



Faculty of Information and Communication Technology

**AN INTEGRATED PRINCIPAL COMPONENT ANALYSIS AND
WEIGHTED APRIORI-T ALGORITHM FOR IMBALANCED DATA
ROOT CAUSE ANALYSIS**

Ong Phaik Ling

Master of Science in Information and Communication Technology

2016

**AN INTEGRATED PRINCIPAL COMPONENT ANALYSIS AND WEIGHTED
APRIORI-T ALGORITHM FOR IMBALANCED DATA ROOT CAUSE
ANALYSIS**

ONG PHAIK LING

**A thesis submitted
in fulfillment of the requirements for the degree of Master of Science in Information
and Communication Technology**

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2016

DECLARATION

I declare that this thesis entitled “An Integrated Principal Component Analysis and Weighted Apriori-T Algorithm for Imbalanced Data Root Cause Analysis” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature :

Name :

Date :

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in term of scope and quality for the award of Master of Science in Information and Communication Technology.

Signature :

Supervisor Name :

Date :

DEDICATION

To my beloved parents, Mr. Ong Beng San and Mrs. Lim Sew Lean, your love and support are my greatest inspiration upon accomplish this study.

To my dearest supervisors, Associate Professor Dr. Choo Yun Huoy and Associate Professor Dr. Azah Kamilah Muda for being responsible, receptive and always by my side to encourage and motivate me.

To my dear friend, especially Liew Siaw Hong for your support and motivation throughout this study.

ABSTRACT

Root Cause Analysis (RCA) is often used in manufacturing analysis to prevent the reoccurrence of undesired events. Association rule mining (ARM) was introduced in RCA to extract frequently occur patterns, interesting correlations, associations or casual structures among items in the database. However, frequent pattern mining (FPM) using Apriori-like algorithms and support-confidence framework suffers from the myth of rare item problem in nature. This has greatly reduced the performance of RCA, especially in manufacturing domain, where existence of imbalanced data is a norm in a production plant. In addition, exponential growth of data causes high computational costs in Apriori-like algorithms. Hence, this research aims to propose a two stage FPM, integrating Principal Component Analysis (PCA) and Weighted Apriori-T (PCA-WAT) algorithm to address these problems. PCA is used to generate item weight by considering maximally distributed covariance to normalise the effect of rare items. Using PCA, significant rare item will have a higher weight while less significant high occurrence item will have a lower weight. On the other hand, Apriori-T with indexing enumeration tree is used for low cost FPM. A semiconductor manufacturing case study with Work In Progress data and true alarm data is used to proof the proposed algorithm. The proposed PCA-WAT algorithm is benchmarked with the Apriori and Apriori-T algorithms. Comparison analysis on weighted support has been performed to evaluate the capability of PCA in normalising item's support value. The experimental results have proven that PCA is able to normalise the item support value and reduce the influence of imbalance data in FPM. Both quality and performance measure are used as performance measurement. The quality measures aim to compare the frequent itemsets and interesting rules generated across different support and confidence thresholds, ranging from 5% to 20%, and 10% to 90% respectively. The rules validation involves a business analyst from the related field. The domain expert has verified that the generated rules are able to explain the contributing factors towards failure analysis. However, significant rare rules are not easily discovered because the normalised weighted support values are generally lower compared to the original support values. The performance measures aim to compare the execution time in second (s) and the execution Random Access Memory (RAM) in megabyte (MB). The experiment results proven that the implementation of Apriori-T has lowered the computational cost by at least 90% of computation time and 35.33% of computation RAM as compared to Apriori. The primary contribution of this study is to propose a two-stage FPM to perform RCA in manufacturing domain with the existence of imbalanced dataset. In conclusion, the proposed algorithm is able to overcome the rare item issue by implementing covariance based support value normalization and high computational costs issue by implementing indexing enumeration tree structure. Future work of this study should focus on rule interpretation to generate more human understandable rule by novice in data mining. In addition, suitable support and confidence thresholds are needed after the normalisation process to better discover the significant rare itemset.

