Data Mining to Recognize Fail Parts in Manufacturing Process

Wanida Kanarkard[1], **Danaipong Chetchotsak**[2], **Daranee Hormdee**[3], **Rod Adams**[4], and **Neil Davey**[5], Non-members

ABSTRACT

In many manufacturing processes, some key process parameters have very strong relationship with the normal or various faulty products of finished products. The abnormal changes of these process parameters could result in various categories of faulty products. In this paper, a data mining model is developed for on-line intelligent monitoring and diagnosis of the manufacturing processes. In the proposed model, an Apriori learning rules developed for monitoring the manufacturing process and recognizing faulty quality of the products being produced. In addition, this algorithm is developed to discover the causal relationship between manufacturing parameters and product quality. These extracted rules are applied for diagnosis of the manufacturing process; provide guidelines on improving the product quality.

Therefore, the data mining system provides abnormal warnings, reveals assignable cause(s), and helps operators optimally set the process parameters. The proposed model is successfully applied to an assembly line in hard disk drive process, which improves the product quality and saves manufacturing cost.

Keywords: Fault Diagnosis, Association Rule Learning, Data Mining, Hard Disk Drive Process, Manufacturing

1. INTRODUCTION

Every industrial manufacturing application requires a suitable monitoring and diagnosis system for its processes in order to identify any decrease in product quality and then provide an accurate diagnosis report for generation of fault products.

Data mining techniques have become widespread in manufacturing. Moreover, various rules may be obtained using data mining techniques, and only a small number of these rules may be selected for implementation due, at least in part, to limitations of budget and resources. Association rule mining differs from traditional machine learning techniques by permitting decision makers to pick from the many potential models that can be supported by the data (Webb & Zhang, 2005). Generally, association rule mining discovers all rules that meet certain sets of criteria or constraints, such as minimum support and minimum confidence, rather than generating a single model that best matches the data.

Evaluating the interestingness or usefulness of association rules is important in data mining. In many manufacturing applications, it is necessary to rank rules from data mining due to the number of quality rules (Tan & Kumar, 2000) and manufacturing resource constraint (Choi, Ahn, & Kim, 2005). Selecting the more valuable rules for implementation increases the possibility of success in data mining.

Furthermore, the data mining model is appropriate to be used in some manufacturing systems, where the complex and nonlinear relationship between process parameters and the quality categories of the final products is existed. To illustrate the usefulness and effectiveness of the proposed system, its application in a quality monitoring and diagnosis system of an assembly line in hard disk drive process is illustrated in this paper.

In this study, we developed an Apriori learning rule algorithm to extract rules from the manufacturing process, which are used to express the causal relationships between process parameters and product output measures. Moreover, these rules are used for monitoring abnormal parameters in manufacturing systems. The proposed system not only pinpoints the quality problems of the manufacturing process, but also reveals the causal relationships of why they occur and how to prevent them. The successful implementation of the developed methodology in an industrial case demonstrates its effectiveness and potential applicability for these manufacturing processes, where there existed the causal relationships between the process parameters and product output measures.

The other parts of this paper are organized as follows: Section 2 presents the data mining and mining association rules system. An application of the proposed system in a manufacturing system is presented in Section 3. Finally, conclusions are provided in Sec-

Manuscript received on May 27, 2008 ; revised on January 30, 2009.

[1,3] The authors are with Department of Computer Engineering, Khon Kaen University, Khon Kaen, Thailand, E-mail: wanida@kku.ac.th and darhor@kku.ac.th

[2] The author is with Department of Industrial Engineering, Khon Kaen University, Khon Kaen, Thailand , E-mail: cdanai@kku.ac.th

[4,5] The authors are with Department of Computer Science, University of Hertfordshire, Hatfield, United Kingdom , E-mail: r.g.adams@herts.ac.uk and n.davey@herts.ac.uk

tion 4.

