Jurnal Teknologi

Full Paper

A MANUFACTURING FAILURE ROOT CAUSE ANALYSIS IN IMBALANCE DATA SET USING PCA WEIGHTED ASSOCIATION RULE MINING

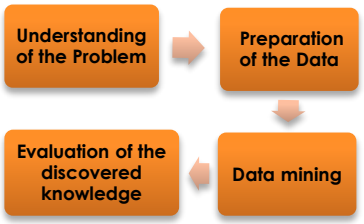
Phaik-Ling Ong, Yun-Huoy Choo*, Azah Kamilah Muda

Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia

*Corresponding author
huoy@utem.edu.my

Article history
Received
15 May 2015
Received in revised form
1 July 2015
Accepted
11 August 2015

Graphical abstract



Abstract

Root cause analysis is key issue for manufacturing processes. It has been a very challenging problem due to the increasing level of complexity and huge number of operational aspects in manufacturing systems. Association rule mining (ARM) which aids in root cause analysis was introduced to extract interesting correlations, frequent patterns, associations or casual structures among items in the transactional database. Although ARM was proven outstanding in many application domains, not many researches were focusing on solving rare items problem in imbalance dataset. The existence of imbalanced dataset in manufacturing environment make the classical ARM fails to extract interesting pattern in an efficient way. Weighted association rule mining (WARM) overcomes the rare items problem by assigning weights to items. The goal of using weighted support is to make use of the weight in the mining process and priorities the selection of the selection of targeted itemsets according to their significance, rather than frequency alone. However, the development of a suitable weight assignment scheme has been an important issue. In this research, we proposed principal component analysis (PCA) to automate the weight in WARM. The result shows that PCA-WARM is capable in capturing pattern from the data of industrial process. These patterns are proven able to explain industrial failure.

Keywords: ARM, root cause analysis, WARM, PCA

1.0 INTRODUCTION

The dawn of the industrial revolution has affected both companies and countries to be transformative. At earlier stage, tremendous prosperity and profitability can be achieved with power. However, after World Wide II, the economy has been driven by manufacturing resulting in creation of high paid middle class job. More recently, the widespread growth of digital information, proliferation of bilateral and multilateral trade agreements, physical and financial infrastructure and significant change in geopolitical relations between East and West have contributed to the rapid globalization of manufacturing [1]. Earlier research [1–3] 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 [1]. Moreover, the tight financial determines what happened, why it happened and how to eliminate it

so it will not happen again. Root cause analysis (RCA) for quality-related problem is a key issue in quality and productivity improvement in manufacturing [4]. RCA is a process of analysis to define the problem, understand the causal mechanism underlying transition from desirable to undesirable condition, and to identify the root cause of problem in order to keep the problem from recurring [5]. Unfortunately, root cause analysis is a very challenging problem [4].

The advancement of information technology and sensor technology make RCA process even harder as most of the manufacturing companies, regardless sizes, usually operate in data-rich environments [5-6]. It is because huge volume of high dimensional data in manufacturing databases make manual or statistical analysis of data impractical [6-7]. Consequently lead to a situation of "rich data but poor information", also known as drowning in data yet starving for knowledge although we live in an information age [8]. As a result, there is a need to discover knowledge from data using more efficient way which is intelligent and automated data analysis methodologies.

The rest of the paper is organized as follows. In the next section, we discussed literature review of RCA using traditional method. Drawbacks of RCA using traditional methods were pinpointed and lead to RCA using knowledge discovery. Later in section 3, a thorough literature review on data mining, ARM and WARM has been done. Experimental processes of this paper were being discussed in section 4. Experimental results were displayed in section 5 and lastly followed by conclusion in section 6.

2.0 ROOT CAUSE ANALYSIS USING TRADITIONAL METHOD

Various methods have been implemented by organizations in performing RCA to achieve zero defect manufacturing. Based on extensive literature review, some existing RCA tools were identified and examined on their relation to the different behavior of RCA [5]. The existing RCA tools that have been studied are Cause-Effect Diagram, Kepner-Tregoe Process, Fault Tree Analysis, Current Reality Tree, 5-Whys, Apollo Root Cause Analysis, Interrelationship Diagram, CATWOE, Barrier Analysis, TRIZ, System Process Improvement Model, Causal Factor Analysis , Event-Causal Analysis, Bayesian Interference, Failure Mode and Effects Analysis, Change Analysis, Rapid Problem Resolution, Common Cause Analysis, Cause-Effect Matrix, Markov Models, Drill-Down Tree, Swim Lane, Value Stream Map, Process Map and statistical test.

According to the finding, although all the studied RCA tools are able to explore reasonably causes, identify special cause variation and address hard issues, however, they are still not capable in solving problem-structuring method [5]. Before offering a solution, it is believe that a RCA tools need to be able to assist in structuring the problem in order to further assist in understanding of the causation of problem. [5] stated that it is more important to have a comprehensive understanding of why the problems occur than just pinpointing a specific root cause.

