University of Cincinnati

Date: 6/20/2013

I, Eric Guang Jye Huang M.S., hereby submit this original work as part of the requirements for the degree of Master of Science in Mechanical Engineering.

It is entitled:

An Improved Fault Detection Methodology for Semiconductor Applications Based on Multi-regime Identification

Student's name: **Eric Guang Jye Huang M.S.**

This work and its defense approved by:

Committee chair: Jay Lee, Ph.D.

Committee member: Yan Chen, PhD

Committee member: David Thompson, Ph.D.

UNIVERSITY OF Cincinnati

6707

An Improved Fault Detection Methodology for Semiconductor Applications Based On Multi-regime Identification

A thesis submitted to the Division of Research
and Advanced Studies of
the University of Cincinnati

Master of Science

In the Department of Mechanical Engineering
of the College of Engineering

2013

by

Eric Huang

B.E. in Mechanical Engineering, Shanghai Jiaotong University (2011)

Committee Chair: Dr. Jay Lee
Committee Member: Dr. David Thompson
Committee Member: Dr. Yan Chen

ABSTRACT

As the technology trends moving forward rapidly in semiconductor manufacturing industry, the importance of prognostics and health management cannot be neglected. Any kind of failure happens during the manufacturing process will cause huge lost of the profit. The traditional human inspection and experience of detecting operating faults is obsolescent because more and more signals are used to control the manufacturing process in semiconductor industry to fit the requirement of product which will make the failure definition becoming more complicated.

Condition Based Monitoring enabled prognostics have been widely accepted by many industries. However, in real deployment, equipment or process fault detection accuracy is still a big challenge. From data-driven modeling point of view, the loss of accuracy comes from several aspects including data quality, individual equipment behavior variation, external input material variation, environment difference, operation condition and even modeling inaccuracy.

Many researches focus on applying new algorithm or improving existing methods to extract information from the data and detect failure of equipment. They made great breakthrough and contribution on improving the fault detection algorithm calculation efficient and accuracy. However, sometimes the low accuracy of fault detection result is because of the data characteristic instead of the algorithm itself. For example, recipe change will affect machine operating status to cause shift and drift in collected signals, which is called multiple regimes. Every regime is one kind of class which contains its specific characteristic. With multiple regimes identification, uncorrelated cycle data can be separated to different groups to avoid the confusion. Considering the learning and classification ability of SOM, it will be applied to

identify multiple regimes. By learning each regime’s pattern, SOM can classify different regimes to reduce the impact of data shift and drift.

The key development in this research is to improve fault detection method based on multiple regimes identification. Three one-class fault detection methods PCA-MSPC, FD-kNN and 1-SVM will be applied in each regime respectively. Due to the reason that the operation will always be aborted immediately when an error is detected, so the quantity of faulty data is usually limited. In order to deal with this issue, one class fault detection method which only needs normal condition data will be applied in this research. Besides, in order to handle different data characteristic, three fault detection methods are applied in the system as comparison.

In the semiconductor manufacturing process case study given in this work, the one-class fault detection system based on multiple regimes identification, named local model, showed superior performance than global model. The faults detected rate is enhanced up to 40% by using local-based fault detection. In this case study, a number of practical concerns were considered including data quantity limitation and multiple regimes issue, and the fault detection comparison result between local and global model for all three methods will also be given.

ACKNOWLEDGMENTS

It is really my honor to work in Center for Intelligence Maintenance Systems under Professor Jay Lee's guidance and support. I would also like to thank all the other members severed my master thesis committee, Dr. David Thompson and Dr. Yan Chen. It has been a great pleasure having our staffs Mr. Patrick Brown and Mr. Michael Lyons to help us arranging all the daily rounds and schedule, and they also help me a lot on my English improvement. Moreover, this work cannot be accomplished without the cooperation and help of my colleagues in the IMS center, including Dr. Edzel Lapira, Mr. David Siegel, Mr. Wenyu Zhao, Mr. Mohammad Rezvani, Ms. Wenjing Jin, Mr. Shanghu Yang, Mr. Hossein Davari Ardakani, Ms. Xiaorui Tong Ms. Christina Lucas, Mr. Shuo Chang, Mr. Chao Jin, Mr. Dustin Brisset, Mr. Zhe Shi, Ms. Hung-An Kao, Mr. Zongchang Liu, Mr. Behrad Baqheri, Ms. Meihui Wang, Ms. Fally Titikpina, Dr. Lei Zhang, Dr. Jiangtian Yang, Dr. Hongmei Liu, Dr. Chein Chung Su and all of my friends who offer me a great place to study and support me to focus my research work. Last but not least, I do appreciate the love and support of my big brother Xi-Jie and my parents Wen-Bo and Dai-Jun. They are the reason make me strong and successful.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background

In the past, semiconductor practice etching and deposition process rely on empiric operation to guarantee availability with limited understanding of physics and chemistry [1]. Production yield and profit are difficult to be improved. However, as the state of the art develops, semiconductor industry has huge needs of manufacturing production process information and monitoring control for enhancing yield and profit to satisfy costumer's demand. As a matter of fact, in current semiconductor industry, well developed monitoring systems are able to collect a large amount of process data, from academia, many researches focus on applying algorithms to extract information from the data and detect failure of equipment ahead.

Otherwise, over the past few years, the wafer diameter increases dramatically from 8 in to 12 in (200 mm to 300 mm) and the thickness shrinks to below 1 mm because of the development of semiconductor manufacturing technique [2]. As the result, the high aspect ratio wafers caused huge pressure on productivity because the manufacturing process becomes much more complex which will increase manufacturing cycle time [3]. Based on the requirement of information for decision making to operate production line more efficiently [4], an appropriate maintenance schedule to increase overall equipment effectiveness (OEE) is the key issue in modern semiconductor industry. Most semiconductor manufacturers invest great amount of money trying to figure out the method of improvement of the OEE during manufacturing operation [5]. ERP (Enterprise Resource Planning) and MES (Manufacturing Execution System) are the two popular systems helping companies to schedule labor hour, production demand, and system maintenance to increase profit and reduce downtime [6, 7]. In order to calculate the true OEE by implementing decision making system, real-time data collection is necessary to prove the reliability and accuracy.

Semiconductor manufacturing process consists of different complex steps shown in Figure 1.1. Within those sophisticate procedures, etching is one of the most important process which will significantly affect the wafer yield [8]. Presently, OEE based decision making systems made tremendous effort on improvement of production process. However, it is still insufficient and hard to really increase manufacturing efficiency, product quality, and yield by scheduling fixed cycle maintenance since that will be wasting lots of time and money [9]. Fixed cycle maintenance cannot detect sudden defects which will cause unexpected downtime, and, furthermore, it costs much more than the actual necessity to inspect machines with little requirement. Due to the reason mentioned above, plenty of researchers developed different categories algorithms, such us statistical and data-driven, of fault detection to achieve preventive maintenance to reduce downtime.

The identity of those methods should be capable of handling high-dimension, noisy, and highly correlated data [10]. Statistical algorithm is a popular method which is widely used and developed for fault detection implementation in many different areas for over ten years, but the result sometimes is deficient when analyzing the data without specific characteristics [11]. Under this certain circumstance, the fault detection accuracy is low and will be missing many critical failure modes, which will cause losing tremendous profit due to late maintenance. Data-driven based method will be more accurate in some cases because of the ability of process learning, but the process is complex and needs huge data to train and update the model which decreases the analysis speed. The ratio of accuracy and non-delay monitoring needs to be balanced for yield and profit improvement.

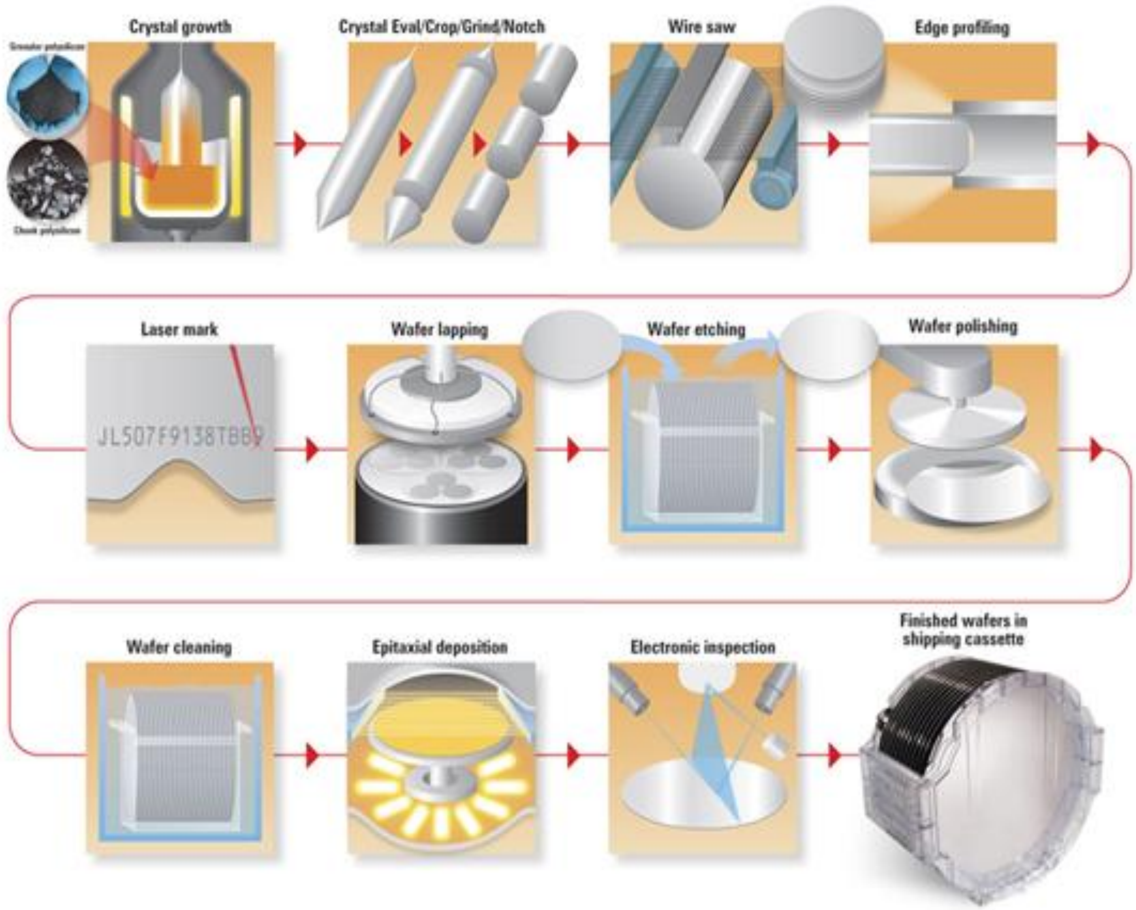


Figure 1.1 Semiconductor Wafer Manufacturing Process[12]

1.2 Motivation

In practice, due to the fact that failure modes will be detected in a short time after error happens, so data quantity and quality is limited regardless of the sensors sensitivity and flexibility. Only a short period of anomaly data can be collected by data acquisition (DAQ) system because of the machine emergency shutdown. This strategy for guaranteeing constant wafer quality and supply causes data availability limitation. Insufficient anomaly data is difficult to be used for model training by many fault detection algorithms. The multi-class based fault detection algorithm needs as many as faulty data for model training to learn the failure pattern for result accuracy improvement. Those methods are not applicable for practical purpose because of the limitation.

Besides, some key factors, such as machine working condition shift, data variance, and time length inconsistence caused by recipe change and different materials during manufacturing process, will affect the performance of failure detection in real semiconductor industry. In order to solve this issue, more efforts are made to improve the detection accuracy by applying different techniques. Some data pre-processing methods, such as frequency filter, dynamic time warping (DTW) and trajectory mean shift, etc, are applied before fault detection modeling to improve algorithms precision [13, 14], and the result showed more accurate and reliable. However, most of these data pre-processing methods require human intervention [15, 16], which will make automation process more difficult. A comprehensive method embedded in fault detection system for real-time process monitoring and better monitoring accuracy will be future research main goal to gain high profit and manufacturing efficiency. Traditionally, automation machine health monitoring can only be applied precisely based on engineer's experience to process the raw signal under most of the conditions. As the issue mentioned previously, human intervention based data pre-processing is the gap of achieving automatic process monitoring.

Imbalanced data and experiment noise which impact algorithms performance are gradually increasing issues [17], so in order to deal with the problem, feature selection will be a apparent need before training the algorithm model for fault detection . For example, in bioinformatics area, data imbalanced is a huge challenge because of the large dimensionality and small sample size. Hundreds of thousands of genes' characteristic are collected as features, but only small number of samples comparing to features are used for analysis [18]. Feature selection for microarray analysis becomes more important to solve this situation. Likewise, data collected from real world semiconductor manufacturing production line has large quantity [19], wherefore, feature selection for dimension reduction becomes more significant as the result of increasing measurements of process variables, like pressure, temperature, voltage, current, power, flow rate, etc, controlled and monitored to satisfy sophisticated and meticulous manufacturing process [20].

Last but not least, signal variance based on manufacturing process condition shift and drift sometimes mislead algorithms detecting failure modes. In wafer etching process, for example, when TCP power is re-adjusted for different process recipe and strategy, the change in the metric will be greater than the normal variance in the controlled parameters [21]. The distance between same signal distribution caused by machine condition change which has multi-regime characteristic is confusing for fault detection model to distinguish the category of abnormal signal [22]. Multiple regime detection and classification will be the key factor to enhance the accuracy of fault detection method.

Existing approaches are limited on precision of fault detection due to the fact of data availability limitation, non-automatic data pre-processing, large data and feature quantity, and multi-regime issue, so this research focuses on reducing the impact of these issues to achieve semiconductor manufacturing near-zero downtime.

1.3 Research Objectives and Tasks

Recent research in semiconductor industry more focuses on comparing results from different algorithms instead of data characteristic itself, so developing data-driven methods to differentiate data variance caused by equipment degradation and process drifting, and then identify equipment degradation by a Global-Local Classification method based on one class fault detection algorithm is the objective of the paper.

This research aims to improve fault detection precision by differentiation of data pattern variation caused by changing condition and system degradation. As to reach this research goal, the following research task will be accomplished:

- Task 1: Build a accurate local-based fault detection methodology based on multiple regimes identification
- Task 2: Investigate data pattern characteristics, which cause by machine operation condition changing and degradation respectively for semiconductor process
- Task 3: Develop methods to identify/calculate the data characteristics and extract potential features and apply dimension reduction method to select useful features for model training
- Task 4: Apply one-class approach for precise fault detection to handle limited faulty data issue
- Task 5: Benchmarking:
    - Global and local model based fault detection methods
    - PCA, KNN and 1-SVM methods for local models based fault detection

1.4 Thesis Organization

This paper is divided into 6 chapters to explain the research idea and philosophy which will be represented on the area of feature extraction, feature selection, classification for multi-regime detection, fault detection, and result comparison.

Chapter 2 shows the research people have done in semiconductor area. Some previous systematic methods were developed for detecting failure mode by analyzing several different kinds of process data. As a result of different data characteristic, the appropriate method was implemented to deal with different conditions, such as time inconsistent, variance, noise, etc. Also, practical applications are given as the case study for proving algorithms feasibility and accuracy in each papers.

Chapter 3 introduces the overall approach structure and all the proposed algorithms used in this research. A comprehensive understanding of the combination method of multiple regimes identification and fault detection is given in this chapter. The definition of Global-local model based on Self-organizing Map and three proposed algorithms detail information, one-class Support Vector Machine, k-nearest Neighbor and Principle Components Analysis are also included.

Chapter 4 gives a comprehensive description of multiple regimes identification. The main focus in this chapter is introducing the multiple regimes identification based on Self-organizing Map method. Based on the number of recipes run through the machine, SOM can learn the data pattern from training data and classify the coming testing data. This method helps to solve the confusion of regimes and faults. By showing a case study of bearing failure diagnosis, the multiple regime identification capability of SOM is proved.

Chapter 5 introduces the importance of data preparation including de-nosing, statistic calculation, feature selection, and fault detection. Every steps are very significant handling issues buried in data to increase the algorithm performance and transform the raw data to useful information. The detail information of all the algorithms will be given in this chapter.

Chapter 6 gives a semiconductor manufacturing case study using the methods mentioned in previous chapters. Starting from data preparation to model building, the procedures of fault detection system are systematically introduced step by step. The detail information about how to identify multiple regimes and build three different fault detection methods respectively are also given. The final comparison result of three different fault detection algorithms based on global and local model are listed in the table.

Chapter 7 summaries the ideas and finding of the research in this thesis. The future recommendation of the work will be given as well.

2. LITERATURE REVIEW

2.1 Method of Distinguishing Data Pattern for Multiple Regimes and Equipment Failure

Long distance between two data distribution collected from the same signal is often assumed as the failure pattern if one of the data distribution is collected from normal condition [23, 24]. Lots of algorithms detect fault, calculate confidence value and diagnose failure mode based on distribution pattern. However, the pattern sometimes is caused by manufacturing process shift and drift, different incoming materials, process monitoring sensors drifting, etc, instead of error mode [25]. Tool state shift are part of preventive maintenance when engineer replaces parts and clean machines [26]. After the maintenance process, the same signal will show different pattern because of machine condition change, which will be wrongly detected by some fault detection algorithm as an abnormal performance [27, 28]. This pseudo fault condition will affect algorithm's precision, and then unnecessary maintenance is scheduled to waste money. When this type of situation happens frequently, downtime probability will be increasing to reduce yield and profit.

