

---

# Mining Warranty Data in Manufacturing Industry

---



The University of Michigan-Dearborn  
Henry W. Patton Center for Engineering  
Education and Practice

---

Henry W. Patton Center for Engineering  
Education and Practice  
Final Technical Report

# Mining Warranty Data in Manufacturing Industry

(Project #2005/1)

By:

Jirachai Buddhakulsomsiri, Assistant Professor  
Armen Zakarian, Associate Professor  
Department of Industrial and Manufacturing Systems Engineering

## Table of Contents

Synopsis .....	iii
1. Background .....	1
2. Objectives .....	1
3. Approach .....	2
4. Results .....	8
5. Conclusions.....	15
6. Impact	
Educational .....	15
Industrial .....	16
7. Acknowledgments .....	16
8. References.....	16

## Synopsis

Many industries, including the automotive industry, face the task of improving product quality while minimizing warranty costs. Product quality problems are monitored during the warranty period through the claims filed against the products. This process generates large volumes of data records, which are typically stored in a warranty database or data warehouse. By analyzing the warranty claims, companies improve product development processes, advance product design, and make modifications to their manufacturing and assembly systems, with the aim to improve product quality and reduce warranty costs. The acquisition, preprocessing, and analysis of warranty data for complex products produced in large volumes (e.g., automobiles) is not a trivial problem, and requires effective data analysis or mining algorithms. Traditionally, this was the responsibility of data analysts who generally used statistical techniques and tools. However, the scope of this activity has recently changed. For example, widespread usage of computers and networking technologies and the introduction of new data acquisition systems in automotive companies, dealerships, and repair shops has allowed automotive original equipment manufacturers (OEM) to capture millions of transactions and create large electronic databases, which store manufacturing, product, and sales-related data. These data can be analyzed to develop useful knowledge, such as trends, patterns, or causal relationships between product and warranty claim attributes that may enable companies to effectively spot defects, discover the root causes of problems, and formulate an action plan to remedy the problem and improve quality. Acquisition, representation and proper interpretation of such knowledge may lead to significant savings in warranty costs and attainment of product goodwill.

In this research a new association rule generation algorithm for warranty data analysis is developed. The algorithm uses elementary set concept and database manipulation techniques to develop useful relationships between product attributes and causes of failure. These relationships (knowledge) are represented using IF-THEN association rules, where the IF portion of the rule includes set of attributes representing product features (e.g., production date, repair date, mileage at repair, transmission, engine type and so on), and THEN portion of the rule includes a set of attributes that represent the decision outcome (e.g., problem-related labor code). Once association rules are developed, the algorithm applies a statistical analysis technique to evaluate the significance of each rule. The rules that pass the significance test are reported in a solution. An application of the association rule generation algorithm is presented with a data mining case study from the automotive industry. The knowledge (rules) extracted from the automotive warranty data are used to identify root causes of a particular warranty problem or develop useful conclusions.

This research also developed a sequential pattern mining algorithm for extracting hidden knowledge from a large amount of warranty data. The algorithm uses elementary set concept and database manipulation techniques to search for patterns or relationships among occurrences of warranty claims over time. Significant patterns provide knowledge of one (or more) product failure that leads to future product faults.

## 1. Background

Many industries, including the automotive industry, face the task of improving product quality while minimizing warranty costs. Product quality is a by-product of the effectiveness of product development processes and the production systems that are used to develop, manufacture, and assemble the product. Therefore, product quality can be improved through continuous improvement in product design and development processes and in the development of robust manufacturing and assembly systems. However, no matter how well a product is designed and manufactured, it may fail in the usage environment, either by chance or by some assignable causes. When product fails within a certain time (warranty) period, the warranty is the manufacturer's assurance to a buyer that the product will be repaired without cost. In a service environment where dealers are more likely to replace than repair, the cost of component failure during the warranty period can easily equal three to ten times the supplier's unit price (Baird 2000, Feng et al. 2001, Cali 1993). Consequently, companies invest a significant amount of time and resources to monitor, document, and evaluate warranty problems. Product quality problems are monitored during the warranty period through the claims filed against the products. This process generates large volumes of data records, which are typically stored in a warranty database or data warehouse. By analyzing the warranty claims, companies improve product development processes, advance product design, and make modifications to their manufacturing and assembly systems, with the aim to improve product quality and reduce warranty costs. The acquisition, preprocessing, and analysis of warranty data of complex products produced in large volumes (e.g., automobiles) is not a trivial problem, and requires effective data analysis or mining algorithms. Traditionally, this was the responsibility of data analysts, who generally used statistical techniques and tools. However, the scope of this activity has recently changed. For example, widespread usage of computers and networking technologies and introduction of new data acquisition systems in automotive companies, dealerships, and repair shops has allowed automotive OEM to capture millions of transactions and create large electronic databases, which store manufacturing, product, and sales-related data. These data can be analyzed to develop useful knowledge, such as trends, patterns, or causal relationships between product and warranty claim attributes that may enable companies to effectively spot defects, discover the root causes of problems, and formulate an action plan to remedy the problem and improve quality. Acquisition, representation and proper interpretation of such knowledge may lead to significant savings in warranty costs and attainment of product goodwill.

## 2. Objectives

The goal of this research is to develop a data mining algorithm and a software tool to assist quality and warranty data analysts in manufacturing companies in gaining and extracting knowledge about their quality issues from warranty data. The three major goals of this project are as follows:

- Develop a data mining methodology for extracting knowledge from warranty data.
- Develop a software tool that implements the developed algorithms to assist manufacturing engineers or data analysts in manufacturing organizations to perform data mining tasks.
- Apply the algorithms to warranty data obtained from industry to validate the models.

