Using Data Mining Technology to Design an Intelligent Quality Analysis Control System for Semiconductor Packaging Industry

RUEY-SHUN CHEN* AND RUEY-CHYI WU
Institute of Information Management
National Chiao Tung University
1001, Ta Hsueh Road, Hsinchu City,
TAIWAN

Abstract: - The aims of this paper is to depict an intelligent quality analysis control system involved in using data warehouse, data mining, decision tree, and Bayesian classification analysis to discover the main inconsistency reasons in the manufacturing process of semiconductor packaging plants and compare the correctness of classification analysis of the two methods, so as to set up an intelligent quality analysis control system providing an efficiency tool for analyzing problems, with a view to identifying the causes of problems, making decision immediately, and eventually reducing the cycle time taken to solve quality-related problems.

The contributions of this research are illustrated as follows. Predictions made by the target group by means of decision tree analysis are more accurate than those made by Bayesian classification, indicating that decision tree analysis is an effective mean of classification analysis in semiconductor packaging quality problems, whereas evaluation of feasible methods by data warehouse, and data mining followed by establishment of the basis for a quality analysis system environment, that is characteristic of knowledge sharing may be applied to analysis of the quality problems in all corporation.

Key-Words: - Data mining, Data warehouse, Decision tree, Bayesian, OLAP.

1 Introduction
With the keen competition among enterprises, it is vital to catch up rivals' competition strategies and fulfill customers’ needs to take proper problem-solving steps timely. However, owing to a huge amount of information, finding unearthed information and identifying clients’ responses to any problems with products are usually time-consuming [1] [9]. Customers will be satisfied to a greater extent, only if the staff acquires intended or required information more efficiently and swiftly. Some key issues are urgent to be built in a useful quality control system, i.e., providing useful information, constructing know-how that integrating data warehouse and data mining, analyzing the data about any previous problems with product quality through the creation of a database, discovering the reasons of problems with the products timely, and solving the problems quickly.

Enterprises engaged in keen competition attach great importance to improve products quality; whoever obtains accurate information faster, and make decisions immediately, than rivals do, will have a chance of being successful [11]. Therefore, to discover the reasons of product quality problems and cope with the problems, one has to use those information systems with quality control functions to explore and solve the problems.

In a word, an intelligent quality analysis control system based on data warehouse and data mining is to be constructed with a view to achieving the following goals.

1. Discovering the main reasons that cause the problems of products quality and developing applicable methods.
2. Decreasing the product cycle time and increasing the total improvement rate of products.

2 Literature Review
2.1 Data Warehouse
A data warehouse involves not only all data required but applications essential to data processing. These applications include the data conversion into the data warehousing applications from external media. The data can be categorized as (1) fact data, (2) metadata, (3) dimension data, and (4) aggregation data, and the application software can be categorized as (1) load manager, (2) data warehouse manager, and (3) query manager [10] [11].
2.2 Data Mining
According to Cabena, Hadjinian, Stadler, Verhees and Zamasi [2], data mining is a process extracting effective information, which is unknown previously, from a large database for executives to make critical decisions. Besides, Frawley, Paitetsky-Shapiro, and Mathews [3] define data mining as a process exploring in database unobvious, implicit, unprecedented information which may be useful. Thus data mining is to use specific techniques, generalize and organize data from database and then excavate unknown, hidden information for executives to make decisions.

2.3 Semiconductor Packaging Process
A semiconductor manufacturing process consists of IC design, mask production, wafer fabrication, wafer packaging and testing. The whole manufacturing process can also be divided into front-end process and back-end process; wafer packaging and wafer testing are included in back-end process. A semiconductor packaging process consists of four stages, shown as Fig. 1. Different products undergo different processes and packaging patterns, which will then influence the processing methods. A processing method may be designated based on customer special requests [5].