Roughly, there are two types of ways to classify different regimes to avoid the wrong features introduced by recipe change. Some researches focused on Principle Component Analysis (PCA) algorithm trying to build Global-local model to reduce multiple regimes effect in semiconductor industry [29]. Due to the mean shift and covariance in process data collected under different machine condition, PCA can successfully separate the data to several groups, one for each of the experiments [22]. This circumstance shows that model based on all the data will include larger region of multivariable space as normal than single groups model. Other than the direct classification method, regimes diversity can be decreased using PCs, T square and SPE as input features extracted by PCA algorithm [30, 31]. Because of the characteristic of PCs, fault detection algorithms can avoid the dissimilarity generated by process shift. This method can achieve both of the purposes of distinguishing data pattern for multiple regimes and equipment failure and statistic calculation for feature extraction [32].

Otherwise, as equipment has multiple operation condition, appropriate modeling techniques with the capability of handling condition data with multiple regime will be a very efficient and powerful tool. Self-organizing Maps (SOM) is used to classify different groups is a and is also capable to learn and organize information without being given the corresponding label [33]. SOM will learn the characteristic of input training data and organize neurons with similar features next to each other on the map. After the structure is trained using the baseline, the best matching unit (BMU) will be defined when the new data is inputted. The distance between the new data and BMU, called minimum quantization error (MQE), will be used to access the performance of machines. Large distance means the condition is highly dissimilar from the normal condition, and small distance can be treated as minor failure or healthy condition depends on the threshold. This SOM-MQE shows very good result on wind turbine [34] and alternator application [35].

2.2 Signal Processing Methods in Semi-conductor industry

Data pre-processing is an important link in PHM system for improving the performance of process monitoring algorithm and can have significant impact for both continuous and batch process data [11]. Time length inconsistency, noise, variance, etc, issues are the obstacles of precise fault detection methods. For example, during the wafer manufacturing process, number of product will be changed based on different order. The operating time is varied day by day because of different circumstances. Without solving these problems, the results might vary greatly by using the same algorithm to analyze similar data source.

Recent researches are more implementing Dynamic Time Warping (DTW) and mean shift method to align unequal time sequence which are non-linearly distorted in the time domain to determine the measure of their independent similarity of certain non-linear variations in the time domain [30]. After synchronizing the trajectory of different batch process data to match the reference step trajectory, the trajectory will follow the unimodal distribution which reduces the dissimilarity of two batch process data from the same signal source [13].

Wavelet based analysis, Fourier transform (FT), frequency filter are very popular frequency based methods of signal processing with many applicants in physics and engineering [36, 37]. Process data usually includes noises generated by semiconductor manufacturing procedure factors in facility, so wavelet based analysis are commonly used for de-noising to improve data resolution [38, 39]. Simple high and low pass frequency filter can also be effectively to handle some more apparent multi-sources frequency signal [40]. Without the step of removing noise, training data will include more information than we expect which always disturb the performance of upcoming fault detection process.

2.3 Feature Extraction and Dimension Reduction

Feature extracted from raw signals is a very critical part to ensure the robustness and accuracy of fault detection algorithm result. However, the impact of features vary depending on the data characteristic. Most of the papers invest great deal of time and effort on testing different statistical features, such as maximum, minimum, mean, RMS, standard deviation, etc [41]. If the features which can be simply calculated obviously reflect the characteristic of raw data by visualization, then they will satisfy the requirement of fault detection algorithm. Most of the time, not all of the features trend are so obvious to reflect the data pattern or some of the feature will have similar pattern as other variables, so feature selection method needs to be applied to eliminate redundant features.

Feature categorisation can be divided into four sections which are irrelevant, weakly relevant, strongly relevant, and redundant, and the criteria of relevance and redundancy are elaborate from the discriminatory power [42]. Fisher criterion has the nature of calculating the score of each features to show the relation between two classes. The relevancy of features will be defined by the distance between the same two features from normal and faulty data respectively. The number of highest feature score, which means the level of relevancy, will be selected depends on the requirement [43]. PCA is a powerful linear scheme for compressing a set of high dimensional vectors into lower dimensional vectors. PCs of PCA does not only have the ability to be used to classify the data to different groups as previous mentioned but also can be treated as features which already reduced multiple regimes impact for model training [44, 45].

High level of variables monitored by sensors causes huge pressure on computing, so dimension reduction is used to eliminate redundant variables to keep the operation effectiveness [20]. Signals which play the same role or with zero variation as potential features can be removed since they have limited contribution for algorithm.

2.4 Fault Detection Method for Semiconductor industry applications

The interest of fault detection implemented in semiconductor industry is due to the near-zero defects tolerance on wafer production. Each failure should be immediately detected when it happens to avoid huge profit loss and reduce wafer fraction defective. Fault detection method can be divided to two categories which are traditional statistical method and data mining techniques [46].

Traditional statistical analysis method uses statistical properties of a data time series (mean, range, variance, correlation, etc) to characterize the behavior of a system [47]. For distinguishing the pattern of equipment degradation, appropriate variables need to be found by using multivariable statistical analysis, and then the information will be shown after the statistical model is built to set the threshold and plot the result on 2-D graph. PCA and Statistical Process Control (SPC) are very common methods used in semiconductor research area to improve wafer quality, manufacturing process efficiency and machine performance [22, 40]. A false alarm will be set to detect "out of control" signals to control the manufacturing process and monitor machinery condition [48].

Data-driven method can also be roughly categorized to supervised and unsupervised learning methods based on whether the data being labeled or not. Supervised learning is a machine learning task of building a function from labeled training data set. The data for training purpose consist of a vector of input data and the corresponding output value which is used to label the input data for different categories. After the supervised learning method analyzes the training data and produces a inferred function which can be a regression model or classifier, this learning model should predict the right output value for the testing data based on the previous training process. For example, during the learning process of supervised machine learning based for fault detection, such as Neural Network (NN), and Bayesian-based analysis, parameters will be adjusted to optimize the model in most cases [49, 50]. This adaptive system is more flexible to fit all different situations' requirement. For other purpose like fault diagnosis, Decision Tree will be a good method for classifying different failure modes [51]. By changing the model structure during the learning process, the result will be more precise and reliable.

On the contrary, unsupervised learning is trying to reveal the hidden layer from the unlabeled training data. Since all the examples are unlabeled, no value will be given to evaluate the potential results. However, from another point of view, part of the unsupervised learning methods can be implemented under much more situations in real world industry based on this characteristic, because there is no need to understand faulty pattern before training the model for this type of algorithms. One class SVM (1-SVM), K-Nearest-Neighbor (KNN), and Support Vector Data Description (SVDD) [52] are developed for the purpose of solving insufficient failure data issue [53]. Normal data will be classified as a group and projected onto 2-D plot, so the baseline of normal condition can be decided based on the percentage of normal data being included under the threshold to detect outliers. Due to the reason that only normal data is necessary for modeling, these methods will be very helpful under the limited anomaly data situation.

2.5 Research Gaps

Fault detection method is very critical to be developed and implemented in real world semiconductor manufacturing to improve wafers' quality and yield. Most of the researches made big contribution on this area by enhancing algorithm accuracy and developing novel methods. However, some of the points they didn't consider comprehensively.

- More and more novel methods are developing every year to solve different kinds of problem to improve the fault detection algorithm performance, but the real time monitoring capability will be a more critical part in practice.
- Existing algorithms improvement and novel methods development are very important, but they ignored the practical questions that how these techniques are applied to real world semiconductor manufacturing environment.

Further significant works will be done to solve the problems mentioned above by comparing the benefit of implementing one class based global and local model.

3. OVERVIEW OF LOCAL-BASED FAULT DETECTION SYSTEM

3.1 Overview

Real-time fault detection method is useful for previous finding of system failure to avoid the yield lost and low quality. However, in reality, most of the data collected from system includes information of machine operating condition change which will increase the difficulty of fault detection. For example, in the machining process, different operation strategies are applied to fulfill the requirement of different orders. Machine condition, materials, and environment could be the facts that cause different data pattern which is called multiple regimes. This multiple regimes nature included in collected data will confuse the fault detection algorithm because of the distribution distance generated by different manufacturing strategies instead of system failures. Without distinguishing those characteristic of the regimes hidden in data before applying fault detection method, the detection result might have lower accuracy and be infirm most of the time.

In order to avoid this situation, fault detection model needs to be applied on each regime respectively. PCA is a popular feature selection method which naturally comes with regimes identification ability used by many researches [22, 54]. The data which is projected to 2D graph can show different clusters to identify regimes based on the number of PCs. However, in some cases, learning-based clustering method has better performance on the data with big quantity based on its learning ability [33]. When the available training data quantity increases , the classification result can be improved. Besides, learning-based clustering method can identify the multiple regime issues automatically by learning the data pattern without prior engineer knowledge.

Therefore, the objective of this study is to establish the framework for a local-based fault detection method. During the data preparation process, which will be detail introduced in chapter 4, each specific fault detection model will be built under certain characteristic based on regime identification to fit different requirements. This local-based fault detection model can reduce the interruption caused by operation cycles which have limited relevance to each other.

3.2 Approach Framework

The overall fault detection method approach has two modes: training and testing. Training process is constituted by two critical parts which are data preparation and fault detection model building. The purpose of training is to reduce the impact of data variation and noise buried in the data and also learns the data pattern to build the training model. By doing this, selected features, multiple regime identification model and local-based fault detection model are built for new coming data in the testing process. New data will go through the same procedure as training process using the training features and models and be transformed into visualized information.

During the training process, when the raw data first comes into the system, it will be assumed under normal condition. In this situation, machine needs to be maintained or checked without any kind of failures before doing the training process. After the training dataset is created, segmentation, feature extraction, and feature selection will be applied to pre-process the data. The purpose of data pre-processing is removing noise and outlier, and mining useful information buried in data to increase the method accuracy. The retained features will be used for building multiple regime identification training model to learn the data pattern to distinguish different regimes. After each regime is defined, its nature will be learned as well to prepare for the classification in testing process. Data pre-processing and multiple regime identification constitute the data preparation process to reduce the variance which will affect the performance of fault detection algorithm.

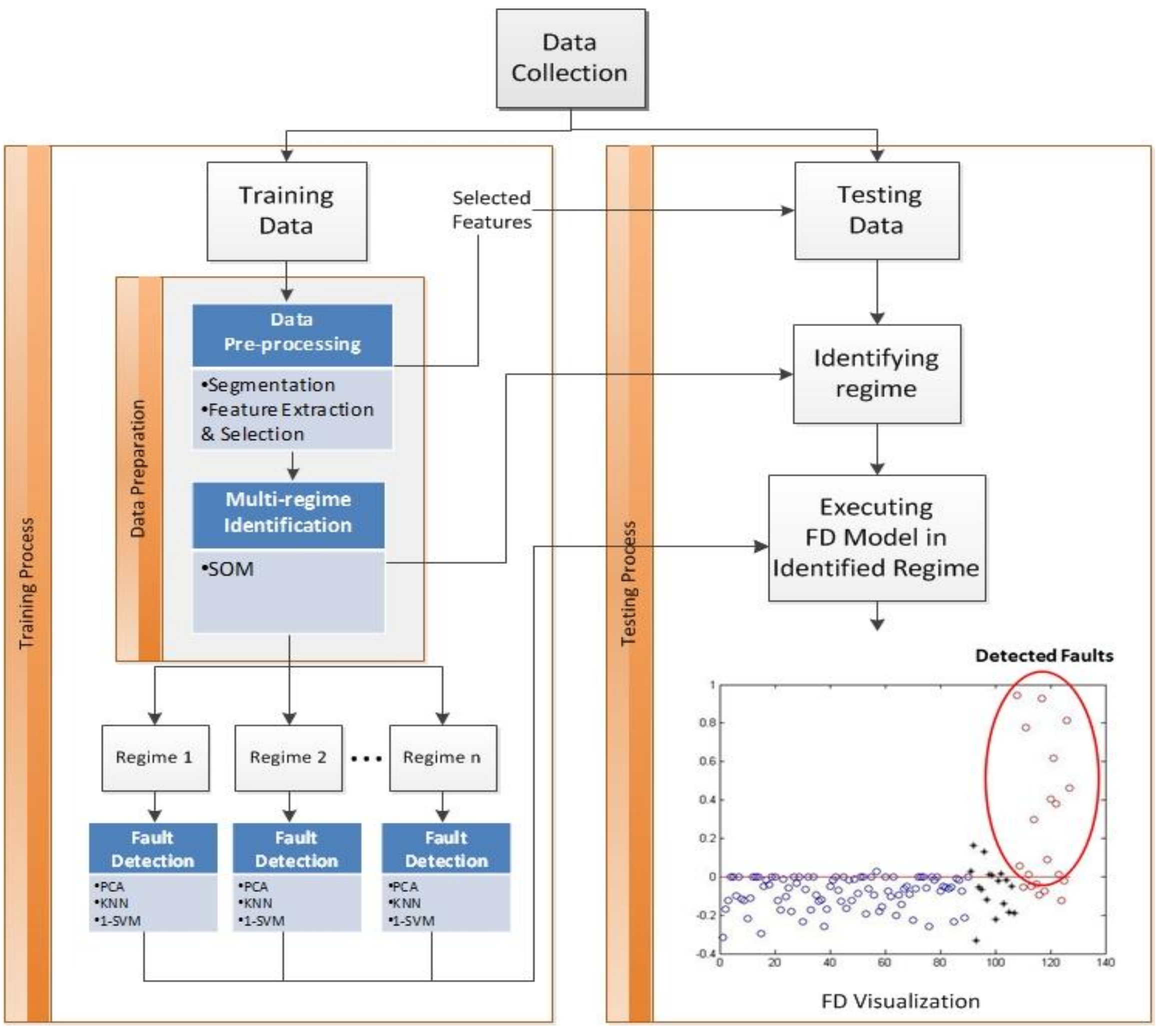


Figure 3.1 Fault Detection System Framework

Once all the regimes are defined, for each regime, three fault detection algorithms Principal Components Analysis, K-nearest Neighbor Algorithm, and one class Support Vector Machine are trained respectively using the same training data with selected features to build fault detection model. These training models include each specific regime pattern, so the redundant information in global model will be removed to increase algorithm performance by data classification.

After the training process is finished, the new coming data mixed with normal and abnormal condition is defined as testing data used to monitor machine operating condition. The preparation process will be applied on testing data using the same selected features and multiple regime identification model to identify different regimes. When each regime is identified, the detected faults result will be given by applying fault detection algorithm to show the visualized information. Values calculated by fault detection algorithm over the threshold will be detected and system will send out an alert to notify the engineer. The detected faults can be the suggestion of maintenance to avoid severe and unrecoverable destruct.

3.3 Data Preparation and Model Building

Model training process is constituted by two main parts, data preparation and local-based fault detection. The result from each part will be stored for testing new data. Training process is applied to minimize the data variation and noise, extract useful information from the raw data, reduce the impact of multiple regimes, and implement fault detection model training to learn data pattern. Every step is significantly applied to pre-process the data to ensure the best performance of fault detection algorithm detecting the failures.

Data preparation consists of two procedures which are Data Pre-processing and Multiple Regime Identification. These two are very critical actions to reduce the impact of noise, redundant features, correlated data and covariance caused by different operation condition.

The scope of data preparation includes the following aspects:

- De-noising: Reduce noise of raw data and remove outliers. In most cases, this will be the first step before feature extraction.
- Feature extraction: Extract useful information buried in raw data to present data pattern.
- Dimension reduction: Lower the dimension to increase the computation speed and remove redundant variables. It can be divided to variable selection and instance selection.
- Multiple regime identification: Machine condition changed by different recipes sometimes will cause the normal status being wrongly detected. Classifying different regimes before applying fault detection method is a very significant step to reduce the effect of operation condition change.

Selected features and regime identification model will be stored for testing purpose after data preparation process to pre-process the testing data and identify regime. After all the regimes are classified, fault detection training model will be built in each regime based on their own characteristic respectively. This local-based model method greatly reduces the irrelevance between two distributions which belong to different classes, and the performance and calculation efficiency of fault detection are expected to be enhanced. In order to handle different data characteristic, three fault detection models PCA-MSPC, 1-SVM, and FD-kNN are built in this system, and these fault detection training models will be used for testing process as well.

3.4 Fault Detection Model Testing and Visualization

Once the training process is done, retained features, regime identification model, and fault detection models are implemented to feature selection, regime identification, and fault detection for new testing data respectively. The same features will be selected based on the index from the training dataset. By applying the regime identification training model on the new testing data with selected features, all the testing data will be classified to their most fitted regime based on the data pattern. Fault detection methods will be applied in each regime to obtain the local faults after all the classification are finalized.

Although the fault detection method is useful to offer information for engineers to analyze the condition of the manufacturing process, the communication between the engineer and operator to define the level of issue is an important part in the factory system [29]. In some cases, machine anomaly information can be easily caught by engineer and shut down decision will be directly made and sent to the process line. However, most of the cases, it is impossible to intervene the process line on every alerts from the machine. The purpose of the fault detection method is to alert engineers and operators the abnormal condition as soon as possible to fix the issue. Therefore, it is very significant to convert raw data to meaningful information that engineers can quickly understand and report the situation. By applying PCA-MSPC, kNN, and 1-SVM, meaningless raw data can be transferred to useful machine condition to summarize the state of the process. Result will be shown on a 2-D graph with threshold based on the confidence level of the data. Every data point above the threshold will be detected as abnormality to alert user to take appropriate action. This is an easy understanding alert system for engineers to quickly digest the result and also very useful information for maintenance to avoid any unrecoverable destruct.