2. DATA MINING AND MINING ASSOCIATION RULES

Data mining, also referred to as 'knowledge discovery', means the process of extracting nontrivial, implicit, previously unknown and potentially useful information from databases (Han & Kamber, 2001). Depending on the types of knowledge derived, mining approaches may be classified as finding association rules (Agrawal & Srikant, 1994; Brin, Motwani,Ullman, & Tsur, 1997; Wur & Leu, 1999), classification rules (Cheeseman & Stutz, 1996), clustering rules (Zhang, Ramakrishnan, & Livny, 1996) and others (Catledge & Pitkow, 1995; Han & Kamber, 2001). The most commonly seen is finding association rules in transaction databases.

Conceptually, an association rule indicates that the occurrence of certain items in a transaction would imply the occurrence of other items in the same transaction (Agrawal et al., 1993).

A. Apriori algorithm

The Apriori algorithm is a state of the art algorithm most of the association rule algorithms are somewhat variations of this algorithm (Agrawal et al., 1993). The Apriori algorithm works iteratively. It first finds the set of large 1-item sets, and then set of 2- itemsets, and so on. The number of scan over the transaction database is as many as the length of the maximal item set. Apriori is based on the following fact: The simple but powerful observation leads to the generation of a smaller candidate set using the set of large item sets found in the previous iteration. The Apriori algorithm presented in Agrawal and Srikant (1994) is given as follows:

```
Apriori()
  L1 ={large 1-itemsets}
  k = 2
  while Lk−1 ≠ 0 do
  begin
      Ck = apriori_gen(Lk−1)
      for all transactions t in D do
      begin
          C^t = subset(Ck, t)
          for all candiate c C^t do
          c.count = c.count+1
      end
  Lk = {c ϵ Ck —c.count ≥ minsu}
  k = k + 1
  end
```

Apriori first scans the transaction databases D in order to count the support of each item i in I, and determines the set of large 1- itemsets. Then, iteration is performed for each of the computation of the set of 2-itemsets, 3-itemsets, and so on. The k$^t h$ iteration consists of two steps (Rushing, Ranganath, Hinke, & Graves, 2002):

• Generate the candidate set C_k from the set of large (k-1)-itemsets, L_{k-1}.

• Scan the database in order to compute the support of each candidate itemset in C_k. The candidate generation algorithm is given as follows:

```
Apriori_gen(Lk−1)
Ck =0
for all itemsets X ϵ Lk−1 and Y ϵ Lk−1 do
  if X1=Y1X_{k−2}=Y_{k−2} ^ ... ^ X_{k−1} ¡ Y_{k−1} then begin
    C = X1X2 ...X_{k−1}Y_{k−1}
    add C to Ck
end
```

delete candidate itemsets in Ck whose any subset is not in L_{k-1}

The candidate generation procedure computes the set of potentially large k-itemsets from the set of large (k-1)-itemsets. A new candidate k-itemset is generated from two large (k-1)-itemsets if their first (k-2) items are the same. The candidate set Ck is a superset of the large k-itemsets. The candidate set is guaranteed to include all possible large k-itemsets because of the fact that all subsets of a large itemset are also large. Since all large itemsets in L_{k-1} are checked for contribution to candidate itemset, the candidate set C_k is certainly a superset of large k-itemsets.

After the candidates are generated, their counts must be computed in order to determine which of them are large. This counting step is really important in the efficiency of the algorithm, because the set of the candidate itemsets may be possibly large. Apriori handles this problem by employing a hash tree for storing the candidate. The candidate generation algorithm is used to find the candidate itemsets contained in a transaction using this hash tree structure. For each transaction T in the transaction database D, the candidates contained in T are found using the hash tree, and then their counts are incremented. After examining all transaction in D, the ones that are large are inserted into L_k (Karabatak et al., 2006).

For association rules like X ⇒ Y, two criteria are jointly used for rule evaluation as follows:

Support: The support, s, is the percentage of transactions that contain X$\bigcup$Y (Agrawal et al., 1993). It takes the form

$$s = P(X \bigcup Y)$$

Confidence: The confidence, c, is the ratio of the percentage of transactions that contain X $\bigcup$ Y to the percentage of transactions that contain X (Agrawal et al., 1993). It takes the form

$$c = \frac{P(X \bigcup Y)}{P(X)} = P(Y|X)$$

3. RESULTS

Due to the complexity and nonlinearity of the relationship between process parameters and the product quality, it is very difficult to model the relationship effectively and accurately by using common mathematics methods. Artificial Neural Networks (ANNs) is very effective for modeling the relationship with complexity and nonlinear between input space and output space (Bishop, C.M., 1995). ANNs have been used widely for monitoring process changes, and can predict what is expected to happen in the process (Guh, R.S., 2005; Wang, C.H. et. al, 2007).