### 3. Approach

#### 3.1 Source and Characteristics of Warranty Data

This research focuses on warranty data analysis in automotive industry. Typically, the automotive warranty data is gathered from two different sources: 1) manufacturing and assembly plants, and 2) automobile dealerships and repair shops (see Figure 1). The warranty data collection process starts at the manufacturing and assembly plants. Here product manufacturing and assembly information (e.g., product identification number, production date, product options, plant identification, supplier-related data) is collected and stored in the plant database. Once a product is sold, sales transaction data (e.g., product identification number, sales date) is captured at the automobile dealership. When a product quality problem occurs within the warranty period and the customer requests a repair, a warranty claim is initiated at the dealership or repair shop. Product warranty information collected at this level includes repair-related labor code, repair date, mileage-at-repair, labor and part costs, and so on. The warranty data collected at the automotive dealerships and repair shops is then transferred to the company's claims processing department for claim acceptance or rejection. Production, sales, and repair-related data in accepted claims are combined and stored in a warranty data warehouse that is centrally located and used by quality engineering experts to perform data analysis and provide recommendations for design and manufacturing improvements. Warranty data collection and integration processes include some formidable challenges. One may see from Figure 1 that a claims processing department may need to collect and integrate data from various different sources. For example, in the automotive field, each OEM has hundreds of different plants and thousands of different dealerships and repair shops. Each of these entities (plants, dealerships, repair shops) collects and stores data differently and the data integration into one unified data warehouse can be a challenging task. It needs to be emphasized that without effective and accurate integration of data from these various sources, the extraction of meaningful knowledge can be limited.

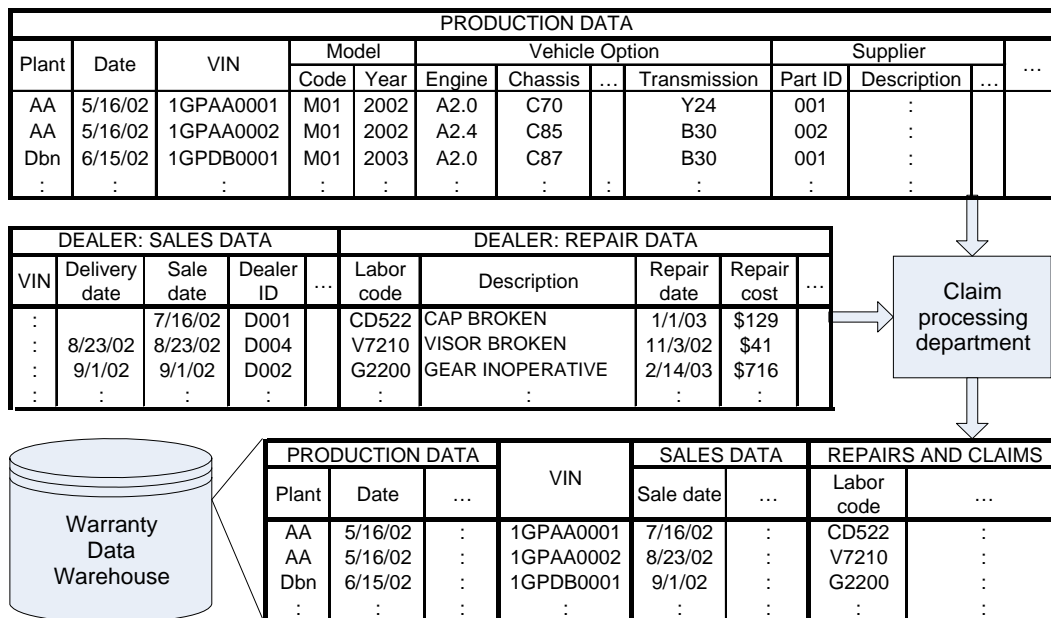


Figure 1. Three different sources of warranty data.

Analysis of warranty data also presents a difficult challenge for several other reasons. First, the characteristics of warranty data attributes (or data fields) can be one of the following three types: temporal (e.g., production date, sales date, repair dates), numerical (e.g., mileage at repair, labor cost, part cost), and categorical (e.g., labor code, model and engine types). In addition, complex products such as automobiles may have hundreds of different attributes. Since they are produced in large volumes, their unified warranty database may include several millions of data records. Therefore, the warranty data analysis methods not only should be able to cope with these various data types, but should also be effective in analyzing large data sets.

Data mining can be used to discover hidden knowledge from such large warranty data sets. Based on the type of knowledge that can be mined, data mining problems in warranty analysis can be classified in to the following categories (Han and Kamber, 2001):

*Association rule mining* exposes interesting relationships among a large set of data objects by showing attribute-value conditions that occur together frequently (Zhang and Zhou, 2004). Association rules obtained from the warranty data set may reveal, for example, the association between the product attributes (e.g., production date and/or option content) and the warranty problem (e.g., labor code). The association between product attributes can be represented with IF-THEN decision rules. In this decision rule, the IF portion of the rule includes a set of attributes representing the product features (e.g., production date, transmission type) and THEN portion of the rule includes a set of attributes that represent the decision outcome (e.g., specific labor code).

*Classification and prediction* involves identifying models that can describe variability within a warranty data set and predict unknown values of a specific attribute of interest (e.g., total warranty cost for the next period) based on the known values of other attributes (e.g., production data, expected product life time, and so on). Several studies on warranty cost and product reliability analysis implement such an approach using statistical tools and techniques.