3.5 Conclusion

In this local-based fault detection system, the raw data will be transformed to useful information for user to understand the machine operation condition step by step. In training process, data characteristic and pattern can be learned by identification and fault detection training model is expected to increase the calculation efficiency and enhance the robustness of algorithms during the testing process.

Final fault detection visualization overall result will be given by combining all the local faults detected in each regime. The visualization plot is very intuitive to let engineers to catch the condition of machine operation status and alarm notification to make the decision for predictive maintenance. Engineer can work with operator to fix the issue immediately based on this function with accurate detection result.

4. MULTIPLE REGIME IDENTIFICATION

4.1 Terminology

Many applications have multiple operation level during the normal daily running. For rotating applications, elevator lifting system for example, different number of passengers inside the cabin will affect the operation status of lifting system. The output signals, such as voltage, current, power consumption, etc, will be different based on the load change. For manufacturing area, one machine has several orders waiting in sequence, so the machine condition always varies in order to fit different recipe requirement.

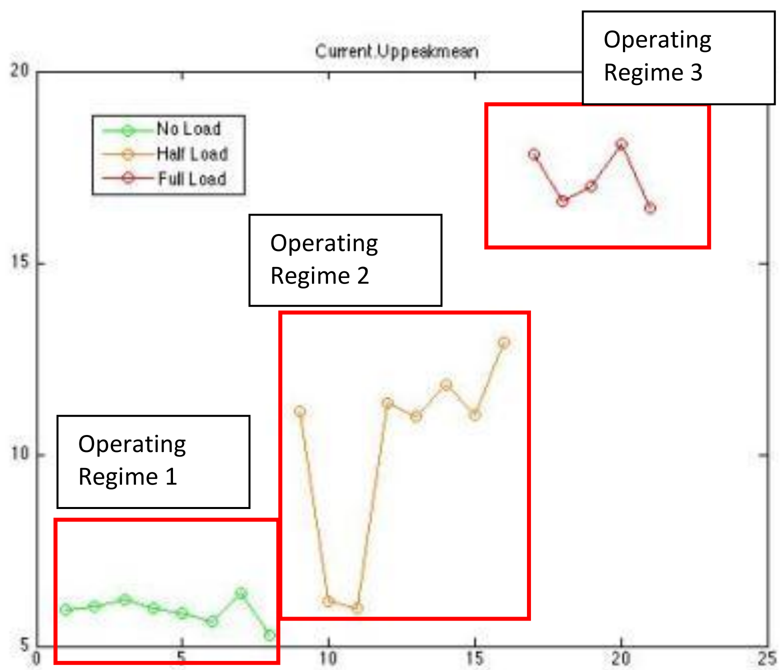


Figure 4.1 Multiple Regimes

These examples show the common problem on how to differentiate that signal change is caused by operation regime or system fault. Figure 3 clearly illustrate this conception that different level of operation status will present distinct data pattern. It is important to classify raw data to each group to reduce the inaccuracy result caused by multiple regimes before applying any fault detection method.

By extracting features as training input from raw data, self-organizing map (SOM) will be applied to learn the data pattern from each regime during training process. After finishing the training process, SOM model is built for analyzing the testing data. In this process, highly uncorrelated data caused by machine condition change will be separated to different groups.

4.2 Multiple Regimes Identification

The purpose of multiple regimes identification is used to decrease the impact of regimes. By learning the data characteristic, the gap between normal condition and faulty condition can be defined as different regimes or operation faults.

Self-organizing Map is an outstanding algorithm can easily handle this issue. In the training process, SOM algorithm will learn the data pattern from all the groups respectively based on the regime generated by recipe shift. With the learned data pattern, SOM can divide the coming test data to each group. This process helps to reduce the effect of regimes to the performance of fault detection algorithm. The detail information about how SOM works to reach multiple regimes identification will be introduced in next section.

4.2.1 Self-organizing Map Introduction

When the input data is coming, N-dimension input data can be denoted by: $X_i = [x_{i1}, x_{i1}, \dots x_{i1}]^T$,$i = 1,2, \dots, t$,$t$ is the number of input samples. The weight vector of each neuron j on the map has the same dimension as the input space and can be represented by: $W_j = [w_{j2}, w_{j2}, \dots w_{j2}]^T, j = 1,2, \dots, m$, m is the number of neurons (nodes) on the map. The goal for SOM training is to move the representative neuron node to the regions of the vector space that are dense in the input vector V. When an input vector of V is presented to the algorithm, all representatives (neuron nodes) compete with each other. The winner, which lies closer to V, as well as its neighboring nodes are updated so as to move towards V. It is a selection and competitive learning process.

This training occurs in several steps and over much iteration:

1. Each node's weights are initialized.
2. A vector is chosen at random from the set of training data and presented to the Lattice (map).
3. Every node is examined to calculate which one's weight is most like the input vector. The winning node is commonly known as the Best Matching Unit (BMU).
4. The radius of the neighborhood of the BMU is now calculated. This is a value that starts large, typically set to the 'radius' of the lattice, but diminishes each time-step. Any nodes found within this radius are deemed to be inside the BMU's neighborhood.
5. Each neighboring node's (the nodes found in step 4) weights are adjusted to make them more like the input vector. Weight vector adjusts according to the equation

$$W(a+1) = W(a) + L(a)[V(a) - W(a)]$$

where a represents the time-step and L is a variable called learning rate and will decrease over time. Basically, what this equation is saying is that the newly adjusted weight for the node is equal to the old weight (W), plus a fraction of the difference (L) between the old weight and the input vector V.

6. Repeat step 2 for N iterations or stop earlier when criteria $|W(t+1) - W(t)| < \varepsilon$ has been met

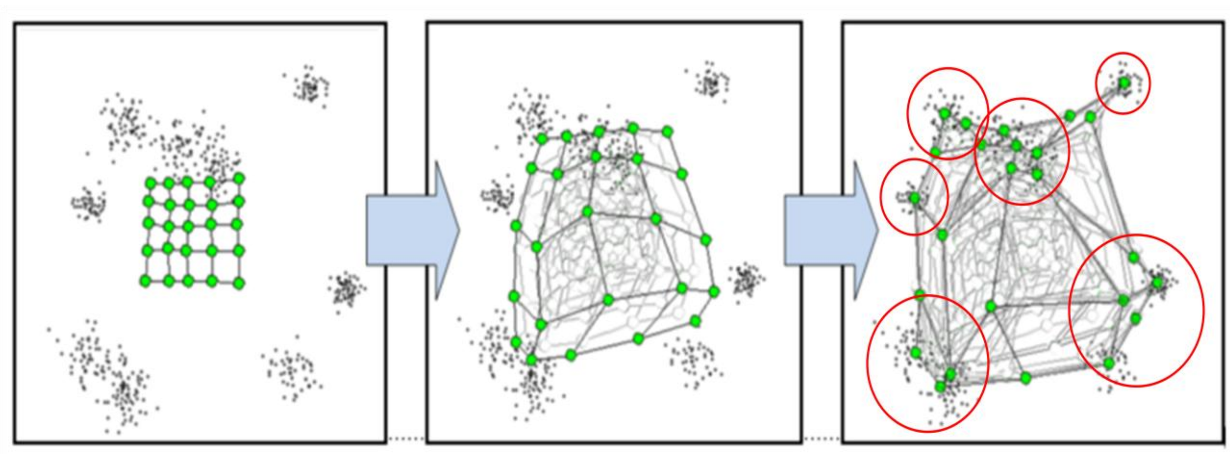


Figure 4.2 SOM Training Process [55]

The above graph shows the SOM training process. Green dots are neuron nodes and black dots are input vectors. After training, Neuron Nodes move to the input vectors. High dimension vectors cannot be viewed directly, SOM can serve as a supplement tool to see the relationship of high dimension input data in 2D lattice (map).

Each red circle shows different characters of each group. Neuron nodes will learn the characteristic of input data, and similar neuron nodes will converge to generate a group to build the training model.

4.2.2 Regime Partitioning Example

With dimension reduction, it is easy for us humans to see relationships between vast amounts of data. The U-Matrix can represent the self organizing map (SOM) and the Euclidean distance between the neighboring neurons is showed as a continuous color gradient from darkest to lightest arbitrary tints.

To demonstrate the classification ability of SOM, here is a simple example data which is a public dataset consists of four measurements from 150 Iris flowers: 50 Iris-setosa, 50 Iris-versicolor and 50 Iris-virginica. The measurements are length and width of sepal and petal leaves. [56]

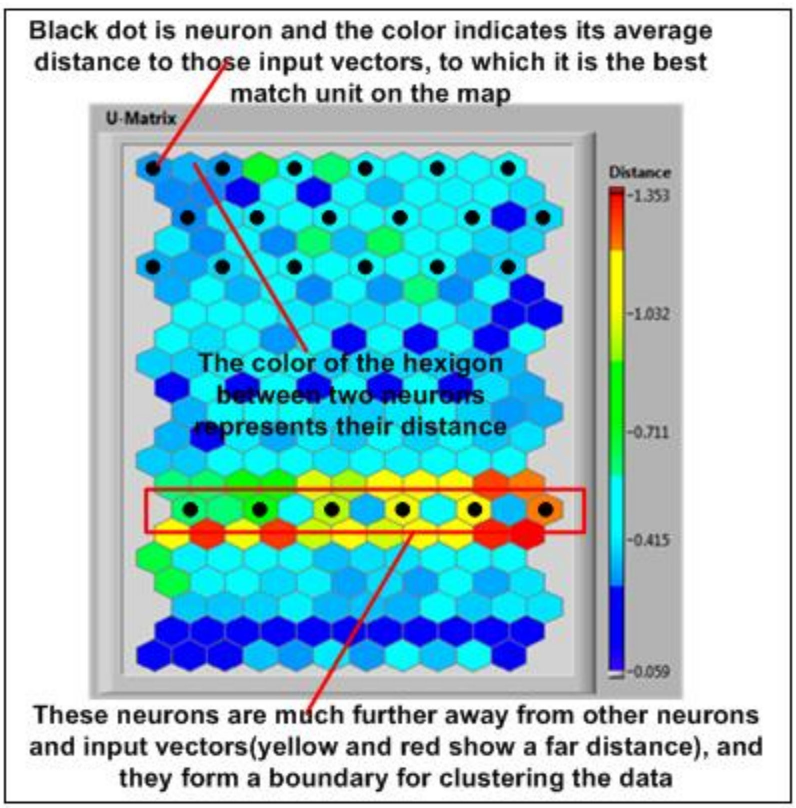


Figure 4.3 U matrix for SOM

This demo shows how to use SOM to clustering this 150 samples and how to test with new data by using the Label and Hit Plot. From the U-matrix it is easy to see that the bottom three rows of neurons form a clear cluster and the top seven rows form another cluster. By looking at the labels, it is immediately seen that this corresponds to the Setosa subspecies. The two other subspecies Versicolor and Virginica form the other cluster. The U matrix shows no clear separation between them. The Setosa subspecies exhibits small petals and short but wide sepals. The separating factor between Versicolor and Virginica is that the latter has bigger leaves.

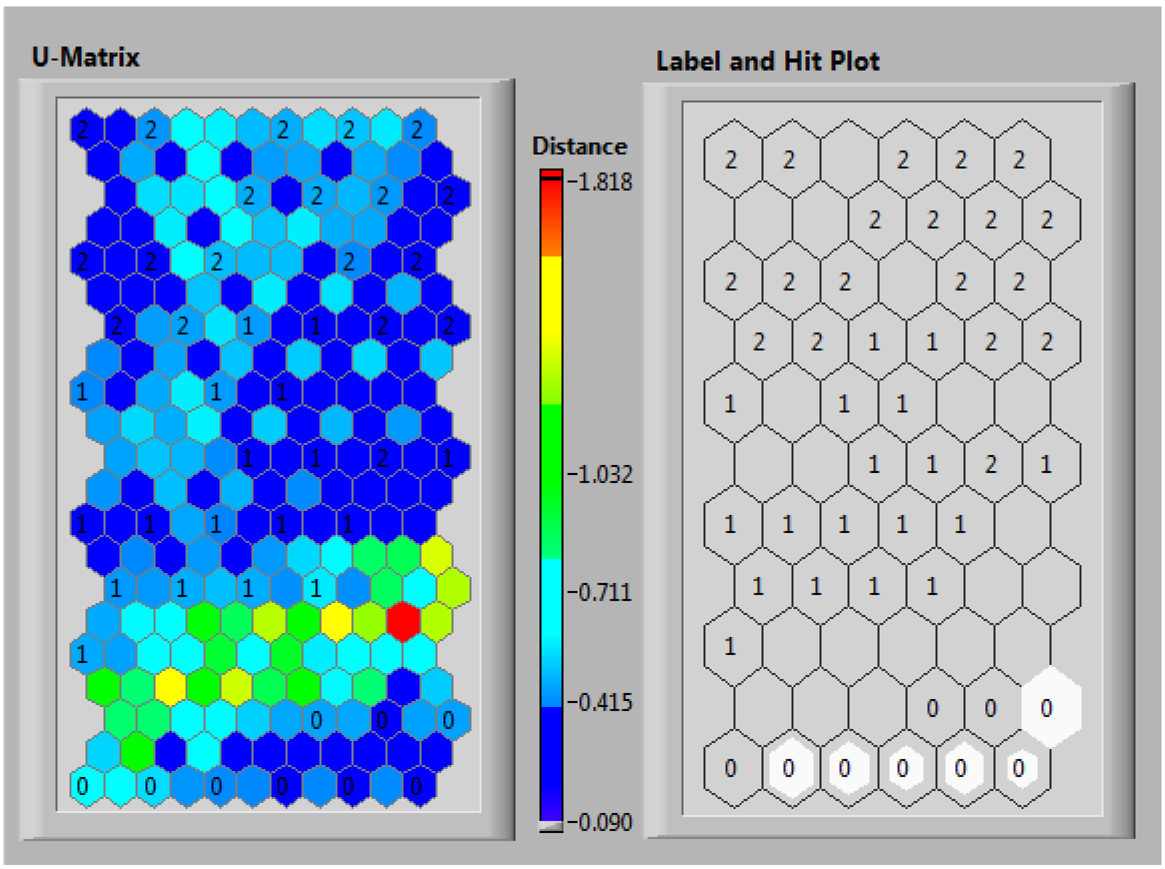


Figure 4.4 SOM Classification Result

4.3 Case Study

4.3.1 Data Description

Rolling element bearing is an indispensable component in rotary machinery systems, and quite usually a critical one that costs significant expense and time when it fails. To monitor bearing health condition, various techniques have been applied including vibration analysis, oil debris and acoustic emission. Among these techniques, vibration analysis is currently the most established with various signal processing methods employed to analyze the fault signatures in high-resolution vibration waveform. An inevitable challenge for bearing monitoring is the multiple possible failure modes and their combinations, which requires advanced analytical methods to extract most relevant information and identify failure modes with high accuracy.

To evaluate analytical methods for bearing fault diagnosis, a test-bed is set up to collect vibration data under different bearing fault conditions. A SKF 32208 tapered roller bearing (TRB) is used and an accelerometer is installed on the orthogonal direction of its housing to measure the axial vibration. A PCM3178-HG-B DAQ card, embedded in Advantech UNO-2160 box, is used to convert analog waveform to digital data.

During the tests, seven (7) bearings with induced defects and a new bearing with no defect are used in turn to generate data under eight different conditions which means there are eight regimes included in the data. The bearing models are the same, thus same specification and geometry parameters. The fault conditions are roller defect, inner-race defect, outer-race defect, the three combinations of any two independent faults and the combination of all three.

4.3.2 Bearing Failure Diagnosis

The task for this research is to develop a systematic method to identify the regimes and diagnose the specific failure modes when bearing is known degraded. The first module is feature extraction that extracts time domain, frequency domain and other features from raw vibration data. Dimension reduction algorithms can be used to downsize the extracted feature set. Eventually classification algorithms are used to learn from labeled training data, and diagnose unknown condition based on testing data.

There are eight different bearing conditions, including normal, roller fault, outer race fault, inner fault, and some fault combination are included in the bearing data. The matrix has several rows which are bearing features extracted from test-bed. By checking the label, user can understand the relationship among the bearing data. Data collected from different fault mode allocate in different region of the map. The testing data hit to the number 7 fault mode means the algorithm detected that the bearing probably has the type of fault.

In the Figure 4.6, by using the bearing data, it shows both the U matrix plot and Label and Hit plot. The U matrix plot has the distance information represented by different colors, but the label and hit plot doesn't have color hexagon to represent distance between neighboring neurons. By checking the boundaries set based on the U matrix bright color hexagons, label and hit plot helps user to check the ability of SOM classification on this bearing data and view the distribution of each groups' input data. Through hitting testing data, label and hit plot can diagnose all the faults type by classifying them to different clusters.

