the Fig.5.8, we find that the displacement vibration occurred in horizontal (X) and vertical displacement (Z) more frequently than in perpendicular displacement(Y),and the GPS Y value fluctuates within a certain range. Fig.5.8 and Fig.5.10 demonstrate a few number of outliers.

Overall, the number of outliers is low, which the outlier rate is less than 0.0005%.

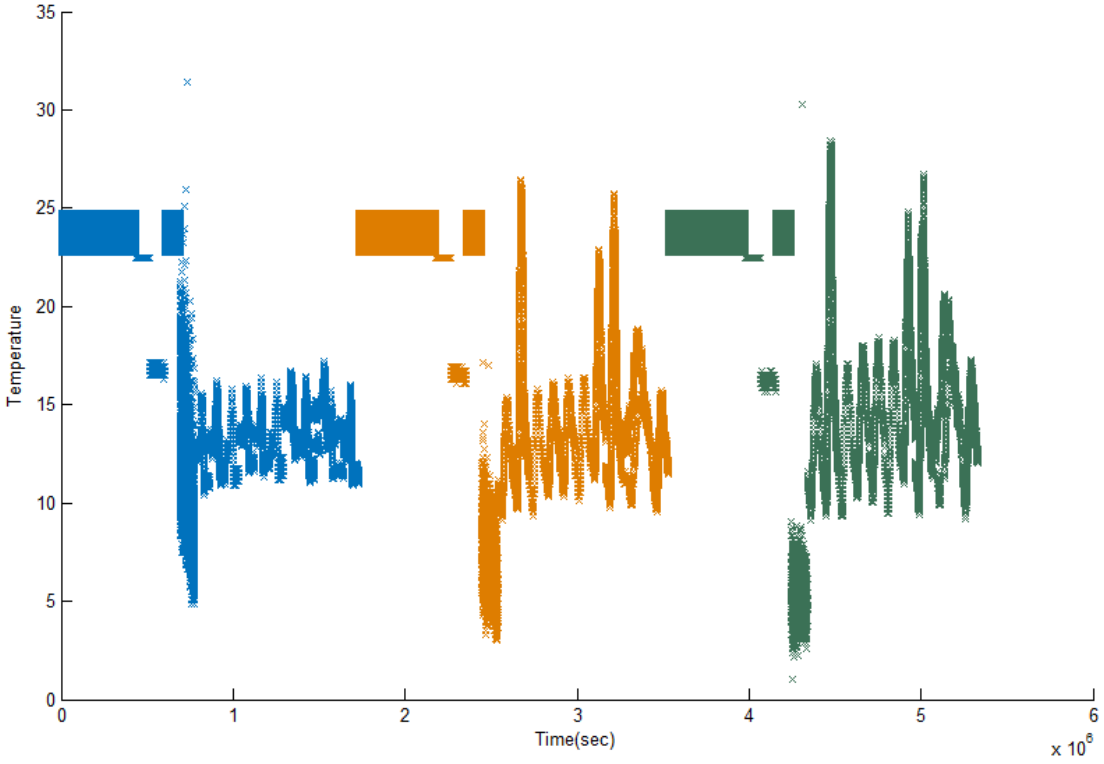


FIGURE 5.7: 3-day road surface temperature data : blue dots represent the first day data; orange dots represent the second day data; the green dots represent the third day data

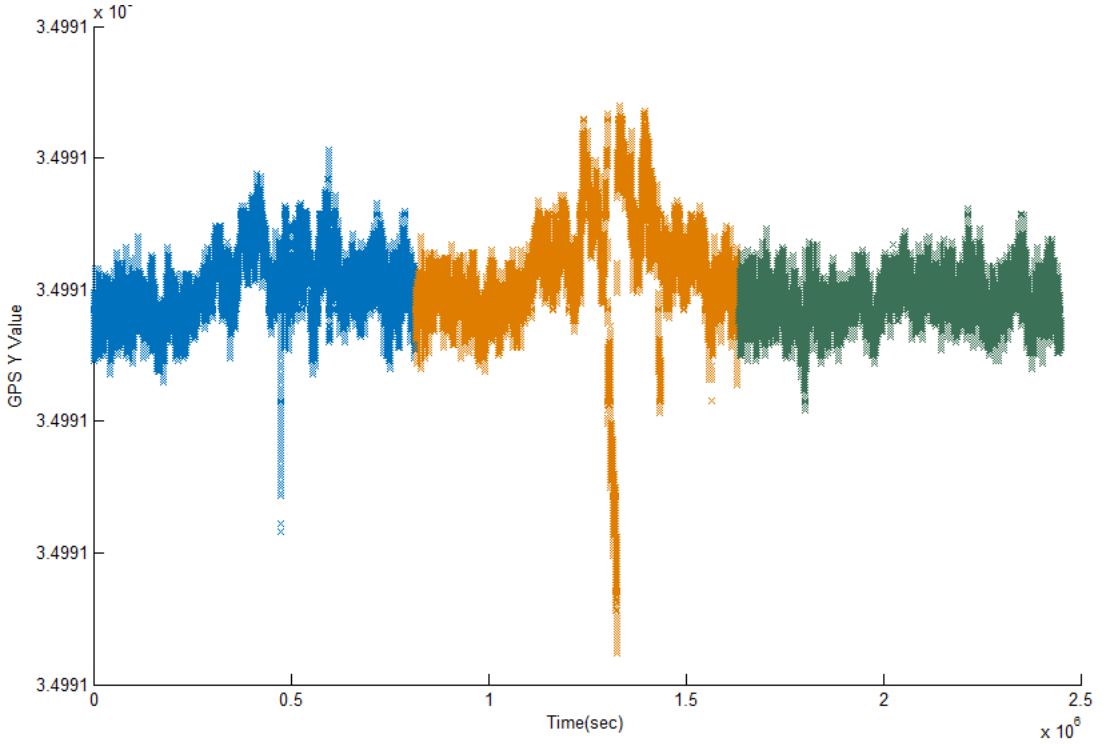


FIGURE 5.8: GPS X value : blue dots represent the first day data; orange dots represent the second day data; green dots represent the third day data

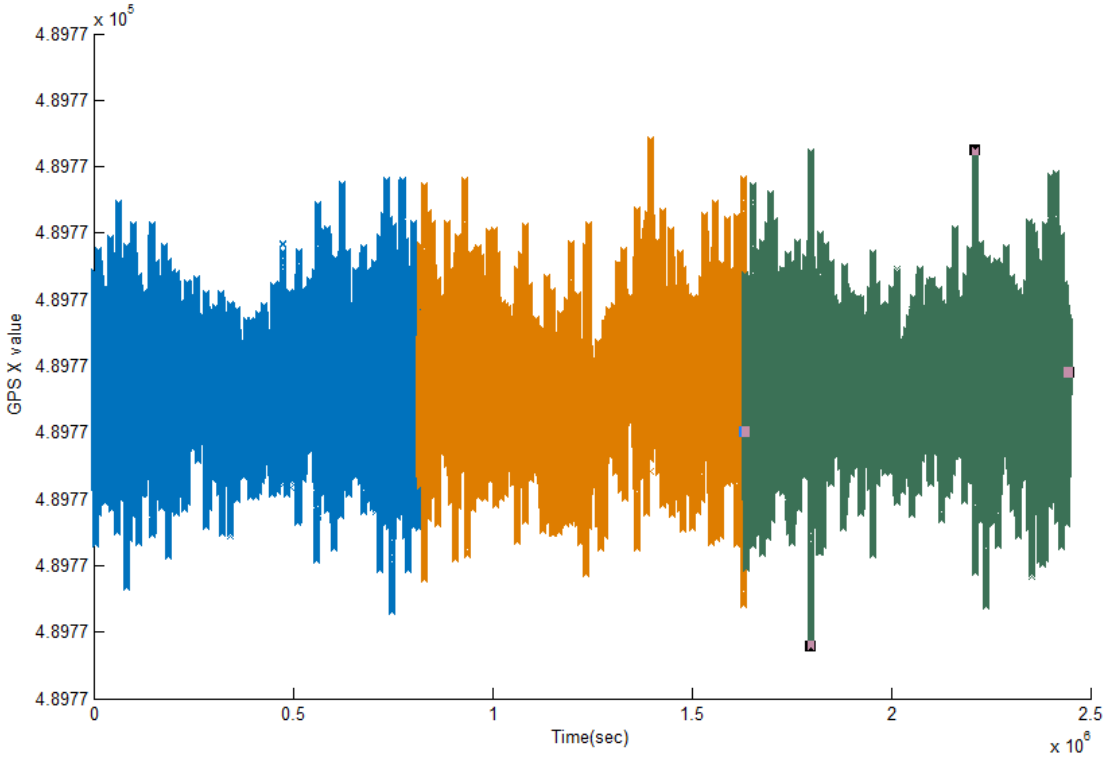


FIGURE 5.9: GPS Y value : blue dots represent the first day data; orange dots represent the second day data; green dots represent the third day data

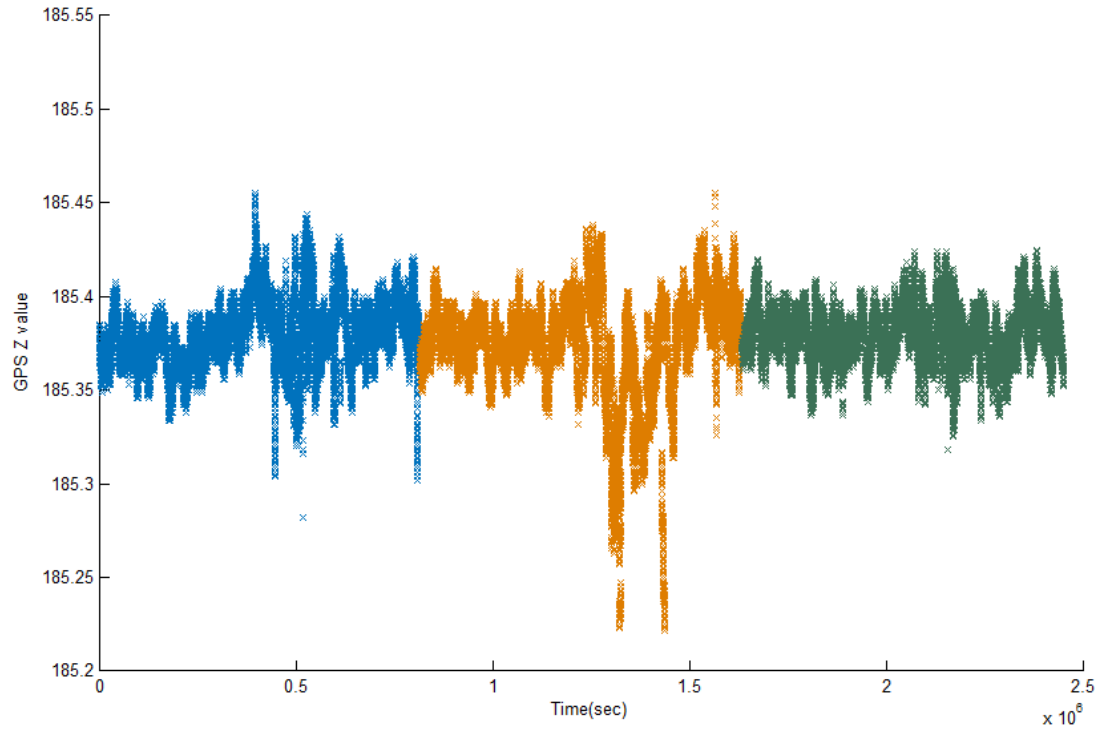


FIGURE 5.10: GPS Z value : blue dots represent the first day data; orange dots represent the second day data; green dots represent the third day data

In the previous experiment, we can see from our method that if the outlier rate is small, it is easy to detect the anomalies. We add extra outliers into the raw data to test the performance of our MEK method. From these scatters, we can observe that different data has different features:

- Wind speed data : distribution is changes over the time
- Temperature data : temperature is changes periodically
- GPS data : GPS Y value is stable; anomalies occur more frequently in GPS X and Z dimensions

5.4.2.2 Result

In this experiment, we compared our MEK method with the methods of multi-dimensional data cube (MDC) [120], multi-dimensional intrusion (MID) detection [121], and unsupervised clustering(UC) [122]. We evaluated these methods in terms of accuracy, sensitivity and specificity.

Fig.5.11 shows the result of the comparison. Our method shows good performance compared with the other methods, especially for sensitivity and specificity. Although MDC, MID and UC are time-series based multi-dimensional anomaly detection methods, they fail to show good performance under in the high-frequency data environment. In particular, the sensitivities of MDC and MID are low which is dangerous in SHM problem. Any false negative is a potential risk which could lead to a public safety issue, and as a result, high sensitivity is the first priority in SHM problems. The expectations of MDC, MID and UC are low. For instance,

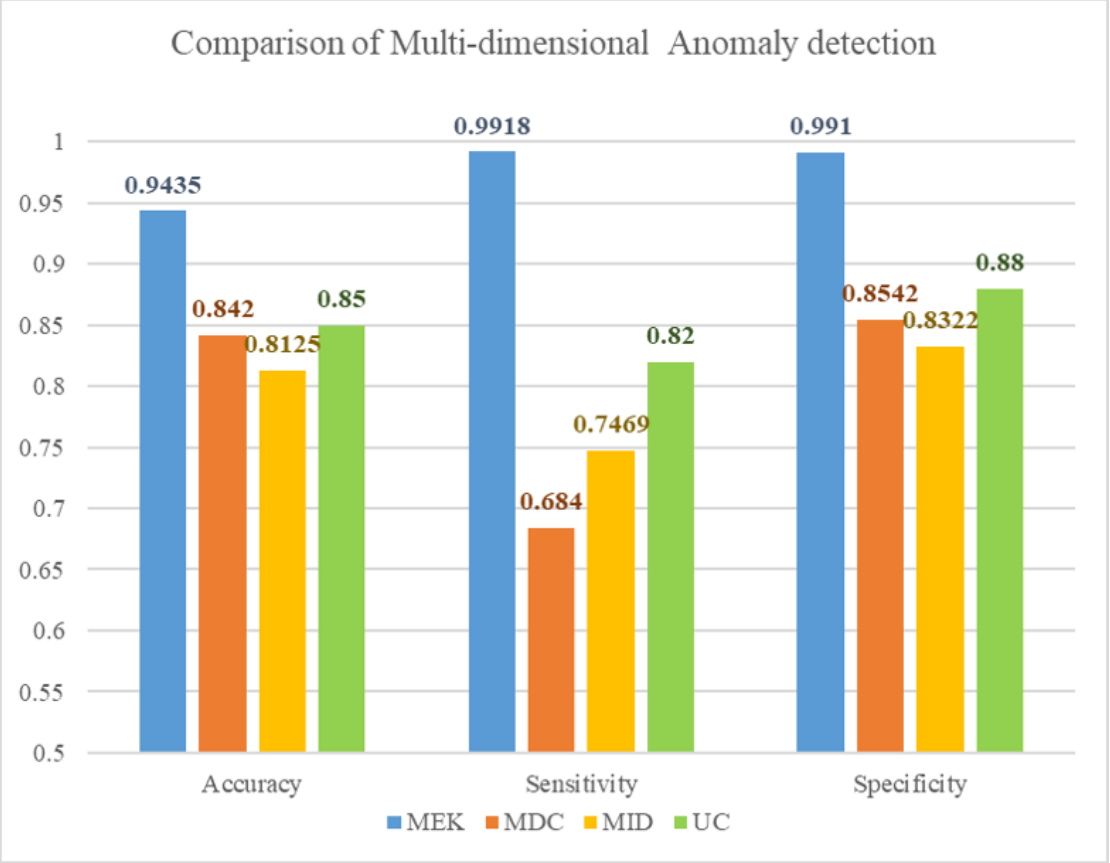


FIGURE 5.11: Comparison Results using SHM datasets

if there are 100,000 data points and around 15% of the data are recognised incorrectly, any false negative could have a catastrophic impact on both public safety and local economy. To guarantee the public safety, it is important to maintain a high level of sensitivity to avoid the false negative. Our MEK method can maintain high-level sensitivity and accuracy for detecting anomalies in multi-dimensions. In conclusion, our method has a superior performance in terms of accuracy, sensitivity and specificity in SHM datasets. Moreover, it can handle multi-dimensional data with different features.