*Sequential pattern analysis* looks for patterns or relationships between attributes of the warranty claims that occur over time. This type of analysis is of major interest to quality experts in the automotive industry, as it allows one to identify temporal relationships between various product faults. For example, the results of such analysis may reveal a pattern of occurrence where one product failure leads to another product fault at a later time.

*Text mining* is used in product warranty analysis to evaluate large volumes of textual data that include customer feedback and technician reports. Text mining attempts to identify patterns in the text and predict outcomes that describe product quality problems. Typically, knowledge is mined by observing a growth in the text cluster over time, or sudden increase in the text cluster size. The knowledge extracted from text mining can serve as an early problem detection system and may alert data analysts to perform more detailed analysis on the detected pattern.

*Data visualization* methods analyze and represent data in graphical form so it can be easily understood and interpreted by humans (Zhang and Zhou, 2004). Graphical displays of data (i.e., Pareto charts,

correlation matrices, and so on) allow analysts to interact with the data, gain problem insight, and interpret information represented in the data that is overwhelming in size and complexity. Data visualization methods are widely used in industry for warranty analysis.

This research focuses on association rule mining in warranty data. Thus, difficulties that one may encounter when trying to generate association rules from a warranty data set is discussed next. Scalability of the mining algorithm is important, as warranty data sets usually contain a large number of data objects and may require considerable computation efforts. Furthermore, the dimensionality of the algorithm is of importance as well, since warranty data sets may contain many potential condition and decision attributes that may result in many association rules. Finally, the characteristic (discrete value) of the decision attribute of interest (i.e., problem-related labor code) limits the type of data mining techniques that can be considered. In addition, labor code in warranty data may assume a large number of possible outcomes that may also lead to the generation of a large number of association rules. Therefore, selection of a proper mining technique must take these factors into account, so that meaningful association rules can be generated from a large data set within a reasonable computation time.

### 3.2 Association Rule Generation Algorithm

In this section a new association rule generation algorithm for warranty data analysis is presented. The algorithm uses the elementary set concept and database manipulation techniques to develop useful relationships between product attributes and causes of failure (i.e., repair labor codes). These relationships are represented in the form of IF/THEN rules, where the IF portion of the rule includes the set of attributes representing product features (e.g., production date, repair date, mileage at repair, transmission, engine type, and so on), and the THEN portion of the rule includes the set of attributes that represent decision outcome (e.g., specific labor code). Once association rules are developed, the algorithm applies statistical analysis techniques to evaluate the significance of each rule. The rules that pass the significance test are reported in a solution. Before the steps of the association rule generation algorithm are presented, the following notation is introduced.

*Notation:*

$C = \{c_1, c_2, \dots, c_n\}$	Condition attributes set
$D = \{d_1, d_2, \dots, d_m\}$	Decision attributes set
Elementary set	Set of objects that have the same values for the attributes in set C or D
$C_i$	Elementary set of C, where $i = 1, \dots, p$
$D_j$	Elementary set of D, where $j = 1, \dots, q$
$V(C_i, c_k)$	Value of attribute $c_k$ in the elementary set $C_i$
$V(D_j, d_l)$	Value of attribute $d_l$ in the elementary set $D_j$
$X_{ij}$	Intersection of elementary sets $C_i$ and $D_j$
$f(r, a_j)$	Value of attribute $a_j$ for object $r$
$P$	Percent of objects in an elementary set of condition attribute set that correspond to a rule

Q	Percent of objects in an elementary set of decision attribute set that correspond to a rule
N	The total number of objects in the data set
Relative Strength (RS)	Percent of objects that correspond to a rule

*Steps of the Association Rules Generation Algorithm*

- Step 1. Initialize  $C = \{c_1, c_2, \dots, c_n\}$ ;  $D = \{d_1, d_2, \dots, d_m\}$  and define  $P'$ ,  $Q'$ , and  $RS'$  as thresholds for  $P$ ,  $Q$ , and  $RS$ , respectively.
- Step 2. Determine  $X_{ij} = C_i \cap D_j$  for each  $i = 1, \dots, p$  and  $j = 1, \dots, q$ .
- Step 3. For each  $X_{ij} \neq \emptyset$  generate a rule:  
 IF  $c_1 = V(C_i, c_1)$  AND ... AND  $c_n = V(C_i, c_n)$ ,  
 THEN  $d_1 = V(D_j, d_1)$  AND ... AND  $d_m = V(D_j, d_m)$  [ $P, Q, RS$ ],  
 where  $P = |X_{ij}| / |C_i|$ ;  $Q = |X_{ij}| / |D_j|$ ;  $RS = |X_{ij}| / N$ .
- Step 4. Discard rules obtained in Step 3 for which  $P < P'$  or  $Q < Q'$  or  $RS < RS'$ .
- Step 5. For each of the remaining rules, perform statistical evaluation using a chi-square  $X^2$  statistic.  
 For each rule obtained from Step 4, set up a 2x2 table to test  $H_0$ . The association rule is significant, as shown in Table 1 below.

Table 1. A 2x2 table for chi-square test.

IF \ THEN	$d_l = V(D_j, d_l), \forall l=1, \dots, m$	At least one $d_l \neq V(D_j, d_l)$	Total
$c_k = V(C_i, c_k), \forall k=1, \dots, n$	$n_{11} =  X_{ij} $	$n_{12} =  C_i  -  X_{ij} $	$n_{1\bullet} =  C_i $
At least one $c_k \neq V(C_i, c_k)$	$n_{21} =  D_j  -  X_{ij} $	$n_{22} = N -  C_i  -  D_j  +  X_{ij} $	$n_{2\bullet} = N -  C_i $
Total	$n_{\bullet 1} =  D_j $	$n_{\bullet 2} = N -  D_j $	$n_{\bullet\bullet} = N$

Using the information from the table above, calculate the chi-square statistic with one degree of freedom for the rule as follows:

$$X^2 = \frac{n_{\bullet\bullet} (n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}$$

IF p-value  $P(X^2 > \chi_{1,\alpha}^2) < \alpha$ , where  $\alpha$  is the significance level of the test, THEN report the rule (i.e., the rule is statistically significant at  $\alpha$  level of confidence). Otherwise, go to Step 5.