1. Stage of wafer cleaning/mounting/saw: Wafer income inspection, Wafer mounting, Wafer Sawing/Cleaning, Post Saw Inspection, and sampling inspection by the QC (Quality Control) division.
2. Stage of die bonding/wire bonding/curing: Die Bonding, i.e. to glue dies on the lead frame one after one; Epoxy Curing, which places the semi-finished goods in an oven for curing; Wire Bonding and Post bond inspection, and QC inspection.
3. Stage of molding/marking: Molding, Backside Marking, which means to have a mark on the bottom of an IC; Post mold cure, that sends an IC semi-finished goods for curing once more; Trimming/Dejunking, trimming the pins on leads; Solder Plating, soldering external pins on ICs; and QC inspection.
4. Stage of forming/packing/storage: Top Marking, which marks on the front side of ICs; Post Mold Cure, which sends IC semi-finished goods to oven for curing again; Forming/Singulation; Final Visual Inspection; QC inspection and then should be inspected by the QC division for packing compliance and conformity with customer requirements prior to storage.

Fig. 1. Semiconductor packaging process.

3 Problem Definition
As the manufacturing industry always uses CAR (corrective action request) to record quality-related problems raised by customers, lacking an effective way to make the most of information, it results in wasteful time and unnecessary cost on investigations and analysis when the problem reoccurs. Furthermore, data is often large in size and complicated, the personnel in charge of quality-related problems can hardly identify the discrepancy factor or generalize the characteristics or types of the problems rapidly or correctly. For these two main reasons, the challenge required to be settled in the study lies in the conclusion of the problems arising in connection with the semiconductor manufacturing plant. There are cases applying data mining in a number of literature reviews, including manufacturing, financing and telecommunications. Among data mining tools, the common classification methods are Decision Tree, Naïve Bayesian, and Neural Network. In applied telecommunications cases, Naïve Bayesian is better than Decision Tree in prediction effect [4]. Thus the study is intended to use the Decision Tree and Naïve Bayesian, and collect the previous CAR data for analysis, and identify which algorithm is superior in application to the semiconductor packaging industry.

4 System Design
4.1 System Design Framework
The complete design framework for intelligent quality analysis control system is illustrated in Fig. 2. The
function of each design element is shown in the following steps.

1. Experts, domain knowledge: Determine which goals to achieve with data mining for relevant data collection, data pre-processing, selection of data attributes and data mining methods.

2. Data collection: Converse historical data from the existing systems into the processing area.

3. Data standardization: Standardize the data type to ensure the consistency between subsequently collected data and pre-processed data.

4. Data preprocess: Proceed with data integration, conversion, extraction, and cleaning.

5. Data warehouse, OLAP (Online Analytical Process): Reduced data search time is the key to the whole process of data mining. Hence data warehousing is applied to address these challenges. Besides, OLAP operations in data cubes include rollup, drilldown, slice, dice, and pivot.

6. Select attributes: The proper attribute for specific analysis target is determined by the expert in the domain concerned, because either insufficient or excessive attributes cannot achieve correct analysis results.

7. Decision tree data mining engine: A data mining engine is the core in the system framework. The most commonly used methods are classification, clustering, and sequential patterns, etc.

8. Bayesian data mining engine: The design of the data mining engine is vital, as the data mining engine serves as the nucleus of the entire framework.

9. Results evaluation: Tremendous mined data and patterns may exist; the mined result can be more available and interpretable through parameter setup.

10. Results display: The mined result may be presented by user preference.

11. Knowledge base: The knowledge base, which stores expert expertise and the rules available after data mining, can be updated from time to time to be the basis for various decision-making supports.

Fig. 2. Complete design framework for intelligent quality analysis control system.

4.2 The Integration of Quality Analysis Control Systems and Data Warehouses

4.2.1 Establishment of data warehouse architecture

This research utilizes the top to bottom model to establish a data warehouse system. The database of intelligent quality analysis control system serves as the origin of data, by extracting and transforming data, an integral and unified data warehouse system may be established. Data marts and data warehouses have a one-sided relationship, in which data from a data warehouse flows into a data mart. This process can be divided into three levels from top to bottom: operational data, data warehouse and intelligent application, as shown in Fig. 3 [12].