5.4.3 Long-term performance of MEK

5.4.3.1 Experiment design and procedure

To prove that our MEK method can work properly in a long-term scenario, we conducted an experiment using very large datasets to test its performance. In this experiment, we used wind speed data, road surface temperature and GPS data, which contained a various number of data instances from 1,000,000 to 10,000,000 data instances, with all datasets are synchronized to one data instance per second(1 Hz sampling rate). Because the outlier rate was low at only around 0.0003%, we introduced extra outliers into the raw datasets so that the outlier rate reached to around 9% to 10%. The procedure of injecting outliers into raw datasets is same as in the previous experiment which described in section 5.4.2.1. Because the performance of our MEK method was superior to that of with other methods in previous experiments, we only demonstrate the accuracy of MEK along with the size of the dataset.

5.4.3.2 Result

Fig.5.12 shows the result of the long-term performance of the MEK method. With the increasing size of datasets, the performance of MEK method declines slightly from around 94 % accuracy to around 92% accuracy. For the smaller data sizes (1 million to 3 million), the accuracy is maintained at around 94%. From 3 million data instances the accuracy began to decrease to around 93%. From 6 million to 10 million data instances, the accuracy stabilised at around 92% accuracy. In

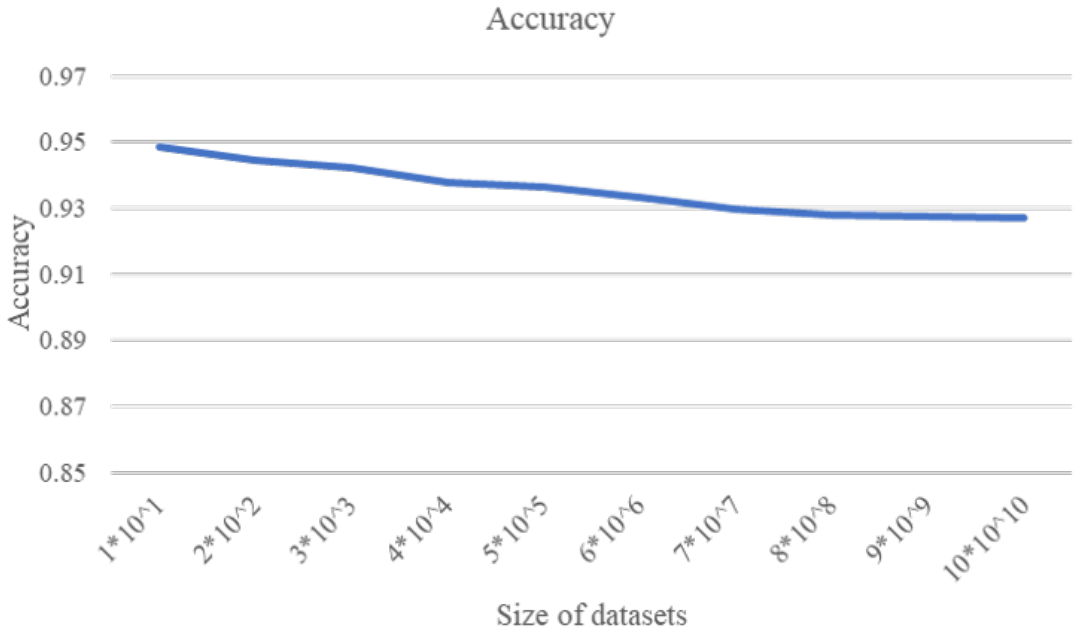


FIGURE 5.12: Long-term performance of MEK

sum up, our MEK method displayed a stable and good performance under large volume dataset scenarios over a long-term period. From the result of the long-term experiment, our MEK could maintain the accuracy of around 93% to 92% for long-term monitoring projects.

5.5 Limitation

Our MEK method is capable of detecting anomalies with around 93% accuracy in the high-frequency data environment. However, it has some limitations. Currently, the OAF is defined according to the scenario of application and background or based on sensitivity requirements. The OAF threshold is a constant lacking the capacity for automatic adjustment. From the above experiments, we found our method had a poorer performance under a small-size dataset scenario.

In future work, we intend to focus on an adaptive AF threshold definition and on enhancing the performance of the MEK method to overcome the stated limitations.

5.6 Summary

Dynamically changing environments present a challenge in online data stream anomaly detection. Our proposed MEK method adopts a sliding window principle and ensemble analysis to capture anomalies in data streams. We tested our method using public UCI datasets and practical SHM data and the results showed our method to be capable of detecting anomalies in data streams in high-frequency and high-volatility environments. From our experiments, we found that our method is significantly better than other methods. Moreover, our MEK method would enable structural engineers to monitor the structural health status in real-time. In addition, our MEK method can maintain around 93% to 92% accuracy in long-term monitoring. Moreover, our MEK method can handle heterogeneous and characterized data types efficiently. There are some limitations: our MEK method is restricted to adjusting itself automatically; its performance drops when the data environment is characterized by small-size. These are our challenge in work that is underway.

The main material of this chapter has been submitted to IEEE international conference on data mining (ICDM 2018) , and it has been under review by ICDM 2018.

Chapter 6

A hybrid intelligent framework for structural health monitoring

In the Chapter 3, 4 and 5 we proposed and develop methods for online anomaly detection for SHM data. Real time anomalies detected are the rare data behavioural patterns that are the indication of underlying structure's operating conditions. In practice, however, we also need a holistic method to assess the health status of a structure based on various sensor data types behaviours. Consequently, we are motivated to propose an integrated framework, incorporated with methods of anomaly detection to evaluate the health status of a bridge from multiple dimensional perspective. Hence our hybrid intelligent framework work in Chapter 6 represents an extended linkage from chapter 3, 4 and 5 for structural health monitoring system which should have the capability of computing a composite health index. The composite health index is a convenient indicator for use by bridge safety operational decision maker as a guide to judge the bridge health

status. The health composite index is derived from the consensus (also known as aggregation) algorithm's outcome which is computed based on historical and real time inputs from various sensor data types.

6.1 Motivation

We proposed a structural health monitoring framework that is based on the hybrid intelligent systems to compute composite structure health index as structure health indicator in this chapter. With our proposed health structure monitor framework, it provides a decision-level analysis for SHM. In contrast, recent studies focus on signal processing, mechanical modelling and computer-based systems, which are not robust enough to provide decision-level analysis confronted with heterogeneous data sources. Therefore, a hybrid intelligent system is required to alleviate possibilities of data conflicts, lack of conciseness and incompleteness that arise from the heterogeneous data sources to compute composite structural health index, which is essential for risk analysis. Our proposed hybrid intelligence framework is based on hybrid adaptive resonant theory, neural network (superior learning from dynamic data) and adaptive fuzzy inference (superior reasoning with dynamic fuzzy rules derived from time series data formulated membership functions) systems. The outcome of our proposed framework is a structural composite health index, an aggregated index via optimally weighted variables. We implemented our framework using a set of GPS and wind velocity sensors output from a real bridge to evaluate our framework. These sensors data used in our study are collected from

long cable-bridge. Our case study showed that our proposed intelligent data fusing framework is capable of reporting the structural status correctly.

In addition, because the practical data collected from the real bridge, most of the time data are normal. We used finite element model (FEM) to produce data to test our method in order to give a benchmark to determine what level is considered as healthy or normal.

6.2 Related work

There are many SHM methods using different techniques. Early methods employed signal processing techniques to find the abnormal signal, for example, if a single member in a structure was damaged, the fundamental frequency would be altered. However, this method was only able to detect whether the civil infrastructure was damaged. Other researcher proposed method was able to find the damage length and location in the structure. These damage detection techniques are all based on the frequency changes [126] [127], for example, a change in nature fundamental frequency implies there exists one or more member of a structure had been damaged. Therefore, these methods are vulnerable to environmental noises easily. In addition, the damage caused by material degradation, acid corrosion, man-made factor and etc. often fail to be detected by these methods. These methods are only concerned with the pixel-level structural health.

Applications of heterogeneous types of sensors have been introduced to improve the structural health monitoring. The sensor technology includes vast categories,

such as Mircoelectromechanical System (MEMS), LIDAR, infrared thermography and so forth. Various types of sensors form a sensor network for collecting data. Data fusion technique is introduced to overcome problems on data aggregation due to heterogeneous sources and characteristics of data for extracting predictive-to-prescriptive structure health analytic. Consequently, how to design and implement the data fusion is the critical component for SHM. There are many categories to classify different types of data fusions. Generally, data fusion has three levels which are pixel-level, eigen-level and decision level, each level is responsible for different specific tasks. Recent studies have employed data fusion to monitoring bridge structure, these studies proposed different data fusion techniques for SHM. [128] is capable of handling mono-type sensor data, which is not adequate enough for monitoring civil infrastructures. [129] [130] [131] [132] although have capabilities of collecting heterogeneous sensor types, however, these proposed methods are at pixel-level or eigen-level, where each has lack of delusional analytic and information. Our proposed framework is able to deal with heterogeneous data sources inclusive of decision-level information.

Multi-agent system (MAS) based SHM [33] system is an approach for large-scale structure health monitoring. It deployed different agents (design and develop by agent framework) to accomplish specific tasks. Typically, a MAS composes four types of agents: data monitoring agents, data interpretation agents, damage diagnostic agents and information layer agents. Each category of the agent has its sub-agent to execute specific tasks. The data monitoring agent is responsible to

monitoring signals of specific locations or members in structures; data interpretation agent is designated for signal processing, such as signal smoothing, peak value extraction, denoising and so forth; damage diagnostic agent is responsible to detect and locate the damage. The advantage of MAS is its scalability and its effectiveness had been evaluated by [133]. Although MAS is able to handle with large-scale structures and heterogeneous data sources, MAS lacks decisional information to suggest the maintenance and inspection routines.

6.3 Preliminary

6.3.1 Adaptive Resonance Theory(ART) [1] [2]

It is ubiquitous that in the real world we are always facing expectations, how do we cope with that? How do we recognize the familiar facts and absorb these familiar facts to our knowledge quickly? To learn things adaptively and achieved the balance between the adaptive and stability is important for real-time applications. Namely, these problems are called stability-plasticity dilemma. The adaptive resonance theory (ART) is designated to autonomously adaptive to handle those problems. The ART theory contributed to the self-organized neural networks. There are many categories of neural networks, such as feed-forward neural network, back-propagation neural network, these types of neural networks are called multilayer perceptron (MLP). These different types of neural networks have different architectures and mechanisms to learn from the data. Generally, MLP assumes the inputs are independent to each other. Thus it is extremely sensitive to the

input. If any new input comes to the neural network, it would often overwrite the past learning. In our case, the sensor keeps collecting data, MLP neural network is not a proper solution under this circumstance. ART self-organized neural networks is the solution in our framework. In the real circumstances, facts are unpredictable, ART is capable of learning things adaptively by maintaining a good balance between adaptive and stable.

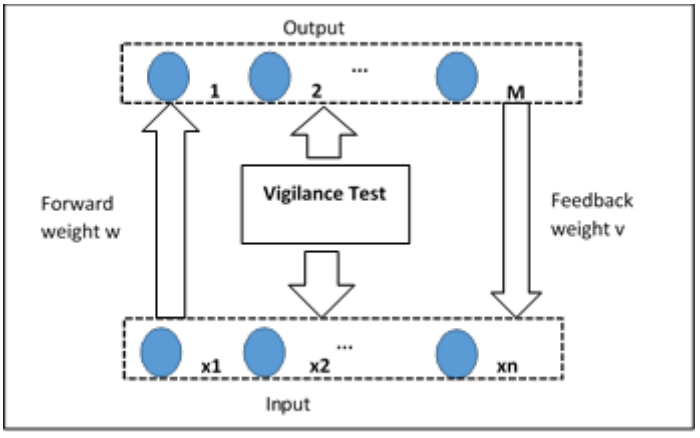


FIGURE 6.1: Structure of ART self-organized neural network

Fig. 6.1 shows the structure of ART self-organized neural network. It classifies the input variables into different neurons by computing the forward weight and feedback weight. When a new neuron is set up when failed in a vigilance test, and the test process ends when it has no more neurons to test. If an input is greater than the pre-defined threshold, the weights are updated according to the input. The detail of ART self-organized neural network is introduced in Section 6.4.2. Assumed, we can consider each different type of sensor is a variable, it could be classified into different neurons. After the training, we can use the weights as the parameters to leverage the relationships of different variables for further processing. The outcomes of our proposed framework is a structure health index,

which composites by the leveraged input variables. Therefore, ART self-organized neural network is critically important in our proposed framework.

6.4 Hybrid intelligence system for SHM

Our proposed framework employs hybrid artificial intelligence systems which are neural networks and fuzzy expert systems. In our framework, the neural network is ART self-organized neural network, which is unsupervised learning. Because the sensors are located in the severe circumstances, it is not a straightforward task to predict the expected data. After the ART self-organized neural network training procedure, the output would be the weighted data, which is used as the input for the fuzzy inference system. The fuzzy inference system is the next core component in our framework, which composites different types of data into the structure health index. Fig. 6.2 shows the overview of our proposed framework. Generally, there are three components which are: pre-processing, neural network and fuzzy inference system. Neural network component and Fuzzy Inference system are relying on the output of the previous component.