The association rule generation algorithm starts with the initialization of the sets  $C$  and  $D$  and user defined thresholds  $P'$ ,  $Q'$ , and  $RS'$ . Each non-empty intersection of the elementary sets  $C_i$  and  $D_j$  obtained in Step 2 is represented with a single IF-THEN decision rule in Step 3. In this decision rule the IF portion of the rule includes a set of attributes representing the conditions of the rule and the THEN portion of the rule includes a set of attributes representing decisions. For each rule, the parameters  $P$ ,  $Q$ , and  $RS$  are also evaluated in this step. Rules obtained in Step 3 are further screened in Step 4 using the user-defined thresholds of rule parameters  $P'$ ,  $Q'$ , and  $RS'$ . Finally, in Step 5, the rules that satisfy Step 4 conditions are evaluated using a statistical analysis technique. To illustrate the steps of decision rules extraction

algorithm and how parameters  $P$ ,  $Q$ , and  $RS$  are used to analyze the rule, consider a sample warranty data shown in Table 2.

Assume a data analyst applies the association rule generation algorithm to explore possible relationships between the product attributes “Engine” and “Labor Code” using the data set in Table 2. In Step 1 of the algorithm, sets  $C = \{\text{Engine}\}$  and  $D = \{\text{Labor Code}\}$  are initialized and user-defined thresholds  $P' = 50\%$ ,  $Q' = 50\%$ , and  $RS' = 25\%$  are identified. In Step 2, the elementary sets  $C$  and  $D$  are determined and their corresponding intersections are calculated:

Table 2. Sample warranty data set for the illustrative example.

Object No.	Engine	Transmission	Production Year	Labor Code
1	B	2	2000	3
2	B	2	2001	1
3	A	2	2001	3
4	A	2	2000	3
5	B	1	2000	2
6	A	1	2002	2
7	A	1	2002	3
8	A	2	2002	3
9	B	1	2000	2
10	B	2	2002	2
11	A	1	2001	3
12	B	2	2001	1

$$C_1 = \{1, 2, 5, 9, 10, 12\}, C_2 = \{3, 4, 6, 7, 8, 11\}$$

$$D_1 = \{1, 3, 4, 7, 8, 11\}, D_2 = \{2, 12\}, D_3 = \{5, 6, 9, 10\}$$

$$X_{11} = C_1 \cap D_1 = \{1\}, X_{12} = C_1 \cap D_2 = \{2, 12\}, X_{13} = C_1 \cap D_3 = \{5, 9, 10\},$$

$$X_{21} = C_2 \cap D_1 = \{3, 4, 7, 8, 11\}, X_{22} = C_2 \cap D_2 = \emptyset, X_{23} = C_2 \cap D_3 = \{6\}$$

Also in this step, the values of the elementary sets are determined:

$$V(C_1, \text{“Engine”}) = B, V(C_2, \text{“Engine”}) = A$$

$$V(D_1, \text{“Labor code”}) = 3, V(D_2, \text{“Labor code”}) = 1, \text{ and } V(D_3, \text{“Labor code”}) = 2$$

In Step 3 of the algorithm, the decision rules are generated and the values of parameters  $P$ ,  $Q$ , and  $RS$  are calculated for each rule:

Rule 1: **IF** Engine = B **THEN** Labor Code = 3. [ $P = 16.67\%$ ,  $Q = 16.67\%$ ,  $RS = 8.33\%$ ]

Rule 2: **IF** Engine = B **THEN** Labor Code = 1. [ $P = 33.33\%$ ,  $Q = 100\%$ ,  $RS = 16.67\%$ ]

Rule 3: **IF** Engine = B **THEN** Labor Code = 2. [ $P = 50\%$ ,  $Q = 75\%$ ,  $RS = 25\%$ ]

Rule 4: **IF** Engine = A **THEN** Labor Code = 3. [ $P = 83.33\%$ ,  $Q = 83.33\%$ ,  $RS = 41.67\%$ ]

Rule 5: **IF** Engine = A **THEN** Labor Code = 2. [ $P = 16.67\%$ ,  $Q = 25\%$ ,  $RS = 8.33\%$ ]

Figure 2. Association rules generated from the sample data set.

In Figure 2, Rule 1 corresponds to objects 1 and 4 in Table 2, and Rule 2 corresponds to objects 2 and 12. The value of  $P = 33.33\%$  of Rule 1 indicates that one-third of the objects in the data set that have the condition Engine = B are covered by this rule. The value of  $Q = 100\%$  in the same rule indicates that all the objects with the decision Labor Code = 1 are covered by the rule.  $RS = 16.67\%$  indicates that only 2 out of 12 objects correspond to the rule.

In Step 4, Rules 1, 2 and 5 are discarded because at least one of the rule support parameters has a value lower than the thresholds, whereas Rules 3 and 4 are kept as potential strong rules. Finally, in Step 5, the chi-square test reveals that at  $\alpha = 0.05$ , Rule 4 is significant with  $X^2 = 5.33$  and p-value = 0.02, while Rule 3 is not significant, with  $X^2 = 1.5$  and p-value = 0.22. Table 3 provides some details on how in Step 5 of the algorithm the data in Table 2 is tabulated into 2x2 tables for Rules 3 and 4.