However, it cannot provide explicit and comprehensible rules for fault diagnosis and recovery. On the other hand, the Apriori learning rule can extract such useful rules to solve this problem in manufacturing processes. Therefore, considering these critical needs for computational efficiency and effectiveness in real-time process monitoring and diagnosis, a data mining model is proposed to implement on-line monitoring and diagnosis tasks by using the Apriori.

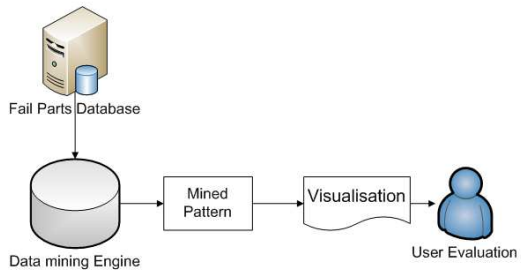

Fig.1: *Major components in fail parts recognition system in hard disk drive assembly line*

The current practice of analyzing data by engineers was that they had to spent at least 2 hours every morning for manually finding the bad parameters that affect the fail parts. This passive analysis was not effective enough to find out the root of errors so that the problem could be solved in time. It causes the action that is necessary to be implemented to reduce the defect, maybe too late. Therefore this proposed system was designed to serve the needs in which the analysis of data could be done automatically. The proposed model can be applied to an automated manufacturing process where observations are collected automatically from manufacturing processes, monitored and analyzed by a computer-based information system in real-time, without human intervention (Fig.1).

Engineers could schedule the time so that the mining system could retrieve the fail parts database automatically according to the time settings and mined the data real time in order to find the affect parameters. The Fig.2 illustrates the time setting in which the mining system could be run automatically. The proposed system took less than 2 minutes for 30,000 records of 58 parameters. This daily analysis was automatically updating fail parts association rules when the new scheduling was invoked. This process not only saves the time of analysis but also provide the problem solving in time.

Fig.2: *Real time scheduling mining system*

In this study, the data mining model is developed for on-line monitoring and diagnosis of fault products in manufacturing systems, where there is the strong relationship between process parameters and product quality (See Fig.3).

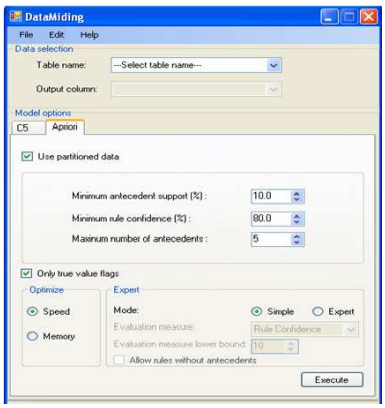

Fig.3: *Real time data mining model*

A total of 30,000 examples of 58 manufacturing process parameters and product quality (fail or pass) were collected from a local HDD maker in Thailand. Those attributes are the characteristics and parameters of each particular product measured during the manufacturing processes.

With the data mining results from the association rules, the strong and weak interrelations among different attributes can provide suggestions and references for fail parts in hard disk drive assembly line. After repeated testing within 58 attributes, the minimum threshold of support and confidence is 25% and 90% and the meaningful association rules are sum up to twenty-nine. (see Table 1) and the web diagram of fail parts association (see Fig. 4).