6.4.1 Pre-processing

Pre-processing component is responsible to clean and sanitize the raw data collected from sensors. The main tasks include denoising and integration (may be optional depending on sensor type). There is a vast number of denoising methods but selecting the proper method for denoising is skilful. Wavelet denoising , for

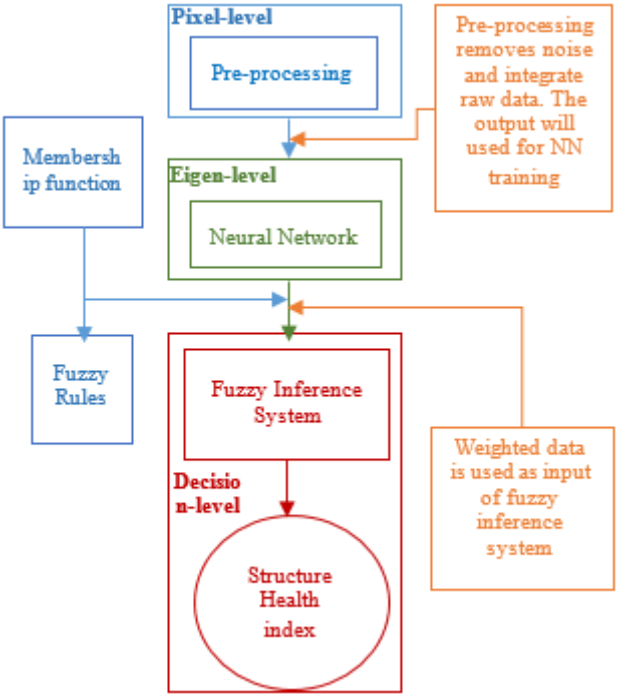


FIGURE 6.2: Hybrid intelligence Framework for SHM

instance, has been applied by many industries and studies and could be a good candidate of choice. There are many members in wavelet family which can be used for denoising . In our case study, based on the wavelet denoising we developed a real-time wavelet denoising .

Integration is another optional process to integrate the data to meet a specific requirement. For instance, a data with 1 Hz sampling rate can be integrated into an hourly. Generally, integration techniques are basic statistics, such as mean and mode.

6.4.2 Neural Network

Fig. 6.3 shows the process of ART Neural Network training, which is one of the core components in our proposed framework, it used the pre-processed data as

input to training the neurons with the weights. Once the training has completed, we used the weights to leverage the pre-processed data for further processing. Fig. 6.3 shows the process of the ART Neural network. At the initial stage, the weight w is set in Equation 6.1:

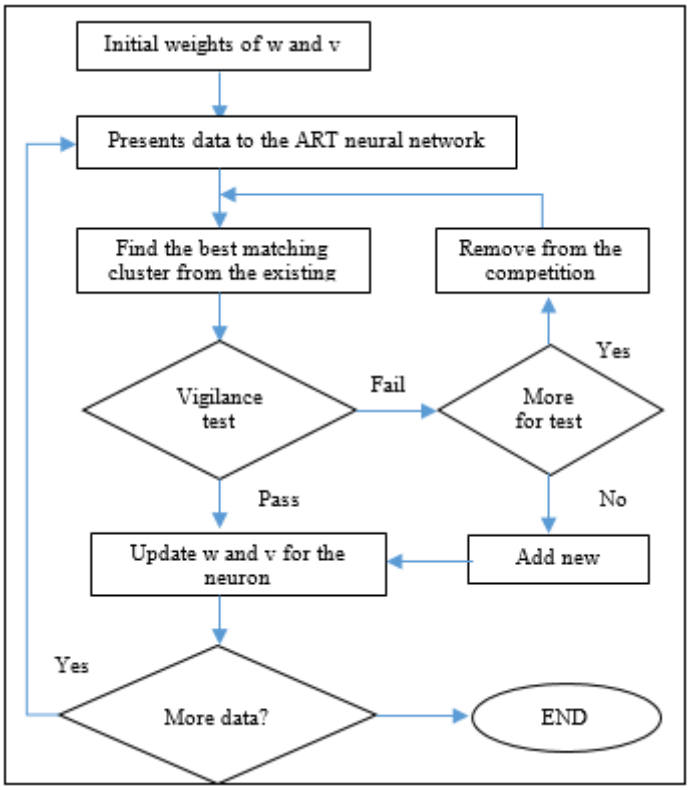


FIGURE 6.3: Process of Self-organized neural network

$$w = \frac{1}{||x|| + 1} \tag{6.1}$$

where x is the input variable (pattern). The v is set as 1 as default. To find the best matching neurons, we computed a "matching score" which reflects the degree of similarity of a data/patter. The score is defined in equation 6.2:

$$y_j = \sum_{i=1}^{N} w_{ij} x_i \tag{6.2}$$

where N is the number of input, j is the sequence number of existing neurons. After the competition, the winner neuron is the neuron with the highest matching score. After the finding the matching neurons, we tested is this pattern close enough to the best matching neuron. Thus we ran a vigilance test to ensure the neuron pattern is the matching. The vigilance test is defined in equation 6.3:

$$\frac{v_j x}{x} > \rho \tag{6.3}$$

where ρ is the pre-defined vigilance factor which $0 < \rho < 1$. The higher vigilance factor, the higher accuracy of the neurons. If x passes the test, the weight w and v are updated by Euqation 6.4 and 6.5 respectively.

$$w_{ij} = \frac{v_{ij} x_i}{0.5 + \sum_{i=1}^{N} v_{ij} x_i} \tag{6.4}$$

$$v_{ij} = v'_{ij} x_i \tag{6.5}$$

where the v'_{ij} is the old v_{ij} containing old weight values. If the pattern fails the test, and there is no more neurons can be used for the test, then a new neuron is added and the weight w and v are updated after adding a new neuron. After the training process is completed, we used the weight w to leverage the pre-processed data to weighted data for fuzzy inference system to compute composite the health index.

6.4.3 Fuzzy Inference System

Fuzzy inference system is another core component in our proposed framework. It maps the given inputs to an output using the fuzzy set theory. Conventional logic or Boolean logic uses the sharp distinction, which differentiates thing by true or false without any vague or imprecision. Unlike Boolean logic, fuzzy logic is multi-valued by defining membership fuzzy sets. The fuzzy set is different from the conventional set. Assuming we have conventions set X and x is an element in X, which can be denoted $x \in X$. However, the fuzzy set would consider x would be either belong to $X(x \in X)$, or does not belong to $X(x \notin X$. Assuming wind velocity is over 20m/s is considering as strong wind. If a measurement is 19.9 m/s which is classified to the non-strong wind by Boolean logic. In fuzzy logic, it would consider as 0.99 strong wind. (If we defined 20 is the strong wind as 1)[134].

By defining fuzzy rules, we are able to build our fuzzy inference system. Generally, there are two categories of fuzzy inference techniques which are Mammdani-style inference and Sugeno-style inference. The difference between these two styles is the defuzzification step. Sugeno-style inference is more computationally efficient than Mammdani-style inference. There are 4 steps to build a fuzzy inference system :

1. Defining input variables
2. Constructing membership fuctions (fuzzy set) for each variables and the output membership function.
3. Defining the fuzzy rules base

4. Encoding input variables, fuzzy sets and fuzzy rules and procedures to perform fuzzy inference

Fuzzy inference has following 4 steps:

1. Fuzzification

   Measuring the degree of crisp input form fuzzy sets for each input.

2. Rule evaluation

   According to the rules to compute the rules consequent. AND operation is equivalent to MAX operation. OR operation is equivalent to MIN operation.

3. Aggregation of rule consequent

   Aggregate the rules consequents computed from the previous step to form an aggregated fuzzy set.

4. Defuzzification The difference between Mammadani-style and Sugeno-style is here. Mammadani-style is used to compute the centroid which is defined in 6.6.

$$HealthIndex(HI) = \frac{\sum_{x=a}^{b} \mu_A(x)^x}{\sum_{x=a}^{b} A(x)} \tag{6.6}$$

where $\mu_{A(x)}$ is the fuzzy sets, x is consequent value.

The Sugeno-style defuzzification is different from Mammadani-style, it computes the weighted average of consequents values from rule evaluation.

In our framework, we used mammadani-style because it widely adopted to capture expert knowledge. The result of the defuzification is the structure index. In addition, unlike conventional fuzzy inference system, we can define our membership function by two options. One is conventional pre-defined membership function according to expert knowledge, another one is the adaptive membership function, which changes their membership function according to the maximum and minimum data.

6.5 Empirical Evaluation

6.5.1 Practical data

To evaluate our framework, we used the data collected from a cable bridge with one-month wind velocity and GPS displacement data (i.e 2,592,000 GPS and 259,200 wind speed observations) to evaluate our framework. We explain each component in our framework by using our data.

6.5.1.1 Data pre-processing

Fig. 6.4 shows the deployment of sensors networks and networks topology. Sensors mounted on the bridge send the data to monitor centre via optical fibre switch. These data are stored in the data server, any workstation can retrieve the raw data from their server. These raw data need to be pre-processed since these data are collected from bridge directly, which contains massive noise. The sampling

rate of GPS data is 10Hz, which collects 10 data every second. Wind velocity data are collected every 1 minutes. Therefore, we also need to synchronize these two datasets for further processing. In our case, we used wavelet to denoise our data first (with sym family). After the denoising, we have to synchronize (integrate) the data into the same time domain. Because the difference between each collected data is small and the sampling rate is 10Hz, we used mean of each 10 data to synchronize our GPS data.

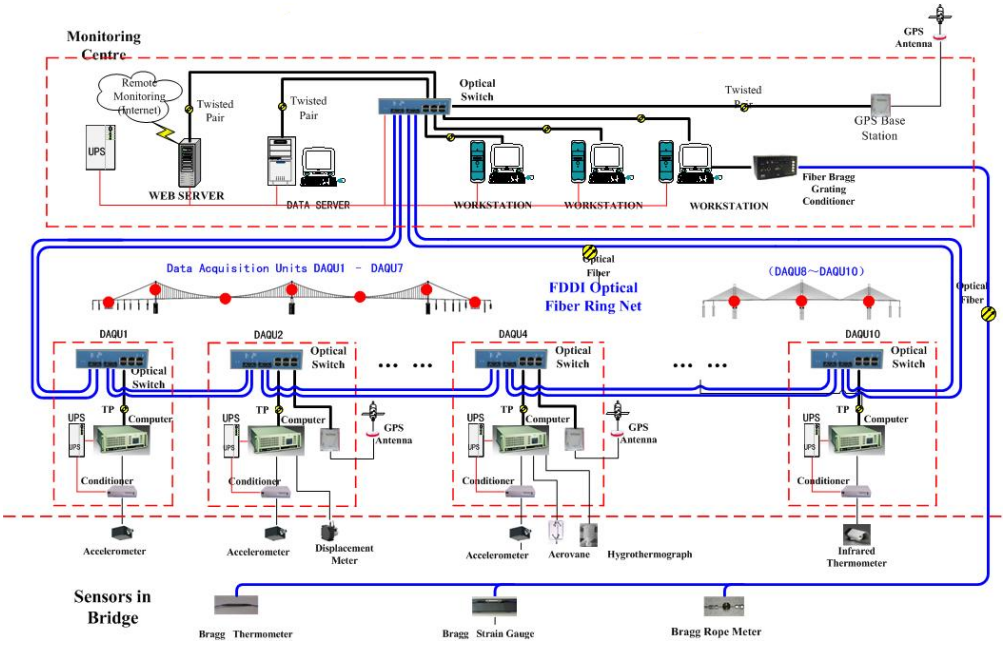


FIGURE 6.4: The network topology of bridge structure health

6.5.1.2 Neural Network

Once the pre-processing data are ready, we used these data for training our ART neural network. We set different vigilance factor to see the difference of final result. The higher vigilance factor indicates the higher accuracy, the results are different.

Vigilance Factor	**Number of Neurons**	**Weights**
0.6	2	0.032787 0.019802;
0.7	2	0.032787; 0.021505;
0.8	3	0.032787; 0.025974; 0.020619;
0.9	4	0.032787; 0.028986; 0.025974; 0.021505;
0.9	4	0.021505; 0.020619; 0.019802; 0.019417;

TABLE 6.1: Result of ART Neural Network

In our case study, we trained the neural network with different vigilance factor, the higher vigilance factor the higher accuracy. It is obvious to observe that (Table 6.1) with the increase in vigilance factor, the number of neurons is increasing. The weights of vigilance factor 0.6 and 0.7 are same, with the increase in vigilance factor, the weight is changing. The high vigilance factor does not indicate the more accurate model, it would result in the over-fitting problem.

6.5.1.3 Fuzzy Inference system

After leveraging the pre-defined data with weight computed by ART neuron network, we used the leveraged data as input for fuzzy inference system. We only have wind velocity and GPS data, consequently we only have two input variables. The next step is to define the membership function. Generally, the membership function of wind velocity is based on the expert knowledge which is Beaufort scale.In our case, according to the Beaufort Scale we have four fuzzy sets for wind

velocity which are: light, normal, strong and very strong. The GPS displacement is dynamic membership function which depends on the period of GPS data, it would be computed dynamically according to the existing data. In this case, the membership function of GPS has 3 categories which are light, normal and strong. We defined four fuzzy sets for output membership, which are safe (80-100), normal (55-85), risky (20-60) and highly risky (0-30).

To perform fuzzy inference, we need fuzzy rules. In this case, the fuzzy rules we had defined:

- If the wind is light AND GPS is light. Then the output is safe
- If the wind is normal AND GPS is light. Then the output is safe
- If the wind is strong AND GPS is light. Then the output is Normal
- If the wind is very strong AND GPS is light the then output is Risky
- If the wind is light AND GPS is normal. Then the output is safe
- If the wind is normal AND GPS is normal. Then the output is normal
- If the wind is strong AND GPS is normal. Then the output is normal
- If the wind is very strong AND GPS is normal. Then the output is risky
- If the wind is light AND GPS is strong. Then the output is risky
- If the wind is normal AND GPS is strong. Then the output is risky
- If the wind is strong and GPS is strong. Then the output is highly risky

- If the wind is very strong and GPS is strong. Then the output is highly risky.

6.5.1.4 Result of practical data

	GPS Input		Wind Input		Health Index	
Vigilance Factor	Max	Min	Max	Min	Max	Min
0.6	185805.5415	185805.5472	0.0065	0.5481	92.4	92.1
0.7	185805.5415	185805.5472	0.0065	0.5481	92.4	92.1
0.8	264672.8549	264672.863	0.0093	0.7808	87.4	76.1
0.9	386005.4542	386005.466	0.0136	1.1387	84.6	73.4
0.99	287400.2377	287400.2466	0.0101	0.8478	82	72.8

TABLE 6.2: Result of Health Index

For each vigilance factor, the scale of the membership function is adjusted according to the weights we computed from ART neural network. All the input values select the max and min value from that month. Because the weights of different vigilance factors are different, the leveraged input values are different.