Table 3. 2x2 tables generated for Rules 3 and 4 and used in chi-square tests.

<b>Rule 3</b>	Labor code = 2	Labor code $\neq$ 2	Total
Engine = B	3	3	6
Engine $\neq$ B	1	5	6
Total	4	8	12
<b>Rule 4</b>	Labor code = 3	Labor code $\neq$ 3	Total
Engine = A	5	1	6
Engine $\neq$ A	1	5	6
Total	6	6	12

It needs to be emphasized that for an association rule that includes at least two decision attributes, of which only one is significantly associated with the condition attribute(s), the chi-square test may show that the rule is significant. This is similar to a problem where the chi-square test is used to find an association between two factors with more than two levels. A small p-value indicates that the two factors are associated without providing knowledge as to which level(s) of which factor describes the association between them. To address this problem, the user, for example, may break a significant 2-to-2 rules into four 1-to-1 rules and look at each rule individually to determine where the association is coming from.

In the association rule generation algorithm, the evaluation of elementary sets and their intersections requires considerable computation time. The algorithm assumes that the data table is stored in a flat file and processes it row-by-row. Such an approach may require considerable computation time when used to analyze a large data set. However, in most warranty data mining applications, data is kept in a database environment where one may use efficient set-oriented database operations, such as *projection* and *cardinality* (Garcia-Molina, 2001), to perform the computations. For example, for a table with attributes  $A = \{A_1, A_2, \dots, A_n\}$  the *projection* operation into attributes  $A' \subseteq A$  allows quick generation of all possible combinations of values for attributes  $A'$  and evaluation of the number of rows in the original table that have those attribute values. The *projection* operation can be effectively used to generate all the possible non-empty intersections of elementary sets in the association rule generation algorithm. To

illustrate its application, consider the data in Table 2 and the results of the *projection* operation for the attributes “Engine” and “Labor Code,” shown in Table 4.

Table 4. Results of the projection operation.

Engine	Labor Code	Cardinality
A	2	1
A	3	5
B	1	2
B	2	3
B	3	1

It can be easily checked that the combinations  $\{A, 2\}$ ,  $\{A, 3\}$ ,  $\{B, 1\}$ ,  $\{B, 2\}$ , and  $\{B, 3\}$  are the unique value combination for the attributes “Engine” and “Labor Code” in Table 2. The column “Cardinality” in Table 4 shows the number of rows in Table 2 that have the specified combination of attribute values. Each unique combination of attribute values corresponds to a non-empty intersection of elementary sets of the attributes “Engine” and “Labor Code,” and to a rule in Figure 2. Therefore, all the rules shown in Figure 2 can be generated from the rows of projection Table 4. For example, one may see that Rules 1 and 2 can be obtained using the fifth and third row of the projection table, respectively.

Thus, in the association rule generation algorithm, instead of computing the elementary sets of  $C$  and  $D$  and using the set intersections to generate the rules, one can obtain the same rules from the *projection* of the original table into attributes  $C \cup D$ . Since this approach utilizes highly efficient set-oriented database operations, the rule extraction process is much faster and scalable to the large datasets. Therefore, Steps 2 and 3 of the original algorithm can be modified to take advantage of the projection operation shown next:

Step 2'. Generate *projection* of the original table into attributes  $C \cup D$ .

Step 3'. For each row  $r$  in the resulting table, generate a rule of the form:

IF  $c_1 = f(r, c_1)$  AND ... AND  $c_n = f(r, c_n)$ ,

THEN  $d_1 = f(r, d_1)$  AND ... AND  $d_m = f(r, d_m)$  [P, Q, RS],

where  $P = |C_r \cap D_r| / |C_r|$ ;  $Q = |C_r \cap D_r| / |D_r|$ ;  $RS = |C_r \cap D_r| / N$ ;  $S = |C_r \cap D_r|$ ,

$C_r$  and  $D_r$  are the elementary sets of  $C$  and  $D$ , respectively, and

$|C_r \cap D_r|$ ,  $|C_r|$ , and  $|D_r|$  are obtained from the *projection* and *cardinality* of the data table to attributes  $C \cup D$ ,  $C$ , and  $D$ , respectively.

## 4. Results

In this section, the application of the association rule generation algorithm is presented with a data mining case study from the automotive industry. The emphases are on data collection, preprocessing, and analysis of computation results rather than on the algorithm itself.

### 4.1. Data source and data preprocessing

The automotive warranty data sets used in this study had been collected over a two-year period and contain 684,038 records (objects) of warranty claims for a specific vehicle model. Each object contains 88 attributes and vehicle-related problems are represented by 2,238 different labor codes with a total warranty cost of \$83,130,943. First, data preprocessing is performed to identify the completeness and usefulness of the attributes in the data set. In this step, attributes with missing and null values, redundant attributes, and attributes that are considered irrelevant (e.g., attributes representing the dealership location and type) are eliminated to improve the computation time of the algorithm and the usefulness of the results. Also in this step, the attributes with continuous values (e.g., mileage-at-repair) are clustered using the K-means clustering algorithm (Johnson and Wichern, 1998). Typically, the number of clusters is determined by the data analyst based on his or her expertise. However, in this research, Beale's F-type statistic (Johnson, 1998) is used for selecting the appropriate number of clusters for the mileage-at-repair attribute. The pseudo F-statistic is as follows:

$$F\text{-Statistic} = \frac{(W_2 - W_1)/((N - c_2)k_2 - (N - c_1)k_1)}{W_1/(N - c_1)k_1}$$

where  $c_1$  and  $c_2$  are two different numbers of clusters considered, and  $c_1 > c_2$ ,  $k_1 = c_1^{-2/p}$ ,  $k_2 = c_2^{-2/p}$ ;  $p$  is the dimensionality of the data objects to be clustered;  $W_1$  and  $W_2$  are the corresponding sums of squares within the clusters, computed from distances between the objects and their cluster means; and  $N$  is the total number of objects.