Besides that, existing RCA tools are lack of system perspective [5]. Their failure in capturing non-linear causal mechanism restricts them in finding a single absolute cause and thus added to the myth of RCA [5]. Since the existing RCA tools incapable in observing non-linear relationship, hence, the interrelatedness among causal factor also cannot be considered [5]. In addition, existing RCA tools which only address hard issues and neglected soft issues reflect that existing RCA tools inadequate in capturing whole picture of a problem [5].

Statistical Process Control (SPC) [11-12] and Design of Experiments (DOE) [13] are very common statistical methods to detect and analyze variations in manufacturing process. Unfortunately, the extensive use of SPC and DOE do not ensure high yields at the end of the process. Although SPC and DOE are necessary for the control of inline parameter, however, they are insufficient for reducing defect as one can observe significant yield losses [4].

With the advancement of information technology and sensor technology, most of the manufacturing companies, regardless sizes, usually operate in data-rich environments [5-6]. These data are usually huge and high dimensional. Consequently, It is impractical to use traditional RCA tools as most of the traditional methods are theoretical tool that required man force to interpret the causes manually [8-9,14]. Furthermore, [15] illustrated traditional tool as technology which follow a classic technology s-curve as shown in Figure 1. Traditional analysis approaches are believed to produce diminishing returns(the tail of the technology s-curve) in respect to the growth in available data [15]. Therefore, the inability of these traditional methods to sufficiently reduce defects necessitates the search for more efficient methods.

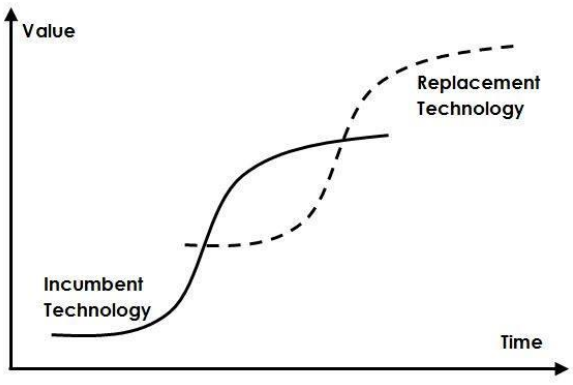


Figure 1 Technology s-curve progression

2.1 Root Cause Analysis using Knowledge Discovery

Knowledge discovery (KD) refers to a nontrivial multi-step process to extract valid, novel, potentially useful and ultimately understandable pattern from data [10,16-17]where later the pattern can be identified in the form of useful or interesting knowledge [8]. KD often associated with databases where it first interact with databases, carries out a search of patterns or relationships and finally produces pieces of meaningful knowledge [8]. Figure 2 shows the process of knowledge discovery.

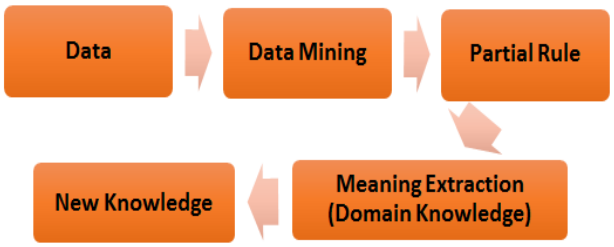


Figure 2 Process of knowledge discovery

2.1.1 Data

Data are defined as "facts and statistics collected together for reference or analysis" according to the New Oxford American Dictionary. More recently, data are used to refer to fact that are stored and shared electronically, in database or other computer applications [18]. In today's world, each piece of data is important especially in the field of manufacturing where data are treated as product. These data may be related to planning and control, design, performances, products, machine, maintenance, logistic, inventories, materials, assembly and might include the associations, trends, patterns and dependencies of the company, process and product and others [8,19].

Data are one of the important steps in KD. In a black box process, computation is data manipulation. In simpler way, the input data will be transforming into output data. In other word, low quality of data will affect the performance of DM and indirectly affect the KD. In addition, aaccording to [20], the quality of decision can be affected by the quality of data. Hence, the quality of quality of data mining (DM) is highly dependant on the quality of training data [21-23]. The electronically processed data can be perfect which mean that they are precise, integrally known and correct. Otherwise, they are imperfect. Imperfection in data can be broadly classify into three independent dimensions which are imprecision, uncertainty and incoherence [20]. Since quality of data will affect the quality of data mining, therefore it is recommended to assess the data quality before implement into any research. Quality of data can be both subjectively and objectively assessed [24]. The detail of how to assess the quality of data can be found in [24].

Noise, missing values and class imbalance are the factors that influence the quality of data [21]. Among all the factors, class imbalance problem has been receive huge attention from manufacturing researcher [25–28]. In manufacturing environment, a sample of data are normally in imbalanced distribution as during inspection of product, batches that pass the quality assurance examinations are far more than batches that are fail [29]. Apart from that, cases whereby the majority of the data collected from manufacturing systems that exhibit normal operating behaviors while as the occurrence of encountering faulty operating condition is limited added to the issue of imbalance data [30].

