VALIDATING CAUSAL EFFECT RULES USING CLUSTER-BASED COHORT STUDY

MUHAMMAD NAUFAL IMAN

MASTER OF COMPUTER SCIENCE
(SOFTWARE ENGINEERING AND INTELLIGENCE)

2017

Faculty of Information and Communication Technology

VALIDATING CAUSAL EFFECT RULES USING CLUSTER-BASED COHORT STUDY

Muhammad Naufal Iman

Master of Computer Science in Software Engineering and Intelligence

2017

VALIDATING CAUSAL EFFECT RULES USING CLUSTER-BASED COHORT STUDY

MUHAMMAD NAUFAL IMAN

**A thesis submitted
in fulfillment of the requirements for the degree of Master of Computer Science
in Software Engineering and Intelligence**

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

DECLARATION

I declare that this thesis entitled “Validating Causal Effect Rules Using Cluster-Based Cohort Study” 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 : Muhammad Naufal Iman

Date : …………………………

APPROVAL

I hereby declare that I have read this dissertation / report and in my opinion this dissertation / report is sufficient in terms of scope and quality as a partial fulfillment of Master of Computer Science (Software Engineering and Intelligence).

Signature : ………………………………….

Supervisor Name : Assoc. Prof. Dr. Choo Yun Huoy

Date : ………………………………….

DEDICATION

To my beloved father and mother

“Bapak Supriyono and Ibu Wunipah”

To my beloved sister

“Yan Melasari”

To my beloved family

“Keluarga Besar Alm. Bpk. Tjarmo’in and Keluarga Besar Alm. Bpk. Karil Prawiro Sukarto”

To all my beloved friends

“Teman-teman FT Udinus, Teman-teman BKB69 and Konco Saklawase”

To my beloved country

“Indonesia”

ABSTRACT

Mining association rules from massive amount of data in the database is interesting for many industries especially for root cause analysis. Many techniques have been introduced to identify causal effect root cause using association rules mining framework, such as the support-confidence and support-lift framework. However, verifying and validating causal effect root causes usually involve an expert from the business domain. This has increased the complexity and time taken in the rule mining process. Hence, this study proposed the use of cohort study approach to statistically verify the generated causal effect root cause by Apriori association rules mining technique. The study follows the experimental methodology in testing and validating the proposed cohort study approach. The project had also studied on the partitioning technique in cohort study approach. The proposed cluster-based partitioning technique using k-mean clustering was compared with the manual partitioning technique through experimental results analysis. The data used in the experiments were taken from a semi-conductor manufacturer in Melaka. The data involve true alarm of failure detection collected from the business intelligence reporting unit. The results have shown positive results on root cause validation using k-mean partitioning cohort study. The manual partitioning cohort study has generated 107 rules while the k-means partitioned cohort study produced 49 rules. Only 8 rules appeared in both approached. Thus, we can conclude that the 8 rules generated by both approaches are definite causal effect rules, while the others are to be further confirmed by domain expert. In summary, cohort study approach can be used for validating a causal effect rules to a certain extend. Manual partitioning to create different cohort data can be done only if there is sufficient knowledge about the data. In the other hand, K-Means clustering technique can be used to partition the raw data into different cohorts for further validation. The limitation of this work lies on the validation of generated root causes with the domain expert due to time constraints. So, the future work in this study should focus on the domain expert validation. Besides, the use of lift standardization and thresholding should also be concerned for it is believed to be able to improve the results of generated causal effect rules.