When Beale's F-Statistic  $> F_{\nu_1, \nu_2}$ , where  $\nu_1 = (N - c_2)k_2 - (N - c_1)k_1$  and  $\nu_2 = (N - c_1)k_1$ , one would prefer  $c_1$  over  $c_2$ . Using Beale's F-statistic, it is determined that the appropriate number of clusters for the attribute mileage-at-repair is fourteen.

The data preprocessing step eliminates all but ten attributes from the initial data set. These ten attributes, along with their number of possible outcomes, are shown in Table 5.

Table 5. Ten attributes used in the association rule generation algorithm.

Attribute	Number of possible values
Chassis package	5
Engine	6
Engine family	28
Fuel economy code	2213
Merchandizing model code	20
Mileage-at-repair	14
Production month-year	35
Transmission	4
Vehicle series	10
Labor code	2238

#### 4.2. Analysis of Computation Results

Once the data preprocessing is complete, the association rule generation algorithm is used to obtain rules from the data. The algorithm is implemented in a C#.NET programming environment and uses an Oracle 9i database server to organize and manipulate the data. Furthermore, a Pentium 4 (1.7 GHz, 256 Mb of RAM) personal computer is used to analyze the data. Nine out of ten remaining attributes that describe vehicle features are used as condition attributes, and the labor code that describes potential product problems is used as decision attribute in the mining algorithm.

Table 6. Labor codes with the highest warranty costs.

Rank	Labor code	# of data objects	Warranty cost	Number of rules generated (for $\alpha = 0.05$ )				
				P'=5% Q'=5%	P'=10% Q'=10%	P'=10% Q'=20%	P'=10% Q'=30%	P'=10% Q'=50%
1	C9200	1,870	\$ 4,027,262	0	0	0	0	0
2	P9945	3,203	\$ 2,924,698	2	0	0	0	0
3	Y7640	3,623	\$ 2,515,307	30	2	0	0	0
4	K2800	54,464	\$ 2,443,604	996	63	14	5	0
5	F0330	2,911	\$ 2,264,247	0	0	0	0	0
6	D3600	4,453	\$ 1,840,058	0	0	0	0	0
7	F7785	316	\$ 1,394,255	0	0	0	0	0
8	G0170	10,134	\$ 1,310,867	8	0	0	0	0
9	G5720	1,775	\$ 1,225,698	0	0	0	0	0
10	D7510	14,183	\$ 1,017,420	1,860	375	206	50	13
Total		96,932	\$ 20,963,416					

All the possible IF/THEN association rules (i.e., 1 to 1 through 9 to 1 rules) are considered. Due to the large quantity of the data and the many possible decision outcomes, an overwhelming number of association rules are generated. However, since the goal here is to find association rules that describe the major portion of the warranty costs, focus should not only be on significant rules, but also on rules that explain product problems that have a large dollar amount tied up in them. Table 6 shows the top ten labor codes in terms of warranty costs, the number of data objects associated with each labor code, and the number of significant rules generated for each of the ten labor codes assuming different threshold values for rule parameters. Ten labor codes shown in Table 6 account for approximately 25% of the total vehicle warranty costs, and therefore the analysis concentrates on finding strong association rules that may explain the root causes of vehicle problems represented by these labor codes.

The columns of Table 6 represent the number of rules extracted for each of the ten labor codes assuming various thresholds for rule support parameters, such as P' = 5% and 10%; Q' = 5%, 10%, 20%, 30%, and 50%; and  $\alpha = 0.05$ . Note that the parameter RS' is not used for rule filtering in Step 4 of the algorithm.

Because of the large number of objects in the data set, it is unlikely that any rule can correspond to a large percentage of data records. Instead, in this example, the number of data objects (QTY) is used as the rule filtering parameter and is set to  $QTY = 50$  (i.e., a rule with  $QTY \leq 50$  is dropped from further consideration). Also,  $P'$  is set from 5-10% (i.e., small values) as the data set includes a large number of decision outcomes for condition attributes (see Table 5), thus making it unlikely for rules to have a high value of  $P$ .

From the results presented in Table 6 one may see that the algorithm is able to generate statistically significant rules for the labor codes P9945, Y7640, K2800, and D7510. Examples of the extracted rules, which are categorized based on the knowledge contained in them, are shown in Figures 3, 4, and 7, along with rule strengths. Six different rules presented in Figure 3 capture the relationships between vehicle features and labor codes. Rule 1 in Figure 3 relates labor code K2800 to vehicle transmission T05. This rule accounts for 90.67% of the total cost associated with this warranty problem (\$2,215,722 out of possible \$2,443,603). The rule parameters indicate that 9.09% ( $P=9.09\%$ ) of all vehicles with the transmission T05 have a problem explained by labor code K2800, and 90.88% ( $Q=90.88\%$ ) of all vehicles with this problem have the transmission T05. The total number of data objects (i.e., transactions) that support this rule is 49,496, which is 7.24% of the total objects. Similarly, according to Rule 2, 8.99% of all vehicles with the engine E02 have the problem described by the labor code D7510, and 89.59% of all vehicles with this problem have the engine E02.