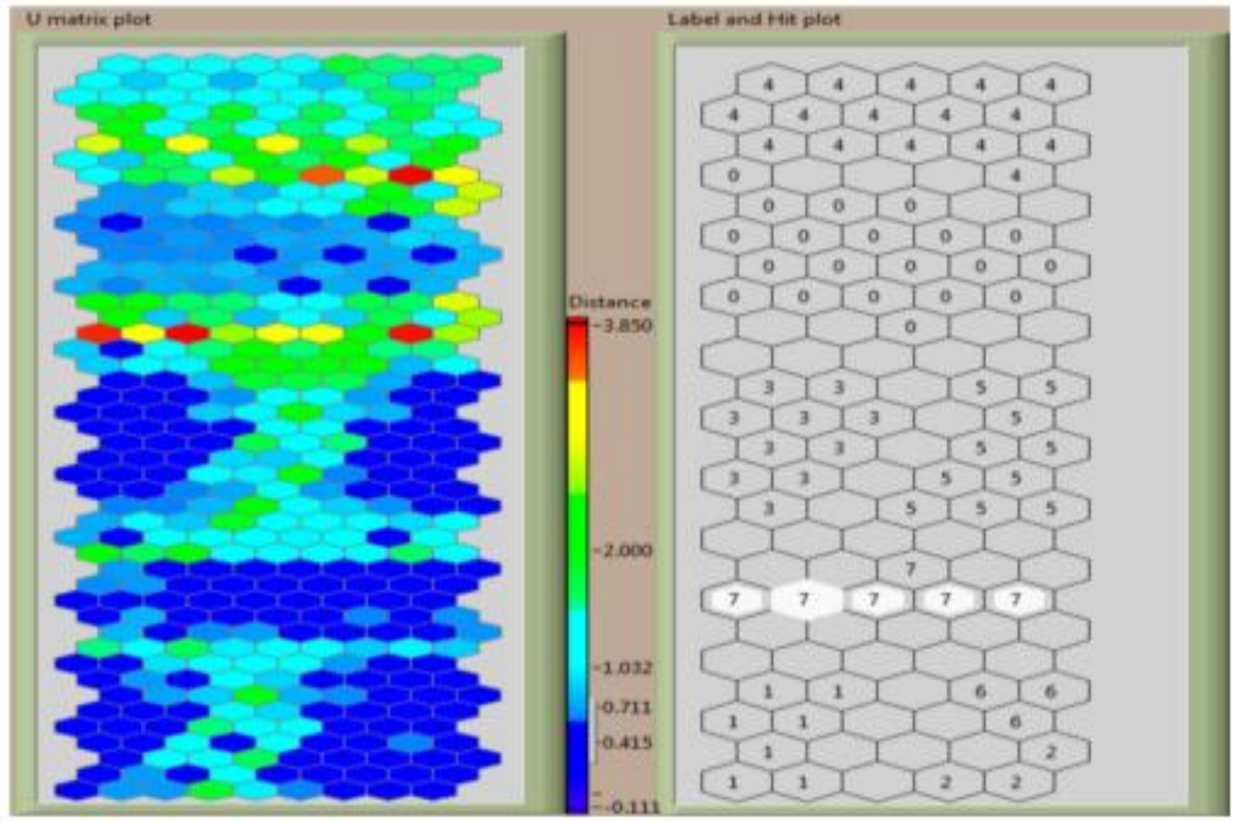


Figure 4.5 Bearing Faults Diagnosis: Label and Hit plot and U matrix plot

By matching the recent feature to one of the feature clusters that corresponds to different fault modes, when the new testing data is coming, SOM model will classify data with different pattern to each appropriate group based on previous training process.

4.4 Conclusion

SOM is an easy understanding and clear visualization method for multiple regimes identification. It works well in clustering data, so multiple regime can be easily identified based on its learning based ability. Due to the learning characteristic, clustering result can be improved by inputting more training data. It will help the model to learn more information from raw data to deal with classification problem. By applying this multiple regime identification method, local fault detection model can be built to avoid the noise existing in global model.

However, during the calculating process, many maps need to be constructed in order to get one final good map. To achieve a good clustering result, one important step is feature selection and information fusion. Since the classification accuracy of SOM cannot be quantified, it is hard to achieve an optimal classification result through combining SOM and some feature selection techniques. Also, SOM is very computationally expensive; therefore the feature selection method will be very significant to improve the method performance.

5 PROPOSED METHODOLOGIES INTRODUCTION

5.1 Data Preparation

As discussed in Chapter 3, the approach framework for fault detection includes one essential part which is named as data preparation before applying local fault detection model. Data preparation consists of two procedures which are Data Pre-processing and Multiple Regime Identification. These two are very critical actions to reduce the impact of redundant features, correlated data and covariance caused by different operation condition.

5.1.1 De-noising

De-noising is the process of removing noise or outlier from raw data which will affect the performance of fault detection algorithm. In order to alleviate the impact, a simple moving average (SMA) is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles. SMA is the unweighted running average of the previous n datum points that can ensure the variations in the average are aligned with the variations in the data rather than being shifted in time.

$$SMA = \frac{X_M + X_{M-1} + \cdots + X_{M-(n-1)}}{n}$$

One nature of the SMA is that if the dataset has a periodic fluctuation, it will eliminate the data variation when an SMA of that period is applied because the average always contains one complete cycle. However, a perfectly regular cycle seldom exists in real world. For a number of applications it is advantageous to avoid the shifting induced by using only "past" data. Therefore a central moving average can be computed using data equally spaced either side of the point in the series where the average is calculated. This requires to use an odd number of datum points in the sample window.

5.1.2 Statistics Calculation

Statistics calculation (a.k.a feature extraction) is one of the most significant processes in many fault detection algorithms. It's capability of extracting useful information and reduce redundant from the original data to aid fast processing and accurate fault detection. The decision of features extraction method always depends on the data characteristic. For different situation, features vary in different dataset. Some simple statistics such as maximum, minimum, average, standard deviation, etc, can be applied to most of the cases to find the peak value or level difference. However, these features are not suitable sometimes to deal with more complicated data pattern.

The Root Mean Square (RMS), also known as the quadratic mean, is a statistical measure of the magnitude of a varying quantity. The RMS value of a set of values or a continuous-time waveform is the square root of the arithmetic average of the squares of the original values or the square of the function that defines the continuous waveform. In the case of a set of n values $\{x_1, x_2, \dots, x_n\}$, the RMS values is given by this formula:

$$x_{rms} = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2))}$$

RMS is often used as a synonym for standard deviation when they refer to the square root of the mean squared deviation of a signal from a given baseline or fit. It is especially useful when the signals are fluctuating between positive and negative.

5.1.3 Dimension Reduction Methods

Dimension reduction has two categories which are instance selection and variable selection (a.k.a feature selection, sensor selection). The former method includes clustering and classification algorithm to separate raw data to different groups. The detail information of instance selection will be given in next section. The latter method is more commonly used to reduce the number of random variables under consideration [57]. It is capable of selecting a subset of variables and maximize relevance and minimize redundancy.

Feature Selection

Instance Selection

Samples	Var1	Var2	Var3	Var4	Var5	...	...	...
1	-0.201	1.209	0.848	-0.074	0.593	...	...	...
2	-0.211	0.645	0.917	0.552	0.458	...	...	...
3	0.744	0.598	0.378	0.861	-0.466	...	...	...
4	0.411	1.293	0.607	1.792	0.285	...	...	...
5	0.402	0.097	-0.082	0.166	-0.397	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...

Figure 5.1 Dimension Reduction

Fisher criterion is a simple and effective classification algorithm that performs one-dimensional vector discriminant result. It projects high-dimensional data to a line which can maximize the distance between the two distributions' mean and also minimize the variance within each distribution. The following equation defines the Fisher criterion:

$$J_{f_1}(P,Q) = \frac{\left|\mu_{P,f_1} - \mu_{Q,f_1}\right|^2}{\sigma^2_{P,f_1} + \sigma^2_{P,f_2}}$$

Where μ is the distribution average, σ^2 is the variance, and the subscripts P and Q represent the two conditions of the dataset, normal and abnormal. Fisher criterion method has great parallels to linear perceptrons. By computing fisher discriminant score for each variable, the variable with high scores will be selected as useful features.

5.2 Fault Detection Approaches

The operation will always be aborted immediately when an error is detected, so this the quantity of faulty data is insufficient. In order to deal with this issue, one class fault detection methods will be applied in this research. One class means only normal data is needed to train the model and all the faults can be detected based on the baseline set by normal data. The capability of this type of algorithm is proved in chapter 2 by previous researches.

Three fault detection methods Principal Component Analysis, One-Class Support Vector Machine, and FD k-Nearest Neighbor Rule are benchmarked in the system and the fault detection comparison result will also be given in this section. Some of the methods have higher capability to handle nonlinear data, and some of the methods have better performance on dealing with Gaussian distribution. Due to the reason, three different fault detection methods will be compared to choose the best.

5.2.1 Principal Component Analysis

Principal components analysis (PCA) is a popular algorithm which is usually used to simplify a dataset by projecting high-dimension to a 2-D plot [8]. Moreover, it is denoted as a linear transformation which picks a new coordinate for the data set such that the greatest variance by any projection of the data set comes to lie on the first axis (then called the first principal component), the second greatest variance on the second axis, and so on. The matrix after transformation will have a diagonal covariance matrix, which means new matrixes column vectors are not correlated with each other. This makes it easy to eliminate the influence of noise or redundant variables on the determinant variables. Figure 5.1 gives one simple example.

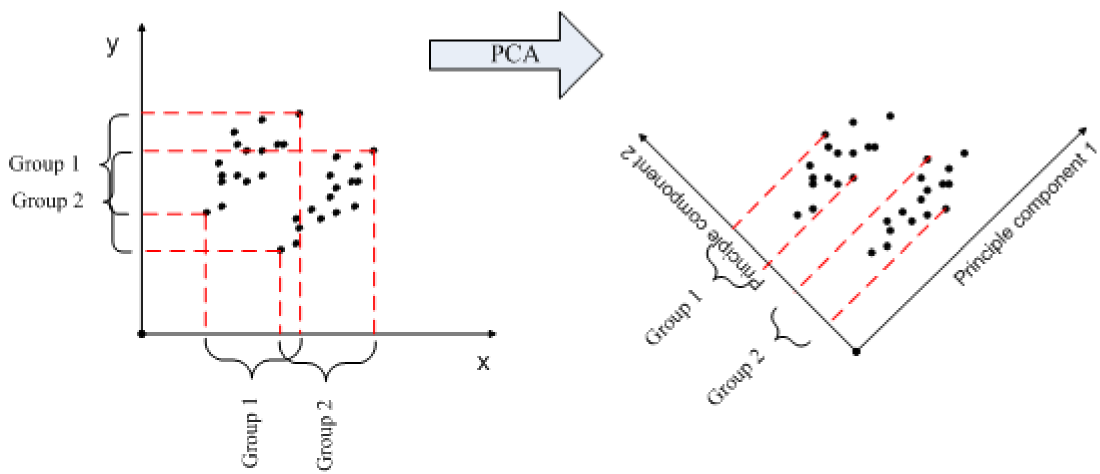


Figure 5.2 Principle component analysis

This is a data set consists of two groups of data. The left graph shows a scatter plot of the data. Its hard to identify the two groups by only using either x variable or y variable, since x and y are correlated. The duplicated information they carried weaken their ability to classify the data by themselves. The right graph shows result after PCA transformation. Data has been projected to new coordinate system with PC1 and PC2 as basis. PC1 and PC2 are completely uncorrelated and the two groups can be easily identified by setting a dividing threshold for PC2.

In order to compute the PCA analysis for a new sample vector of n variables, $v \epsilon \Re^m$ denotes the variable vector. Assuming that there are N variables for each signals, a data matrix

$X\epsilon\Re^{m\times n}$is built with M rows representing the samples. The matrix X needs to be standardized to zero mean and unit variance for covariance-based and correlation-based PCA respectively. The matrix can be decomposed into a loading matrix P and a score matrix T,

$$X = TV^T + \tilde{X} = TV^T + \tilde{T}\tilde{V}^T = [T\ \tilde{T}][V\ \tilde{V}]^T \equiv \bar{T}\bar{V}^T$$

where $\tilde{X} = \tilde{T}\tilde{V}^T$ stands for residual matrix, $\bar{T} = [T\ \tilde{T}]$ and $\bar{V} = [V\ \tilde{V}]$. The symmetric covariance matrix K can be represented as:

$$K = \sum_{i=1}^{n} \lambda_i \vec{v_i}\vec{v_i}^T \approx \frac{1}{N-1} X^T X = [V\ \bar{V}]\Lambda[V\ \bar{V}]^T$$

Where

$$\Lambda = \frac{1}{N-1}[T\ \bar{T}]^T[T\ \bar{T}] = diag(\lambda_1, \lambda_2, \dots \lambda_n)$$

$$V = [\vec{v}_1, \vec{v}_2, \dots \vec{v}_i]$$

and

$$\lambda_i = \frac{1}{N-1} t_i{}^T t_i = var\{t_i\}$$

t_i is the ith column of T and λ_i are the eigenvalues of the covariance matrix in descending order. Once the model is built, the original feature set can be expressed as a sum of the data projection into the principal component space (PCS) and the residual subspace (RS):

$$\hat{x} = PP^T x \in S_p$$

$$\tilde{x} = \tilde{P}\tilde{P}^T x = (I - PP^T)x \in S_r$$

Since S_p and S_r are orthogonal,

$$\hat{x}^T\tilde{x} = 0$$

and

$$\hat{x} + \tilde{x} = x$$

Because K is positive semi-definite, all its eigenvalues are real and no less than zero, thus λ_i>0. λ_i depicts the amount of the covariance matrix energy projected in the direction of the corresponding eigenvector $\vec{v}_1$.

When it exists a high degree of correlation among the components of $\vec{X}$, only a few of the eigenvalues Λ in account for most of the energy in the matrix K. Sort the eigenvalues in descending order so that $\lambda_1 > \lambda_2 >, \dots, > \lambda_n$ Choose the largest p eigenvalues that constitute a majority of $\sum_{i=1}^{r} \lambda_i$, and choose the eigenvectors accordingly.

$$\Lambda_p = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$$

$$V_p = [\vec{v}_1, \vec{v}_2, \dots \vec{v}_i]$$

Then

$$K = [V_p\ \overline{V_p}]\Lambda_p[V_p\ \overline{V_p}]^T$$

Depending on the amount of cumulative variance to be included, choose the appropriate number of PCs. For example, if the threshold is 95%, the formula will be

$$0.95 < \frac{\sum_{i=1}^{A} \lambda_i}{b}$$

Where A is the minimum number of selected features, λ_i is the ith largest eigenvalue and b is the overall number of eigenvalues. When the number of selected PCs is determined, the new

V_A can also be created from the egienvectors corresponding the several largest A eigenvalues in V_p.

The matrix V_A is used for transforming the original raw data to get Principal Components. The principal components of new data can be calculated trough the following equation

$$P_A = M_n V_A$$

Where M_n is the raw data directly collected from sensors or controller, and P_A is an $n \times A$ matrix which includes A selected Principal Components with n samples. By converting the features into Principal Components which will be used to monitor the system operating condition, more variation will be contained in Principal Components and important information is retained.

5.2.1.1 Hotelling's T^2 Statistic

Although PCA is useful for reducing the number of variables which require monitoring, it still cannot show the variance in the data by just one single value. Therefore, a distance-based method, Hotelling's T^2 Statistic, is used to convert a Principal Components to a single value to summarize the condition of machine. The Hotelling's T^2 statistic measures the variations in each principal component space as the sum of normalized squared scores,

$$T^2 = t_i \Lambda^{-1} t_i^T = x^T P \Lambda^{-1} P^T x$$

The loadings matrix P and the eigenvalues Λ are the first l values, where l is the number of selected principal components in the model. Assuming the process is under normal condition and the data follow a multivariate normal distribution, the T^2 statistic is related to an F distribution based on the covariance and population average which are estimated from data.

$$\frac{N(N-1)}{l(N^2-1)} T^2 \sim F_{l,N-1}$$

Where $F_{l,N-1}$ is an F distribution with l and N-1 value as the degrees of freedom. By setting up the threshold by using a given significant level α, the process can be considered under normal condition if:

$$T^2 \leq T_\alpha^2 \equiv \frac{l(N^2-1)}{N(N-1)} F_{l,N-1;\alpha}$$

If the mean is accurately calculated and only the covariance needs to be estimated from the data, the T^2 upper control limits can be defined as the following equation:

$$T_\alpha^2 \equiv \frac{l(N-1)}{(N-1)} F_{l,N-1;\alpha}$$

The only difference between the above two equations is a factor of (N+1)/N. If the number of data points, N, is large enough, the T^2 value can be approximated with a χ^2 distribution with l value as the degree of freedom and

$$T_\alpha^2 = \chi_{l;\alpha}^2$$

In reality, process monitoring typically has large N value, so the χ^2 upper control limit is suitable and often implemented in the process monitoring.

5.2.1.2 Squared Prediction Error

The Squared Prediction Error (SPE) statistic is based on the error between the raw data and the PCA model. It measures the projection of the sample vector on the residual space to present how consistent the sample and model are and how accurately the PCA model can recreate the data from PCs.

$$SPE \equiv \|\tilde{x}\|^2 = \|(I - PP^T)x\|^2$$

If

$$SPE \leq \delta_\alpha^2$$

The process can be considered under normal condition. Where δ_α^2 denotes the control limit for SPE with a α confidence level. The SPE control limit was developed by Jackson and Mudholkar [58],

$$\delta_\alpha^2 = \theta_1 \left(\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{{\theta_1}^2} \right)^{1/h_0}$$

where

$$\theta_i = \sum_{j=l+1}^{m} \lambda_j^i , \quad i = 1,2,3$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3{\theta_2}^2}$$

l is the number of selected principal components and c_α is the normal deviate based on the upper 1- α percentile. It assumes x follows a multivariate normal distribution, and the approximation for this distribution is made to derive the control limit and only be available when θ_1 is very big. Otherwise, the number of selected principal components in the model will not affect the result.