Fig. 3. Data warehouse system established through the top to bottom model.

4.2.2 Design a method to integrate quality analysis control systems and data warehouses

The integration of intelligent quality analysis control system and data warehouses involves the collection of different types of data from their original sources. Therefore, it is common to encounter the problems of inconsistent, incomplete and duplicate data. This data is then placed in a data staging area where it undergoes such processes as the collection, selection, cleaning, transformation, combination, removal of duplicates indexing, etc. Next, the data is stored within a presentation server. Organized data stored in the presentation server can be used directly for the user to query data. At this point, users can carry out search tasks. The system framework is shown in Fig. 4 [6].
4.2.3 Establishment of data warehouse schema
This research project utilizes the starflake schema in designing the schema for data warehouse. This schema is based upon the manufacturing fact table, time, order, lot, product and quality dimension table shown in Fig. 5. The quality dimension table composes of a machine, a material, a method and a man dimension table.

4.2.4 Establishment of multidimensional model
In order to provide an even wider range of search capabilities, this research uses the four dimensions of man, machine, method and material to construct a four-dimensional data cube model. After completing the construction of data cubes, it is possible to integrate decision-making analysis and the data mining system [13]. OLAP technology is able to blend together people’s observations and intelligence within the data mining system, thus improving the speed and depth at which data is excavated.

4.3 Establishment of Data Mining System
4.3.1 Classification and predictions procedures
By collecting tuples within the database, defining them systematically on the basis of the specific target of analysis, searching out common characteristics and establishing a class process, classification is accomplished. Furthermore, previously classified historical data may be utilized to anticipate which class each item belongs to. Data classification is basically comprised of the two-step process: learning and classification [7] [8].

4.3.2 Design of decision tree data mining engine
When building a decision tree engine, we adopt the decision tree algorithm of SQL server 2005 for calculation [4]. The computation steps are provided below [8].

**Step 1.** Prepare previously classified training data.

**Step 2.** Establish a decision tree node. Determine whether or not this node is a leaf node, or calculate information gain for the test attribute. The calculation method is shown below in steps 3-5.

**Step 3.** The expected information of the classified data samples selected for calculation: Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (for i = 1,…,m). Let s_i be the number of samples of S in class C_i. The expected information needed to classify is given by

\[
I(S_1, S_2, ..., S_m) = \sum_{i=1}^{m} P_i \log_2(P_i)
\]

Where P_i is the probability that an arbitrary sample belongs to class C_i and is estimated by s_i/s.

**Step 4.** The expected information of the test attribute selected for calculation: Let attribute A have v distinct values, {a_1,a_2,...,a_v}. Attribute A can be used to partition S into v subsets, {S_1,S_2,...,S_v}, where S_j contains those samples in S that have value a_j of A. Let s_j be the number of samples of class C_i in a subset S_j. The entropy, or expected information based on the partitioning into subsets by A, is given by

\[
E(A) = \sum_{j=1}^{v} \frac{S_j}{S} I(S_1, S_2, ..., S_m)
\]

\[
I(S_1, S_2, ..., S_m) = \sum_{j=1}^{v} P_j \log_2(P_j), P_j = s_j/S_j
\]

Where P_j is the probability that a sample in S_j belongs to class C_j.

**Step 5.** Calculate the information gain of the selected test attribute: The encoding information that would be gained by branching on A is

\[
\text{Gain Ratio (A)} = \frac{I(S_1, S_2, ..., S_m) - E(A)}{\text{Split Gains}}
\]

**Step 6.** Repeat steps 2-5 until the information gain of the test attributes are completely calculated.

**Step 7.** Select the test attribute with the highest information gain to act as the node of partition for the decision tree.

**Step 8.** To complete set up of the decision tree, follow this sequence of steps to find test attribute.
The na"ive Bayesian classifier, or simple Bayesian classifier, works as follows [8].