The nature of imbalance data sets fall into two cases which are the data are naturally imbalance and the data are too expensive to obtain data of the minority class for learning [31]. In real time application such as manufacturing process, large amount of data is generated with skewed distribution [32]. A data set said to be highly skewed if sample from one class is in higher number than other [31-32]. In an imbalance data set, a class is said to be major class if the number of instances is more and vice versa [33]. Solutions to the class imbalance problem were proposed to both data or external and algorithm or internal level. At the data level, these solution include different form of sampling such as oversampling, undersampling and combination of oversampling and undersampling [33]. However, solution at data level will cause information loss. On the other hand, solution proposed at algorithm level including adjust the cost of various class, adjust probabilistic of various class, adjust decision threshold and recognition based rather than discrimination based [33]. Solutions at algorithm level were being implemented and will discuss in detail at the following.

3.0 DATA MINING

One of the most important processes in knowledge discovery is data mining. Data mining has evolved into a mainstream technology due to two complementary and yet antagonistic phenomena which are the deluge of data and starvation of knowledge as shown in Figure 3. According to [35], data mining can be counted as logical succession to information technology. The evolution of information technology initiated the data collection in 60's [35]. Later the first relational database management system was develop in 70's to store the collected data [35]. During the 80's the enhancement of data access technique with widely available language is being implement globally [35]. Consequently, the management of data is better with the development of data warehouse and decision support system.

These development aids in manipulation of data from all sources, data from multiple layer (dynamic) and analysis of data. However, the developments of data warehouse and decision support system alone do not provide a satisfactory solution to solve the problem that we are facing nowadays in the era of information age. The problem of rich data but poor information requires advanced data analysis tools led to data mining [35]. Data mining is flourished rapidly and widely applied in all areas or field ever since it was developed due to as stated below:

1. In contract to decision support system, data mining is computer driven which mean that data mining can be fully automated. It is important especially the volume of data has grown exponentially and has exceeded human ability to interpret.
2. Data mining can solve the query formulation problem.
3. Data mining confronts the visualization and understanding of large datasets efficiently.

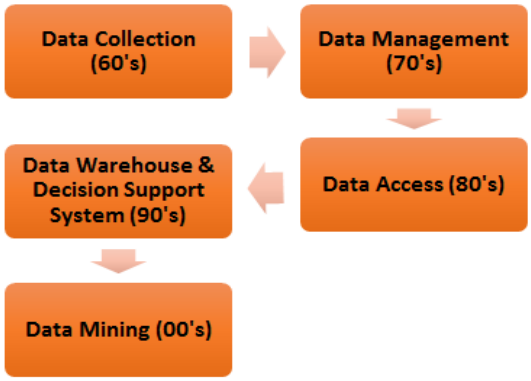


Figure 3 Technology Evolution towards Data Mining

Before data mining is implementing, it is important to identify the type of knowledge to be mined as it will determine the type of data mining. Based on the task of data mining, there are several categories of data mining technology. These technologies include anomaly detection, association rule mining (ARM), clustering, classification, regression and summarization [34-35]. However, ARM was used in this research to discover interesting relations between variables in large databases. ARM is the most important association rule mining technology to determine frequent patterns from transaction data besides generate association rules from frequent pattern [38]. It is also believed that the use of ARM in frequent pattern captured from industrial process can provide useful knowledge to explain industrial failure and hence aid in root cause analysis [39].

3.1 Association Rule Mining

ARM or also known as frequent pattern mining was first introduced in 1993 by Agrawal with the purpose of discovery interesting correlations, frequent patterns, association or casual structure among sets of item in a given transactional dataset or other data repositories [40]. The relationships among itemsets are based on the co-occurrence instead of their inherent properties. These associations are normally expressed in the form of association rule. The idea or motivation for seeking association rule came from market basket analysis which is used to examine customer behavior. Generally there are two steps involved in the process of association rule mining. The first step is to find the entire frequent pattern or itemsets that satisfy minimum support and the second step is to generate all association rules that satisfy the minimum confidence from the frequent pattern.

Most of the ARM approach adopted Apriori-like (classical) approach. There is less explored area in association mining is infrequent itemset mining. However, infrequent itemset mining is important as in real world application especially in manufacturing environment, the data are normally skewed or imbalance when dealing with defect product. With the classical Apriori-like algorithm, it is not efficient to generate all rules that are interesting with the appearance of imbalance data.

Classical Apriori-like algorithm will bias toward the majority class and ignore the minority class in the data set. In order to find rules in minority class, the threshold of minimum support have to be lowered. As result, plenty of patterns or rules will be generated. With the situation of abundant frequent patterns or rules, classical apriori-like algoritms may suffer from trival cost as most of the items in database would be allowed to participate in itemset generation and affected the scalability of the algorithm [40]. Besides that, it is costly to handle a huge number of candidate items when generating rules [40]. Apart from that, it is time consuming to repeatedly scan the database and check support of the candidate itemsets generated [40]. In addition, the complexity of the computation will increase exponentially with regard to the number of itemsets generated [40]. Hence, there is a need to improve classical Apriori-like algorithm to suit better when dealing with imbalanced data sets.

3.2 Weighted Association Rule Mining

Numerous efforts have been done to overcome the limitation of classical ARM. According to [41], many of the improved algorithms replace an item's support with a weighted form of support. Weighted association rule mining (WARM) was first given a formal definition by [42]. In this approach, each item is assigned a weight to reflect its importance. The higher the priority, the higher the weight assigned to the item. However, issues on how to develop a suitable weight scheme arise with the concept of WARM. There have been various approaches in dealing with finding a suitable weight scheme over the years. These efforts can be broadly categorized

into the domain approach where item weights are assigned solely on the basis of domain knowledge and the other in which item weights are inferred on the basis of interactions between items in a transactional database.