In the mining among the 58 fail parts attribute items, this study explores that the meaningful item

Table 1: *ASSOCIATION RULES OF HARD DISK DRIVE ASSEMBLY LINE BY FAIL PARTS (MINSUP = 25%; MINCONF = 90%)*

Rule	Consequent	Antecedent	Support%	Confidence%	Lift
1	Grade = Fail	SSEC = 10	35.560	97.279	0.999
2	Grade = Fail	SSEC = 13	39.021	97.352	1.000
3	Grade = Fail	MMX_Cell = 8	43.406	97.166	0.998
4	Grade = Fail	A-PrimeMC = 1	47.928	97.490	1.001
5	Grade = Fail	FCA = 2	48.486	97.271	0.999
6	Grade = Fail	FCA = 1	49.486	97.492	1.001
7	Grade = Fail	A-PrimeMC = 2	52.072	97.281	0.999
8	Grade = Fail	MMX_Cell = 5	54.188	97.548	1.002
9	Grade = Fail	BLKL_Line = 4	99.273	97.389	1.000
10	Grade = Fail	SSEC = 10 and BLKL_Line = 4	35.036	97.300	0.999
11	Grade = Fail	SSEC = 13 and BLKL_Line = 4	39.021	97.352	1.000
12	Grade = Fail	MMX_Cell = 8 and BLKL_Line = 4	42.892	97.170	0.998
13	Grade = Fail	A-PrimeMC = 1 and FCA = 1	25.735	97.536	1.002
14	Grade = Fail	A-PrimeMC = 1 and MMX_Cell = 5	25.692	97.510	1.001
15	Grade = Fail	A-PrimeMC = 1 and BLKL_Line = 4	47.616	97.497	1.001
16	Grade = Fail	FCA = 2 and A-PrimeMC = 2	27.534	97.160	0.998
17	Grade = Fail	FCA = 2 and MMX_Cell = 5	25.413	97.354	1.000
18	Grade = Fail	FCA = 2 and BLKL_Line = 4	47.758	97.287	0.999
19	Grade = Fail	FCA = 1 and MMX_Cell = 5	27.999	97.735	1.004
20	Grade = Fail	FCA = 1 and BLKL_Line = 4	49.486	97.492	1.001
21	Grade = Fail	A-PrimeMC = 2 and MMX_Cell = 5	28.496	97.583	1.002
22	Grade = Fail	A-PrimeMC = 2 and BLKL_Line = 4	51.657	97.290	0.999
23	Grade = Fail	MMX_Cell = 5 and BLKL_Line = 4	53.975	97.559	1.002
24	Grade = Fail	A-PrimeMC = 1 and FCA = 1 and BLKL_Line = 4	25.735	97.536	1.002
25	Grade = Fail	A-PrimeMC = 1 and MMX_Cell = 5 and BLKL_Line = 4	25.692	97.510	1.001
26	Grade = Fail	FCA = 2 and A-PrimeMC = 2 and BLKL_Line = 4	27.119	97.177	0.998
27	Grade = Fail	FCA = 2 and MMX_Cell = 5 and BLKL_Line = 4	25.200	97.375	1.000
28	Grade = Fail	FCA = 1 and MMX_Cell = 5 and BLKL_Line = 4	27.999	97.735	1.004
29	Grade = Fail	A-PrimeMC = 2 and MMX_Cell = 5 and BLKL_Line = 4	28.283	97.603	1.002

depth has three levels and the highest lift value is Rule19 and Rule28. Five important attributes that affect the quality of the hard disk are identified. Moreover, the data mining system also presents the relation among these attributes. Using Rule19 and Rule28 as an example, the Rule19 mining results (i.e. "FCA" and "MMX_Cell") and the Rule28 mining results (i.e. "FCA", "MMX_Cell" and "BLKL_Line") have strong interrelations among 58 attributes and have the maximum lift value of 1.004. According to the association rules in Table 1, the maximum lift and confidence value are Rule28 and fail parts attributes are "FCA", "MMX_Cell" and "BLKL_Line".

Association map display is a multi-attribute mapping approach, which allows the analyst to see the overall picture of a problem in an attribute space from analysis results. When analysis results from different attributes are integrated together on a map, this means that the analyst can initiate strategic and tactical plans with complete information or knowledge for decision-making or problem solving. This paper presents the effected parameters of fail parts maps with the association rules.

This map shows that different knowledge patterns and rules are extracted for recognize fail parts in hard disk drive assembly suggestions and solutions. It can be seen that mapping is an alternative approach, which can be integrated with the data mining approach for multi-dimensional data analysis. In addition, this map illustrates a visualization data analysis result, and this is what data mining approach works to develop. Therefore, this paper suggests that the map display approach could be implemented for multi-attribute data analysis in other research problems.