ABSTRAK

Perlombongan peraturan sekutuan daripada jumlah data yang besar di dalam pangkalan data adalah menarik kepada pelbagai industri terutamanya untuk tujuan penganalisa punca. Pelbagai teknik telah diperkenalkan untuk mengenal pasti punca akar dengan menggunakan rangka kerja perlombongan peraturan sekutuan, sebagai contoh rangka kerja support-confidence dan support-lift. Walau bagaimanapun, mengesahkan dan mengesahkan kesan sebab serta akibat punca biasanya melibatkan pakar dari domain perniagaan. Ini telah merumitkan dan meningkatkan masa yang diambil utnuk memproses perlombongan peraturan. Oleh itu, kajian ini mencadangkan penggunaan pendekatan kajian kohort dariapda analisa statistik untuk mengesahkan punca akar yang dihasilakn oleh teknik perlombongan persatuan sekutuan Apriori. Kajian ini adalah berdasarkan kepada kaedah eksperimen untuk pengujian dan pengesahan pendekatan kajian kohort yang dicadangkan ini. Projek ini juga telah mengkaji teknik pembahagian dalam pendekatan kajian kohort. Teknik pembahagian berdasarkan konsep kelompok k-means telah dibandingkan dengan teknik pembahagian manual di dalam penganalisa hasil eksperimen. Data yang digunakan dalam eksperimen ini adalah data dari sebuah kilang pembuatan semi-konduktor di Melaka. Data yang digunakan merupakan data true alarm daripada hasil pengesan kerosakan yang dikumpulkan untuk tujuan pelaporan unit kecerdasan perniagaan. Keputusan telah menunjukkan hasil yang positif kepada pendekatan kajian kohort menggunakan teknik pengelompokan k-Means. Sebanyak 107 peraturan telah dihasilkan dengan menggunakan pembahagian manual di dalam kajian kohort, manakala 49 peraturan telah dihasilakn oleh kaedah kajian kohort berdasarkan pembahagian pengkelompok k-Means. Di antaranya, hanya sebanyak 8 peraturan yang sama telah dihasilkan oleh kedua-dua kaedah ini.Maka, kita boleh membuat kesimpulan bahawa 8 peraturan yang sama yang dihasilkan oleh kedua-dua kaedah tersebut adalah peraturan punca akar yang sebenar, manakala, peraturan-peraturan yang lain perlu disemak kesahihannya oleh pakar bidang. Kesimpulannya, kaedah kajian kohort boleh digunakan untuk mengesahkan peraturan punca akar sehingga tahap tertentu. Pembahagian manual untuk menghasilkan data kohort yang berlainan boleh digunakan sekiranya terdapat cukup pengetahuan terhadap data yang ingin dibuat pembahagian. Sebaliknya, kaedah pengkelompok k-Means boelh digunakan untuk membahagikan data asal kepada beberapa kohort untuk pengesahan yang seterusnya. Kekangan masa menyebabkan pengesahan peraturan bersama pakar bidang tidak dapat dijalankan. Selain daripada itu, penggunaan lift yang selaras serta thresholding perlu diambil perhatian kerana ianya dipercayai dapat memperbaiki keputusan peraturan punca akar yang dihasilkan.

ACKNOWLEDGEMENTS

Alhamdullilahi Robbil'Alamin, first of all, I would like to express my deepest praise and gratitude to Allah SWT who has given his blessing and mercy to complete this thesis. Second, I would like to express my deepest appreciation to my beloved supervisor Associate Professor Dr. Choo Yun Huoy for her guidance, suggestions, advices and encouragement throughout this study. I also would like to express my appreciation to Dean Faculty Information and Communication Technology, Prof. Dr. Burairah Bin Hussin and also Prof. Nanna Suryana Herman for their support and help towards my postgraduate study. I am especially grateful to Ministry of Education and Culture Indonesia, Ministry of Research Technology and Higher Education Indonesia and also Dian Nuswantoro Universisty (UDINUS) for providing the scholarship Beasiswa Unggulan by which I could have stated this master studying in UTeM Malaysia. Last but not least, thanks to my all friends, Syaifur, Tosan, Maulana, Galuh, Wendy, Anisa, Masduki, Aditya. Finally, my deepest gratitude goes to my parents and family for their love, prayers and encouragement. To those who indirectly contributed to this research, my proud thanks for their kindness.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1. Research Background

Root cause analysis is a design process to investigate and to categorize an event. Simply stated, Root cause analysis is a designing tool to help people to identify not only what and how an event occurred but also why it happened (Rooney and Van den Heuvel, 2004). According to Dalal and Chhillar (2013), root cause analysis (RCA) is the one of the popular identifying process method today, it can give some correction when we do a wrong identification and give solution such as preventive action to avoid the mistake occurrences in the future. So, from this point, we can get the main cause of why the manufacturing cannot get the maximum production by using Root Cause Analysis and get the effect of the causal by Association rule mining. Therefore, it requires an analysis of the data and their relation using association rule mining such as the data root problem, effect problems, etc. Then, those all parts are need to be united and assembled with apriori algorithm.

Association rules mining is a one of data mining popular technique, ARM also a technique to investigate the relationships of items and attributes in data. Association mining produces interpretable and actionable results, in the form of item sets or rules, with computed values of interestingness measurement, such as support, confidence and lift(Nur, 2011).

Mining an association rules from huge data becomes interesting for many industries which can help in decision making process in business strategy. The techniques for discovering association rules from the data have traditionally focused on identifying relationships between items, which show some aspects of human behavior e.g. buying behavior for determining items that customers buy together(Kumar and Chezian, 2012). According to Cui et al. (2005), clustering data is the process grouping with similar characteristic and information into several clusters. Clustering algorithms can be applied into many case problems such as data analysis exploratory in data mining in which the implementation of the problem machine learning has an objective like reducing the computation, training process complexity, and even improving the performance.

One of the algorithm in applying data mining technique for association rules mining in root cause analysis is apriori. It is a basic algorithm for the determination of frequent itemset for association rules in the type of boolean. The support and confidence as the benchmark and measurement for association rules mining is one of the association analysis stage that gives an effect to attract many researchers to produce and identify it. The support is a useful information of the corresponding itemset that appears in the dataset. The confidence is to measure statistical condition of the interestingness of the association rule. As the name of techniques mentioned above, apriori algorithm uses the prior knowledge properties which process more information in order to find a frequent itemset.