Rule 3 is a strong rule that describes relationships between multiple vehicle features (i.e., engine and engine family) and labor code D7510. One may see that Rule 3 is a somewhat extended version of Rule 2, as it includes one additional vehicle feature (i.e., engine family = EF0080) in the condition (IF) portion of the rule. By comparing Rules 2 and 3, it is apparent that there is a trade-off between the number of condition attributes in the rule, its  $Q$  value, and the dollar amount explained by the rule. When a new condition attribute (i.e., engine family) is added to the rule, the rule then becomes more unique, and the value of parameter  $Q$  decreases (see Rule 2 versus Rule 3 in Figure 3). Interesting conclusions can be drawn when one analyzes the results presented in Rules 3 and 3' simultaneously. Rule 3 indicates that vehicles equipped with the engine E02 that belong to the engine family EF0080 had a significant number of problems described by the labor code D7510. Rule 3' indicates that vehicles that had the same engine E02 that were part of an engine family other than EF0080 had a relatively small number of similar problems. In other words, an examination of Rules 3 and 3' indicates that the problem described by the labor code D7510 may be more unique to vehicles that have engine E02 and also belong to engine family EF0080. Also, by comparing warranty costs covered by Rules 2 and 3, one may see that 96% of the warranty costs (i.e., \$876,067 out of \$910,820) associated with the labor code D7510 are for the vehicles that have engine options covered in the condition portion of Rule 3. Finally, Rule 4 extends Rule 3 to include one additional vehicle feature (i.e., transmission T05) in the condition (IF) portion of the rule. Rule 4 indicates that the problem described by the labor code D7510 is more unique to vehicles that have transmission T05, engine E02, and belong to engine family EF0080. In fact, warranty cost comparison of Rules 4 and 4' indicate that when vehicles with the above engine options are not equipped with the transmission T05, the vehicle problem described by the labor code D7510 is minimal. Finally, comparison

of the warranty costs covered by Rules 2 and 4 indicates that 85% of the warranty costs (i.e., \$771,375 out of \$910,820) associated with the labor code D7510 are for vehicles that have engine and transmission options covered in the condition portion of Rule 4. From the problem root cause analysis standpoint, strong rules with a larger number of condition attributes are preferable, as they identify unique combinations of vehicle options that may develop a problem described by the labor code.

Rule 1:	<b>IF</b> Transmission = T05 <b>THEN</b> Labor code = K2800. [P = 9.09%, Q = 90.88%, RS = 7.24%, QTY = 49,496, TC = \$2,215,722]
Rule 2:	<b>IF</b> Engine = E02 <b>THEN</b> Labor Code = D7510. [P = 8.99%, Q = 89.59%, RS = 1.86%, QTY = 12,707, TC = \$910,820]
Rule 3:	<b>IF</b> Engine = E02 AND Engine Family = EF0080 <b>THEN</b> Labor Code = D7510. [P = 8.96%, Q = 86.13%, RS = 1.79%, QTY = 12,216, TC = \$876,068]
Rule 3':	<b>IF</b> Engine = E02 AND Engine Family = Others <b>THEN</b> Labor Code = D7510. [P = 9.78%, Q = 3.46%, RS = 0.07%, QTY = 491, TC = \$34,752]
Rule 4:	<b>IF</b> Engine = E02 AND Engine Family = EF0080 AND Transmission = T05 <b>THEN</b> Labor Code = D7510. [P = 9.15%, Q = 75.77%, RS = 1.57%, QTY = 10,746, TC = \$771,375]
Rule 4':	<b>IF</b> Engine = E02 AND Engine Family = EF0080 AND Transmission = Others <b>THEN</b> Labor Code = D7510. [P = 7.72%, Q = 10.36%, RS = 0.21%, QTY = 1,470, TC = \$104,693]

Figure 3. Association between vehicle feature(s) and labor code.

Figure 4 summarizes strong rules that capture interesting relationships between vehicle features, production periods, and a labor code. Rules 5, 6, 7, and 8 and Figure 5 show how association captured in Rule 1 between transmission T05 and labor code K2800 is spread out over several production periods. Relatively small Q values of these types of rules can be explained by the fact that they are calculated based on the total number of vehicles that developed a problem explained by the labor code K2800. However, when production period is also added to the decision attributes and considered in the THEN portion of the rules (see Rule 5', 6', 7', and 8'), the Q values for these rules are calculated only based on the number of vehicles with the labor code K2800 that are produced within that production period. The latter yields Q = 93.02%, 93.56%, 92.23%, and 95.24% for December, March, June, and September, respectively, and indicates strong association between transmission T05 and labor code K2800 within each production month. Figure 6 shows Q value distribution of rules that relate transmission T05 and each of the 33 production periods in the data set to a labor code K2800 (i.e., rules similar to 5, 6, 7, 8 in Figure 5). In other words, Figure 5 shows how the value of Q = 90.88% from Rule 1 is distributed among 33 production periods. From Figure 5, one can see that Q peaks during the initial stages of the production

and then decreases steadily, which reflects the decrease in the number of occurrences of labor code K2800.

Figure 6 shows a Q value trend of rules for which production period is included as part of the decision attributes (i.e., rules similar to Rules 5', 6', 7' and 8' in Figure 4). From Figure 6, one may see that the trend of Q is stable throughout the production periods. Results presented in Figure 6 indicate that regardless of the number of occurrences, labor code K2800 and transmission T05 are closely associated. Furthermore, from the above discussion and results presented in Figure 6, one may also conclude that the root cause of the problem described by K2800 is not due to production issues, as the Q value is very stable based on month-to-month production. However, if one were to notice large fluctuations in Q in Figure 6, this may be an indicator that there were some production issues during that time period that may have contributed to the problem K2800.