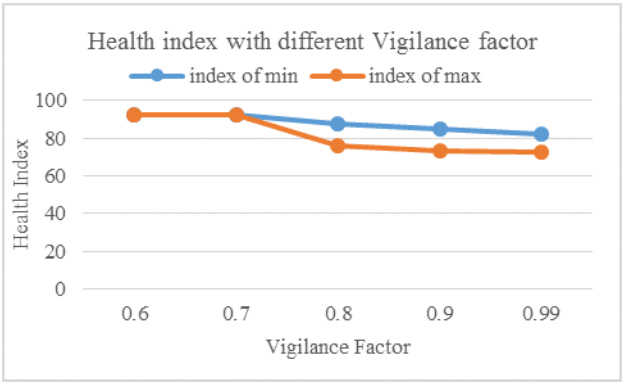


FIGURE 6.5: The network topology of bridge structure health

With the increase of vigilance factor, the index of max is approaching to 82 and the index of min is approaching to around 71 (Fig. 6.5 and Table 6.2). Although the difference between index of max and index of min is around 11, the overall

result suggests the structural health is in the normal or safe range, which can be observed from defined output fuzzy set. The structure health index computed by our framework matched with the official bridge health report. From this perspective, our proposed structural health index is reasonably reliable. According to the structural health index, the operator is able to make a decision to the bridge. If the structure health index gives a risky indicator (30-60), an inspection needs to be arranged to inspect the structure for maintenance. If the structure health index decline to highly risky, this bridge needs to stop service for guarantying the public safety. The operator can determine the specific actions and decisions. Although we proposed a method to evaluate a health status of a structure, the arbitrary to determine the degree of health status is not clear. In Section 6.5.2 we conducted experiments to discover an arbitrary value of a SHI using simulation data.

6.5.2 Result of simulation data

It is hard to collect abnormal data from the practical bridge, so we used finite element analysis to simulate the practical bridge structure to produce a massive amount of data to determine the benchmark health index and the arbitrary value.

6.5.2.1 FEM model and Data

Finite element model (FEM) is the most widely used model in civil engineering research , especially for structure analysis. To generate the simulation data from FEM, we simulated the structure according to the practical data. We used ANSYS

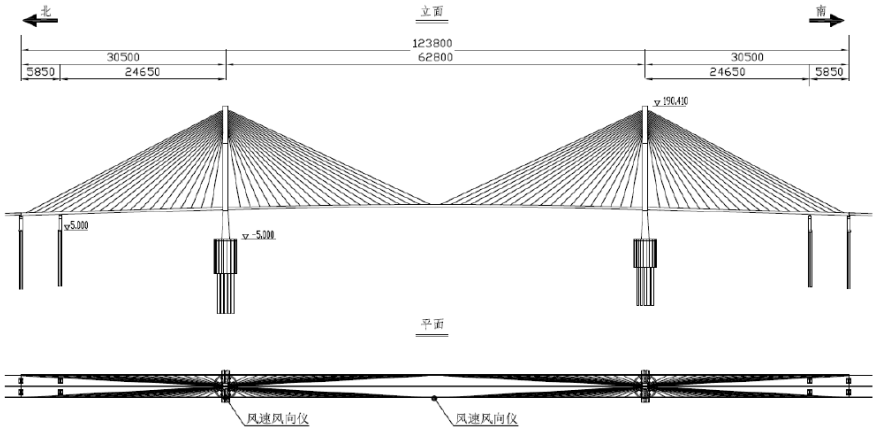


FIGURE 6.6: Side view and top view of real bridge structure

to build our FEM, since it equipped with comprehensive and abundant finite element analysis tools. There are three main components that we need to model:

1. Spine Beam : In spine beam model, bridge towers and pipers are deployed by beam element in ANSYS.

2. Deck : The deck of bridge are modelled by BEAM4 element (ANSYS tool), which is uniaxial element equipping with tension, compression, torsion and bending abilities.

3. Cable : Cable is modelled by LINK10 (ANSYS tool) element with tensile stress capability.

Based on the real bridge structure parameter(Fig.6.6), we used FEM tool build a same bridge structure which shows in Fig.6.7.

After bridge modelling, we conducted a probabilistic analysis for calculating the bridge response, which simulates the process of sensor collecting data from a bridge. Monte Carlo simulation (MCS) is used to simulate the uncertainty of

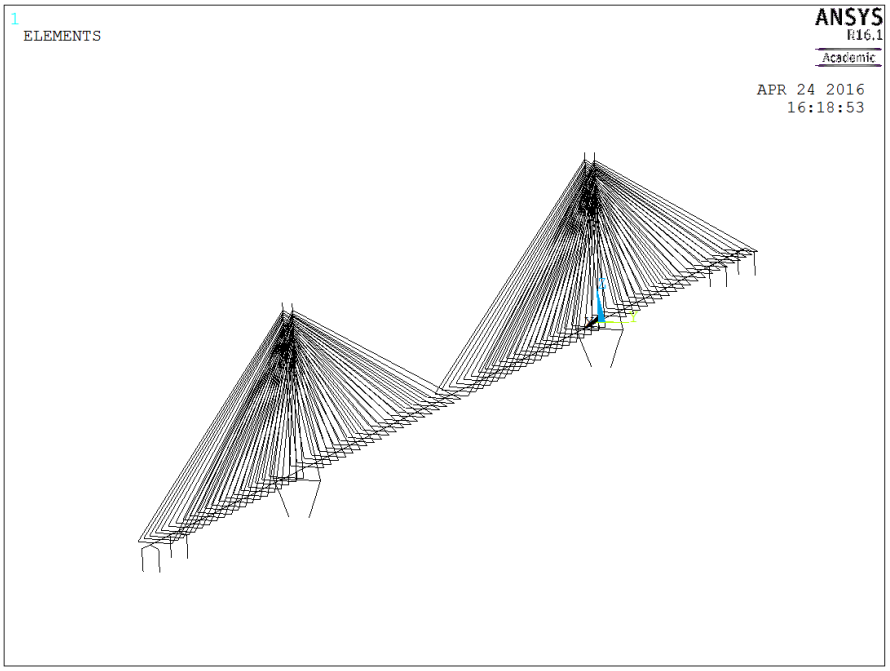


FIGURE 6.7: 3D side view of bridge model

the environmental situation. The simulation result is considered as physical observations which collecting from the sensor. In this simulation, we only consider three type of measurements: 1. Middle Vertical displacement (From beam) 2. Axle stress of Middle span (From deck) 3. Cable textile stress (From cable). In total, we sampled 10,000 data point from each type of measurement.

6.5.2.2 Structural health index computation

All the data generated from the FEM simulation have the same process with the second and third steps in our proposed framework. The fuzzy rules are defined according to the distribution of each type of measurement. The principle of rules definition includes:

- If one of mid displacement, cable stress or beam stress is 'low', the health index (HI) is 'safe'

- If one of mid displacement, cable stress or beam stress is 'average', the HI is 'normal'
- If one of mid displacement, cable stress or beam stress is 'high', the HI is 'risky'
- If all of measurements are 'low', then HI is 'safe'
- If all of measurements are 'average', then HI is 'normal'
- If all of measurements are 'high', then HI is 'risky'

6.5.2.3 Result of simulation data

Fig.6.8 shows the distribution of HI with 100,000 data point of each type of measurement. From Fig. 6.8 demonstrates that most of the cases the HI is ranging from 0.60 to 0.8, which contains 99,871 cases with approximate 99% proportion. This result indicates that under most of the scenarios, the health structural is 'normal' and 'safe'. This result is also similar to practical data, where most HI are above 70 under 'normal' or 'safe' status. Consequently, we defined any health index above 70 is considered as a normal status of a structure.

6.6 Summary

Our proposed hybrid intelligence structure health monitoring framework is able to provide a real-time monitoring with a decision-level suggestion. It employed ART neural network for handling with quick changing circumstances and dynamic fuzzy

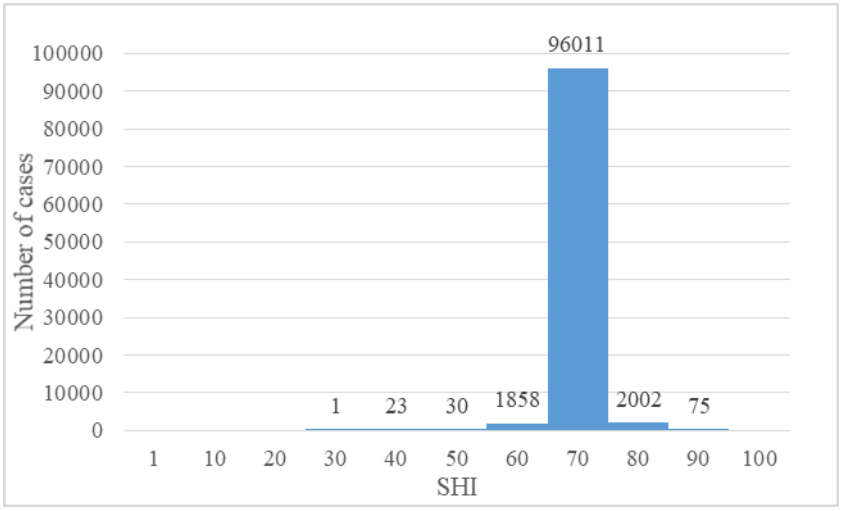


FIGURE 6.8: Health index of simulation data

inference system helps to tune the membership function adaptively. Through the bridge case study, we evaluated our framework is reliable. In addition, our result is consistent with official bridge health report. Unlike convention SHM systems, it capable to handle with heterogeneous data and decision-level suggestions by using the structure health index as an indicator. According to this structure health index, effective maintenance and inspection can be planned and arranged. In addition, we also use FEM to simulate a massive amount of data to set the benchmark health index at 70. The main material of this chapter has been published in IEEE international conference on industrial electronics and applications (ICIEA 2016).

- Publication : D.Sun, V.Lee and Y. Lu, "An intelligent data fusion framework for structural health monitoring", proceeding of International Conference on Industrial Electronics and Applications (ICIEA) ERA Rank A Conference, June 2016, Hefei. Proceeding of IEEE ICIEA pp.49-54.

Chapter 7

Conclusion and future work

With the tremendous development of machine learning and data mining techniques in the recent decades, civil engineers have attempted to apply these techniques to SHM. Current SHM deploys various types of sensors on a structure, which generate a massive amount of data each day. Extracting valuable information can help civil engineers to identify risky cases of a structure. Online analysis of anomaly detection or heterogeneous data analysis helps them to evaluate the health status of a structure based on the sensor data.

This PhD project was aimed to solve the problem of online anomaly detection and heterogeneous data analysis for SHM. I proposed method which has been verified with extensive empirical evaluation using public UCI datasets and practical SHM datasets.

7.1 Contribution

In this thesis, I propose four methods to solve the two research question (in Chapter 1). In this context, I have made following contributions :

1. A proposed sample entropy gradient method for comparing the similarity between two time-series data. It can be used for anomaly detection and measuring similarity of two time-series datasets (Research Question 1)

2. A proposed ensemble kernel (EK) for online high-frequency data anomaly detection. It detects anomalies for a single dimensional data stream efficiently with a stable performance. (Research Question 1)

3. A proposed multi-dimensional ensemble kernel method for online high-frequency data anomaly detection. It detects anomalies for the multi-dimensional data stream with a stable performance. (Research Question 1 and 2)

4. A proposed structural health monitoring evaluation method using the neural network and fuzzy inference system. It introduces the health index(HI) as a health status indicator of a structure.

To answer the first research question, I proposes SEG, EK and MEK methods. SEG method is a similarity comparison based method, which computes the distance between the incoming data with benchmark data to identify anomalies. To achieve a good result, SEG relies on a good quality benchmark dataset. Consequently, I propose EK mefthod to overcome this problem. EM method applied

ensemble analysis, kernel density estimator and sliding windows to detect anomalies for high-frequency SHM data. EK can maintain a stable performance for high-frequency SHM data. However, EK method is only used for single-dimension anomaly detection, I extended EK to MEK to used on multi-dimension anomaly detection. Except to analyze abnormal data, how to analyse heterogeneous data to support decision making for MEK is also important. MEK can be used for heterogeneous data analysis because the OFA indicates the overall health status of a structure. However, OFA is extremely sensitive, any abnormal behaviour could result in a low OFA score. Hybrid intelligence system is an alternative solution to analyzing heterogeneous data. It gives a health index of a structure to indicate the health status of a structure. It is an easy-understand index to help civil engineers to develop effective maintenance plans and disaster plans.

7.2 Future work

Based on this project, there are some possible potential directions :

1. Reinforcement learning (RL) is also a potential method for heterogeneous SHM data analysis. To apply RL to SHM, we could treat each sensor as an object with a number of states. By connecting all the states, a Markov chain for SHM can be formed. Based on this Markov chain I could try different RL algorithm to see the effectiveness of RL.

2. There are many algorithms in RL, I can try different RL algorithms to SHM data to compare the performance of them. Moreover, we can also compare

RL on SHM with our hybrid intelligence system to investigate which algorithm is more effective on SHM problems.

3. The recurrent neural network is also an extensively used method for natural language process problems. It specialized in processing sequence to sequence data. Most of SHM data are time-series data, which is a type of sequence-to-sequence data. We can investigate the effectiveness of RNN for SHM.

4. My current project concentrated on prediction and health status evaluation. Applying new machine learning and data mining technique on damage detection and damage localization of SHM is also an interesting area. Applying conventional neural network on damage detection can help civil engineers to detect damage to a structure.

Appendix A

Appendix A

Appendix A presents the detail of financial case study using main international stock market indices. Computing the similarity between stock market indices using SEGs sheds light on a country's share market trend, which is an important indicator of the country's economic status. The use of SEG avoids the need for prior knowledge of tolerance parameter setting. Consequently, the SEG based approach is more scalable and can be more relevant to more application domains where the domain-specific knowledge needed to set tolerance parameters is not easily accessible.

The study also contributes to practice in terms of predictive/trend information regarding individual and institutional investment communities that is not readily available in the CSE-based approach. The case study results assist long-term investors to make informed decisions on which stock markets to follow for each period. For more risk-seeking short-term investors, intra-period gradients can be

regarded as an alternative benchmark for long-term and short-term moving average trading indicators.