**Step 1.** Each data sample is represented by an n-dimensional feature vector, \( X = (x_1, x_2, \ldots, x_n) \), depicting n measurements made on the sample from n attributes, respectively, \( A_1, A_2, \ldots, A_n \).

**Step 2.** Suppose that there are m classes, \( C_1, C_2, \ldots, C_m \). Given an unknown data sample, \( X \) i.e., having no class label, the classifier will predict that \( X \). That is, the na"ive Bayesian classifier assigns an unknown sample \( X \) to the class \( C_i \) if and only if

\[
P(C_j|X) \quad \text{for } 1 \leq j \leq m, j \neq i.
\]

Thus we maximize \( P(C_i|X) \). The class \( C_i \) for which \( P(C_i|X) \) is maximized is called the maximum posteriori hypothesis. By bayes theorem,

\[
P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}
\]

**Step 3.** As \( P(X) \) is constant for all classes, only \( P(X|C_i)P(C_i) \) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, \( P(C_1) = P(C_2) = \ldots = P(C_m) \), and we would therefore maximize \( P(X|C_i) \). Otherwise, we maximize \( P(X|C_i)P(C_i) \). Note that the class prior probabilities may be estimated by \( P(C_i) = s_i/s \), where \( s_i \) is the number of training samples of class \( C_i \), and \( s \) is the total number of training samples.

**Step 4.** Given data sets with many attributes, it would be extremely computationally expensive to compute \( P(X|C_i) \). In order to reduce computation in evaluation \( P(X|C_i) \), the na"ive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample, that is, there are no dependence relationships among the attributes. Thus,

\[
P(X|C_i) = \prod_{i=1}^{n} P(x_i|C_i)
\]

The probabilities \( P(x_1|C_i) \), \( P(x_2|C_i) \), \ldots, \( P(x_n|C_i) \) can be estimated from the training samples, where

\[
\text{If } A_k \text{ is categorical, then } P(x_k|C_i) = s_{ik}/s_i \quad \text{where } s_{ik} \text{ is the number of training samples of class } C_i \text{ having the value } x_k \text{ for } A_k \text{, and } s_i \text{ is the number of training samples belonging to } C_i.
\]

**Step 5.** In order to classify an unknown sample \( X \), \( P(X|C_i)P(C_i) \) is evaluated for each class \( C_i \). Sample \( X \) is then assigned to the class \( C_i \) if and only if

\[
P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \quad \text{for } 1 \leq j \leq m, j \neq i.
\]

In other words, it is assigned to the class \( C_i \) for which \( P(X|C_i)P(C_i) \) is the maximum.

### 4.3.4 Establishment of visual figure of data mining

The operational process and knowledge rules excavated by data mining can be displayed through visual figure. A graphical user interface for a data mining engine is composed of the functional elements, including data collection and data mining query composition, presentation of discovered patterns, manipulation of data mining primitives, interactive multilevel mining, on-line help manuals, indexed search, debugging, and other interactive graphical facilities.

### 5 System Implement

#### 5.1 System Implementation Environment

The system’s environment includes Data Warehouse Server, Data Mining Server, Web Server, and Data Mining and intelligent quality analysis control front-end personal computers.

#### 5.2 Data Aggregation and Generation

In this stage, we diagnose and analyze quality data. We proceed to collect data and select data items from the desired analysis scope for follow-up analysis.

Data attributes are defined as man, machine, material and method. During the analysis process, we should summarize the data and convert them into a consistent format for data mining. And then we assign numbers (e.g. 0, 1, 2…) to attribute variables and quality problem types. "0" represents attribute variables with "no error": rank them in a proper order. Similarly, quality problem (judgment) types are numbered 1, 2, 3… Please refer to Table 1, the contrast table of numbers of attribute variables and quality problems.
5.3 Set Up of Data Warehouse

The data warehouse system is a starflake framework centered on the manufacturing fact table with the association time, order, lot, product, and quality dimension table. In the system, the quality problem dimension table includes multiple dimensions. The schemas of essential data are shown in Tables 2, 3.