ABSTRAK

Analisis punca (RCA) selalu digunakan dalam analisa pembuatan untuk mengelakkan pengulangan kejadian yang tidak diingini. Perlombongan petua sekutuan (ARM) telah diperkenalkan pada RCA untuk mendapatkan corak yang kerap berlaku, berkorelasi menarik, sekutu atau berstruktur kasual di dalam pangkalan data. Namun begitu, algoritma “frequent pattern mining” (FPM) seperti Apriori yang menggunakan “support-confidence framework” sukar mengenali item berkekerapan rendah yang penting. Ini menyebabkan prestasi RCA merosot, terutamanya di dalam bidang pembuatan yang lazim menghasilkan data tidak seimbang. Selain itu, algoritma Apriori juga mengalami masalah peninggian kos komputasi apabila data semakin berkembang. Oleh itu, kajian ini mencadangkan dua peringkat FPM yang mengintegrasikan Analisis Komponen Utama (PCA) dan Wajaran Apriori-T (WAT) algoritma untuk menyelesaikan masalah-masalah tersebut. PCA digunakan untuk menjana pemberat item bagi menormalkan pengaruh item yang berkekerapan rendah berdasarkan taburan kovarian maksimum. Dengan menggunakan PCA, itemset penting tetapi berkekerapan rendah akan mempunyai pemberat yang lebih tinggi dan sebaliknya. Sementara itu, Apriori-T dengan index pembancian pokok digunakan bagi mengurangkan kos komputasi. Data “Work In Progress” dan “true alarm” daripada industri semikonduktor pembuatan telah digunakan untuk perbandingan keupayaan algoritma-algoritma PCA-WAT, Apriori-T dan Apriori. Hasil penggunaan PCA menunjukkan bahawa pemberat item yang diperuntukkan oleh PCA dapat menormalkan nilai “support” item dan mengurangkan pengaruh data yang tidak seimbang di dalam FPM. Pengukuran prestasi dan kualiti telah digunakan sebagai ukuran prestasi dalam kajian ini. Ukuran kualiti membandingkan hasil set item berkekerapan tinggi dan petua yang menarik, merentasi pelbagai “support threshold” daripada 5%-20% dan “confidence threshold” daripada 10%-90%. Pakar bidang telah mengesahkan bahawa petua yang dihasilkan dapat menjelaskan faktor-faktor yang terlibat di dalam analisis kecacatan. Namun begitu, petua berkekerapan rendah yang penting didapati sukar dijana kerana nilai pemberat “support” telah menjadi lebih rendah berbanding yang asal selepas proses normalisasi. Ukuran prestasi membandingkan penggunaan masa (s) dan memori akses acak (Mb). Algoritma Apriori-T terbukti dapat mengurangkan sebanyak 90% penggunaan masa dan 35.33% memori berbanding dengan algoritma Apriori. Sumbangan utama kajian ini adalah cadangan FPM dua peringkat untuk set data yang tidak seimbang bagi melaksanakan RCA. Kesimpulannya, nilai “support” berdasarkan kovarian dapat meninggikan kebarangkalian penemuan itemset penting tetapi berkekerapan rendah manakala index pembancian pokok dapat mengurangkan kos komputasi. Kajian seterusnya boleh memfokus pada penafsiran hasil petua janaan yang lebih senang difahami terutama kepada pengguna bukan dalam bidang perlombongan data. Di samping itu, cadangan “threshold” yang sesuai untuk nilai “support” dan nilai “confidence” perlu dilakukan selepas proses penormalan untuk menyenangkan penemuan itemset penting tetapi berkekerapan rendah.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my main supervisor- Associate Professor Dr. Choo Yun Huoy and my co-supervisor- Associate Professor Dr.Azah Kamilah Muda from the Faculty of Information and Communication Technology Universiti Teknikal Malaysia Melaka (UTeM) for their useful comments, remarks, assistance, guidance and encouragement throughout this study. Without them, I could not have done this study successfully.

Furthermore, I would also like to take this opportunity to express my sincere acknowledgement to domain expert from semiconductor manufacturing especially Dr. Jonathan Chang, Lee Ching Foong and Jacky Tan Teck Hsiung for being supportive throughout this study.

Besides that, a special thanks to UTeM for providing scholarship- myBrainUTeM. Without the scholarship, I could not have started this study.

Last but not least, an honourable mention goes to my beloved parents, my lovely friends for their understanding and support.