When a fault occurs on the machine, the faulty sample vector consists of normal portion superimposed with failure portion. The SPE will exceed δ_α^2 and the fault will be detected by the system to send out an alert.

5.2.2 Support Vector Machine

Support Vector Machines is a good supervised learning algorithm which is usually used for regression and classification. The common factor of SVM is the use of a method which is known as the "kernel trick" to apply linear classification method to non-linear classification issue [59].

Suppose we want to classify some data points into two classes. Often people have interest in classifying data as part of a machine learning process. These data points may not necessarily be points in R2 but may be multidimensional points. We are interested in the ability of separating them by a hyperplane (a generalization of a plane in three dimensional space or more than three dimensions). As we examine a hyperplane, this form of classification is known as linear classification. We also want to choose a hyperplane that separates the data points "neatly", with maximum distance to the closest data point from both classes -- this distance is called the margin. We desire this property since if we add another data point to the points we already have; we can more accurately classify the new point since the separation between the two classes is greater.

The form's data points are considered:

$$\{(x_1, c_1), (x_2, c_2), \dots (x_n, c_n)\}$$

The c_i is either 1 or -1 --- this constant represents that the point x_i belongs to which class. This as training data, which denotes the correct classification which we would like the SVM to eventually distinguish, by means of the dividing hyperplane, which takes the form

$$w \cdot x - b = 0$$

We are expecting the hyperplane to maximize the distance from the dividing hyperplane and also to avoid any points existing between these two hyperplane. By using geometry, we find the distance between the hyperplanes being $2/|\boldsymbol{w}|$, so we want to minimize$|\boldsymbol{w}|$. In order to exclude the data points, all i need to be ensured either

$$w \cdot x_i - b \geq 1$$

or

$$w \cdot x_i - b \leq -1$$

This can also be represented as:

$$c_i(w \cdot x_i - b) \geq 1 \qquad 1 \leq i \leq n$$

The issue now is to minimize $|\boldsymbol{w}|$ subject to the above threshold, and it is a quadratic programming (QP) optimization issue. Normally, it is difficult to find the linear separation hyperplane in the original space of the data points because all the points are “squeezed” in a small dimension. Therefore our strategy is to try to project the data into a higher dimension space, to sparse them; and increase the possibility of linear separation. This result can be showed in Figure 5.2

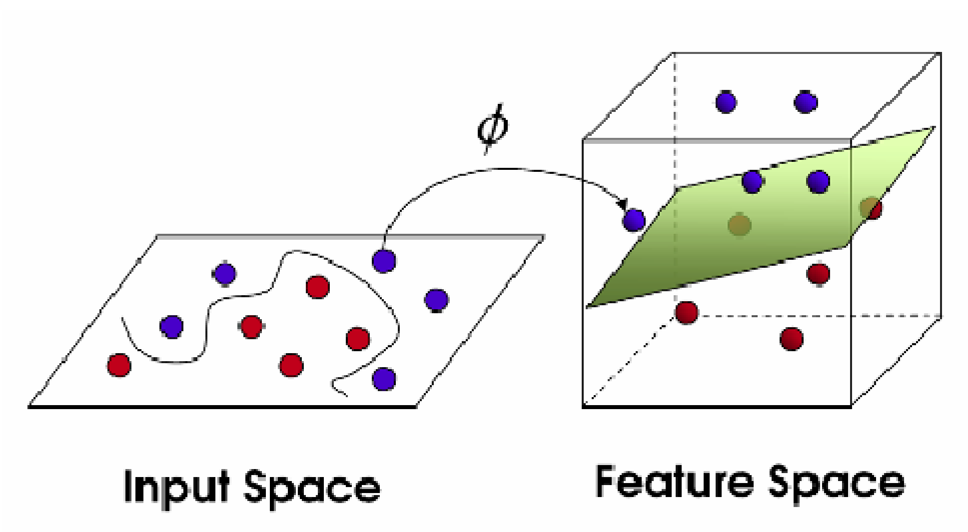


Figure 5.3 SVM motivation

Such projection function (ϕ) can be implicitly defined by a kernel function (k) which could adopt several different forms, some common kernel functions are list as follow:

- Polynomial (homogeneous): $k(x, x') = (x \cdot x')^d$
- Polynomial (inhomogeneous): $k(x, x') = (x \cdot x' + 1)^d$
- Gaussian RBF: $k(x, x') = exp\left(-\frac{\|x-x'\|}{2\sigma^2}\right)$
- Sigmoid: $k(x, x') = \tanh(kx \cdot x' + c)\,, k > 0\ and\ c < 0$

5.2.2.1 One-Class Support Vector Machine

The one-class SVM algorithm proposed by Schölkopf et al. [60], is extended from the original support vector machine methodology. It is usually used to handle one-class classification problem that only one type of data is inputted as training data to build the model. Essentially, the objective of one-class classification technique is novelty detection or outlier detection. In the context of fault detection, all the possible faulty samples will be separated from normal samples based on the training model. The main difference between traditional support vector machines and one-class support vector machines is that the former one needs both normal operating data and faulty operating data to train the model for classification whereas the latter one only requires normal operating data. By being given the normal operating training samples, one-class SVM calculates the boundary that includes most of the training data points. If the testing dataset is classified into this boundary, it is categorized as normal condition, or it will be categorized as an outlier or failure condition. Figure 5.3 clearly describes the basic knowledge of one-class support vector machine. The '+' label represents the normal data.

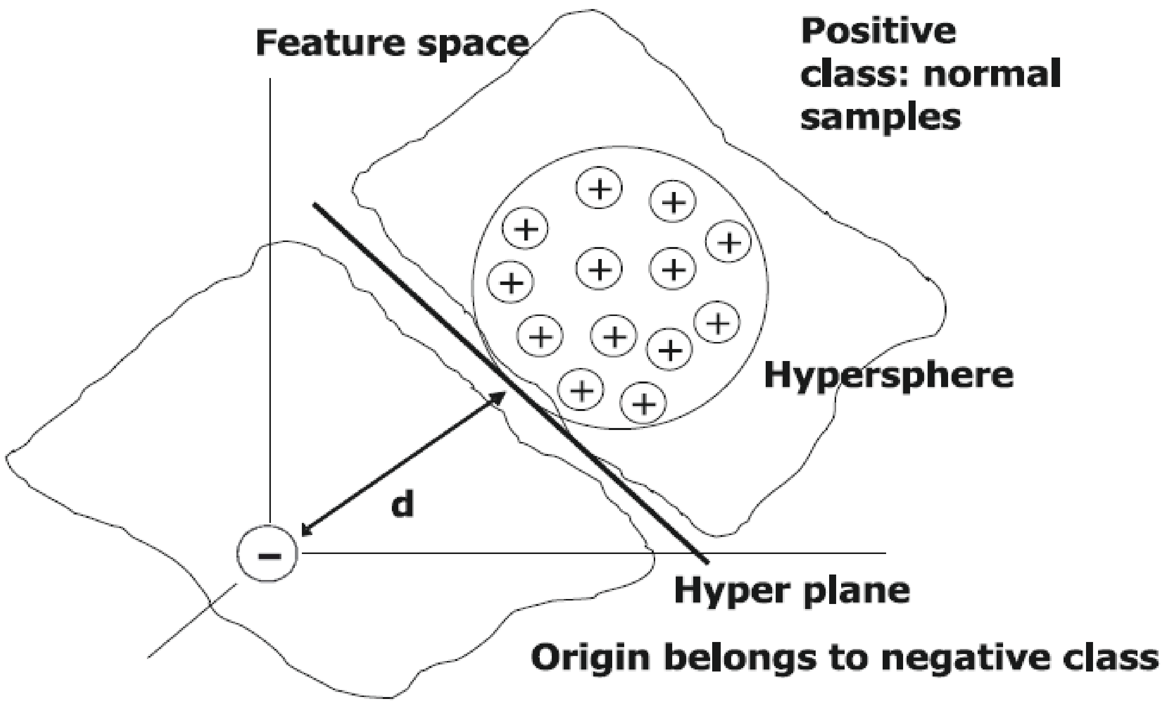


Figure 5.4 Graphical illustration of one-class SVM [26]

There is an assumption that the origin in the feature space is belongs to faulty or negative class. Therefore, the objective will be maximizing the distance between the origin and the normal data class. However, most of the cases, it is difficult to perfectly separate the normal data from the origin, so a slack of variable υ, which is the maximum fraction of false alarms being allowed in the training data, is defined.

Consider $x_i \in \Re^n$, i=1,...,m as a single class training dataset, and each of the x_i is associated to a label $y_i \in \{1\}$. As mentioned earlier, the input space are mapped to a higher dimension feature space by applying projection function (ϕ) where a hyper plane is built to separate normal samples from the origin. This can be written as follows,

$$min \frac{1}{2} \|w\|^2$$

subject to

$$w \cdot \phi(x_i) + b \geq 0$$

However, most of the cases, it is difficult to perfectly separate the normal data from the origin, so a slack of variable ξ_i and $v \in (0,1)$ is defined to solve this problem. The modified optimization equation is written as follows,

$$min_{w,\xi_i,b} \frac{1}{2} \|w\|^2 + \frac{1}{vm} \sum_{i=1}^{m} \xi_i + b$$

subject to

$$w \cdot \phi(x_i) + b + \xi_i \geq 0, \quad \xi_i \geq 0 \quad i = 1,2, \dots, m$$

Lagrange multipliers $\alpha_i \geq 0$ and $\eta_i \geq 0$ are introduced and Lagrangian is formed based on the earlier case of binary SVMs. The partial derivatives of the Lagrangian formula are set as follow,

$$L(w, \xi, b) = \frac{1}{2} \|w\|^2 + \frac{1}{vm} \sum_{i=1}^{m} \xi_i + b - \sum_{i=1}^{m} \alpha_i (w \cdot \phi(x_i) + b + \xi_i) - \sum_{i=1}^{m} \eta_i \xi_i$$

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{m} \alpha_i \, \phi(x_i)$$

$$\frac{\partial L}{\partial \xi_i} = 0 \rightarrow \alpha_i = \frac{1}{vm} - \eta_i, \quad \eta_i \in (0,1)$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{m} \alpha_i = 1$$

Once the kernel function $K(x, y) = \phi(x) \cdot \phi(y)$ is chosen appropriately, the dual form of the Lagrangian can be written as follow,

$$min_{\alpha' s} \alpha_i \alpha_j K\left(x_i, x_j\right)$$

subject to

$$0 \leq \alpha_i \leq 1, \quad \sum_{i=1}^{m} \alpha_i = \nu m$$

After the optimization problem is solved, w and b are subsequently calculated. In order to estimate the distance metric, the measure of the perpendicular distance in the feature space of a given sample vector x from the boundary is considered,

$$F(x) = w \cdot \phi(x) + b = \sum_{i=1}^{m} \alpha_i \ \phi(x_i) \cdot \phi(x) + b = \sum_{i=1}^{m} \alpha_i K(x_i, x) + b$$

α_i and b are the model parameters used for algorithm performance optimization purpose. Therefore, the preliminary understanding of the knowledge of mapping function $\phi(x)$ is not needed. If F(x) is greater than or equal to zero, the sample points are within the boundary which means it is classified as normal operating condition. If F(x) is lesser than zero, the sample points lie outside the margin which will be classified as machine failure mode. The F(x) is the output of one-class SVM fault detection method.

5.2.3 k-Nearest Neighbor Rule

In pattern recognition, the k-nearest Neighbor (kNN) rule is a classification method to classify a new sample point by computing the sum of distance to predefined number of nearest neighboring training samples in the feature space [54]. The input data for kNN is an unlabeled which means it classifies sample data based on their similarity between testing data and training data. kNN has been used in many applications in the areas of pattern recognition, data mining, diagnosis, etc. In many cases, the kNN rule has been used for fault diagnosis or classification. However, it is difficult to identify and characterize potential failures and include them in the training data, so the application of fault diagnosis based on the kNN rule will be very limited.

By using given unlabeled sample data x, the kNN rule finds the nearest predefined number of labeled sample data from training dataset based on the distance metric calculation. Several different distance metrics have been proposed for kNN algorithm, but the most common distance metric is still the Euclidean distance. Once the x's k-nearest neighbors are found, there are two ways of voting for kNN to determine its label based on the distance metric. One voting scheme is majority voting, and another is weight-sum voting. In majority voting, the decision of the class based on the number of k-nearest neighbor training samples lie within the range. In the weight-sum voting, every distance between the x and each training sample will be considered to weight each vote based on the rule that near samples should be counted more than far samples.

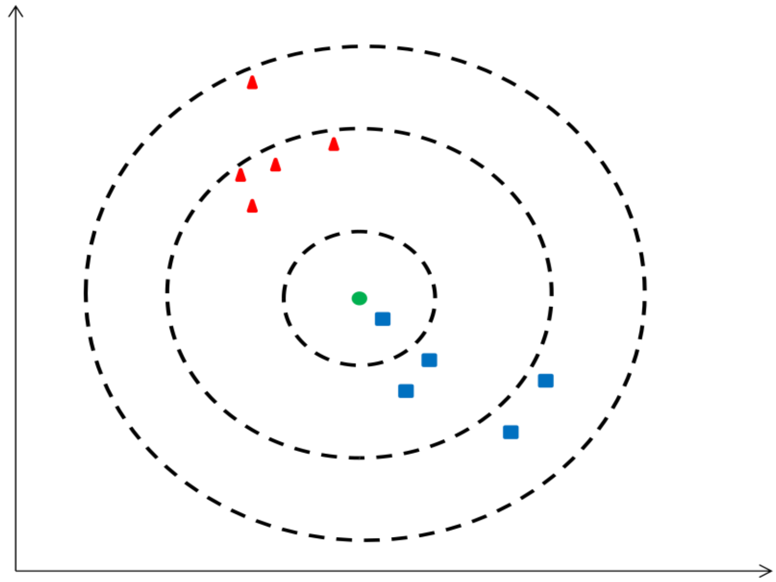


Figure 5.5 Example with two features [28]

In figure 5.4, red triangles and blue squares represent two training samples from class X and Y respectively. The green circle in the center is x, the sample needs to be classified. The three dash circles show different selection of parameter k, 1, 3, and 5. If the majority voting is applied, x will be classified to class Y when k is equal to 1; x is class X when k is equal to 3; there is no decision when k is equal to 5. In contrast, if the weight-sum voting is applied where each vote will be weighted by $1/_{d^2}$ (d is the Euclidean distance), then x will be always classified as class Y when k is equal to 1,3, or 5.

5.2.3.1 Fault Detection Based on K-Nearest Neighbor Rule

The kNN rule classifies sample data based on their similarity between testing data and training data. In this case, the training data needs to be constituted by labeled normal samples and various faulty samples. The incoming testing data will be classified to normal or one type of known failure using the kNN classification method. The main difference between the traditional kNN fault diagnosis and kNN based fault detection is that there is no need to label or define the data in advance for kNN fault detection. As the requirement comes, the traditional kNN rule will be modified to for fault detection.

The fault detection method using the kNN rule is based on the concept of the similarity between the normal sample trajectory from testing data and the normal samples trajectories from training data. On the other hand, the abnormal sample trajectory should have some deviation which is different from the pattern of normal samples from training data. In conclusion, the sum of abnormal sample's distances to the nearest neighboring training samples should be greater than the sum of normal sample's distances to the nearest neighboring training samples. The threshold for detecting faulty samples can be defined with certain confidence level by determining the sum of the distribution of normal samples' to their nearest neighboring samples in training data. Once the incoming samples value is below the threshold, it will be classified as normal condition; otherwise, the sample will be detected as faulty condition.

During the fault detection model building, each sample in training data set needs to find their k- nearest neighbors. The kNN squared distance of sample i (D_i^2) which is defined as the sum of squared distances of sample i to its k-nearest neighbors, will be calculated for each sample.

$$D_i^2 = \sum_{j=1}^{k} d_{ij}^2$$

Where d_{ij}^2 denotes the squared Euclidean distance between sample i to its jth nearest neighbor.

After the distance equation is built, the threshold D_{α}^{2} with confidence level α can be estimated based on the squared Euclidean distance. For example, a 99% confidence limit can be estimated as the value for the situation that 99% of the calibration samples are under the threshold. For an incoming undefined testing sample x, the fault detection will first find its k-nearest neighbors from the training data. When those nearest neighbors are found, the calculation of x's kNN square distance D_{x}^{2} will be given. If D_{x}^{2} is greater than D_{α}^{2}, it is classified as a failure. Otherwise, it is detected as normal condition.

5.3 Conclusion

Most of the fault detection methods need both normal and abnormal data to build the training model. However, in reality, the quantity of abnormal data is usually limited from the system because the operation will be aborted once an alert comes out. Due to the reason, the one-class fault detection method, which only normal data is needed for model training, will be more suitable to fit the real world requirement and strict limitation. One-class fault detection method makes no assumption about the linearity of the process, so the development of this type of method naturally can handle process nonlinearity and multimodal environment. These three methods will be applied on the real semiconductor manufacturing data to compare their performance in chapter 6.