I used NASDAQ (US), ASX (Australia), Nikkei (Japan), and SSE(China, Shanghai) indices. Before analyses of the stock markets, we present some general descriptive statistics (Table A.1) and we calculated return of investment on stock market indices for each period. In this study we had two categories of comparison: Markets of developed countries vs. markets of developed countries and markets of developed countries vs. markets of developing counties. For the markets of developed countries vs. developed countries, we compared:

1. NASDAQ (the U.S.) vs. ASX (Australia, AU)
2. NASDAQ vs. Nikkei (Japan, JP)

For the markets of developed countries vs. markets of developing countries, we compare:

1. SSE(China, CN) vs. Nikkei
2. SSE vs. NASDAQ

A.1 Data

All finance data were downloaded from Yahoo Finance from March 1, 2000 to September 1, 2015. The share market indices were the ASX, the NASDAQ, the

		2000-2005	2005-2010	2010-2015
US	Mean	1767.7655	1678.4605	2982.673749
	Std. Dev.	884.5083	237.68674	825.0104527
	Median	1462.33	1701.35	2745.594971
	Skewness	1.6036745	-0.3389005	0.486650265
	Kurtosis	1.4316458	-0.1381977	-1.02718499
	ROI	-182.14%	17.28%	55.43%
AU	Mean	3305.7746	4962.0811	4891.673185
	Std. Dev.	3305.7746	4962.0811	4891.673185
	Median	3293.8999	4905.3999	4836.5
	Skewness	1.0685965	0.1510853	0.124857388
	Kurtosis	1.885322	-0.7476011	-1.10991605
	ROI	23.84%	10.70%	8.04%
JP	Mean	11731.658	13383.284	12540.52507
	Std. Dev.	2733.3931	3114.9897	3660.303776
	Median	11069.01	13500.46	10836.63965
	Skewness	1.1896763	-0.2071418	0.67490437
	Kurtosis	1.108512	-1.2395513	-0.75692183
	ROI	-70.46%	-15.81%	44.00%
CN	Mean	1660.6365	2588.5378	2589.475144
	Std. Dev.	261.13985	1256.1459	597.9855436
	Median	1608.51	2419.78	2389.37
	Skewness	0.4978542	0.7431734	1.80502304
	Kurtosis	-0.8067264	-0.2155683	3.580943959
	ROI	-30.25%	57.79%	2.49%

Table A.1: International stock indices descriptive statistics

Nikkei, and the SSE. Share market index datasets include many data types, such as open, close, volume, highest, lowest, and adjusted close. In this study, we used only the closing price over the given period.

A.2 Financial analysis

Figure A.1 shows the entropy gradients of the NASDAQ and the ASX indices. It is evident that the ASX index is correlated with the NASDAQ index, especially over the first six sequences (the first 180 trading days); sequence 21 to sequence 34

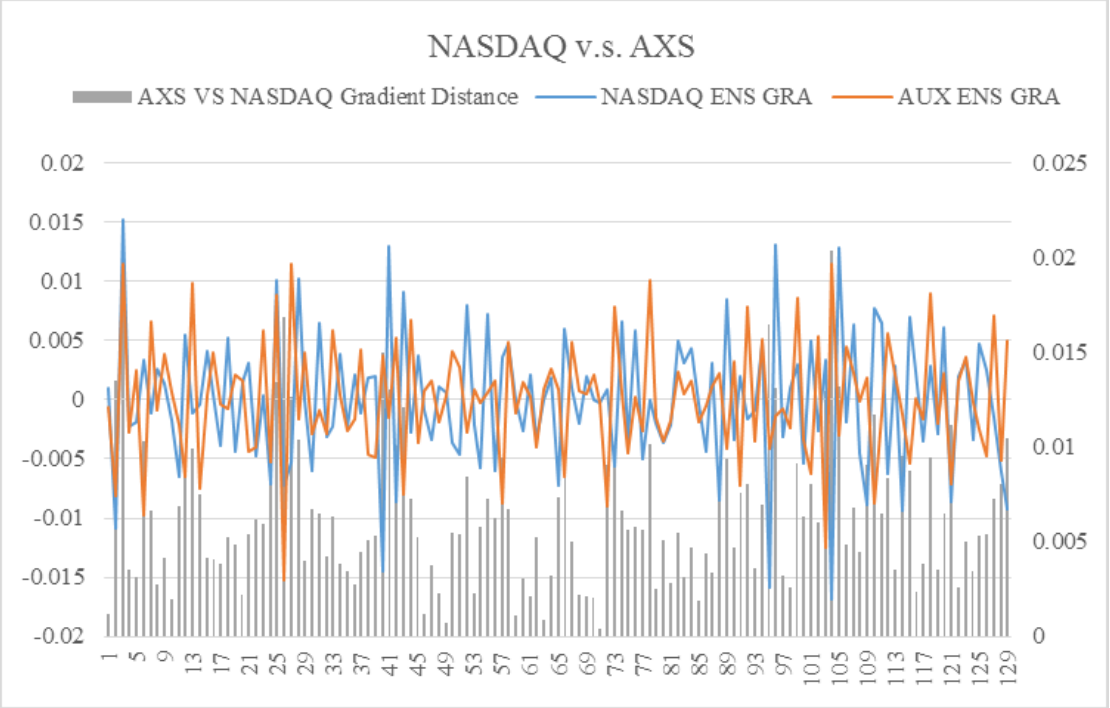


FIGURE A.1: NASDAQ SEG and AXS SEG

and sequence 74 to 85 are also highly correlated. From the political perspective, Australia is an ally of the U.S., sharing information and interacting closely [135]. Also, branches of many international corporations are located in Australia. Reference [136] provides a detailed report of corporations financed by the U.S. From this report, JP Morgan Chase and Citibank were two major investors in some of Australia's largest corporations including the banking, mining, and retailing sectors and so forth. These two corporations accounted for at least 11% of the shares of those companies (details can be found in the report). In banking industry especially, JPMorgan Chase and Citibank held at least 30% of the shares of four of the most important Australian banks (Commonwealth, National Australia Bank, Westpac, and ANZ). Thus the U.S. has a tremendous impact on the Australian financial market. Consequently, the ASX index correlates with the NASDAQ index over many periods. Fig.A.1 demonstrates the Euclidean distance between the ASX

index entropy gradient and the NASDAQ index entropy gradient as well by the grey bar chart. The gradient distance reflects that the difference between the two entropy gradients was small, ranging from 0.0011 to 0.02. Most of the distances between these two entropy gradients are around 0.01. Fig. A.5 (which shows the summary of Five-year entropy gradient Euclidean distances) shows that Australia had the smallest distance from the U.S. (except in the first period). Thus, we can conclude that the ASX is correlated with the NASDAQ over this period.

Figure A.2 shows the Nikkei and the NASDAQ entropy gradients. Generally, in Fig. A.2, these two indices are not highly correlated, but they do correlate in some periods. Reference [137] study indicated that there was no evidence to prove that the Japanese financial market index was independent of the U.S. financial market. These two entropy gradients show less correlation than Australian and the U.S. index. From [137], reference [138] concluded that the U.S. capital has drifted away since October of 1987, which is also reflected by the lower correlation between the two markets from 2000. In some periods, however, some correlations exist, such as sequence 42 to 47 and sequence 79 to 86, where the entropy gradients are consistent. But those correlations last only for short periods. Fig. A.5 shows the distance between these two entropy gradients. Overall, the entropy gradient distance (Fig.A.2) displays a large fluctuation over this period. In the sequence 37 to 105 the fluctuation is relatively large. Moreover, from Figure 5 we can observe that the distance between Japan and the U.S. is relatively large.

Figures A.3 illustrates the entropy gradient and Euclidean distance between China's

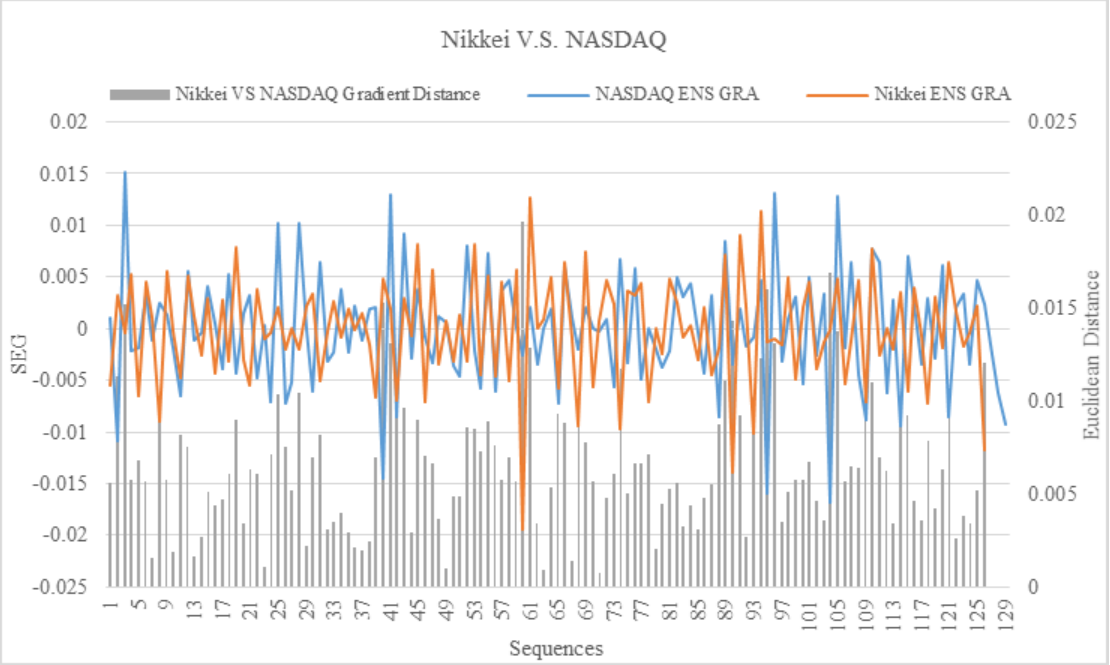


FIGURE A.2: NASDAQ SEG and Nikkei SEG

SSE index and the U.S. NASDAQ index. Overall, it is difficult to observe any correlation between these two indices. However, a few correlations can be found between the two entropy gradients (Figure A.3), from sequence 11 to sequence 13, sequence 16 to 18, sequence 73 to sequence 76, and so forth. These correlations last for only a short time. Reference [139] showed that the U.S. financial market had a low influence on China's financial market, an observation that is reflected by our approach as well. In Fig. A.3, we can observe that the entropy gradient distance ranges from 0.005 to 0.025, and most of the distances remain around 0.01 to 0.015. Overall, the distance between the two markets showed a large fluctuation over this 15 year period, which also implies less correlation over the period. Figure 5 also reflects that the 5-year-period distance is large, indicating a weak correlation over this period.

Figures A.4 depicts the China and Japan entropy gradients and their entropy

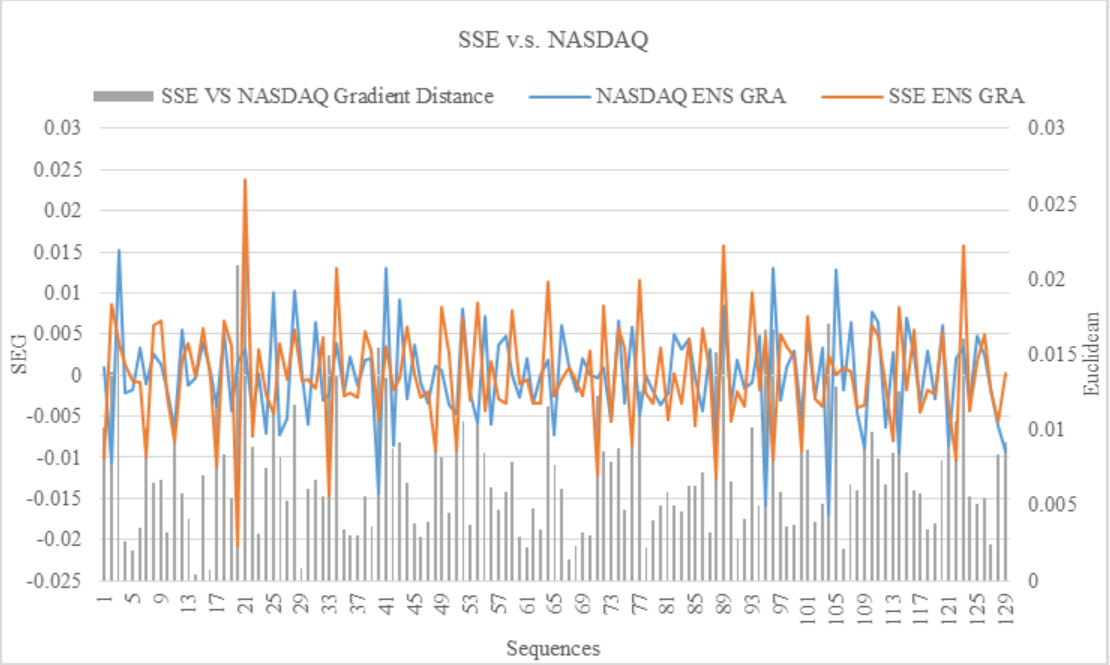


FIGURE A.3: SSE SEG and NASDAQ SEG

gradient distances, respectively. Overall, it is difficult to observe any correlations between the SSE and the Nikkei indices. Some short-term correlations can be discovered from Figure 8, including sequence 10 to 17, sequence 59 to 60, sequence 86 to 91, and sequence 114 to 119. Reference [140] investigated the relationships between stock markets of Asia. The finding was that Japan's stock market had a profound impact on Asian markets, including those of Australia, China, Hong Kong, Malaysia, New Zealand, and Singapore. From Fig. A.4, the distance of entropy gradient fluctuates over the time-series, ranging from 0.0005 to 0.024.

We also calculated the general SEG Euclidean distance of each 5-year period (Fig. A.5). It is evident that the values of the period 2005 to 2010 are the lowest. During that period of the global financial crisis (GFC) involved most stock markets, especially those of the European Union and the U.S.. This can be considered a detrimental factor resulting in the lowest entropy gradient distances among all

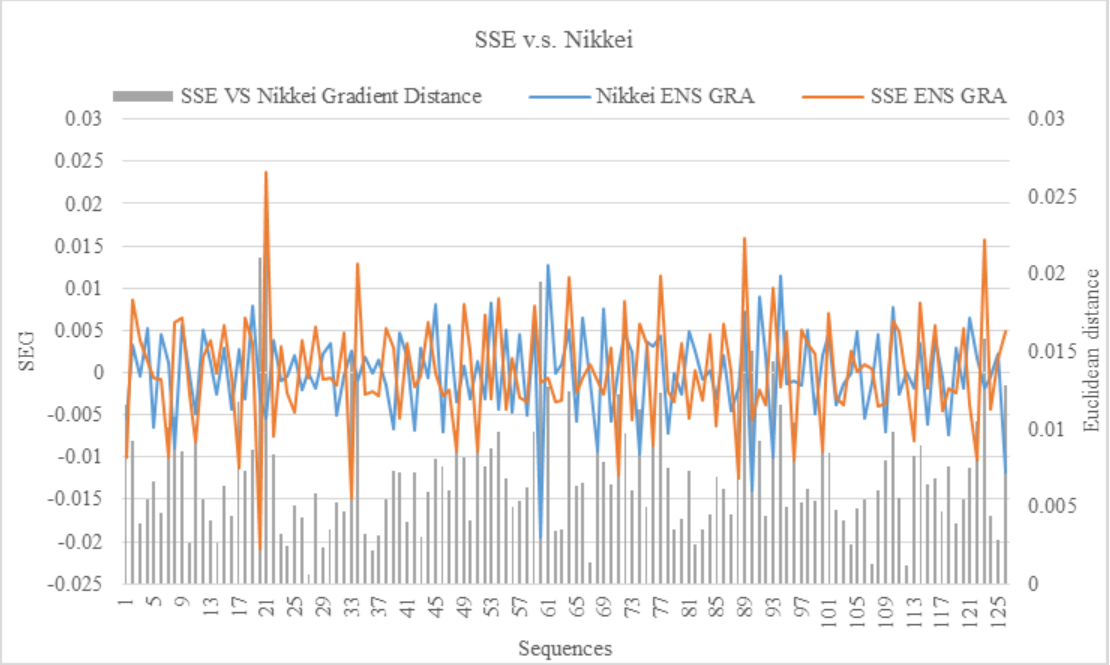


FIGURE A.4: SSE SEG and Nikkei SEG

countries. According to the summary in Fig. A.5, the distance between the SSE and Nikkei financial markets were lowest in the first two periods. Therefore, we can conclude there was a cor-relation between the Chinese and Japanese stock markets.