TABLE OF CONTENTS

	PAGE
DECLARATION	
APPROVAL	
DEDICATION	
ABSTRACT	i
ABSTRAK	ii
ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS	iv
LIST OF TABLES	vii
LIST OF FIGURES	viii
LIST OF APPENDICES	xi
LIST OF ABBREVIATIONS	xii
LIST OF PUBLICATIONS	xv
CHAPTER	
1. INTRODUCTION	1
1.0 Overview	1
1.1 Project Background	2
1.2 Problem Statement	6
1.3 Research Questions	7
1.4 Research Objectives	8
1.5 Research Scope	8
1.6 Research Significance	9
1.7 Expected Output	9
1.8 Thesis Organization	10
1.9 Summary	11
2. LITERATURE REVIEW	12
2.0 Introduction	12
2.1 Globalization of Manufacturing	12
2.2 Root Cause Analysis	16
2.2.1 Root Cause Analysis using non-data mining technique	17
2.2.2 Root Cause Analysis using Knowledge Discovery and Data Mining technique	23
2.3 Knowledge Discovery	29
2.3.1 Knowledge Discovery Process model	30
2.3.1.1 Initial Approaches of KDP	31
2.3.1.2 Central Approaches of KDP	33
2.3.1.3 Other Approaches of KDP	34
2.4 The Importance of Data	35
2.5 Data Mining in Manufacturing	39
2.6 Association Rule Mining	42
2.6.1 Apriori-T	44
2.6.2 Current Trend and Approaches to Rare Item Mining	45
2.6.3 Weighted Association Rule Mining	50
2.6.3.1 Domain based Association Rule Mining	50
2.6.3.2 Data based Association Rule Mining	54

2.7	Principal Component Analysis as Weight Assignment	58
2.8	Summary	62
3.	RESEARCH METHODOLOGY	63
3.0	Introduction	63
3.1	Problem Situation and Solution Concept	63
3.1.1	Problem Situation	63
3.1.2	Solution Concept	64
3.2	Overview of Research Methodology	67
3.2.1	Overall Research Design	67
3.2.2	Investigation Phase	69
3.2.3	Implementation Phase	70
3.2.3.1	Preparation of Data	72
3.2.3.2	Data Mining	73
3.2.3.3	Benchmarking Analysis	74
3.3	Summary	78
4.	SEMICONDUCTOR MANUFACTURING PROCESS AND DATA PREPARATION	79
4.0	Introduction	79
4.1	Semiconductor Manufacturing Process	79
4.2	Raw Data Collection	87
4.3	Preparation of data	90
4.4	Summary	92
5.	PRINCIPAL COMPONENT ANALYSIS WEIGHTED APRIORI-T	93
5.0	Introduction	93
5.1	Research Design	93
5.2	Principal Component Analysis	94
5.2.1	Principal Component	94
5.2.2	Covariance Matrix	95
5.2.3	Eigenvector and Eigenvalue	96
5.2.4	Computing the Principal Components	96
5.2.5	Principal Component Analysis as Weight Assignment Mechanism	98
5.2.6	Principal Component Analysis Process in Determining Item Weight	103
5.3	Association Rule Mining	104
5.3.1	Itemset and Support Count	105
5.3.2	Association Rule	105
5.3.3	Association Rule Discovery	106
5.3.4	Frequent Itemset Generation	106
5.3.5	Rule Generation	108
5.4	Apriori	109
5.4.1	Frequent Itemset Generation in Apriori	110
5.4.1.1	Candidate Generation and Pruning	111
5.4.1.2	Support Counting	112
5.4.2	Rule Generation	112
5.5	Apriori-T	114

5.5.1	Tree Structure in Apriori-T	115
5.5.2	Apriori-T Algorithm	117
5.6	Weighted Apriori-T	123
5.7	Summary	124
6.	EXPERIMENTAL RESULTS AND ANALYSIS	125
6.0	Introduction	125
6.1	Weights Generated by PCA	125
6.2	Comparison Analysis of Support Measure	126
6.3	Quality Measures	128
6.3.1	Analysis of the Number of Frequent Itemset Generated	128
6.3.2	Analysis of the Number of Interesting Rule Generated	130
6.4	Performance Measures	133
6.4.1	Analysis of the Execution Time	133
6.4.2	Analysis of the Execution Memory	135
6.5	Summary	137
7.	CONCLUSION	139
7.0	Introduction	139
7.1	Observation on Strength and Shortcoming	139
7.2	Research Contribution	141
7.3	Propositions for Improvement	143
7.4	Summary	144
	REFERENCES	147
	APPENDICES	163