3.2.1 Domain based Weighted Association Rule Mining

Research work in [43] was among the earliest attempt to deal with traditional ARM problem. In their research, they have developed generalized item support which incorporate both item and transaction weight to obtain maximum flexibility in determining large itemset. Weights were assigned according to expert domain knowledge. They revealed that the assigned weight has the ability to bias the rule generation process to transaction with high importance. Therefore, it is proven that WARM able to prioritize the results accordingly.

Tao *et al.* [44] in their research pinpointed the drawback of implementing weighted schema in traditional ARM. The downward closure property is proven violated with the use of weighted support as an itemsets can consider large even though some of the subsets are not large due to the weighting schema. Therefore, they have revised the downward closure property in their research. They make use of weight for both item and itemsets. The itemsets whose weighted support is larger than threshold were considered as significant itemsets and the result is promising.

In 2004, [45]extend the traditional ARM by applying weight to reflect the interest of each item within the transaction. In their approach, they have proposed a twofold approach. They have first calculated frequent pattern itemsets without considering the weight and later weight is taken in account during process of rule generation. In the rule generation process, they partitioned the weight domain space of each frequent itemsets and identified the popular regions within the domain space. They justify that their approach not only improves the confidence of the rules, but also provides a mechanism for more effective marketing.

Research work in [46] proposed a new algorithm to allocate weight to items based on feature-value pair. In their approach, they exploit multiple correspondence analysis (MCA) to present features in a principal component space to determine relationships among categories and feature-value pairs. Subsequently, they integrate information produced by MCA with percentage of frequency counts of both negative and positive to assign weight for feature-value pairs. Their approach shown promising result compared to well-known algorithm such as decision tree, neural network, naive bayesian, support vector machine and k-nearest neighbor.

Wu and Li [47] in their work employed analytic hierarchy process (AHP) to set the weight of items by associating the degree of relative significance among different values of one attribute and the degree of relative significance among evaluation attributes. Their approach replaced artificial judgment matrix with objective judgment matrix to avoid depending too much on the subjective aspect. According to their study, domain expert has to be taken into consideration to result in accurate weight. In their approach, they are first selected two attributes and made use of subjective judgment to build judgment matrix, then different attribute are multiply by relative significance among different value of selected attribute to calculate weight. With their contribution, it enables the weight to be more understandable and flexible.

Five weighted association rule mining approaches for facility layout problem were applied in [48]. In all the approaches, each item is assigned a weight with respect of user criteria. In their study, five performance measures which are machine utilization, total amount of products produced, cycle time, transfer time and waiting time are used to evaluate the effectiveness of the proposed approach.

Recently, [49] proposed an approach base on high utility rule mining (HURM). In HURM, users are able to define their preference in term of utility function and obtain association rules. Their work also improves utility-based mechanism to tackle with large transaction database. In order to accommodate the benefit in business form, their work also allow user to design quantitatively and utilize three elements which are opportunity, effectiveness and probability. Experimental result shows the proposed approach able to provide users with greater business benefit.

Although some research show promising result with implementing domain knowledge in determining weight in WARM, however, there are a number of researchers [48,50] reveal constraint of domain based approach. According to them, although it is true that some domain research successfully utilize domain knowledge to determine weight, nevertheless, there are cases whereby information is either not available or insufficient to convert raw domain into useable item weight. Besides that, even in domain where domain knowledge are readily available, volatility of data give rise to changes in pattern where in turn reducing the effectiveness of the weight. In addition, there is a possibility whereby knowledge generated only encapsulate known pattern and thus excluding the discovery of unexpected rules. Consequently, it is believe that data driven or automated weighted association rule mining is an important tool in driving the weighted association rule mining process [52].

3.2.2 Data based Weighted Association Rule Mining

In 2008 and 2013, [50] and [53] respectively presented a novel framework in ARM to assign weight without pre-assigned value. In their studies, they converted dataset into a bipartite graph containing items and transactions. Later, weight are

calculated with the adapted HITS model [54] to rank the transaction accordingly. Experimental results shown that the computational cost of link based model are reasonable. The researchers compare the result with four databases and the result proven the proposed approach able to work well with sparse database in finding interesting pattern involving rare items. Despite of the fact, the performance was very similar to apriori according to [54].

Principal component analysis (PCA) was introduced by [52] in a weight assignment mechanism. Their idea is the degree of variation in dataset need to be considered. Eigenvectors and eigenvalue that computed from covariance matrix in PCA were used to rank item and capture the degree of variation respectively. Weights were next assigned to items based on the degree of variation across dataset. Experimental outcome indicate that summary of generated rules with high information could be generated compared to traditional Apriori.