In general, from Figures A.1, A.2 and A.3 we find an interesting trend from sequence 82 to 85, showing an 'M'-shaped fluctuation. This period was in 2008, coinciding with the GFC that had a detrimental impact on international stock markets. These sequences can be relevant to that period since they matched with the time-series, but it is also observed that there is no 'M'-shaped trend in comparisons with the SSE over the same period. Because the U.S. stock market was most influential in international finance, other developed countries were involved in this crisis, as reflected by our approach in sequence 82 to 85. According to Table A.1, in the 2005-2010 period NASDAQ had the lowest means during these

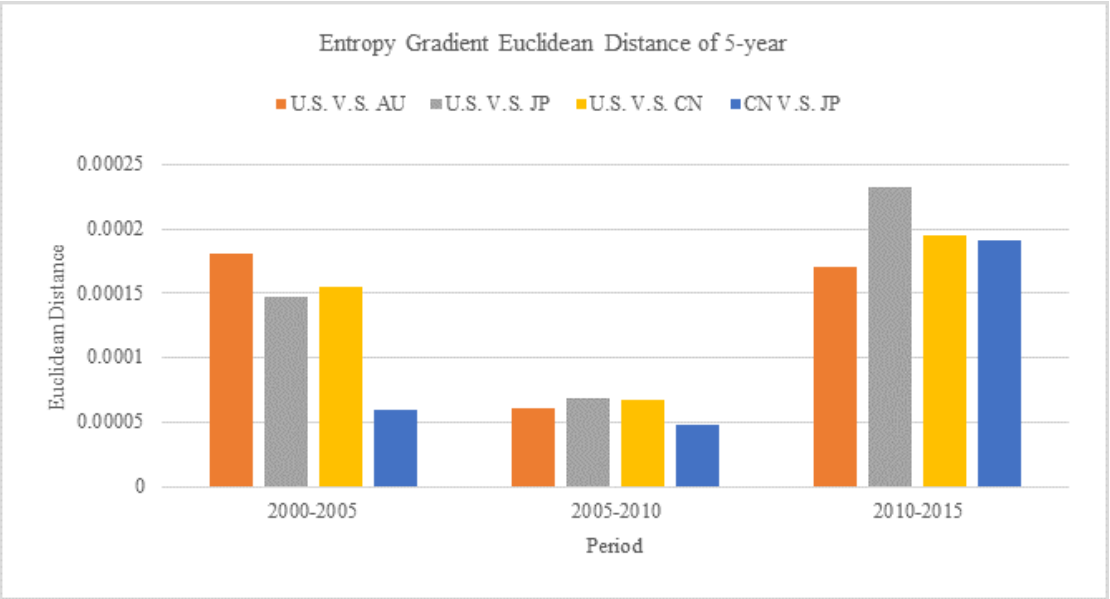


FIGURE A.5: 5-year SEG Euclidean Distance

15 years, since it experienced the GFC in 2008. This financial crisis also affected other countries, which can be indicated by the standard deviation. If, however, we observe the sequence 82 to 85 of entropy gradient of the SSP and the NASDAQ, the 'M' shape is not obvious. This is because China suffered less impact from the GFC [26]. Although the financial market did not have a great impact that does not indicate that the economic impact was slight. A study indicates that export volume declined dramatically after the GFC [140]. Therefore, from this comparison, we are able to declare that our SEG measure can detect correlations between two time-series datasets.

Fig. A.5 provides a summary of the 5-year entropy gradient distances among different countries. The Euclidean distances in the second period are the lowest among these three periods, and are influenced by the GFC in 2008. In the first two periods, China vs. Japan showed the lowest Euclidean distance, implying a strong correlation between these two markets. Australia had a strong correlation

with the U.S. from the second to the third period. The financial markets of Japan and China show less correlation than the other two markets during these periods.

A.3 Comparison of SEG with 30-day-trend

We compare SEG with 30-day-trend to investigate the trend consistency between them. 30-day-trend is one of financial trading indicator used in short-term investment. In this case, we use the linear regression to calculate the trend of each 30-trading-day. By computing the Euclidean distance between SEG and 30-day-trend, we can observe whether these two methods showing a consistent trend. Fig.A.6 shows the result of SEG vs. 30-day-trend of each country. The average of the Euclidean distance of NASDAQ, ASX, Nikkei and SSE are 0.45, 0.50, 0.45 and 0.31 respectively. Overall visualization suggests that in the Figure A.6a, b, c, and d, their approximate overall 30-day-trends are essentially consistent with corresponding SEGs. There are, however, some differences between these two measures occurred with large magnitude fluctuations. For instance, in Fig. A.6a, from sequence 1 to 11, 30-day-trend displays a larger fluctuation during that time, whereas SEG describes a more gentle fluctuation. This difference in fluctuations, due to probably spurious signal in data, has caused the increase of the Euclidean distance. The difference between 30-day-trend and SEG measures is due to their different calculations. 30-day-trend is computed by linear regression of 30-day closing index, whereas SEG is computed by the gradient of two entropies. Hence we can infer that trends described by 30-day-trend and SEG are basically consistent. In other words, our proposed SEG can be used as another financial trading

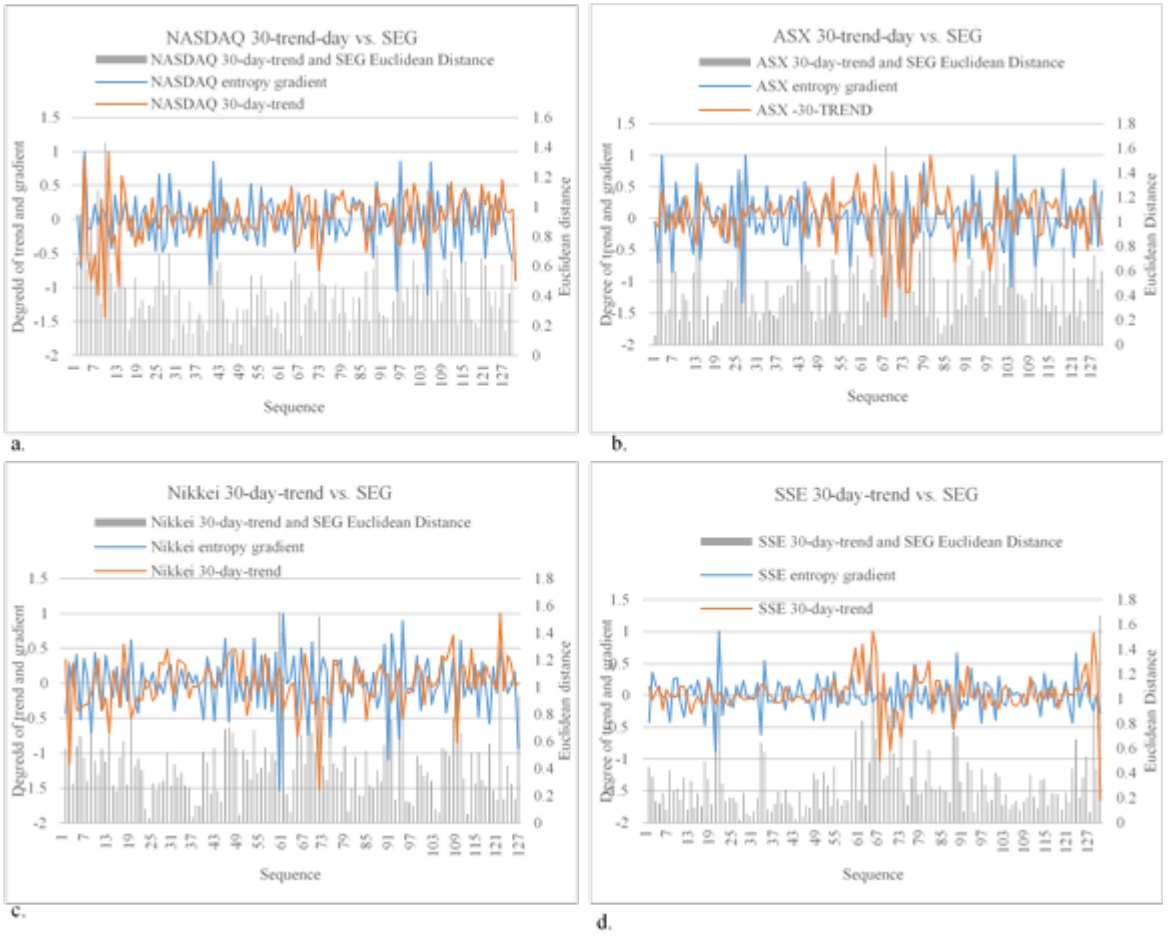


FIGURE A.6: 30-day-trend and SEG

indicator for short-term stock market investment decision making over a specified period.

A.4 SSE and SEG comparison

To investigate the difference between CSE and SEG, we conducted an experiment incorporating international stock market data. All results are shown in Fig. A.7. Each 60trading-days is considered a time unit. Interestingly, if we compare CSE with SEG, asynchrony detected by CSE may be reflected as a highly correlated entropy gradient in SEG figures. For instance, in Fig.A.7, sequence 2 to 4 indicates

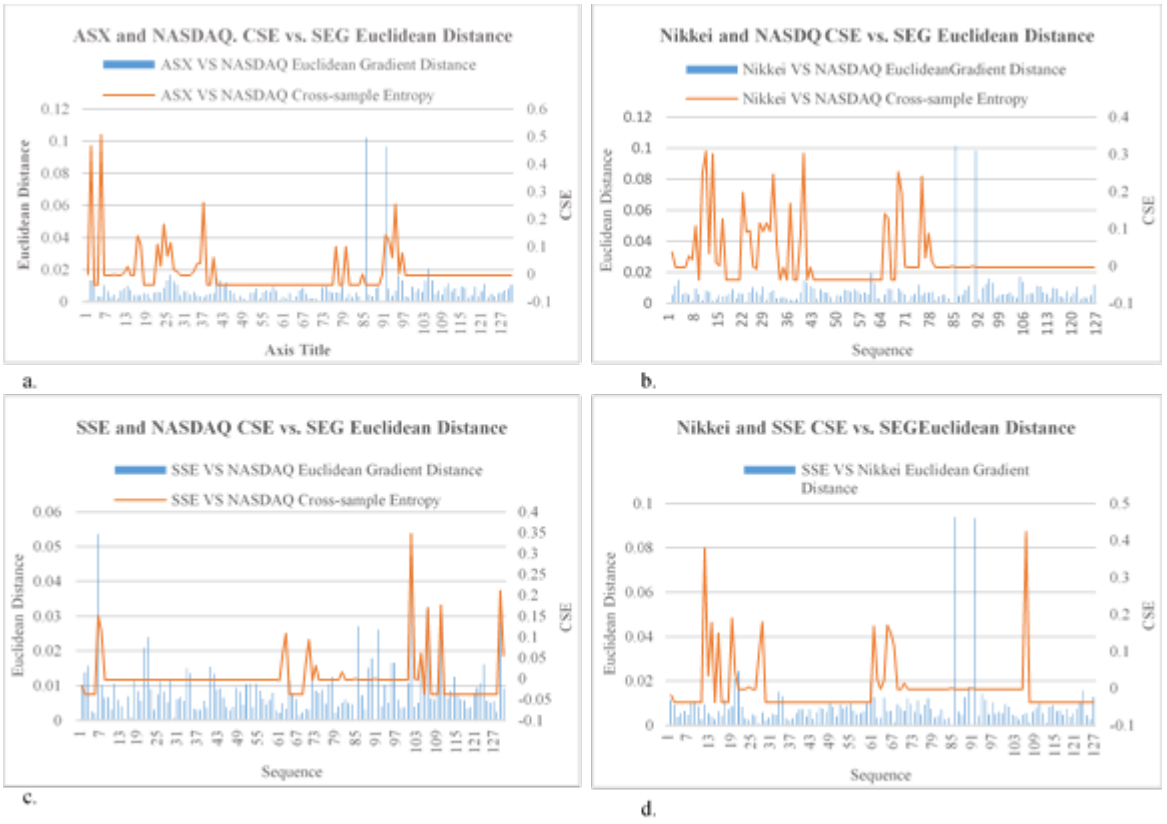


FIGURE A.7: CSE and SEG Euclidean Distance

high CSE values, and sequence 2 to 4 in Fig. A.1 shows a high correlation between the ASX and NASDAQ indices. Sub-figures A.7b, c, and d all display the same phenomena in comparison with sub-figures with Figure 1, 2 and 3. The cause of these phenomena is the different calculation procedures. CSE can identify the difference sensitively, because its aim is to measure similarity and asynchrony. On the other hand, SEG measures the difference of gradients between two entropies of two sequences.

The different computation procedures between CSE and SEG-based computation of similarity lead to multiple interpretations of CSE results. One of the reason is that the two measures are calculated by different procedures, and especially CSE is affected by high volatility of sequences. To investigate this contradiction,

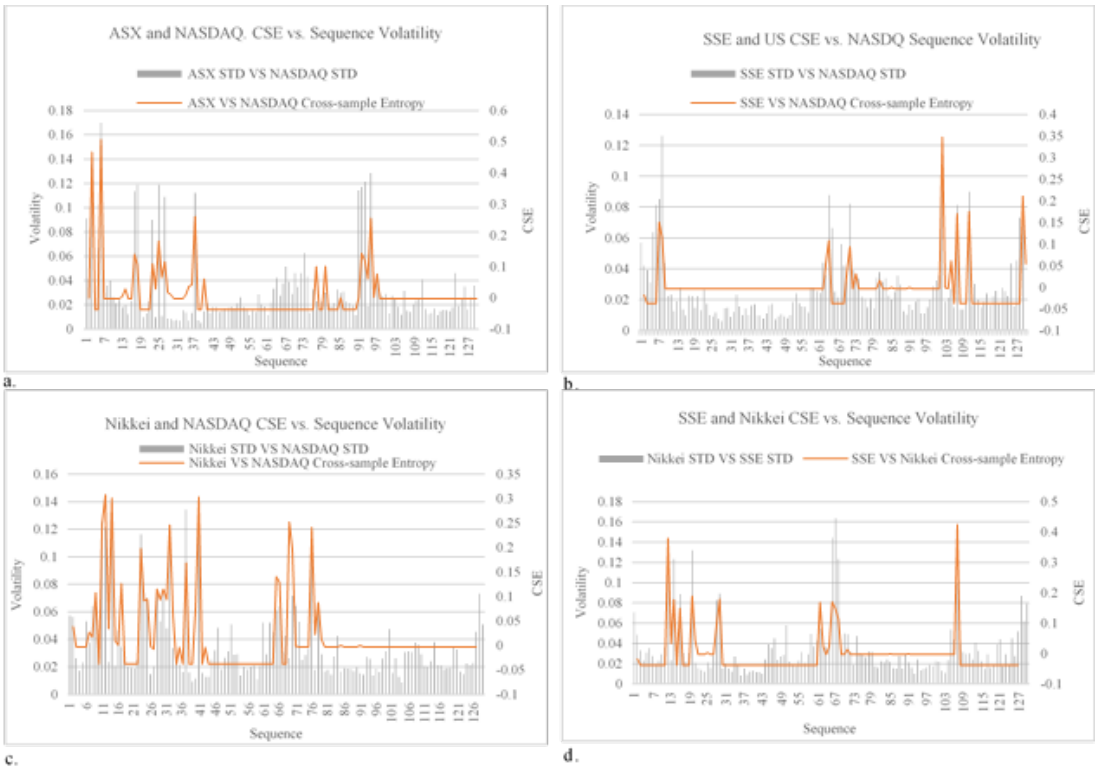


FIGURE A.8: CSE and Sequence Volatility

we compute the volatility (some details are provided in Fig. A.8 and are further explained later) of the sequences to prove that volatility is the cause of the high dissimilarity between the two datasets, but it does not prove that the datasets are uncorrelated. If a sequence has a high CSE value and the data volatility is high, we can conclude that the high CSE is caused by the high volatility of a sequence rather than being contradictory.