LIST OF TABLES

TABLE	TITLE	PAGE
2.1	Comparison of existing qualitative RCA tools based on their functionality	21
2.2	Comparison of existing quantitative RCA tools based on their functionality	22
3.1	Summary of investigation phase	70
3.2	An overall research plan	71
4.1	Example of true alarm raw data	89
4.2	Example of WIP raw data	89
4.3	Example of transformed true alarm data	91
4.4	Example of combined WIP and true alarm data	91
4.5	Example of binary form of combined WIP and true alarm data	91
5.1	An example of manufacturing alarm transaction	102
5.2	Support count by classical ARM	102
5.3	Support count by using weight generated by PCA	103
6.1	Average analysis of algorithms based on execution time (s) at different support threshold	134
6.2	Average analysis of algorithms based on execution memory (mb) at different support threshold	136

LIST OF FIGURES

FIGURE	TITLE	PAGE
2.1	Technology s-curve progression	20
2.2	Chronology diagram of RCA using data mining approach in manufacturing	24
2.3	Evolution of the KDP process model	31
2.4	Cios et al. process model	35
2.5	Technology evolution towards data mining	40
2.6	Structure of Apriori-T	45
2.7	Chronological diagram of literature review on domain based association rule mining	51
2.8	Chronological diagram of literature review on data based association rule mining	55
3.1	Research design	68
3.2	Algorithm of Apriori	75
3.3	Algorithm of Apriori-T	76
3.4	Algorithm of getting execution time	77
3.5	Algorithm of getting execution memory	78
4.1	Front end production processes	81
4.2	Czochralski process	82

4.3	5 main procedures in back-end production	82
4.4	Processes in pre-assembly	83
4.5	Processes in assembly front of line	84
4.6	Different of wire bond and flip chip package	85
4.7	Processes in assembly end of line	86
4.8	Processes in testing and scan & packing	87
5.1	Research design of PCA-WAT algorithm	94
5.2	PCA for data representation	95
5.3	Conceptual dataset showing principal components	100
5.4	Conceptual dataset rotated by its principal components	101
5.5	Block diagram of the PCA processes in determining weight	104
5.6	An itemset lattice structure	107
5.7	Counting the support of candidate itemsets structure	108
5.8	Frequent item generation of apriori algorithm	110
5.9	Lattice structure of pruning candidate rules using confidence measure	113
5.10	Rule generation of Apriori algorithm	113
5.11	Algorithm for Ap-genrules method	114
5.12	Example of data in binary form	115
5.13	Support count of data in parenthesis	116
5.14	Conceptual example of the t-tree data structure	116
5.15	Internal representation of t-tree presented in figure 5.14	116
5.16	Apriori algorithm	117
5.17	The createTotalSupportTree method	118

5.18	The CreatTtreeTopLevel method	118
5.19	The createTtreeLevel2 method	119
5.20	The createTtreeLevelN method	119
5.21	The addSupportToTtreeLevelN method and it's related addSupportToTtreeFindLevel method	120
5.22	The pruneLevelN method	121
5.23	The generateLevelN method and its related generateNextLevel method	122
6.1	Comparative analysis of support and weighted support measure	127
6.2	Comparative analysis of algorithm based on frequent itemset generated	128
6.3	Comparative analysis of algorithm based on number of rule generated at minimum support = 5%	131
6.4	Comparative analysis of algorithms based on number of rule generated at minimum support = 10%	131
6.5	Comparative analysis of algorithms based on number of rule generated at minimum support = 15%	132
6.6	Comparative analysis of algorithms based on number of rule generated at minimum support = 20%	132

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Item weights of all items in the dataset	163
B	Support and weighted support of all items in the dataset	173
C	Result of experiment	184
D	Frequent itemset	187
E	Interesting rules	210
F	Interview report	228