In the same year, valency model was proposed by [51] for weight assignment problem. Their concept is based on the strength of their item connection to adjust the weight of items. They designed a valency model into two different parts. Weight of an item was judged based on its strength of connection and later calculates the weight for items. Experimentation was shown to be significantly better than Apriori. In the following year, they extended the valency model to accommodate in data stream environment [55]. As a result, their extended model was faster in term of execution time while maintaining a promising accuracy as compared to the earlier model. Apart from that, [56] and [52] revised graph based connectivity model to capture interaction between items. The extended scheme was proven able to produce cliques with significantly higher h-confidence value than traditional ARM.

This research induced weight based on PCA. PCA is a mathematical technique that has been widely implemented to determine the relationships between different items in a given dataset [52]. Detailed of PCA weighted ARM is illustrated at the following.

4.0 EXPERIMENT

This section explains the flow of the experiment. The process of the experiment contain four phases which are understanding of the problem, preparation of the data, data mining and evaluation of the discovered knowledge as shown in Figure 4.

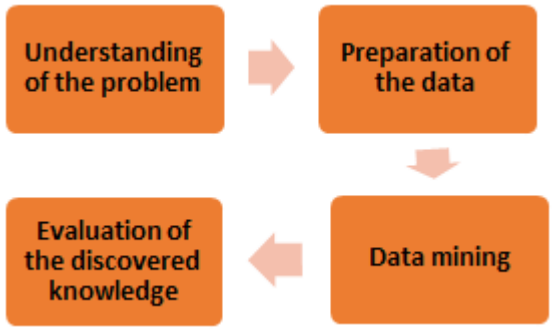


Figure 4 Block diagram of the processes of the experiment

4.1 Understanding of the Problem

This phase involves learning terminology of the specific domain. In this research, understanding of the process of manufacturing in semiconductor is vital as the case study is based on a semiconductor industry. Experts from the related field are gathered to define the problem and determine the project goals together. Current solutions that are implementing to the problem will be discussed to know the weakness of the solutions. Then, a description of the problem including its restrictions will be done. The project goals later will be transformed into data mining goal and initial selection of potential data mining technique will be done. Since the project goal is to find the pattern and correlation among attributes in the data in order to determine the root cause machine-set, the most likely sources of defective products, that cause a low yield situation in a regular manufacturing procedure, association rule mining is being proposed.

4.2 Understanding of the Data

This phase is the one of the key step upon which the success of the entire knowledge discovery process depends. This phase includes collection of sample data, and deciding which data will be needed including its format and size. The domain expert will be gathered again to verify the related data that are needed for the process of data mining. In addition, domain expert will also need to further explain each of the tables and attributes in the data set. Before it is ready to be used, data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values and others.

In this research, in order to reach the goal, data are extracted from work in progress (WIP) and true alarm in a specific time frame. WIP refers to all materials and partly finished products that are at various stages of the production process. True alarm recorded the defect products. These two data sets have been verified by domain expert. Attributes that are relevant have been extracted accordingly.

4.3 Data Mining

This phase is another key step in knowledge discovery process. This step involves usage of the planned data mining techniques and selection of the new ones or improves the planned data mining techniques. In this research, PCA weighted ARM will be used as data mining technique to find the pattern and correlation among attributes in the data. The pattern and rules generated are believed to help in determining the root cause machine-set, the most likely sources of defective products, that causes a low yield situation in a regular manufacturing procedure.

4.3.1 Weighted Association Rule Mining Fundamental

In fundamental WARM, each item or itemset is assigned a weight based on its significance. The higher the priority, the higher the weight assigned to the item. The goal of using weighted support is to make use of the weight in the mining process and priorities the selection of the targeted itemsets according to their significance in dataset rather than frequency alone.

Given a set of items, $= \{i_1, i_2 \dots \dots i_m\}$, a transaction may be defined as a subset of I and a dataset as D of a transaction. An itemset consists of a set of X. The support of X, $sup(X)$, is the proportion of transactions containing X in the dataset. An association rule is expressed in the form $X \rightarrow Y$, where $X \subset I, Y \subset I$, and $X \cap Y = \emptyset$. The rule $X \rightarrow Y$ has coverage of s in the transaction set D, if $s = sup(XY)$, where $sup(XY)$ is the joint support of X and Y. The rule $X \rightarrow Y$ holds in the transaction set D with confidence c where $c = conf(X \rightarrow Y) = sup(XY)/\ sup(X)$. In a weighted ARM, a weight w_i is assigned to each item i, where $0 \leq w_i \leq 1$ reflecting the relative importance of an item over other associated items. The weighted support of an item $i = w_i \times sup(i)$. Similar to traditional ARM, a weighted support threshold and a confidence threshold is assigned to measure the strength of an association rules produced. The weight of an itemset X is $\frac{\varepsilon_{i\epsilon} \times w_i}{|x|} \times sup(X)$.

In WARM, an itemset X is considered to be a frequent itemset if weighted support of the itemset is greater than user defined minimum weighted support (wminsup) threshold as $\frac{\varepsilon_{i\epsilon} \times w_i \times sup(X)}{|x|} \geq wminsup$. The weighted support of $X \rightarrow Y$ is the average weights of itemsets X and Y multiply by their joint support, $sup(XY)$ as $\frac{w_x + w_y}{2} \times sup(XY)$ where w_x and w_y are the weights of itemsets X and Y respectively. A rule $X \rightarrow Y$ is considered interesting if a weighted support is greater than $wminsup$ and the confidence of the rule is greater than or equal to a minimum confidence threshold, $minconf$.