Fig. A.8 shows the result. From the Fig. A.8a, b, c, and d, we find the high degree of CSE has a high volatility value. In Fig. A.8c, for instance, sequence 11 to 16, which has a high degree of CSE with high volatility. Fig. A.8a, b reveal the same situation. Therefore, we can conclude that volatility affects the values of CSE. In our experiments, we investigated the contradiction between CSE and SEG. Our results led to the conclusion that the high CSE value is caused by the

high volatility of the sequences. On the other hand, our results also prove that a high dissimilarity value does not indicate a low correlation between two time-series datasets. In other words, over a particular period, two time-series data may be dissimilar, but that does not imply a lack of correlation.

A.5 Summary

The SEG-based method avoids the need to use a tolerance parameter, allowing the method to be adapted to various cases. Moreover, it allows comparison of time-series datasets/signals for specified time segments with quantified correlations, enabling us to observe correlations over different time segments. Our international share market study demonstrated that the method could identify correlations between two time-series datasets/signals validated through previous financial studies. As shown in previous financial studies, the U.S. had the greatest influence on the international stock market [91]. We also compared CSE and SEG. From that comparison, we found that high volatility affects the result of CSE, whereas SEG was able to detect correlations under the same circumstances. At present, the SEG can only analyses correlations between two time-series datasets/signals. Prediction of future trends or correlations based on current observations remains a problem. We intend to undertake further study to develop a predictive model.

Bibliography

[1] Gail A Carpenter and Stephen Grossberg. The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, 21(3):77–88, 1988.

[2] Teuvo Kohonen. The neural phonetic typewriter. *Computer*, 21(3):11–22, 1988.

[3] Alexandros Labrinidis and Hosagrahar V Jagadish. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12):2032–2033, 2012.

[4] Peter C Chang, Alison Flatau, and SC Liu. Health monitoring of civil infrastructure. *Structural health monitoring*, 2(3):257–267, 2003.

[5] Jinping Ou and Hui Li. Structural health monitoring in mainland china: review and future trends. *Structural Health Monitoring*, 9(3):219–231, 2010.

[6] RC Tennyson, AA Mufti, S Rizkalla, G Tadros, and B Benmokrane. Structural health monitoring of innovative bridges in canada with fiber optic sensors. *Smart materials and Structures*, 10(3):560, 2001.

[7] Mousumi Majumder, Tarun Kumar Gangopadhyay, Ashim Kumar Chakraborty, Kamal Dasgupta, and Dipak Kumar Bhattacharya. Fibre

bragg gratings in structural health monitoring—present status and applications. *Sensors and Actuators A: Physical*, 147(1):150–164, 2008.

[8] Manjusha Ramakrishnan, Ginu Rajan, Yuliya Semenova, and Gerald Farrell. Overview of fiber optic sensor technologies for strain/temperature sensing applications in composite materials. *Sensors*, 16(1):99, 2016.

[9] José Miguel López-Higuera, Luis Rodriguez Cobo, Antonio Quintela Incera, and Adolfo Cobo. Fiber optic sensors in structural health monitoring. *Journal of lightwave technology*, 29(4):587–608, 2011.

[10] Hong-Nan Li, Dong-Sheng Li, and Gang-Bing Song. Recent applications of fiber optic sensors to health monitoring in civil engineering. *Engineering structures*, 26(11):1647–1657, 2004.

[11] JM Ko and YQ Ni. Technology developments in structural health monitoring of large-scale bridges. *Engineering structures*, 27(12):1715–1725, 2005.

[12] Xiaoyuan Wei, Yuan Yang, Wenqing Yao, and Lei Zhang. Pspice modeling of a sandwich piezoelectric ceramic ultrasonic transducer in longitudinal vibration. *Sensors*, 17(10):2253, 2017.

[13] Victor Giurgiutiu, Andrei Zagrai, and Jing Jing Bao. Piezoelectric wafer embedded active sensors for aging aircraft structural health monitoring. *Structural Health Monitoring*, 1(1):41–61, 2002.

[14] G Song, H Gu, YL Mo, TTC Hsu, and H Dhonde. Concrete structural health monitoring using embedded piezoceramic transducers. *Smart Materials and Structures*, 16(4):959, 2007.

[15] YC Ma, YH Yang, JM Li, MW Yang, J Tang, and T Liang. Dynamic and static strain gauge using superimposed fiber bragg gratings. *Measurement Science and Technology*, 23(10):105202, 2012.

[16] Yongbo Dai, Yanju Liu, Jinsong Leng, Gang Deng, and Anand Asundi. A novel time-division multiplexing fiber bragg grating sensor interrogator for structural health monitoring. *Optics and Lasers in Engineering*, 47(10): 1028–1033, 2009.

[17] Wei Liang, Yanyi Huang, Yong Xu, Reginald K Lee, and Amnon Yariv. Highly sensitive fiber bragg grating refractive index sensors. *Applied physics letters*, 86(15):151122, 2005.

[18] S Takeda, Y Aoki, T Ishikawa, N Takeda, and H Kikukawa. Structural health monitoring of composite wing structure during durability test. *Composite structures*, 79(1):133–139, 2007.

[19] Tommy HT Chan, Ling Yu, Hwa-Yaw Tam, Yi-Qing Ni, SY Liu, WH Chung, and LK Cheng. Fiber bragg grating sensors for structural health monitoring of tsing ma bridge: Background and experimental observation. *Engineering structures*, 28(5):648–659, 2006.

[20] A Kerrouche, WJO Boyle, T Sun, and KTV Grattan. Design and in-the-field performance evaluation of compact fbg sensor system for structural health monitoring applications. *Sensors and Actuators A: Physical*, 151(2):107–112, 2009.

[21] Zhi Zhou, Thomas W Graver, Luke Hsu, and Jin-ping Ou. Techniques of advanced fbg sensors: Fabrication, demodulation, encapsulation, and their application in the structural health monitoring of bridges. *Pacific Science Review*, 5(1):116–121, 2003.

[22] Mehmet Celebi. Gps in dynamic monitoring of long-period structures. *Soil Dynamics and Earthquake Engineering*, 20(5-8):477–483, 2000.

[23] Alfredo Knecht and Luca Manetti. Using gps in structural health monitoring. In *Smart Structures and Materials 2001: Sensory Phenomena and Measurement Instrumentation for Smart Structures and Materials*, volume 4328, pages 122–130. International Society for Optics and Photonics, 2001.

[24] Mehmet Celebi and Ahmet Sanli. Gps in pioneering dynamic monitoring of long-period structures. *Earthquake Spectra*, 18(1):47–61, 2002.

[25] Ting-Hua Yi, Hong-Nan Li, and Ming Gu. Recent research and applications of gps-based monitoring technology for high-rise structures. *Structural Control and Health Monitoring*, 20(5):649–670, 2013.

[26] Hoon Sohn. Effects of environmental and operational variability on structural health monitoring. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 365(1851): 539–560, 2007.

[27] Shinae Jang, Hongki Jo, Soojin Cho, Kirill Mechitov, Jennifer A Rice, Sung-Han Sim, Hyung-Jo Jung, Chung Bang Yun, Billie F Spencer Jr, and Gul

Agha. Structural health monitoring of a cable-stayed bridge using smart sensor technology: deployment and evaluation. 2010.

[28] YQ Ni, Y Xia, WY Liao, and JM Ko. Technology innovation in developing the structural health monitoring system for guangzhou new tv tower. *Structural Control and Health Monitoring*, 16(1):73–98, 2009.

[29] MM Reda Taha, A Noureldin, JL Lucero, and TJ Baca. Wavelet transform for structural health monitoring: a compendium of uses and features. *Structural Health Monitoring*, 5(3):267–295, 2006.

[30] Buli Xu and Victor Giurgiutiu. Single mode tuning effects on lamb wave time reversal with piezoelectric wafer active sensors for structural health monitoring. *Journal of Nondestructive Evaluation*, 26(2-4):123–134, 2007.

[31] Lingyu Yu and Zhenhua Tian. Lamb wave structural health monitoring using a hybrid pzt-laser vibrometer approach. *Structural Health Monitoring*, 12(5-6):469–483, 2013.

[32] Hoon Sohn, Gyuhae Park, Jeannette R Wait, Nathan P Limback, and Charles R Farrar. Wavelet-based active sensing for delamination detection in composite structures. *Smart Materials and structures*, 13(1):153, 2003.

[33] Xia Zhao, Shenfang Yuan, Zhenhua Yu, Weisong Ye, and Jun Cao. Designing strategy for multi-agent system based large structural health monitoring. *Expert Systems with Applications*, 34(2):1154–1168, 2008.

[34] Xia Zhao, Shenfang Yuan, Hengbao Zhou, Hongbing Sun, and Lei Qiu. An evaluation on the multi-agent system based structural health monitoring for

large scale structures. *Expert Systems with applications*, 36(3):4900–4914, 2009.

[35] Stefan Bosse and Armin Lechleiter. Structural health and load monitoring with material-embedded sensor networks and self-organizing multi-agent systems. *Procedia Technology*, 15:668–690, 2014.

[36] Shenfang Yuan, Xiaosong Lai, Xia Zhao, Xin Xu, and Liang Zhang. Distributed structural health monitoring system based on smart wireless sensor and multi-agent technology. *Smart Materials and Structures*, 15(1):1, 2005.

[37] Kay Smarsly, Kincho H Law, and Dietrich Hartmann. Implementing a multiagent-based self-managing structural health monitoring system on a wind turbine. In *Proceedings of the 2011 NSF Engineering Research and Innovation Conference. Atlanta, GA, USA*, volume 1, page 2011, 2011.

[38] Keith Worden and Graeme Manson. The application of machine learning to structural health monitoring. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 365 (1851):515–537, 2007.

[39] Michael W Vanik, James L Beck, and SK2000 Au. Bayesian probabilistic approach to structural health monitoring. *Journal of Engineering Mechanics*, 126(7):738–745, 2000.

[40] Jianye Ching and James L Beck. New bayesian model updating algorithm applied to a structural health monitoring benchmark. *Structural Health Monitoring*, 3(4):313–332, 2004.

[41] Hoon Sohn, Charles R Farrar, Norman F Hunter, and Keith Worden. Structural health monitoring using statistical pattern recognition techniques. *Journal of dynamic systems, measurement, and control*, 123(4):706–711, 2001.

[42] Yeesock Kim, Jo Woon Chong, Ki H Chon, and JungMi Kim. Wavelet-based ar–svm for health monitoring of smart structures. *Smart Materials and Structures*, 22(1):015003, 2012.

[43] Vicente Lopes Jr, Gyuhae Park, Harley H Cudney, and Daniel J Inman. Impedance-based structural health monitoring with artificial neural networks. *Journal of Intelligent Material Systems and Structures*, 11(3):206–214, 2000.

[44] Jiyoung Min, Seunghee Park, Chung-Bang Yun, Chang-Geun Lee, and Changgil Lee. Impedance-based structural health monitoring incorporating neural network technique for identification of damage type and severity. *Engineering Structures*, 39:210–220, 2012.

[45] Young-Jin Cha, Wooram Choi, and Oral Büyüköztürk. Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, 32(5):361–378, 2017.

[46] Mathieu Dumoulin. Real-time anomaly detection streaming microservices with h2o and mapr – part 2: Modeling. URL https://mapr.com/blog/real-time-anomaly-detection-2/. [Online; accessed April 27, 2018].

[47] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[48] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pages 77–101. Springer, 2002.

[49] Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 15–27. Springer, 2002.

[50] Ji Zhang and Hai Wang. Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance. *Knowledge and information systems*, 10(3):333–355, 2006.

[51] Edwin M Knox and Raymond T Ng. Algorithms for mining distancebased outliers in large datasets. In *Proceedings of the International Conference on Very Large Data Bases*, pages 392–403. Citeseer, 1998.

[52] Edwin M Knorr and Raymond T Ng. Finding intensional knowledge of distance-based outliers. In *VLDB*, volume 99, pages 211–222, 1999.

[53] Edwin M Knorr, Raymond T Ng, and Vladimir Tucakov. Distance-based outliers: algorithms and applications. *The VLDB Journal—The International Journal on Very Large Data Bases*, 8(3-4):237–253, 2000.

[54] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *ACM Sigmod Record*, volume 29, pages 427–438. ACM, 2000.

[55] Mingxi Wu and Christopher Jermaine. Outlier detection by sampling with accuracy guarantees. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 767–772. ACM, 2006.

[56] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[57] Wen Jin, Anthony KH Tung, and Jiawei Han. Mining top-n local outliers in large databases. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 293–298. ACM, 2001.

[58] Anny Lai-mei Chiu and Ada Wai-chee Fu. Enhancements on local outlier detection. In *Database Engineering and Applications Symposium, 2003. Proceedings. Seventh International*, pages 298–307. IEEE, 2003.

[59] Ville Hautamaki, Ismo Karkkainen, and Pasi Franti. Outlier detection using k-nearest neighbour graph. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 430–433. IEEE, 2004.

[60] Tsuyoshi Idé, Spiros Papadimitriou, and Michail Vlachos. Computing correlation anomaly scores using stochastic nearest neighbors. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 523–528. IEEE, 2007.

[61] Gunnar Ratsch, Sebastian Mika, Bernhard Scholkopf, and K-R Muller. Constructing boosting algorithms from svms: an application to one-class classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1184–1199, 2002.

[62] Volker Roth. Outlier detection with one-class kernel fisher discriminants. In *Advances in Neural Information Processing Systems*, pages 1169–1176, 2005.