LIST OF ABBREVIATIONS

ARM	- Association Rule Mining
BA	- Barrier Analysis
BI	- Bayesian Interference
CA	- Change Analysis
CCA	- Common Cause Analysis
CED	- Cause-Effect Diagram
CEM	- Cause-Effect Matrix
CFA	- Causal Factor Analysis
CRISP-DM	- Cross Industry Standard Process for Data Mining
DDT	- Drill-Down Tree
DM	- Data Mining
DOE	- Design of Experiments
ECC	- Event-Causal Chart
EP	- Emerging Pattern
FPM	Frequent Pattern Mining
FMEA	- Failure Modes and Effects Analysis
FSARM	- Fixed Consequent Association Rule Mining
FTA	- Fault Tree Analysis
GP	- Genetic Programming

HITS	-	Hyperlink-Induced Topic Search
HURM	-	High Utility Rule Mining
ID	-	Interrelationship Diagram
IT	-	Information Technology
KD	-	Knowledge Discovery
KDD	-	Knowledge discovery in Databases
KDP	-	Knowledge discovery process
K-T	-	Kepner-Tregoe Process
MB	-	MegaByte
MCA	-	Multiple Correspondence Analysis
MM	-	Markov Models
MMISR	-	Mining Interesting Imperfectly Sporadic Rules
MSApriori	-	Multiple Support Apriori
PAMMS	-	Probability Apriori Multiple Minimum Support
PCA	-	Principal Component Analysis
PCA-WAT		Principal Component Analysis Weighted Apriori-T
PM	-	Process Map
R&D	-	Research and Development
RBFN	-	Radial Basis Function Network
RCA	-	Root Cause Analysis
RPR	-	Rapid Problem Resolution
RSAA	-	Relative Support Apriori
S	-	Second
SL	-	Swim Lane

SPC	- Statistical Process Control
SPIM	- System Process Improvement Model
ST	- Statistical Test
TRIZ	- Theory of Inventive Problem Solving
VSM	- Value Stream Map
WARM	- Weighted Association Rules Mining
WIP	- Work In Progress

LIST OF PUBLICATIONS

Ong, P.-L., Choo, Y.-H., and Muda, A.K., 2015. A Manufacturing Failure Root Cause Analysis In Imbalance Data Set Using PCA Weighted Association Rule Mining. *Jurnal Teknologi*, 77 (18), pp.103–111.

CHAPTER 1

INTRODUCTION

1.0 Overview

Root Cause Analysis (RCA) is a problem solving method which is used to identify the root causes of problems or faults that cause operating events (Rooney and Heuvel, 2004; Doggett, 2005). Apriori which is data mining technique in Association Rule Mining is introduced as a solution to perform RCA in this study. Although Apriori is proven outstanding in many domain applications, the existence of imbalanced dataset and exponential growth of data in real world application, for example in manufacturing domain, causes Apriori to be inefficient in performing RCA. Many existing techniques have been proposed to overcome the limitation of classical Apriori in imbalanced dataset such as Weighted Apriori, Multi Support Apriori, Adaptive Apriori and etcetera (Koh and Nathan, 2009). Among the proposed techniques, Weighted Apriori is one of the widely used techniques to replace the classical Apriori (Pisalpanus, 2012). However, problem arises on finding a suitable weight assignment method to replace item weight in Weighted Apriori. Therefore, Principal Component Analysis with proven ability to produce reliable weight is proposed. Besides that, the computation cost in Apriori that is proportional with the size of dataset urges the need to implement Apriori-T with proven to have lower computation cost in RCA.

1.1 Project Background

The dawn of the industrial revolution has affected industries in many countries to be transformative. Earlier researches (James et al., 2012; Tohmatsu, 2012; Hausmann and Hidalgo, 2014) confirm that manufacturing has been playing an important role in rising living, creation of high-value job and the growth of economy to nation. Therefore, most of the countries have intensified their effort in building a leading manufacturing field. As a result, the nature of competition between emerging nations, developed nations and between companies have changed. The rapid rise in productive knowledge or the know-how of manufacturing combined with rapidly developing new markets has intensified the competition for both the resources and capabilities necessary for success (Tohmatsu, 2012). Moreover, the tight financial margin that differentiate between success and failure has made manufacturing into a very competitive environment (Choudhary et al., 2009). In the market full with competition, achieving zero-defect products in manufacturing becomes a necessity. It is a common practice for manufacturing to minimize and reduce the number of defects and errors in a process (Wang, 2013).