6. SEMICONDUCTOR CASE STUDY:

FAULT DETECTION METHODOLOGY EVALUATION

In this chapter, the work presents the relation between application of one-class fault detection methods and semiconductor manufacturing process monitoring. In the first section, Data preparation, regime identification, and fault detection will be applied step by step in the system. Only a portion of data is used as training data to for training process. After the data preparation process, regimes classification model SOM is built based on the retained training features. Three different fault detection models PCA, kNN, and one-class SVM are trained in each regime respectively to reduce the noise and increase the algorithm accuracy.

The second section of this chapter implements the rest of the data as testing data on each fault detection algorithm based on multiple regime identification in advance. The final comparison result will be given to evaluate which algorithm is the best suitable for semiconductor batch process monitoring. Both global and local model algorithm performance will be listed to show the impact of application of multiple regime identification.

6.1 Real-time Process Monitoring

In real world, the real-time process monitoring is very important for engineer to react and address the issue before it becomes worse. Traditionally, previous researcher used to collect data off-line or from the experiment to do the analysis for fault detection. The result is usually accurate because the data quantity and quality is good. Both normal and abnormal data will be stored completely cycle by cycle. However, in reality, the data quantity and quality is limited due to the reason that when an alert comes out, the operation will be shut immediately to protect the machine. Once an error happens, less reaction time is the key point to avoid the loss on yield, and, more importantly, the accuracy of the fault detection result will be another significant factor for yield and quality control as well.

6.2 Data Description

The data is collected from an Al stack etch process on the commercially available Lam plasma etch tool at the Texas Instruments Inc. The process goal is to etch the TiN/A1-0.5% Cu/TiN/oxide stack with an inductively coupled BCL_3/Cl_2 plasma [22]. The standard recipe of the manufacturing process includes a series of six steps. The first two steps are for pressure stabilization and gas flow. Step 3 is a short time plasma ignition process. Step 4 and step 5 are the two main steps in the etching process. Step 4 is the etch of the Al layer to terminate at the Al endpoint and step 5 acts as the over-etch for the underlying TiN and oxide layers. Step 6 exhausts the chamber. The stabilization step is followed by the three etching regions: TiN, Al and oxide etch. In this case study, only the data from step 4 and 5 are considered because they contain more useful process information.

Three important sensors, machine state variables, optical emission spectroscopy (OES) and radio-frequency monitoring (RFM), are used in this process for collecting data. However, only machine state variables which are collected in 1 second sample interval are considered in this particular case. That is because, in previous research, machine state variables gave more information about the process and reflected the faults more sensitively in the system [26]. The machine variables used in this case study is listed in Table 6.1.

No.	Variables
1	BCl3 flow
2	Cl2 flow
3	RF bottom power
4	RFB reflected power
5	Endpoint A detector
6	Helium pressure
7	Chamber pressure
8	RF tuner
9	RF load
10	Phase error
11	RF power
12	RF impedance
13	TCP tuner
14	TCP phase error
15	TCP impedance
16	TCP top power
17	TCP reflected power
18	TCP load
19	Vat valve

Table 6.1 Machine Variables Description

This dataset includes 129 wafers, which consists of 108 normal wafers and 21 faulty wafers. These data are collected from a series of three experiments (numbers 29, 31, 33), and 21 faulty wafers are induced intentionally by changing the TCP power, RF power, pressure, Cl_2, or BCl_3 flow rate. Because of the unequal batch duration, the 56th wafer had very few sample points which will be excluded from the analysis. Note that these three experiments were run several weeks apart and data from different experiments has a different mean and somewhat different covariance structure.

In total, there are 107 normal wafers (56^{th} wafer is excluded) and 21 faulty wafers. The 107 normal wafers' data will be randomly divided into a 2:1 ratio for model training and testing.

Two-thirds of normal data are used to form the training and validation set, and the rest of normal data will be mixed with 21 faulty data to test the performance of the algorithm.

6.3 Evaluation of PHM Solutions based on Global-local Model

6.3.1 Description of Global-local model

In modeling procedure, the definition of global model is applying all the data to build fault detection model without any separation, and the local model based on each of the single experiments under different machine conditions. As the idea mentioned in previous chapter, many applications have multiple operation level during the normal daily running. Each operating level can be defined as one regime including specific data pattern. For each regime, three proposed fault detection algorithms Principal Components Analysis, K-nearest Neighbor Algorithm, and one class Support Vector Machine are trained respectively using the same training data. These training models include each specific regime characteristic, so the unknown information in global model will be avoided to increase algorithm performance.

6.3.2 Segmentation

This etch process data includes a series of two steps, step 4 and 5. Because of the different steps, machine condition needs to be change to fit the process requirement. Due to the reason, the variations will be reflected in some of the signals depending on the parameter control. In figure 6.1, the endpoint A detector signal shows a two level difference between step 4 and step 5. This level gap sometimes makes the feature extraction and model building more difficult since the features from one step is not that useful for other steps.

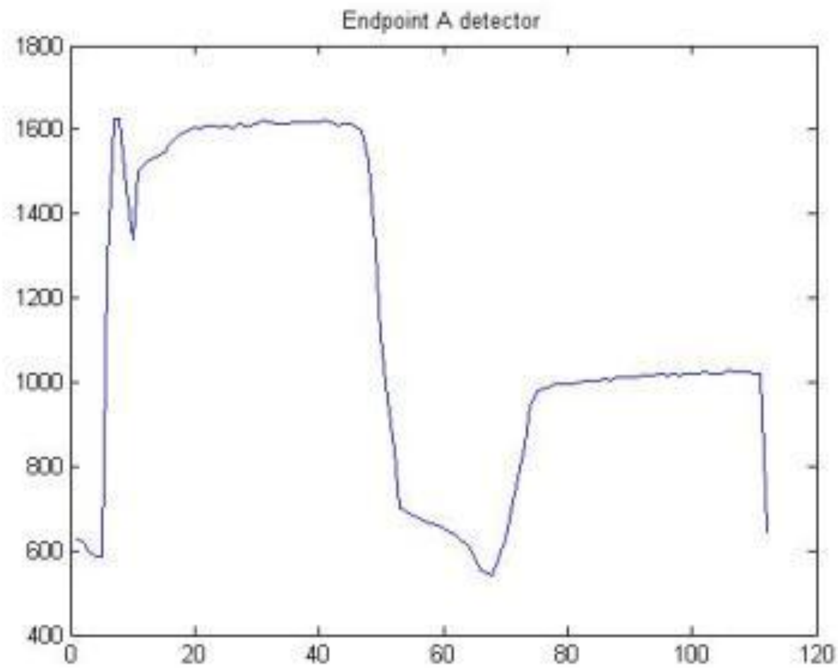


Figure 6.1 Endpoint A Detection Signal

Usually, the obvious level difference can be detected by setting a threshold, but some of the signals have ambiguous boundary to distinguish the process. In this case, the index of each step is offered in the semiconductor process data for segmenting signals. With the index, the two steps 4 and 5 can be easily segmented.

6.3.3 Feature Extraction & Selection

Feature extraction and dimension reduction are two very important aspects of multivariate statistical analysis and can have a significant impact on the overall sensitivity and robustness of the method. Due to the manufacturing process shift caused by aging of the etcher over a clean cycle, differences in the incoming materials and drift in the process monitoring sensors, there are some drift and shift included in the process condition. In order to capture the best data characteristic, mean and covariance statistics are calculated as the features preparation for fault detection algorithms. Because the three experiments were run several weeks apart, the different machine conditions will cause different means and covariance structure in the data from three different experiments. The result show in Figure 6.2.

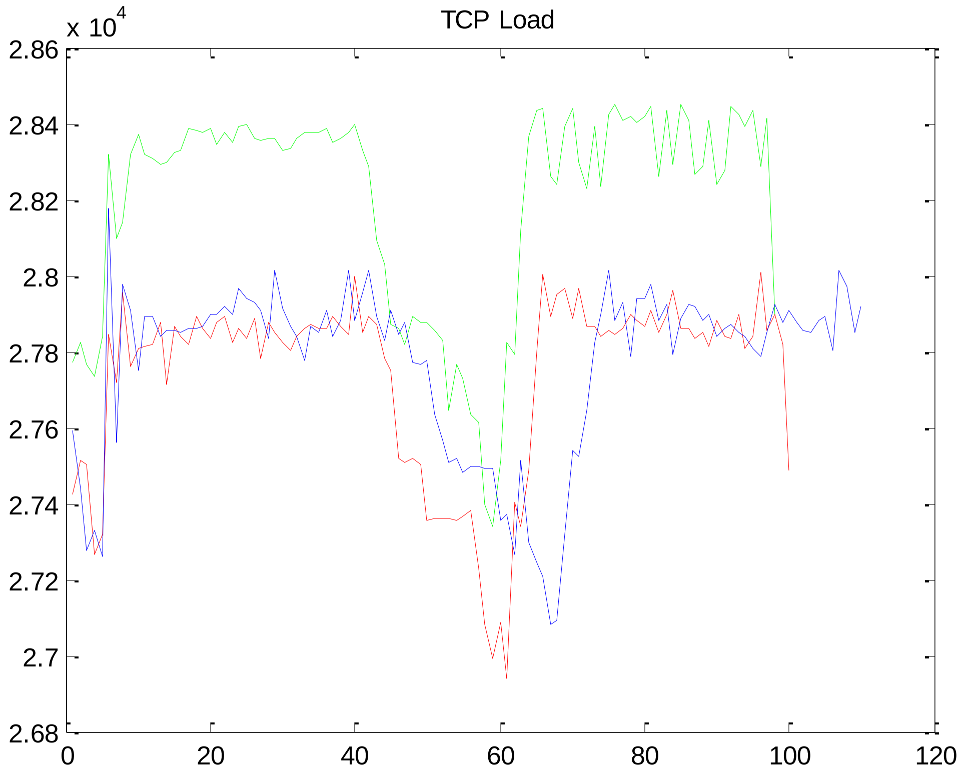


Figure 6.2 Drift shows in TCP Load variable

In general, after feature extraction, feature selection will be applied to reduce the matrix size and strengthen the algorithm performance. It is important to increase the algorithm efficiency by dimension reduction to achieve real-time monitoring. Most of the case, reduce the feature number can also ensure the accuracy of the methods, because some of the features which have slight relevance to the process condition might confuse the algorithm.

However, in this case, feature selection has limited impact on the fault detection result, but it works well on reducing the calculation speed. It might be the reason that all the signals collected from the sensors are related to the process condition. When the data becomes larger with the time elapsing, the great improvement of algorithm processing speed can be expected.

6.3.4 Regime Classification

Multiple regimes is generated by different machine recipes. In this case, the semiconductor etching process data includes three experiments 29, 30, and 31 which were run several weeks apart. As the reason, the machine condition and signal's mean and covariance structure of each experiments are different. These differences are very significant features to capture the data pattern for multiple regimes identification model training, and pre-selected features will also be applied here for the model building.

Self-organizing Map is applied to lean the data pattern to classify different groups. Because of its data mining ability, only a part of the data needs to be used for model training process. In order to prove the greater impact of the regime than faulty data, all the 107 normal wafers data are applied as training data, and all 20 faulty wafers data are used for testing the SOM model. The result is shown in Figure 6.2.

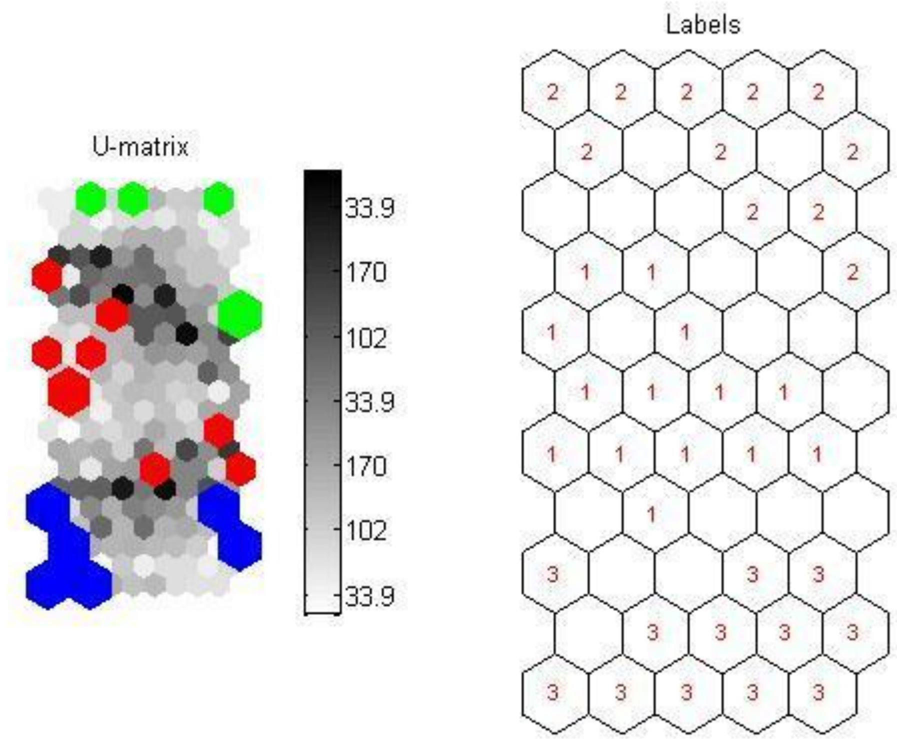


Figure 6.3 SOM Multi-regime Identification

The result in Figure 6.2 shows three different groups classification. On the left side, note that there are two series dark hexagon boundaries which separate the U-matrix to three areas. Each area is one kind of class representing experiments 29, 31 and 33 respectively that the label can be found on the right side based on the 107 wafers data. Experiment 29, 31 and 33 are labeled as 1, 2, and 3 respectively. Three different color hexagons red, green and blue which are located on the map stand for the 20 faulty wafers from three experiments. The colors and labels are used to help for the three groups recognition, so only the two series dark hexagon boundaries will be shown in practice.

6.3.5 Fault Detection Method Based on Global-local Model

In this section, three one-class fault detection algorithms Principal Components Analysis, K-nearest Neighbor Algorithm, and one class Support Vector Machine implemented in semiconductor process data will be introduced. Due to the reason of data quantity limitation in real world, the fault detection method focuses on handling one-class data issue, which means these algorithms only need normal data to build the model for detecting failures. In each algorithm section, one global model and three local models for each experiment are built, and the visualization result is given.

6.3.5.1 Fault Detection kNN

In this case, the result which blue dots are healthy data for training, black dots are healthy data for testing, and green dots are faulty data for testing, is shown in Figure 6.3. In global model, we can see that half of the failures are detected based on the threshold with 95% confidence which means 95% of training data are under this threshold. The confidence number can be changed based on how confidently the user trusts the data.

Although half of the failure mode can be detected by the system, there is still 50% of the chance that faults will be missed to cause unrecoverable damage. Due to the reason, three models are built in three groups respectively to learn different groups' pattern. In local model 1, all of the faults are successfully detected by the system, and local model 2 and 3 missed 4 faults in total. The ratio of detected versus total faults based on local model is 16 of 20. The improvement from 50% to 80% greatly increases the accuracy.

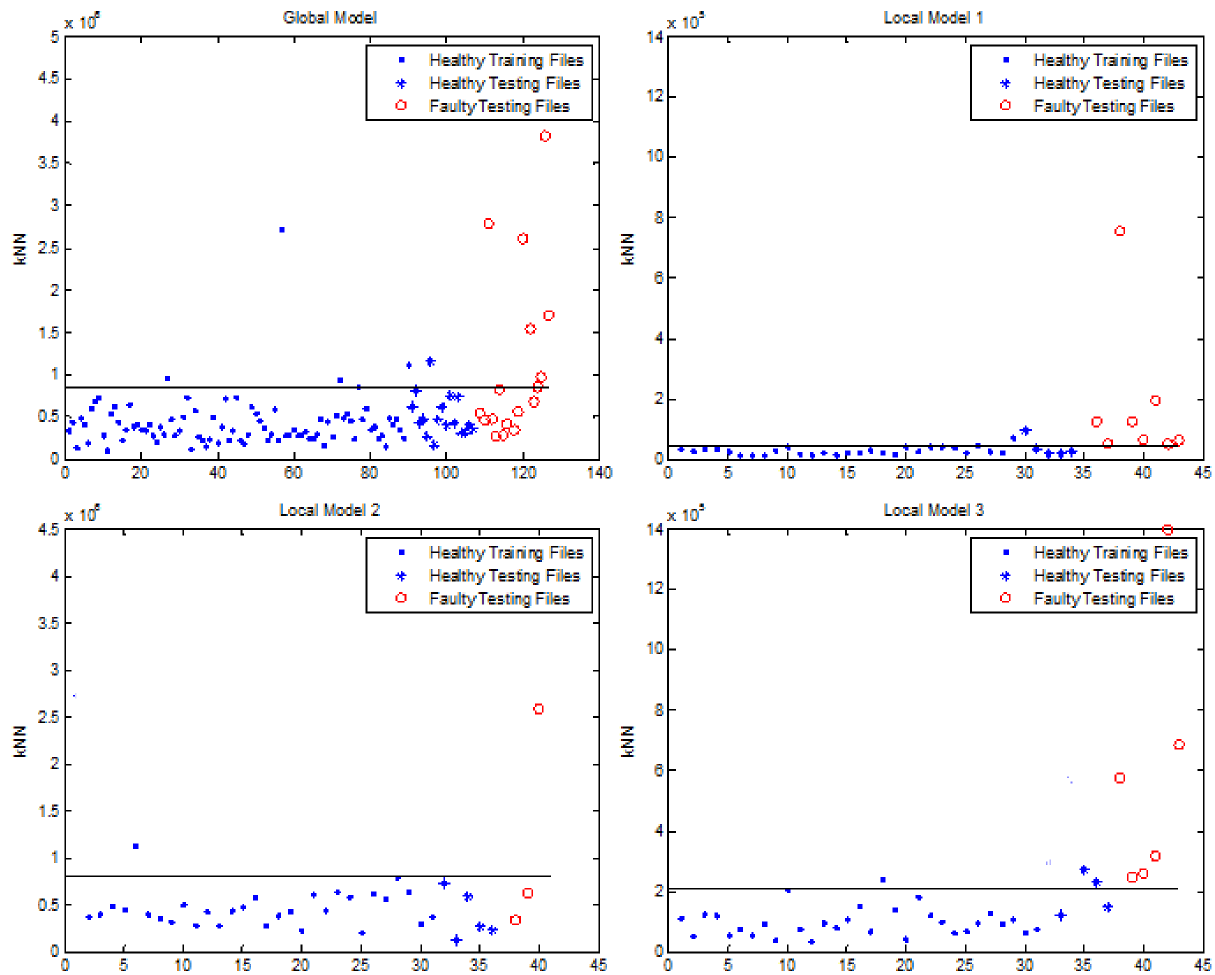


Figure 6.4 kNN Global and Local Model

6.3.5.2 One class SVM

As the plot introduction mentioned in last section, the threshold is also set as 95% confidence and part of the normal data is mixed with all the faulty data as testing data. In the global model, one-class SVM performs slightly better than fault detection kNN, but the accuracy is still below 70%. The system still cannot promise to include most of the faults to avoid the lost caused by undefined errors.