5.4 Analysis of Findings

To establish the database, we have collected the data form January to June in 2005. With the decision tree and Bayesian classification, predictions made by means of decision tree and Bayesian, we have an accuracy of 89% and 83% respectively.

6 Conclusions

The results and contributions of this research are listed as follows.

1. Compared with decision tree and Bayesian classification analysis, predictions made by means of decision tree and Bayesian, we have an accuracy of 89% and 83% respectively.

2. Decision tree algorithm is more effective and appropriate than Bayesian algorithm to analyze the quality problems in the semiconductor packaging industry.

References:


Table 1. Contrast table of numbers of attribute variables and quality judgment (problems)

<table>
<thead>
<tr>
<th>No</th>
<th>Man</th>
<th>Measure</th>
<th>Material</th>
<th>Method</th>
<th>Judgment/Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>NA</td>
<td>Machine</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td>NA</td>
<td>Machine</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>NA</td>
<td>Machine</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>5</td>
<td>NA</td>
<td>Machine</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

1. For classification and analysis with the decision tree, there are 3,389 entries of correct data, representing a success rate of 89%.

2. For classification and analysis with Bayesian algorithm, there are 3,160 entries of correct data, representing a success rate of 83%.

Hence, decision tree method is more effective and accurate than Bayesian to apply to the quality problems in the semiconductor packaging industry. The findings is shown in Table 4.

Table 2. Manufacturing fact table

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Type</th>
<th>Length</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>time_id</td>
<td>int</td>
<td>14</td>
<td>time identification</td>
</tr>
<tr>
<td>order_id</td>
<td>char</td>
<td>16</td>
<td>order identification</td>
</tr>
<tr>
<td>product_id</td>
<td>char</td>
<td>16</td>
<td>product identification</td>
</tr>
<tr>
<td>lot_id</td>
<td>char</td>
<td>16</td>
<td>lot identification</td>
</tr>
<tr>
<td>quality_id</td>
<td>char</td>
<td>16</td>
<td>quality identification</td>
</tr>
<tr>
<td>quantity</td>
<td>int</td>
<td>16</td>
<td>quantity of production</td>
</tr>
</tbody>
</table>

Primary key: time_id, order_id, product_id, lot_id and quality_id

Table 3. Quality dimension table

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Type</th>
<th>Length</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>quality_id</td>
<td>char</td>
<td>16</td>
<td>quality identification</td>
</tr>
<tr>
<td>machine_id</td>
<td>char</td>
<td>20</td>
<td>machine identification</td>
</tr>
<tr>
<td>material_id</td>
<td>char</td>
<td>20</td>
<td>material identification</td>
</tr>
<tr>
<td>method_id</td>
<td>char</td>
<td>20</td>
<td>method identification</td>
</tr>
<tr>
<td>man_id</td>
<td>char</td>
<td>20</td>
<td>man identification</td>
</tr>
</tbody>
</table>

Primary key: quality_id, machine_id, material_id, method_id and man_id

Table 4. Comparisons of data mining results

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Decision Tree</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Sample</td>
<td>15,230</td>
<td>15,230</td>
</tr>
<tr>
<td>Training/testing ratio</td>
<td>3:1</td>
<td>3:1</td>
</tr>
<tr>
<td>Training sample</td>
<td>11,423</td>
<td>11,423</td>
</tr>
<tr>
<td>Testing sample</td>
<td>3,807</td>
<td>3,807</td>
</tr>
<tr>
<td>Correct Data</td>
<td>3,389</td>
<td>3,160</td>
</tr>
<tr>
<td>Accurate rate</td>
<td>89%</td>
<td>83%</td>
</tr>
</tbody>
</table>

Proceedings of the 10th WSEAS International Conference on COMPUTERS, Vouliagmeni, Athens, Greece, July 13-15, 2006 (pp186-192)