4. CONCLUSION

Nowadays, advanced automatic data collection and inspection techniques are now widely adopted in manufacturing industries. In this new industrial manufacturing scenario, a requirement exists for the automation of real-time monitoring and diagnosis implementation. This study proposed a data mining system for online monitoring and diagnosis of manufacturing processes, where some key process parameters (i.e., system inputs) have very strong relationship with the quality of finished products (i.e., system outputs).

In order to provide effective and accurate monitoring and diagnosis, an Apriori Learning rule model is developed. The Apriori can extract accurate and comprehensible rules to help operators accurately diagnosis abnormal processes and optimally set manufacturing process parameters.

Results provided by the Apriori are combined to generate quickly an accurate monitoring and diagnosis report. Once an abnormal manufacturing process occurs and a faulty product is going to be produced, the data mining system can predict this abnormal process change and signal a warning, and report the

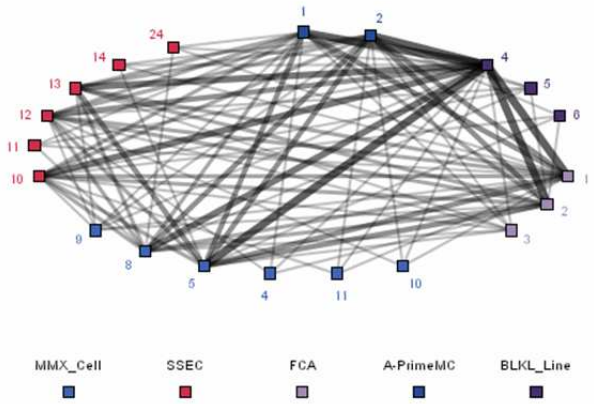

Fig.4: *Diagram of fail parts association in hard disk drive assembly line*

faulty product quality category.

Moreover, the Apriori provides the rules and the reasons for producing good and faulty products. These results will provide additional operational guidelines for operators or engineers to search for the assignable causes and to adjust process parameters to bring the out-of-control process back to the normal state.

Moreover, it is applied successfully into a real-world manufacturing system (hard disk drive assembly line), and the experimental results demonstrate its effectiveness and potential capability for real-world applications. The proposed system not only pinpoints the quality problems of the manufacturing process, but also reveals the causal relationships of why they occur and how to prevent them.

5. ACKNOWLEDGMENT

The authors would like to acknowledge the support from National Electronics and Computer Technology Center, National Science and Technology Development Agency and Industry/University Cooperative Research Center (I/U CRC) in HDD Component, the Faculty of Engineering, Khon Kaen University, Thailand.

References

[1] Agrawal, R., Imielinski, T., & Swami. A., "Mining association rules between sets of items in large database," *In: ACM SIGMOD Conference*, May 1993, pp. 207-216, 1993.

[2] Agrawal, R., &Srikant, R., "Fast algorithm for mining association rules," In: ACM VLDB Conference, September 1994, pp. 487-499, 1994.

[3] Bishop, C.M., *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[4] Brin, S., Motwani, R., Ullman, J. D., & Tsur, S., "Dynamic itemset counting and implication rules for market basket data," *In: ACMSIGMOD Conference*, Tucson, Arizona, USA, pp. 255-264, 1997.

[5] Catledge, L. D.,& Pitkow, J. E., "Characterizing browsing strategies in the WorldWideWeb," *In: Proceedings of Third WWW Conference*, April 1995.

[6] Cheeseman, P., & Stutz, J., *Bayesian classification (AutoClass): theory and results. In U.M. Fayyad, G. Piatetsky-Shaprio, P. Smyth, &R. Uthurusamy (Eds.), Advances in knowledge discovery and data mining* (pp. 153-180). AAAI/MIT Press, pp.153-180, 1996.

[7] Choi, D. H., Ahn, B. S., & Kim, S. H., *Prioritization of association rules in data mining: multiple criteria decision approach. Expert Systems with Applications*, 29(4), pp. 876-878, 2005.