4.3.2 Weighted Association Rule Mining using PCA

PCA is a statistical procedure that makes use of an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Therefore, the first principal component generated from PCA is chosen to be the weight in WARM.

Given a dataset D, the covariance of every pair of items can be expressed in term of covariance matrix M.The matrix P of all possible eigenvector can be derive from $P^{-1}MP = Q$, where Q is diagonal matrix of eigenvalue in M. The first eigenvector $E_1 = (v_1, v_2 \dots \dots v_n)$ will contain a set of numeric values v_i for each item i in the range of $[0,1]$. For a given item i, the value v_i indicates to the extent which the item i captures the variation in D with respect to first principal component. Next, eigenvectors are being modulo to compute magnitude as $V = (|v_1|, |v_2|, \dots, |v_n|)$. Then, the V will be normalised to lie in the range of $[0 \dots 1]$ and result in another vector v_1. Lastly, the weight vector is given by $W = v_1$. Figure 5 shows the step involve in PCA to generate weight to be used in WARM.

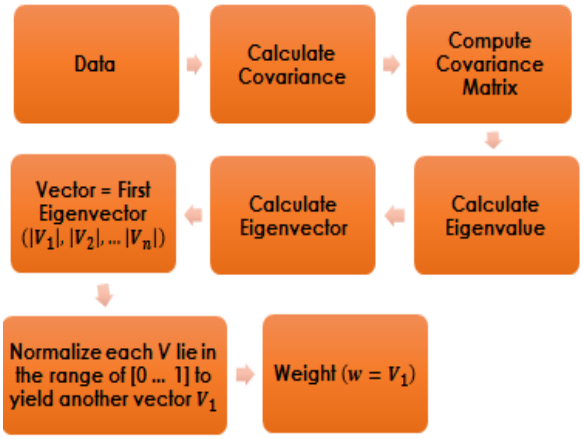


Figure 5 Block diagram of the PCA processes in determining weight

4.4 Evaluation of the Discovered Knowledge

This phase includes understanding of the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge.

5.0 RESULTS AND DISCUSSION

In this section, we report on the result of running PCA WARM using dataset from semiconductor manufacturing. The experimental data were preprocessed as discussed earlier. Figure 6 shows the

sample rules generated from the proposed approach, PCA WARM. Some of the result shown are being replaced with binary number to protect the privacy of the data that provided by company.

Sample of rules generated

1. **Product=9 ==> Error(alarm)= SPC/Inprocess Reject**
2. **Product Type=12 ==> Error(alarm)= SPC/Inprocess Reject**
3. **Process Group=18 ==> Error(alarm)= SPC/Inprocess Reject**
4. **Basic Type=27 ==> Error(alarm)= SPC/Inprocess Reject**
5. **Operator (alarm)= NANI ==> Error(alarm)= SPC/Inprocess Reject**
6. **Lot=64 ==> Error(alarm)= SPC/Inprocess Reject**
7. **Product=9, ProductType=12 ==> Error(alarm)= SPC/Inprocess Reject**
8. **Product=9, ProcessGroup=18 ==> Error(alarm)= SPC/Inprocess Reject**
9. **Product=9, ProcessClass=22 ==> Error(alarm)= SPC/Inprocess Reject**
10. **Product=9, BasicType=27 ==> Error(alarm)= SPC/Inprocess Reject**

Figure 6 Samples of rules generated

According to the result, PCA WARM is capable in capturing pattern from the data of industrial process. These patterns are able to explain industrial failure. As shown in Figure 6, we can know that product number 9 always result in SPC/Inprocess Reject error. Besides that, when operator NANI is on duty, the chances of having SPC/Inprocess Reject error is higher compared to other operators. With this information, the management can take action like provide training to the particular operator. Further analysis also can be done on why the product number 9 always results in the specific error.

6.0 CONCLUSION

In this paper, we have presented PCA-WARM in manufacturing root cause analysis. A classical ARM was adopted with weighted support to overcome the rare items problem by assigning weights to items. The goal of using weighted support is to make use of the weight in the mining process and priorities the selection of targeted itemsets according to their significance, rather than frequency alone. Therefore, it is able to find the association rule even in imbalanced dataset. As we noted earlier in the paper, there is a need to automate the weight as users may be unable to supply weight due to a number of reasons. The implementation of PCA which make use of principle component has proven successful in determining weight in WARM. In conclusion, the proposed technique is able to tackle the issue in imbalanced dataset and extract interesting correlations, frequent patterns, associations or casual structures among items in the transactional database. Patterns or association rules obtained from the proposed approach are able to explain industrial failure.

Acknowledgement

A special thanks to the semiconductor manufacturing company that provided data for this research.