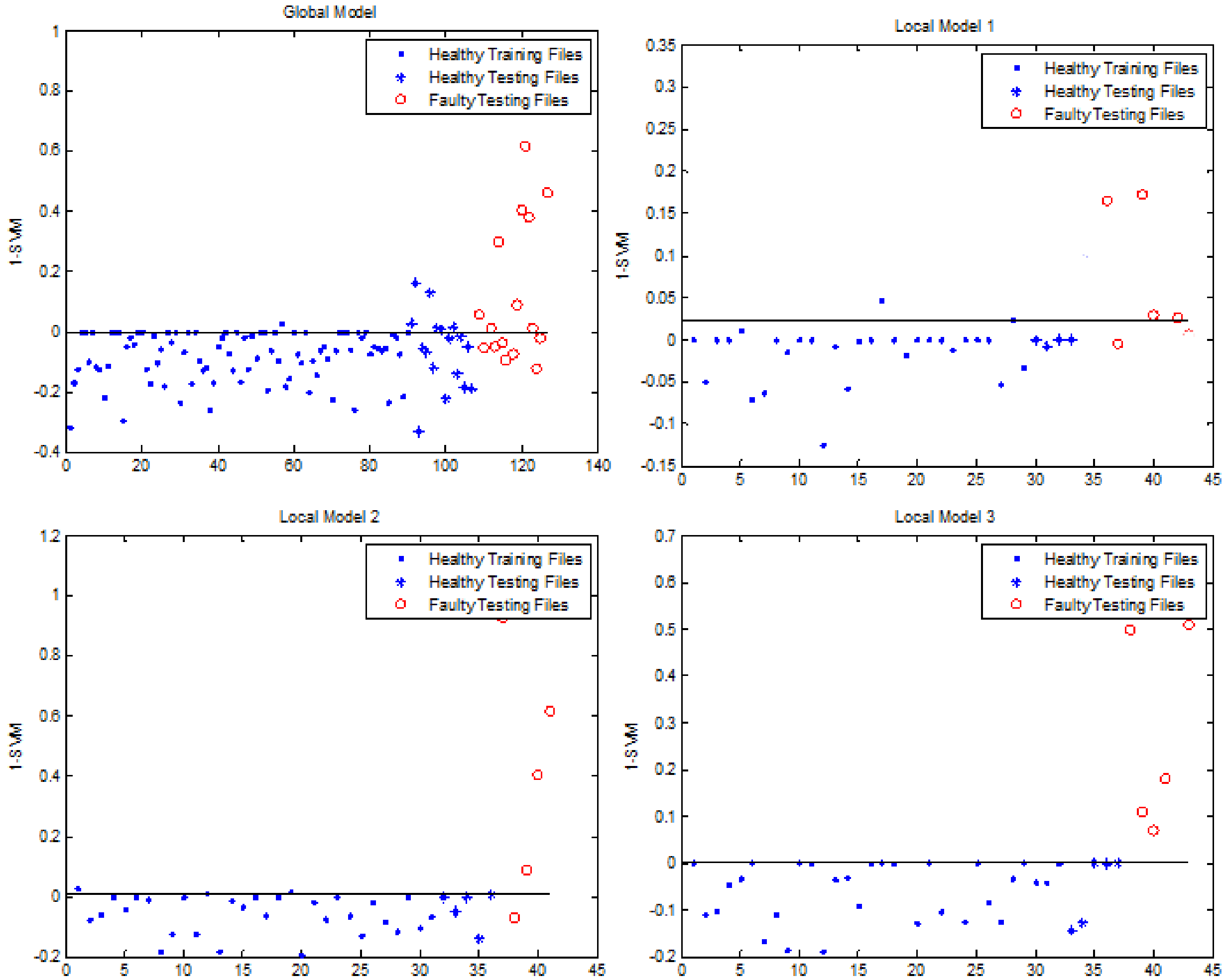
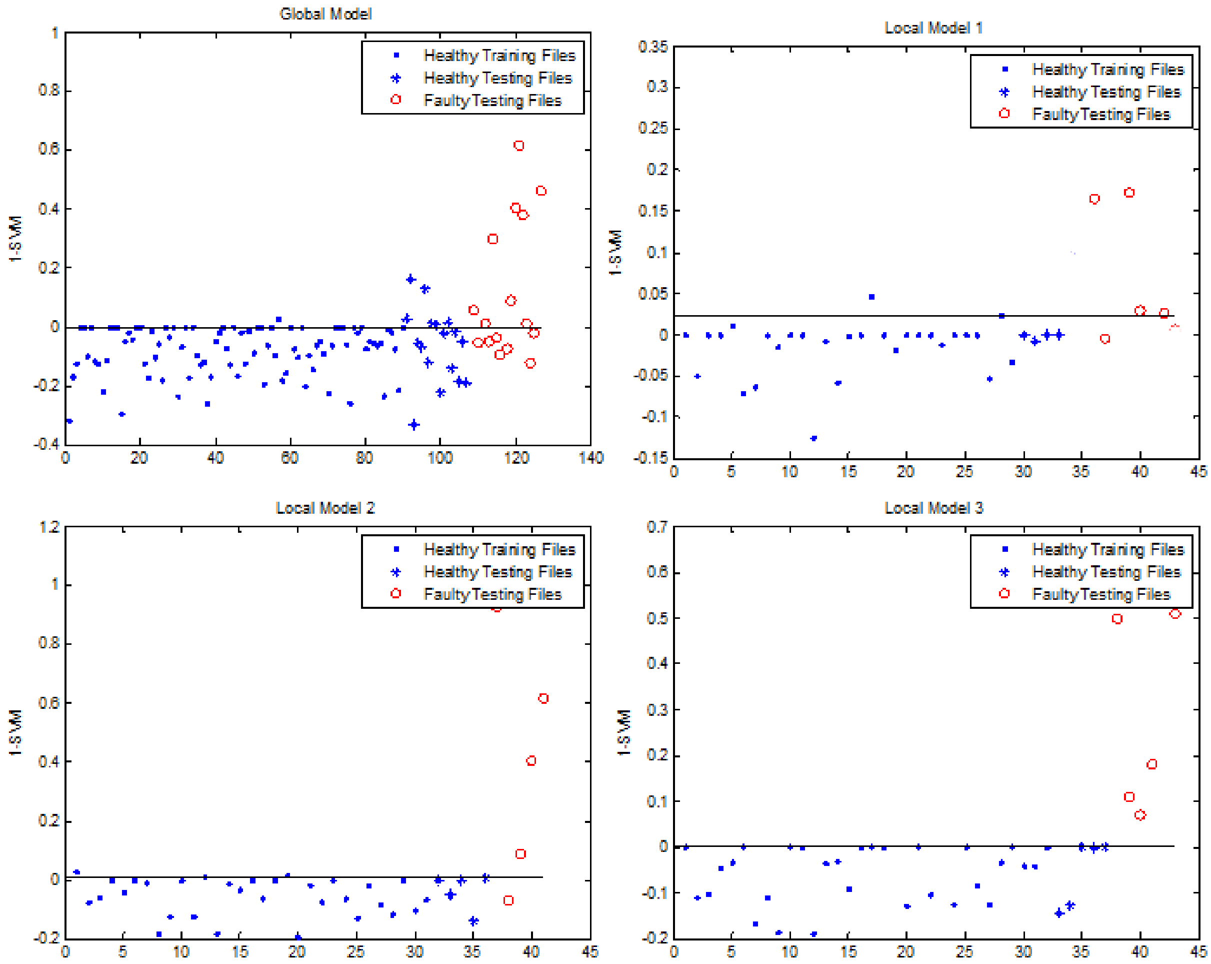


Figure 6.5 1-SVM Global and Local Model

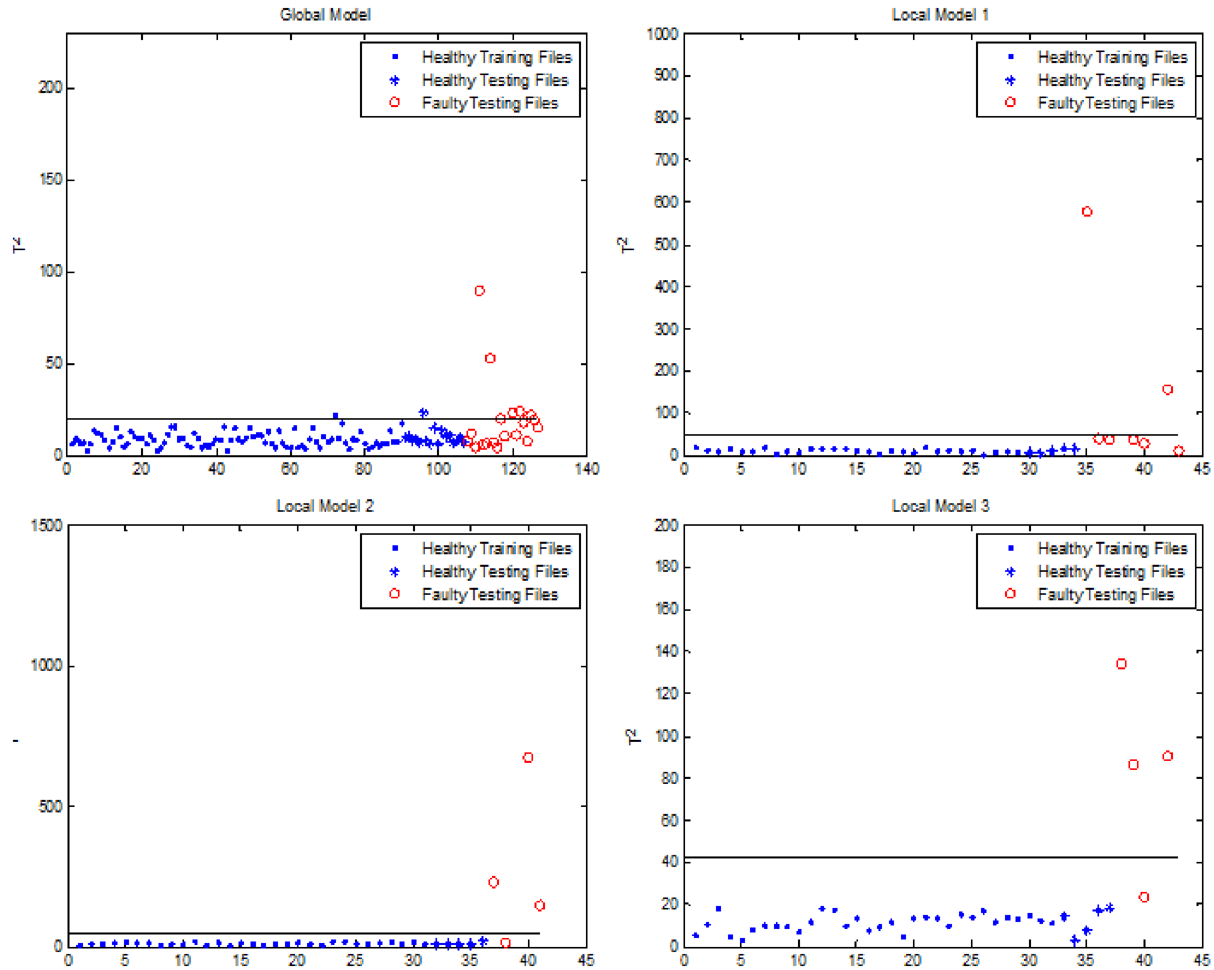
In order to improve the method, local model is applied for one-class SVM as well. There are two faulty points are wrongly detected as normal condition in local model 1 whose performance is worse than kNN. However, local model 2 only missed one point and model 3 detects all the faults perfectly. In total, one-class SVM got 85% accuracy by detecting 17 of the 20 faults. This overall number is slightly greater than kNN and one-class SVM also detects some faults which kNN failed to find.

6.3.5.3 Fault Detection PCA

In fault detection purpose, Principal Components Analysis consists of two statistical calculations, Square Prediction Error and T^2. As the algorithm introduced in chapter 5, PCA is a multivariate statistical fault detection methods. PCA is first applied to detect the faults at the confidence level of 99.5%. Figure 6.6(a) shows the fault detection result based on global model and 3 local models using SPE index and (b) also shows the fault detection result based on global model and 3 local models using T^2 index.



(a)



(b)

Figure 6.6 PCA Global and Local Fault Detection Model: (a) SPE chart and (b) T^2 chart

SPE and T^2 charts together detect 15 faults out of 20 total faults, but there is not complete overlap between the faults detected. We can see that T^2 does not perform very well and the majority of the faults are missed by T^2 in global model. Although the ratio of detected faults is high, the three local models have more outstanding performance than global model. Note that three SPE local models nearly detect all the faults. Only fault 9 is missed, but the result is really close to the threshold.

6.4 Result Comparison

By arranging all faults result which are detected by each methods showing in Figure 6.4 - 6.6, the summary of results of one-class SVM, FD-kNN, PCA SPE and PCA T^2 is shown in Table 6.2. The tick mark against a specific fault and a specific algorithm indicates that the value for that fault using that algorithm exceeds the threshold. In other words, the marked space means that the fault was detected by the method, and the absence means that the fault was not detected by the method.

	Induced Faults	1-SVM Global	1-SVM Local	kNN Global	kNN Local	PCA-SPE Global	PCA-SPE Local	PCA-T^2 Global	PCA-T^2 Local
	TCP +50	√	√	√	√	√	√		√
	RF -12	√	√	√	√		√		
	RF +10	√	√	√	√	√	√		
	Pr +3	√	√	√	√		√	√	√
Group 1	TCP +10	√	√		√		√		
	BCL_3 +5		√	√	√	√	√		
	Pr -2	√	√	√	√	√	√	√	√
	Cl_2 -5		√		√		√		√
	TCP +30			√	√	√	√		
	Cl_2 +5	√	√	√	√		√		√
	RF +8					√	√		
Group 2	BCl_3 -5	√	√			√	√	√	√
	Pr +2	√	√	√	√		√	√	√
	TCP -20	√	√	√	√	√	√		√
	TCP -15	√	√	√	√	√	√	√	√
	Cl_2 -10	√	√		√		√		√
Group 3	RF -12		√	√	√		√		
	BCl_3 +10		√	√	√	√	√	√	√
	Pr +1	√	√	√	√	√	√		√
	TCP +20	√	√	√	√	√	√		√

Table 6.2 Result Comparison

In Table 6.2, the faults which came from different experiments were separated to three groups. The three groups are divided based on the three experiments which were run several weeks apart have drift and shift in all the signals. In each group, global and local model result are given by applying these three methods. As the result shown in each section, it is obvious that the local model result showed superior performance than global model in every group and also in all three algorithms. The total number of faults which local model detected is also more than the total number of faults that global model detected. Otherwise, local model result didn't miss any fault which global model already detected and covered more faults than global model did.

Algorithm	Global Model	Local Model
1-SVM	6.5% (7/107)	0.9% (1/107)
KNN	4.7% (5/107)	4.7% (5/107)
PCA-T^2	1.9% (2/107)	0% (0/107)
PCA-SPE	5.6% (6/107)	0.9% (1/107)

Table 6.3 False Alarm

Although local-based KNN well improved the fault detection result, the false alarm ratio is same as global model and much higher than other three methods. As showed in Table 6.3, 1-SVM, and PCA-SPE greatly reduced the false alarm rate after applying local mode. PCA-T^2 has the worst fault detection result, but it has no false alarm after applying local model.