[8] Guh, R.S. "A hybrid learning-based model for on-line detection and analysis of control chart patterns," *Computers & Industrial Engineering*, Vol. 49, pp. 35-62, 2005.

[9] Han, J., & Kamber, *M. Data mining: concepts and techniques*, Los Altos, CA: Morgan Kaufmann, 2001.

[10] Karabatak, M., Seng?r, A., Ince, M. C., & ve Türkoglu, I. *Association rules for texture classification*, IMS, 2006.

[11] Rushing, J. A., Ranganath, H. S., Hinke, T. H., & Graves, S. J. "Image segmentation using association rule features," *IEEE Transactions on Image Processing*, Vol. 11, pp. 558-566, 2002.

[12] Tan, P. N., & Kumar, V. "Interestingness measures for association patterns: A perspective," *KDD 2000 workshop on post-processing in machine learning and data mining*, Boston, MA, August, 2000.

[13] Wang, C.H, Kuo,W., Qi, H.R. "An integrated approach for process monitoring using wavelet analysis and competitive neural network," *International Journal of Production Research*, Vol. 45, pp. 227-244, 2007.

[14] Webb, G. I., & Zhang, S. "K-optimal rule discovery," *Data Mining and Knowledge Discovery*, Vol. 10(1), pp. 39-79, 2005.

[15] Wur, S. Y., & Leu, Y. "An effective Boolean algorithm for mining association rules in large databases," *In: International Conference on Database Systems for Advanced Applications (DASFAA '99)*, Hsinchu, Taiwan, 1999.

[16] Zhang, T., Ramakrishnan, R., & Livny, M. "BIRCH: an efficient data clustering method for very large databases," *In: ACM SIGMOD International Conference Management of Data*, Montreal, Canada, pp.103-114, 1996.

Wanida Kanarkard is an Associate Professor at Khon Kaen University. She received the B.Eng (2nd class honour) degree in Computer Engineering from Khon Kaen University in 1995 Thailand, M.Sc. (Advanced Computing) from University of London, United Kingdom in 1998 and Ph.D. degree in Computer Engineering from University of Hertfordshire, United Kingdom in 2001. She has been a visiting research fellow at University of Hertfordshire since 2001. Her current research activities are mainly concerned with artificial neural networks, soft computing, intelligent system, data mining and high performance computing.

Danaipong Chetchotsak received the B.Eng degree from Khon Kaen University, Thailand in 1995 and the Ph.D. degree in Industrial Engineering from Wichita State University, KS, USA in 2003. He is an Assistant Professor of dept. of Industrial Engineering, Faculty of Engineering, Khon Kaen University. His research in neural networks and their applications, productivity improvement, reliability and maintenance engineering and simulation has been funded by National Electronics and Computer Technology Center (NECTEC) and I/U CRC in HDD Components, Thailand.

Daranee Hormdee graduated from Khon Kaen University, Thailand, in 1996, obtaining bachelor's degree (B.Eng.) in Computer Engineering. From then, she has joined the University and worked as a Junior Lecturer. In 1998 and 2002 she gained an M.Sc. in Advanced Computer Science and a Ph.D. in Computer Science respectively, both from the University of Manchester, UK. After graduated, she has continued to serve Khon Kaen University, Thailand, as a Lecturer in the Department of Computer Engineering. In 2005, she was awarded an Assistant Professor position. Her research interests include Asynchronous Logic Design, Digital System Design and Embedded Systems.

Rod Adams obtained his BSc and an MSc (by research) in Mathematics from Leicester University in 1968 and 1971. In 1983 he obtained a PhD in Mathematical Logic and in 1987 he obtained an MSc in Computer Science both from Hatfield Polytechnic. In 2005 he was made Professor of Neural Computation at the University of Hertfordshire. His main research interests are in evolutionary and developmental modelling of Neural Networks.

Neil Davey received a BSc degree in Mathematics from the University of Manchester in 1978, a MSc in Mathematical Logic from the University of London in 1979, a MSc in Computing from Brunel University in 1989, and a PhD in Computer Science from University of Hertfordshire in 2004. He is a Lecturer at the University of Hertfordshire, where he does research into Connectionism and Neural Networks.