References

[1] D. Tohmatsu. 2012. The Future of Manufacturing: Opportunities to Drive Economic Growth.
[2] R. Hausmann and C. Hidalgo. 2014. *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. MIT Press.
[3] M. James, S. Jeff, D. Richarcd, S. Gernot, R. Louis, M. Jan, R. Jaana, R. Charles, G. Katy, O. David, and R. Sreenivas. 2012. Manufacturing the Future : The Next Era of Global Growth and Innovation.
[4] L. Rokach and D. Hutter. 2012. Automatic Discovery of the Root Causes for Quality Drift in High Dimensionality Manufacturing Processes. *J. Intell. Manuf.* 23(5): 1915-1930.
[5] H. Yuniarto. 2012. The Shortcomings of Existing Root Cause Analysis Tools. *Proc. World Congr. Eng.* 3.
[6] S. G. He, H. Zhen, A. Wang, and L. Li. 2009. Quality Improvement using Data Mining in Manufacturing Processes. In *Data Mining and Knowledge Discovery in Real Life Applications*, no. February, J. Ponce and A. Karahoca, Eds. I-Tech Education and Publishing. 436.
[7] A. K. Choudhary, J. A. Harding, and M. K. Tiwari. 2008. Data Mining in Manufacturing: A Review Based on the Kind of Knowledge. *J. Intell. Manuf.* 20(5): 501-521.
[8] U. Fayyad and R. Uthurusamy. 1996. Data Mining and Knowledge Discovery in Databases. *Commun. ACM*. 39(11): 24-26.
[9] X. Z. Wang and C. McGreavy. 1998. Automatic Classification for Mining Process Operational Data. *Ind. Eng. Chem. Res.* 37(6): 2215–2222.
[10] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. Kurgan. 1996. *Data Mining : A knowledge Discovery Approach*. Springer.
[11] J. F. Halpin. *Zero Defects: A New Dimension in Quality Assurance*. Mc Graw-Hill.
[12] J. J. Rooney and L. N. Van den Heuvel, 2004. Root Cause Analysis for Beginners. *Qual. Prog.* 45-53.
[13] J. Soenjaya, W. Hsu, M. L. I. Lee, and T. Lee. 2005. Mining Wafer Fabrication: Framework and Challenges. In *Next Generation of Data-Mining Application*. M. M. Kantardzic and J. Zurada., Eds. New York: Wiley-IEEE Press. 17-40.
[14] W. Keqin, T. Shurong, B. Eynard, L. Roucoules, and N. Matta. 2007. Review on Application of Data Mining in Product Design and Manufacturing. In *Fuzzy Systems and Knowledge Discovery FSKD*. 4: 613-618.