6.5 Conclusion

Most of the researches in semiconductor area focusing on improvement of algorithms and invention of new methods. However, improving algorithms or even inventing a new method are time and funds consuming and also very difficult. Due to the reason, this research dedicates on analyzing the data characteristic to reduce the algorithm training difficulty instead of trying to improve the algorithm itself. The combination of multiple regimes identification and fault detection is very efficient and effective to handle machine recipe change problem in practice.

In Table 6.2, this result apparently shows the improved performance of local model of different methods by using semiconductor etching process data. By using SOM for multiple regime identification, the process shift and drift can be detected and used as the metric to classify the data from different experiments. This method effectively reduces the impact of regimes which will confuse the fault detection method to distinguish the difference caused by multiple regimes and faults.

By comparing fault detection result and false alarm ratio, PCA-SPE showed the best result in overall performance. It perfectly detected all the faults and misclassified only one normal condition data based on local model. kNN also had very good result in fault detection, but its false alarm ratio is the highest in all four methods.

7. CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions

This research study aimed to investigate a comprehensive fault detection system, specifically the ability of multiple regimes identification, on the Prognostics and Health Management process. In pursuit of this goal one semiconductor case study was carried out in this work, this semiconductor manufacturing process dataset includes 129 wafers, which consists of 108 normal wafers and 21 faulty wafers. These data are collected from a series of three experiments (numbers 29, 31, 33), and 21 faulty wafers are induced intentionally by changing the TCP power, RF power, pressure, Cl_2, or BCl_3 flow rate. With different conditions in three experiments, specific data pattern could be learned by SOM in the training process to separate the new coming data. In every regime, the PCA-MSPC, FD-kNN, and 1-SVM algorithms were applied and T2, Q, and confidence values were used to quantify the machines health state.

As the result showed in Table 6.2 in chapter 6, the obvious superior result was given to prove the improved performance of local model of different methods by using semiconductor etching process data. With different classed identified by SOM, the process shift and drift can be detected and used as the metric to classify the data from different experiments. This method effectively reduces the impact of regimes which will confuse the method to distinguish the difference caused by multiple regimes and faults. The three algorithms were able to identify over 80% of faulty wafers in overall based on local model which is way more better than the average 50% detection rate based on global model.

7.2 Research Findings

There are a number of key points that can be taken from this research. First and foremost, recipe change is a common strategy in many manufacturing process which will affect machine operating status to cause shift and drift in collected signals is called multiple regimes. Every regime is one kind of class which contains its specific characteristic. Mostly, two different regimes have limited relation and one of the regimes cannot be helpful for fault detection method to detect the faults in other regimes. With multiple regimes identification, uncorrelated data can be separated to different group to avoid the confusion. As the reason mentioned above, considering the learning and classification ability, SOM is choice for handling multiple regimes. It was shown, by learning the data pattern from both normal and abnormal condition, that the training model can near perfectly separate all the testing data. Besides, if more input data are used for training, then the result will be more accurate based on its learning ability.

Second, due to the reason that the issue of limited data quantity is a common problem happened in many manufacturing areas. The operation will always be aborted immediately when an error is detected, so this action will limit the collection quantity of faulty data. In order to deal with this issue, three PCA-MSPC, FD-kNN and 1-SVM one class fault detection methods are applied in this research to handle the data quantity limitation issue. During the training process, only normal data is needed to build the training model. Once the normal condition baseline are set up, the testing data which are above the threshold will be defined as faults.

Last but not least, the key development in this research is the fault detection method improvement based on multiple regimes identification. By combining data pre-processing process, multiple regimes identification and one-class fault detection methods, a comprehensive fault detection system was built in this work. In the case study given in chapter 6, the superior performance of local-based fault detection was proved by comparing the global and local fault detection result with three different one-class algorithms.

7.3 Recommendation for Future Work

In the literature it was stated that while some work has been done in the field of semiconductor, much of that work has been to develop individual algorithms. This work then was meant to perform a preliminary investigation into comparing the fault detection method based on Global model and the Local model methods. As a result there are a number of directions that future work can be taken using this research as a launch pad. First, the result given by testing the three algorithms in this research cannot stand for all the other algorithms. As a direct extension of this work, more new improved one class fault detection methods or other PHM algorithms can be applied in this system to test the performance.

Furthermore, this study dealt with semiconductor manufacturing process, but it could be easily expanded to other industrial areas such as aircraft engines, elevator or any industry with machine manufacturing process. This work mainly focused on the fault detection portion of PHM, but there are still many other tasks that could get benefit from investigating such as this. Fault diagnosis, health assessment, and remaining useful life prediction are all important tools for system health management.

REFERENCES

[1] T. F. Edgar, S. W. Butler, W. J. Campbell, C. Pfeiffer, C. Bode, S. B. Hwang, K. Balakrishnan and J. Hahn, "Automatic control in microelectronics manufacturing: Practices, challenges, and possibilities," *Automatica,* vol. 36, pp. 1567-1603, 2000.

[2] T. Dunn, C. Lee, M. Tronolone and A. Shorey, "Metrology for characterization of wafer thickness uniformity during 3DS-IC processing," in *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd,* 2012, pp. 1239-1244.

[3] E. Coss Jr, "Methods and systems for analyzing process equipment processing variations using sensor data,", 2010.

[4] A. Bleakie and D. Djurdjanovic, "Feature extraction, condition monitoring, and fault modeling in semiconductor manufacturing systems," *Comput. Ind.,* 2012.

[5] M. K. Steve Fulton, "ISMI Consensus Preventive and Predictive Maintenance Vision," *International SEMATECH Manufacturing Initiative,* 2007.

[6] T. Wang and H. Pan, "Improving the OEE and UPH data quality by Automated Data Collection for the semiconductor assembly industry," *Expert Syst. Appl.,* vol. 38, pp. 5764-5773, 2011.

[7] Z. Lee, K. Ying, S. Chen and S. Lin, "Applying PSO-based BPN for predicting the yield rate of DRAM modules produced using defective ICs," *The International Journal of Advanced Manufacturing Technology,* vol. 49, pp. 987-999, 2010.

[8] J. G. Yang, J. Zhang, J. X. Yang and Y. Huang, "A Principal Component Analysis Based Fault Detection Method in Etch Process of Semiconductor Manufacturing," *Key Eng Mat,* vol. 522, pp. 793-798, 2012.

[9] C. Sassenberg, C. Weber, M. Fathi and R. Montino, "A data mining based knowledge management approach for the semiconductor industry," in *Electro/Information Technology, 2009. Eit'09. IEEE International Conference on,* 2009, pp. 72-77.

[10] J. Lee, S. J. Qin and I. Lee, "Fault detection and diagnosis based on modified independent component analysis," *AIChE J.,* vol. 52, pp. 3501-3514, 2006.

[11] J. M. Ko, C. O. Kim, S. J. Lee and J. P. Hong, "Structural Feature-Based Fault-Detection Approach for the Recipes of Similar Products," *Semiconductor Manufacturing, IEEE Transactions on,* vol. 23, pp. 273-283, 2010.

[12] (). *SEMICONDUCTOR WAFER MANUFACTURING*. Available: http://www.memc.com/?view=Semiconductor.

[13] A. Adam, L. C. Chew, M. I. Shapiai, L. W. Jau, Z. Ibrahim and M. Khalid, "A hybrid artificial neural network-naive bayes for solving imbalanced dataset problems in semiconductor manufacturing test process," in *Hybrid Intelligent Systems (HIS), 2011 11th International Conference on,* 2011, pp. 133-138.

[14] Q. P. He and J. Wang, "Statistics pattern analysis: A new process monitoring framework and its application to semiconductor batch processes," *AIChE J.,* vol. 57, pp. 107-121, 2010.

[15] G. S. Hwang and G. Y. Masada, "Methods for improving the reliability of semiconductor fault detection and diagnosis with principal component analysis," 2006.

[16] Y. Jeong, S. Kim and M. K. Jeong, "Automatic identification of defect patterns in semiconductor wafer maps using spatial correlogram and dynamic time warping," *Semiconductor Manufacturing, IEEE Transactions on,* vol. 21, pp. 625-637, 2008.

[17] M. Bensch, M. Schröder, M. Bogdan and W. Rosenstiel, "Feature selection for high-dimensional industrial data," in *Proceedings of European Symposium on Artificial Neural Networks (ESANN),* 2005, .

[18] Y. Saeys, I. Inza and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics,* vol. 23, pp. 2507-2517, 2007.

[19] E. St. Pierre and E. Tuv, "Robust, non-redundant feature selection for yield analysis in semiconductor manufacturing," *Advances in Data Mining.Applications and Theoretical Aspects,* pp. 204-217, 2011.

[20] T. Pan, D. Wong and S. Jang, "Chamber Matching of Semiconductor Manufacturing Process Using Statistical Analysis," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* vol. 42, pp. 571-576, 2012.

[21] D. White, G. Barna, S. Butler, B. Wise and N. Gallagher, "Methodology for robust and sensitive fault detection," in *Electrochem. Soc. Proc,* 1997, pp. 55-79.

[22] B. M. Wise, N. B. Gallagher, S. W. Butler, D. D. White Jr and G. G. Barna, "Comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process," *J. Chemometrics,* vol. 13, pp. 379-396, 1999.

[23] W. Hsu and M. Wang, "Physical discomfort among visual display terminal users in a semiconductor manufacturing company: a study of prevalence and relation to psychosocial and physical/ergonomic factors," *AIHA Journal,* vol. 64, pp. 276-282, 2003.

[24] S. Brahim-Belhouari, A. Bermak, M. Shi and P. C. Chan, "Fast and robust gas identification system using an integrated gas sensor technology and Gaussian mixture models," *Sensors Journal, IEEE,* vol. 5, pp. 1433-1444, 2005.

[25] A. Drozda-Freeman, M. Mclntyre, M. Retersdorf, C. Wooten, X. Song and A. Hesse, "The application and use of an automated spatial pattern recognition (SPR) system in the identification and solving of yield issues in semiconductor manufacturing," in *Advanced Semiconductor Manufacturing Conference, 2007. ASMC 2007. IEEE/SEMI,* 2007, pp. 302-305.

[26] S. Mahadevan and S. L. Shah, "Fault detection and diagnosis in process data using one-class support vector machines," *J. Process Control,* vol. 19, pp. 1627-1639, 2009.

[27] J. Moyne, N. Ward and P. Hawkins, "Leveraging advanced process control (APC) technology in developing predictive maintenance (PdM) systems," in *Advanced Semiconductor Manufacturing Conference (ASMC), 2012 23rd Annual SEMI,* 2012, pp. 221-226.

[28] Q. He and J. Wang, "Fault detection using the k-nearest neighbor rule semiconductor manufacturing processes," IEEE *Semiconductor Manufacturing Transactions on,* vol. 20, 2007.

[29] G. Willson and G. Hwang, "Multivariate fault detection and visualization in the semiconductor industry," 2006.

[30] J. M. Ko and C. O. Kim, "A multivariate parameter trace analysis for online fault detection in a semiconductor etch tool," 2011.

[31] T. Chen and J. Zhang, "On-line multivariate statistical monitoring of batch processes using Gaussian mixture model," *Comput. Chem. Eng.,* vol. 34, pp. 500-507, 2010.

[32] J. Yu, "Fault detection using principal components-based Gaussian mixture model for semiconductor manufacturing processes," *Semiconductor Manufacturing, IEEE Transactions on,* vol. 24, pp. 432-444, 2011.

[33] T. Li and C. Huang, "Defect spatial pattern recognition using a hybrid SOM–SVM approach in semiconductor manufacturing," *Expert Syst. Appl.,* vol. 36, pp. 374-385, 2009.

[34] E. Lapira, D. Brisset, H. Davari Ardakani, D. Siegel and J. Lee, "Wind turbine performance assessment using multi-regime modeling approach," *Renewable Energy,* 2012.

[35] D. Siegel, "Alternator Health Monitoring For Vehicle Applications," 2009.

[36] N. Shankar and Z. Zhong, "Improved segmentation of semiconductor defects using area sieves," *Mach Vision Appl,* vol. 17, pp. 1-7, 2006.

[37] S. Abdullah, A. K. Ariffin, K. E. N. CHE, M. F. Abdullah, A. Jalar and M. F. M. YUNOH, "Failure Analysis of a semiconductor packaging leadframe using the signal processing approach," *International Journal of Modern Physics B,* vol. 24, pp. 175-182, 2010.

[38] K. Jhang, H. Jang, B. Park, J. Ha, I. Park and K. Kim, "Wavelet analysis based deconvolution to improve the resolution of scanning acoustic microscope images for the inspection of thin die layer in semiconductor," *NDT E Int.,* vol. 35, pp. 549-557, 2002.

[39] R. Ganesan, T. K. Das, A. K. Sikder and A. Kumar, "Wavelet-based identification of delamination defect in CMP (Cu-Low k) using nonstationary acoustic emission signal," *Semiconductor Manufacturing, IEEE Transactions on,* vol. 16, pp. 677-685, 2003.

[40] B. E. Goodlin, D. S. Boning, H. H. Sawin and B. M. Wise, "Simultaneous fault detection and classification for semiconductor manufacturing tools," *J. Electrochem. Soc.,* vol. 150, pp. G778-G784, 2003.

[41] K. Kerdprasop and N. Kerdprasop, "Feature selection and boosting techniques to improve fault detection accuracy in the semiconductor manufacturing process," in *Proceedings of the International MultiConference of Engineers and Computer Scientists,* 2011, .

[42] E. R. S. Pierre, E. Tuv and A. Borisov, "Spatial patterns in sort wafer maps and identifying fab tool commonalities," in *Advanced Semiconductor Manufacturing Conference, 2008. ASMC 2008. IEEE/SEMI,* 2008, pp. 268-272.

[43] W. Malina, "On an extended Fisher criterion for feature selection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* pp. 611-614, 1981.

[44] S. A. Lynn, J. Ringwood and N. MacGearailt, "Global and local virtual metrology models for a plasma etch process," *Semiconductor Manufacturing, IEEE Transactions on,* vol. 25, pp. 94-103, 2012.

[45] P. Kang, D. Kim, H. Lee, S. Doh and S. Cho, "Virtual metrology for run-to-run control in semiconductor manufacturing," *Expert Syst. Appl.,* vol. 38, pp. 2508-2522, 2011.

[46] C. Chien, W. Wang and J. Cheng, "Data mining for yield enhancement in semiconductor manufacturing and an empirical study," *Expert Syst. Appl.,* vol. 33, pp. 192-198, 2007.

[47] M. Chen, C. Su, C. Hsu and Y. Liu, "Data transformation in SPC for semiconductor machinery control: A case of monitoring particles," *Int J Prod Res,* vol. 43, pp. 2759-2773, 2005.

[48] J. Yu, "Semiconductor Manufacturing Process Monitoring Using Gaussian Mixture Model and Bayesian Method With Local and Nonlocal Information," *Semiconductor Manufacturing, IEEE Transactions on,* vol. 25, pp. 480-493, 2012.

[49] K. KERDPRASOP and K. Kerdprasop, "Data preparation techniques for improving rare class prediction," in *Proceedings of the 13th WSEAS International Conference on Mathematical Methods, Computational Techniques and Intelligent Systems, and 10th WSEAS International Conference on Non-Linear Analysis, Non-Linear Systems and Chaos, and 7th WSEAS International Conference on Dynamical Systems and Control, and 11th WSEAS International Conference on Wavelet Analysis and Multirate Systems: Recent Researches in Computational Techniques, Non-Linear Systems and Control,* 2011, pp. 204-209.

[50] S. J. Hong and G. S. May, "Neural-network-based sensor fusion of optical emission and mass spectroscopy data for real-time fault detection in reactive ion etching," *Industrial Electronics, IEEE Transactions on,* vol. 52, pp. 1063-1072, 2005.

[51] H. J. Chang, D. S. Song, P. J. Kim and J. Y. Choi, "Spatiotemporal Pattern Modeling for Fault Detection and Classification in Semiconductor Manufacturing," *Semiconductor Manufacturing, IEEE Transactions on,* vol. 25, pp. 72-82, 2012.

[52] K. Kerdprasop and N. Kerdprasop, "Computational Intelligence Techniques to Fault Detection in the Semiconductor Manufacturing Process," *Applied Mechanics and Materials,* vol. 52, pp. 1171-1176, 2011.

[53] H. Senoussi, B. Chebel-Morello, M. Denaï and N. Zerhouni, "Feature selection and categorization to design reliable fault detection systems." *Prognostics and Health Management Society, 2011.,* pp. 257-266, 2011.

[54] Q. P. He and J. Wang, "Large-scale semiconductor process fault detection using a fast pattern recognition-based method," *Semiconductor Manufacturing, IEEE Transactions on,* vol. 23, pp. 194-200, 2010.

[55] (). . Available: http://www.peltarion.com/doc/index.php?title=Image:Animated_SOM_operation.gif.

[56] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics,* vol. 7, 1936.

[57] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science,* vol. 290, pp. 2323-2326, 2000.

[58] J. E. Jackson and G. S. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics,* vol. 21, pp. 341-349, 1979.

[59] D. Kim, P. Kang, S. Cho, H. Lee and S. Doh, "Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing," *Expert Syst. Appl.,* vol. 39, pp. 4075-4083, 2012.

[60] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.,* vol. 13, pp. 1443-1471, 2001.