The most commonly used to find an importance of a rule in association rules mining uses support and confidence measurement. According to Sahoo et al (2015), the support is a useful information of the corresponding itemset that appears in the dataset. The confidence is to measure statistical condition of the interestingness of the association rule. Association rules mining algorithm especially apriori uses support and confidence standard of threshold to produce a large number of rules in association rules mining. The results sometimes do not seem attractive to the researcers. An association rules generation will be valid if it is satisfied with measure of evaluation besides the support and confident. The evaluation process in association rule is needed to make and handle the validation of the result by its interestingness, while the function of lift will fulfil that evaluation measurement (Birant, 2010). Lift is one of the other measure in determining X → Y.

After the causal effect produced by using several steps such as using root cause analysis with association rules mining for the rules generation and lift for the interesting of the rules, we can analyze the manufacturing data and need a validation of root cause analysis in order to make the result of root cause trusted. Nowadays, most of validation techniques are from the people which are the expert of some field. Root cause analysis rules are commonly checked by the experts to make it valid, but an expert usually has a subjective review on some problem. Here, the rules of root cause analysis cannot be used as a good benchmark in which each expert may has a different review for a generation rule. From this background problem, we need an objective validation for root cause analysis rules generation. In this study, the author will use cohort study as validation models and make an analysis and comparison among cohort itself to produce a valid root cause analysis rules.

1.2. Problem Statement

Based on background of the study, issue raised in this study is how to propose the true causal effect of the problem by using validation methods with K-Means clustering method. This study will employ the cohort studies validation method to make the valid and objective root cause analysis rules. Since there are many cases on other study using an expert to validate the rules from root cause analysis but the expert commonly has a subjective review on some problems. Thus, the rules of root cause analysis cannot be applied as a good benchmark in which each expert may has a different review for a generation rule. From this validation method, the root cause analysis rules generation can be used for causal effect of the problem. Because there is no standard for root cause analysis rules validation model, so this study is necessary to propose the model of validation to analyze the problem of causal effect of failure production data. Then, partitioning approach is needed in cohort study and, in this case, uses semi-conductor manufacturing data in Malacca as a sample. It also uses root cause analysis with the association mining to get the success production in manufacturing by using some indicators and implementing these rules.

1.3. Research Question

Based on the background and the problem statement described above, the research question in this study are as follows:

i. How to embed rules generation from root cause analysis to analyze and validate the causal effect?

ii. How to apply association rules mining and causal effect to create root cause analysis?

iii. How to determine K-Means clustering method in cohort study to find RCA?

1.4. Research Objective

Based on the problem stated, research objectives for this research can be derived as follows:

i. To propose cohort study as causal effects validation method.

ii. To propose the K-Means clustering method for cohort partitioning.

iii. To validate the proposed K-Means clustering based cohort study using manufacturing false alarm detection data.

1.5. Research Contribution

The contributions of this study will produce:

i. A validation method for root cause analysis.

ii. A cohort partitioning method based on K-Means clustering.

1.6. Research Scope and Limitation

Based on problem statement above, the problem needs to be restricted. Limitations of the problem in this research are as follows:

i. Data used in this study from semi-conductor manufacturing data in Melaka.

ii. Using failure alarm production data.

iii. This study uses Root Cause Analysis and Association Rules Mining.

iv. Analysis a model of causal effect validation.

v. Using K-Means for cluster-based in this study.

vi. The language used is python.

1.7. Organization of the Thesis

This research provides six chapters for this project report. The report is structured as follows:

Chapter 1

The content of chapter 1 is the introduction of this research. This chapter provides information of main idea about what the researcher will do in this study. Moreover, it also provides a brief outline of research background, problem statement, research question, research objective, scope of study, as well as project report review and chapter summary.

Chapter 2

Chapter 2 is the literature review of this study. This chapter will provide related literature and studies on the research problem. Furthermore, it also describes about some of theoretical aspects such as Root Cause Analysis, Causal Effect Rules Discovery, RCA Validation, RCA in Manufacturing, Technology and chapter summary.

Chapter 3

The content of chapter 3 is the research methodology. This chapter will discuss about the methodology used to answer the research question and achieve the research objectives, including the problem situation and solution, data exploration, overview of the research, research design and the proposed methodology.

Chapter 4

Chapter 4 is the cluster-based cohort techniques. This chapter gives a brief description about the general work of this study to solve the problem. It is the resume from all methods used in this study.

Chapter 5

Chapter 5 is the experiment result and discussion. This chapter will discuss about the experiment to solve the problem and the experiment result using method which is discussed in chapter 4.

Chapter 6

Chapter 6 is the conclusion of this study including research background, problem study, proposed method used, experiments and discussion result of this study. Furthermore, the limitations as well as the suggestion for future work are also provided.

1.8. Chapter Summary

This chapter describes the background of the study which is related to the root cause analysis and association mining to detect the true causal effect. The problem statement described in this research is analyzing root cause to improve manufacturing production. The objective of this study is to propose them as causal effect and validate it using causal effect validation methods to solve the problem statement and to find the true causal effect, k-means for root cause analysis and union of root cause using cohort study. This chapter also provides the outline of the research question, the scope of the study and also the research contribution.