[15] M. Polczynski and A. Kochanski. 2010. Knowledge Discovery and Analysis in Manufacturing. *Qual. Eng.* 22(3): 169-181.
[16] W. J. Frawley, G. Piatetsky-shapiro, and C. J. Matheus. 1992. Knowledge Discovery in Databases : An Overview. *AI Mag.* 13(3): 57-70.
[17] I. Geist. 2002. A Framework for Data Mining and KDD. In *Symposium on Applied computing*. 508-513.
[18] M. West. 2011. *Developing High Quality Data Models*. First edit. Elsevier.
[19] K.-S. Wang. 2013. Towards Zero-Defect Manufacturing (ZDM)—A Data Mining Approach. *Adv. Manuf.* 1(1): 62-74.
[20] G. C. Crisan, C. M. Pintea, and C. Chira 2012. Risk Assesment for Incoherent Data. *Environ. Eng. Manag. J.* 11(12): 2169-2174.
[21] K. Kerdprasop and N. Kerdprasop. 2011. A Data Mining Approach to Automate Fault Detection Model Development in the Semiconductor Manufacturing Process. *Int. J. Mech.* 5(4): 336-344.
[22] R. Blake and P. Mangiameli. 2011. The Effects and Interactions of Data Quality and Problem Complexity on Classification. *J. Data Inf. Qual.* 2(2): 1-28.
[23] J. Stang, T. Hartvigsen, and J. Reitan. 2010. The Effect of Data Quality on Data Mining-Improving Prediction Accuracy by Generic Data Cleansing. In *International Conference on Information Quality ICIQ*.
[24] L. L. Pipino, Y. W. Lee, and R. Y. Wang. 2002. Data Quality Assessment. *Commun. ACM*. 45(4): 211-218.
[25] H. Alhammady and K. Ramamohanarao. 2004. The Application of Emerging Patterns for Improving the Quality of Rare-Class Classification. In *Advances in Knowledge Discovery and Data Mining*, Springer Berlin Heidelberg. 207-211.
[26] G. M. Weiss. 2004. Mining with Rarity : A Unifying Framework. *SIGKDD Explor. Newsl.* 6(1): 7-19.
[27] S. Han, B. Yuan, and W. Liu. 2009. Rare Class Mining : Progress and Prospect. *Chinese Conf. Pattern Recognit.* 1-5.
[28] G. M. Weiss. 2010. Mining with Rare Cases. In *Data Mining and Knowledge Discovery Handbook*. 2nd ed. Springer US, 747-757.
[29] L. Rokach and O. Maimom. 2006. Data Mining for Improving the Quality of Manufacturing: A Feature Set Decomposition Approach. *J. Intell. Manuf.*17(3): 285-299.
[30] C. C. Teck, L. Xiang, Z. Junhong, and D. Woon. 2012. Hybrid Rebalancing Approach to Handle Imbalanced Dataset for Fault Diagnosis in Manufacturing Systems. In *I7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. 1224-1229.
[31] X. Guo, Y. Yin, C. Dong, G. Yang, and Guangtong Zhou, 2008. On the Class Imbalance Problem. *Fourth Int. Conf. Nat. Comput.* 4: 192-201.
[32] L. Rushi, S. D. Snehlata, and M. Latesh. 2013. Class Imbalance Problem in Data Mining : Review. *Int. J. Comput. Sci. Netw.* 2(1).
[33] N. V Chawla, N. Japkowicz, and K. Aleksander. 2004. Editorial : Special Issue on Learning from Imbalanced Data Sets. *ACM Sigkdd Explor. Newsl.* 6(1): 1-6.
[34] S. Wang and X. Yao. 2012. Multiclass Imbalance Problems: Analysis and Potential Solutions. *IEEE Trans. Syst. man, Cybern. Part B, Cybern.* 42(4): 1119-1130.
[35] A. Symeonidis and P. Mitkas. 2005. Data Mining and Knowledge Discovery: A Brief Overview. In *Agent Intelligence Through Data Mining*, United States: Springer.
[36] K. Mehmed. 2011. *Data Mining: Concepts, Models, Methods, and Algorithms*. Second rev. Wiley-Blackwell.
[37] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun. 2011. The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature. *Decis. Support Syst.* 50(3): 559-569.
[38] J. Wu. 2014. Interpretation of Association Rules with Multi-tier Granule Mining. Queensland University of Technology.
[39] Martínez-de-Pisón, F. Javier, E. M.-P. Andrés Sanz, J. Emilio, and C. Dante. 2012. Mining Association Rules from Time Series to Explain Failures in a Hot-Dip Galvanizing Steel Line. *Comput. Ind. Eng.* 63(1): 22-36.
[40] Yun Sing Koh and R. Nathan. 2009. Rare Association Rule Mining: An Overview. In *Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection*. Vol 3., Y. S. Koh, Ed. New York: IGI Global. 320.
[41] S. Pisalpanus. 2012. A *Landmark Model For Assigning Item Weight For Pattern Mining*. Auckland University of Technology.
[42] C. H. Cai, A. W. C. Fu, C. H. Cheng, and W. W. Kwong. 1998. Mining Association Rules with Weighted Items. *Int. Database Eng. Appl. Symp.* 68-77.
[43] G. D. Ramkumar, S. Ranka, and S. Tsur. 1997 Weighted Association Rules: Model and Algorithm. *Proc.ACM SIGKDD*. 1-13.
[44] F. Tao, F. Murtagh, and M. Farid. 2003. Weighted Association Rule Mining Using Weighted Support and Significance Framework. *Proc. ninth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.* 661-666.
[45] W. Wang, J. Yang, and P. Yu. 2004. WAR: Weighted Association Rules for Item Intensities. *Knowl. Inf. Syst.* 6(2): 203-229.
[46] L. Lin and M. L. Shyu. 2010. Weighted Association Rule Mining for Video Semantic Detection. *Int. J. Multimed. Data Eng. Manag.* 1(1): 37-54.
[47] W. Jian and L. X. Ming. 2008. An Effective Algorithm for Mining Weighted Association Rules in Telecommunication Networks. *J. Comput. 3*. 3(10): 20-27.
[48] S. Altuntas and H. Selim. 2012. Facility Layout Using Weighted Association Rule-based Data Mining Algorithms: Evaluation With Simulation. *Expert Syst. Appl.* 39(1): 3-13.
[49] D. Lee, S. H. Park, and S. Moon. 2013. Utility-based Association Rule Mining: A Marketing Solution for Cross-Selling. *Expert Syst. Appl.* 40(7): 2715-2725.
[50] K. Sun and F. Bai. 2008. Mining Weighted Association Rules Without Preassigned Weights. *IEEE Trans. Knowl. Data Eng.* 20(4): 489-495.
[51] Y. S. Koh, R. Pears, and W. Yeap. 2010. Valency Based Weighted Association Rule Mining. *Adv. Knowl. Discov. Data Min*. 274-285.
[52] R. Pears, Y. S. Koh, and G. Dobbie. 2010. EWGen : Automatic Generation of Item Weights for Weighted Association Rule Mining. *Adv. Data Min. Appl*. 36-47.
[53] M. Padmavalli and K. Sreenivasa Rao. 2013. An Efficient Interesting Weighted Association Rule Mining. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* 3(10): 1059-1064.
[54] J. M. Kleinberg. 1999. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*. 46(5): 604-632.
[55] Y. S. Koh, R. Pears, and G. Dobbie. 2011. Automatic Assignment of Item Weights for Pattern Mining on Data Streams. In *Advances in Knowledge Discovery and Data Mining*. vol. 6634. Springer Berlin Heidelberg. 387-398.
[56] R. Pears, Y. S. Koh, G. Dobbie, and W. Yeap. 2013. Weighted Association Rule Mining via a Graph Based Connectivity Model. *Inf. Sci. (Ny)*. 218: 61-84.