Every failure or defect happens for a number of reasons and there is a definite progression of actions and consequences that lead to a failure (Rooney and Heuvel, 2004). According to Vorley (2008), organization often responds to causal factor with short term solutions. Although these short term solutions might help to resolve corresponding problem but constantly rely on quick fixes that require staff to repeat the same task over and over is not an ideal and effective solution (Vorley, 2008). In other word, removing causal factor is not a long term solution as it does not prevent recurrence for the problem. Therefore, quickly identifying root cause machine-sets, the most likely sources of defective products, that causes a low yield situation in a regular manufacturing process has become an essential issues (Chen et al., 2005). According to Dew (1991) and Sproull and Sproull

(2001), identifying and eliminating root cause is of utmost importance. The root cause is defined as the fundamental failure or breakdown of a process which when resolved, can prevent the occurrence of the problem (Rokach and Hutter, 2012; Dalal and Chhillar, 2013). Unfortunately, root cause analysis (RCA) is a very challenging task especially in large scaled dataset (Rokach and Hutter, 2012).

The advancement of information technology and sensor technology intensify the RCA as most of the manufacturing companies, regardless sizes, usually operate in data-rich environments (Choudhary et al., 2009; He et al., 2009). The huge volume of high dimensional data in manufacturing databases make manual or statistical analysis of data impractical (Fayyad and Uthurusamy, 1996; Wang and McGreavy, 1998; Keqin et al., 2007; Choudhary et al., 2009). Consequently lead to a situation of "rich data but poor information" (Cios et al., 1998a; Wang and McGreavy, 1998). Furthermore, Polczynski and Kochanski (2010) illustrated non-data mining techniques as technology which are believed to produce diminishing returns in respect to the growth of data (Polczynski and Kochanski, 2010).

Besides that, 25 existing non data mining RCA tools were identified and examined on their relation to the different behaviour of RCA (Yuniarto, 2012). The findings concluded that the existing RCA tools only pinpoint the specific causes and do not assist in understanding of problem-causation despite of the ability to explore reasonably causes, identify special cause variation and address hard issues (Yuniarto, 2012). Existing RCA tools are also lack of system perspective, their failure in capturing non-linear causal mechanism restricts them in finding a single absolute cause which ignore interrelatedness among causal factor added to the failure of RCA (Yuniarto, 2012). In addition, existing RCA tools which only addressed hard issues and neglected soft issues reflect that existing RCA tools inadequate in capturing whole picture of a problem (Yuniarto, 2012). As a

result, there is a need to discover knowledge from data using more efficient way which is intelligent and automated data analysis methodologies.

Knowledge discovery in databases and data mining (DM) have therefore become extremely important tool in realizing the root cause of the manufacturing problem. With the growth of data mining technology, researchers and practitioners in various aspects of manufacturing have started applying data mining to search for hidden relationships or patterns which might be used to equip their system with new knowledge (Choudhary et al., 2009). In 2006, (Choudhary et al., 2009) clearly indicated the potential scope of data mining in manufacturing to achieve competitive advantages. Besides that, (Polczynski and Kochanski, 2010) knowledge discovery and data mining has emerged as a replacement technology to the non-data mining techniques. Association rule mining (ARM) is a data mining technique for discovering interesting correlations, frequent patterns, association or casual structure among sets of item in a given dataset and normally expressed in the form of association rule. Using ARM algorithm to capture frequent pattern in industrial processes can provide useful knowledge to explain industrial failure and consequently aid in RCA (Martínez-de-Pisón et al., 2012).

Most of the ARM implementations adopt classical Apriori-like approach (Agrawal and Srikant, 1994) to generate interesting rules from frequent patterns mining using the support-confidence framework. Support is a measure on how frequently the item appears in the dataset while confidence is a measure on how strong is the rules generated. Although Apriori has been widely used in many domains, but, the existence of imbalanced data in manufacturing use cases has caused the classical Apriori algorithm fail to extract interesting patterns efficiently. Imbalanced data in manufacturing is normal as batches that passes inspection test are far more than batches that are fail. Besides that, number of errors happen in critical process are far more than other process also lead to the data imbalance