[63] Claudio De Stefano, Carlo Sansone, and Mario Vento. To reject or not to reject: that is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(1):84–94, 2000.

[64] Abdallah Abbey Sebyala, Temitope Olukemi, Lionel Sacks, and Dr Lionel Sacks. Active platform security through intrusion detection using naive bayesian network for anomaly detection. In *London Communications Symposium*. Citeseer, 2002.

[65] Daniel Barbara, Ningning Wu, and Sushil Jajodia. Detecting novel network intrusions using bayes estimators. In *Proceedings of the 2001 SIAM International Conference on Data Mining*, pages 1–17. SIAM, 2001.

[66] Mary M Moya, Mark W Koch, and Larry D Hostetler. One-class classifier networks for target recognition applications. Technical report, Sandia National Labs., Albuquerque, NM (United States), 1993.

[67] Dipankar Dasgupta and Fernando Nino. A comparison of negative and positive selection algorithms in novel pattern detection. In *Systems, man, and cybernetics, 2000 IEEE international conference on*, volume 1, pages 125–130. IEEE, 2000.

[68] Simon Hawkins, Hongxing He, Graham Williams, and Rohan Baxter. Outlier detection using replicator neural networks. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 170–180. Springer, 2002.

[69] Datavisor. Unsupervised analytics: Moving beyond rules engines and learning models. URL https://www.datavisor.com/technical-posts/rules-engines-learning-models-and-beyond/. [Online; accessed April 27, 2018].

[70] Gerhard Münz, Sa Li, and Georg Carle. Traffic anomaly detection using k-means clustering. In *GI/ITG Workshop MMBnet*, 2007.

[71] Rasheda Smith, Alan Bivens, Mark Embrechts, Chandrika Palagiri, and Boleslaw Szymanski. Clustering approaches for anomaly based intrusion detection. *Proceedings of intelligent engineering systems through artificial neural networks*, pages 579–584, 2002.

[72] Zengyou He, Shengchun Deng, and Xiaofei Xu. Outlier detection integrating semantic knowledge. In *International Conference on Web-Age Information Management*, pages 126–131. Springer, 2002.

[73] Ningning Wu and Jing Zhang. Factor analysis based anomaly detection. In *Information Assurance Workshop, 2003. IEEE Systems, Man and Cybernetics Society*, pages 108–115. IEEE, 2003.

[74] Ana M Pires and Carla Santos-Pereira. Using clustering and robust estimators to detect outliers in multivariate data. In *Proceedings of the International Conference on Robust Statistics*, 2005.

[75] Dantong Yu, Gholamhosein Sheikholeslami, and Aidong Zhang. Findout: finding outliers in very large datasets. *Knowledge and Information Systems*, 4(4):387–412, 2002.

[76] Tom A Kuusela, Tuomas T Jartti, Kari UO Tahvanainen, and Timo J Kaila. Nonlinear methods of biosignal analysis in assessing terbutaline-induced heart rate and blood pressure changes. *American Journal of Physiology-Heart and Circulatory Physiology*, 282(2):H773–H781, 2002.

[77] Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000.

[78] Wenbin Shi and Pengjian Shang. Cross-sample entropy statistic as a measure of synchronism and cross-correlation of stock markets. *Nonlinear Dynamics*, 71(3):539–554, 2013.

[79] L Kullmann, Janos Kertész, and K Kaski. Time-dependent cross-correlations between different stock returns: A directed network of influence. *Physical Review E*, 66(2):026125, 2002.

[80] Takayuki Mizuno, Hideki Takayasu, and Misako Takayasu. Correlation networks among currencies. *Physica A: Statistical Mechanics and its Applications*, 364:336–342, 2006.

[81] Douglas E Lake, Joshua S Richman, M Pamela Griffin, and J Randall Moorman. Sample entropy analysis of neonatal heart rate variability. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 283(3):R789–R797, 2002.

[82] Thomas Conlon, Heather J Ruskin, and Martin Crane. Cross-correlation dynamics in financial time series. *Physica A: Statistical Mechanics and its Applications*, 388(5):705–714, 2009.

[83] Steven M Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

[84] Steve Pincus and Burton H Singer. Randomness and degrees of irregularity. *Proceedings of the National Academy of Sciences*, 93(5):2083–2088, 1996.

[85] Peter Grassberger and Itamar Procaccia. Estimation of the kolmogorov entropy from a chaotic signal. *Physical review A*, 28(4):2591, 1983.

[86] J-P Eckmann and David Ruelle. Ergodic theory of chaos and strange attractors. In *The Theory of Chaotic Attractors*, pages 273–312. Springer, 1985.

[87] Steven M Pincus and Ary L Goldberger. Physiological time-series analysis: what does regularity quantify? *American Journal of Physiology-Heart and Circulatory Physiology*, 266(4):H1643–H1656, 1994.

[88] Steven M Pincus, Thomas Mulligan, Ali Iranmanesh, Sylvia Gheorghiu, Michael Godschalk, and Johannes D Veldhuis. Older males secrete luteinizing hormone and testosterone more irregularly, and jointly more asynchronously, than younger males. *Proceedings of the National Academy of Sciences*, 93(24):14100–14105, 1996.

[89] Li-Zhi Liu, Xi-Yuan Qian, and Heng-Yao Lu. Cross-sample entropy of foreign exchange time series. *Physica A: Statistical Mechanics and its Applications*, 389(21):4785–4792, 2010.

[90] Yahoo. Yahoo finance, 2016. URL https://au.finance.yahoo.com/.

[91] Julie Bishop. Australia and the united states in the 21st century. URL https://foreignminister.gov.au/speeches/Pages/2013/jb_sp_131122.aspx?w=tb1CaGpkPX%2FlS0K%2Bg9ZKEg%3D%3D. [Online; accessed May 20, 2018].

[92] Murray Hunter. Who owns corporate australia? URL https://independentaustralia.net/business/business-display/who-owns-corporate-australia,5033. [Online; accessed May 20, 2018].

[93] Charu C Aggarwal. Outlier ensembles: position paper. *ACM SIGKDD Explorations Newsletter*, 14(2):49–58, 2013.

[94] Xuan Hong Dang, Vincent C.S. Lee, Wee Keong Ng, Arridhana Ciptadi, and Kok-Leong Ong. An em-based algorithm for clustering data streams in sliding windows. In *DASFAA*, pages 230–235. Springer, 2009.

[95] Mahsa Salehi, Christopher A Leckie, Masud Moshtaghi, and Tharshan Vaithianathan. A relevance weighted ensemble model for anomaly detection in switching data streams. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 461–473. Springer, 2014.

[96] Edwin M Knox and Raymond T Ng. Algorithms for mining distance based outliers in large datasets. In *Proceedings of the International Conference on Very Large Data Bases*, pages 392–403. Citeseer, 1998.

[97] Dragoljub Pokrajac, Aleksandar Lazarevic, and Longin Jan Latecki. Incremental local outlier detection for data streams. In *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on*, pages 504–515. IEEE, 2007.

[98] Hoang Vu Nguyen, Vivekanand Gopalkrishnan, and Praneeth Namburi. Online outlier detection based on relative neighbourhood dissimilarity. In *WISE*, pages 50–61. Springer, 2008.

[99] Shiblee Sadik and Le Gruenwald. Online outlier detection for data streams. In *Proceedings of the 15th Symposium on International Database Engineering and Applications*, pages 88–96. ACM, 2011.

[100] A. Daneshgar, R. Javadi, and S.B. Shariat Razavi. Clustering and outlier detection using isoperimetric number of trees. *Pattern Recognition*, 46(12): 3371 – 3382, 2013. ISSN 0031-3203.

[101] S.M. Guo, L.C. Chen, and J.S.H. Tsai. A boundary method for outlier detection based on support vector domain description. *Pattern Recognition*, 42(1):77 – 83, 2009. ISSN 0031-3203.

[102] Hossein Moradi Koupaie, Suhaimi Ibrahim, and Javad Hosseinkhani. Outlier detection in stream data by clustering method. *International Journal of Advanced Computer Science and Information Technology*, 2(3):25–34, 2014.

[103] Daniel Barbará, Yi Li, Julia Couto, Jia-Ling Lin, and Sushil Jajodia. Bootstrapping a data mining intrusion detection system. In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 421–425. ACM, 2003.

[104] Emmanuel Müller, Matthias Schiffer, and Thomas Seidl. Statistical selection of relevant subspace projections for outlier ranking. In *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*, pages 434–445. IEEE, 2011.

[105] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 413–422. IEEE, 2008.

[106] Nan Liu and Han Wang. Ensemble based extreme learning machine. *IEEE Signal Processing Letters*, 17(8):754–757, 2010.

[107] Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 157–166. ACM, 2005.

[108] Sanghamitra Bandyopadhyay and Santanu Santra. A genetic approach for efficient outlier detection in projected space. *Pattern Recognition*, 41:1338 – 1349, 2008. ISSN 0031-3203.

[109] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[110] Murray Rosenblatt et al. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[111] Vassiliy A Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, 14(1):153–158, 1969.

[112] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley and Sons, 2015.

[113] Byeong U Park and James S Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85(409):66–72, 1990.

[114] Z. Ouyang, X. Sun, J. Chen, D. Yue, and T. Zhang. Multi-view stacking ensemble for power consumption anomaly detection in the context

of industrial internet of things. *IEEE Access*, PP(99):1–1, 2018. doi: 10.1109/ACCESS.2018.2805908.

[115] Y. Cao, Y. Li, S. Coleman, A. Belatreche, and T. M. McGinnity. Adaptive hidden markov model with anomaly states for price manipulation detection. *IEEE Transactions on Neural Networks and Learning Systems*, 26(2):318–330, Feb 2015. ISSN 2162-237X. doi: 10.1109/TNNLS.2014.2315042.

[116] S. T. Sarasamma and Q. A. Zhu. Min-max hyperellipsoidal clustering for anomaly detection in network security. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(4):887–901, Aug 2006. ISSN 1083-4419. doi: 10.1109/TSMCB.2006.870629.

[117] Jonathan A Silva, Elaine R Faria, Rodrigo C Barros, Eduardo R Hruschka, Andre CPLF De Carvalho, and João Gama. Data stream clustering: A survey. *ACM Computing Surveys (CSUR)*, 46(1):13, 2013.

[118] Peter C Chang, Alison Flatau, and SC Liu. Review paper: health monitoring of civil infrastructure. *Structural health monitoring*, 2(3):257–267, 2003.

[119] Hoon Sohn, Charles R Farrar, Francois M Hemez, Devin D Shunk, Daniel W Stinemates, Brett R Nadler, and Jerry J Czarnecki. A review of structural health monitoring literature: 1996-2001, 2004.

[120] Xiaolei Li and Jiawei Han. Mining approximate top-k subspace anomalies in multi-dimensional time-series data. In *Proceedings of the 33rd international conference on Very large data bases*, pages 447–458. VLDB Endowment, 2007.

[121] Manoj Rameshchandra Thakur and Sugata Sanyal. A multi-dimensional approach towards intrusion detection system. *arXiv preprint arXiv:1205.2340*, 2012.

[122] Alireza Vahdatpour and Majid Sarrafzadeh. Unsupervised discovery of abnormal activity occurrences in multi-dimensional time series, with applications in wearable systems. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 641–652. SIAM, 2010.

[123] H. Ozkan, O. S. Pelvan, and S. S. Kozat. Data imputation through the identification of local anomalies. *IEEE Transactions on Neural Networks and Learning Systems*, 26(10):2381–2395, Oct 2015. ISSN 2162-237X. doi: 10.1109/TNNLS.2014.2382606.

[124] K. J. Hsiao, K. S. Xu, J. Calder, and A. O. Hero. Multicriteria similarity-based anomaly detection using pareto depth analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6):1307–1321, June 2016. ISSN 2162-237X. doi: 10.1109/TNNLS.2015.2466686.

[125] Shebuti Rayana. Odds library, 2016. URL http://odds.cs.stonybrook.edu.

[126] Peter C Chang, Alison Flatau, and SC Liu. Review paper: health monitoring of civil infrastructure. *Structural health monitoring*, 2(3):257–267, 2003.

[127] Shao-Fei Jiang, Da-bao Fu, Chun-Ming Hu, and Zhao-Qi Wu. Damage identification of concrete-filled steel tubular arch bridge using data fusion

based on information allocation theory. *Procedia Engineering*, 15:1705–1710, 2011.

[128] Hoon Sohn, Charles R Farrar, Norman F Hunter, and Keith Worden. Structural health monitoring using statistical pattern recognition techniques. *Journal of dynamic systems, measurement, and control*, 123(4):706–711, 2001.

[129] Steve Lohr. The age of big data. *New York Times*, 11, 2012.

[130] Jong-Woong Park, Sung-Han Sim, and Hyung-Jo Jung. Wireless displacement sensing system for bridges using multi-sensor fusion. *Smart Materials and Structures*, 23(4):045022, 2014.

[131] Hongki Jo and BF Spencer. Multi-metric model-based structural health monitoring. In *SPIE Smart Structures and Materials+ Nondestructive Evaluation and Health Monitoring*, pages 90611F–90611F. International Society for Optics and Photonics, 2014.

[132] Daniele Zonta, Federico Bruschetta, Riccardo Zandonini, Matteo Pozzi, Ming Wang, Yang Zhao, Daniele Inaudi, Daniele Posenato, and Branko Glisic. Analysis of monitoring data from cable-stayed bridge using sensor fusion techniques. In *SPIE Smart Structures and Materials+ Nondestructive Evaluation and Health Monitoring*, pages 869229–869229. International Society for Optics and Photonics, 2013.

[133] Xia Zhao, Shenfang Yuan, Hengbao Zhou, Hongbing Sun, and Lei Qiu. An evaluation on the multi-agent system based structural health monitoring for

large scale structures. *Expert Systems with applications*, 36(3):4900–4914, 2009.

[134] Michael Negnevitsky. *Artificial intelligence: a guide to intelligent systems*. Pearson Education, 2005.

[135] Bala Arshanapalli and John Doukas. International stock market linkages: Evidence from the pre-and post-october 1987 period. *Journal of Banking & Finance*, 17(1):193–208, 1993.

[136] John Wei-Shan Hu, Mei-Yuan Chen, Robert CW Fok, and Bwo-Nung Huang. Causality in volatility and volatility spillover effects between us, japan and four equity markets in the south china growth triangular. *Journal of International Financial Markets, Institutions and Money*, 7(4):351–367, 1997.

[137] Thomas C Chiang and Dazhi Zheng. An empirical analysis of herd behavior in global stock markets. *Journal of Banking & Finance*, 34(8):1911–1921, 2010.

[138] Robert Johnson and Luc Soenen. Asian economic integration and stock market comovement. *Journal of Financial Research*, 25(1):141–157, 2002.

[139] Junmei Fan. Experts: China little affected by us financial crisis. *China. org. cn*, pages 2008–09, 2008.

[140] Linyue Li, Thomas D Willett, and Nan Zhang. The effects of the global financial crisis on china's financial market and macroeconomy. *Economics Research International*, 2012.