- Rul e 5: **IF** Production month = December AND Production year = 2001 AND Transmission = T05 **THEN** Labor code = K2800.  
[P = 15.77%, Q = 7.26%, RS = 0.58%, QTY = 3,956, TC = \$179,489]
- Rul e 5': **IF** Production month = December AND Production year = 2001 AND Transmission = T05 **THEN** Labor code = K2800 AND Production month = December AND Production year = 2001.  
[P = 15.77%, Q = 93.02%, RS = 0.58%, QTY = 3,956, TC = \$179,489]
- Rul e 6: **IF** Production month = March AND Production year = 2002 AND Transmission = T05 **THEN** Labor code = K2800.  
[P = 10.67%, Q = 4.80%, RS = 0.38%, QTY = 2,614, TC = \$117,063]
- Rul e 6': **IF** Production month = March AND Production year = 2002 AND Transmission = T05 **THEN** Labor code = K2800 AND Production month = March AND Production year = 2002.  
[P = 10.67%, Q = 93.56%, RS = 0.38%, QTY = 2,614, TC = \$117,063]
- Rul e 7: **IF** Production month = June AND Production year = 2002 AND Transmission = T05 **THEN** Labor code = K2800.  
[P = 9.20%, Q = 2.81%, RS = 0.22%, QTY = 1,532, TC = \$65,798]
- Rul e 7': **IF** Production month = June AND Production year = 2002 AND Transmission = T05 **THEN** Labor code = K2800 AND Production month = June AND Production year = 2002.  
[P = 9.20%, Q = 92.23%, RS = 0.22%, QTY = 1,532, TC = \$65,798]
- Rul e 8: **IF** Production month = September AND Production year = 2002 AND Transmission = T05 **THEN** Labor code = K2800.  
[P = 9.22%, Q = 4.48%, RS = 0.36%, QTY = 2,439, TC = \$107,047]
- Rul e 8': **IF** Production month = September AND Production year = 2002 AND Transmission = T05 **THEN** Labor code = K2800 AND Production month = September AND Production year = 2002.  
[P = 9.22%, Q = 95.24%, RS = 0.36%, QTY = 2,439, TC = \$107,047]

Figure 4. Association between vehicle feature(s), production period, and labor code.

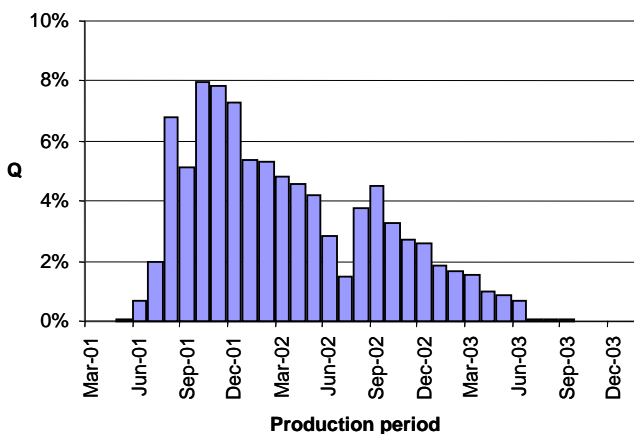


Figure 5. Q value distribution of rules that relate transmission T05 and each of the 33 production periods to labor code K2800.

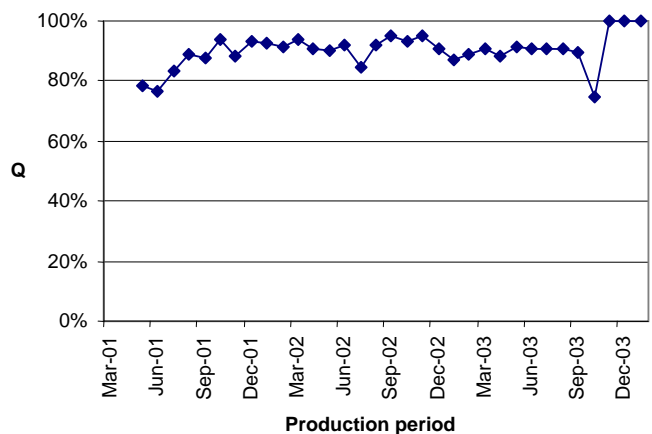


Figure 6. Q value trend of rules for which production period is included as part of a decision attribute.

Rule 9:	<b>IF</b> Mileage-at-repair = (58046, 150000) <b>THEN</b> Labor code = Y7640. [P = 32.29%, Q = 5.69%, RS = 0.03%, QTY = 206, TC = \$147,865]
Rule 10:	<b>IF</b> Mileage-at-repair = (21319, 25073) AND Transmission = T05 <b>THEN</b> Labor code = P9945. [P = 1.30%, Q = 16.45%, RS = 0.08%, QTY = 527, TC = \$483,687]
Rule 11:	<b>IF</b> Mileage-at-repair = (29047, 33223) AND Transmission = T05 <b>THEN</b> Labor code = P9945. [P = 1.46%, Q = 14.27%, RS = 0.07%, QTY = 457, TC = \$433,554]
Rule 12:	<b>IF</b> Mileage-at-repair = (40772, 58000) AND Transmission = T05 <b>THEN</b> Labor code = P9945. [P = 6.59%, Q = 5.78%, RS = 0.03%, QTY = 185, TC = \$133,833]

Figure 7. Association between vehicle feature(s), mileage-at-repair, and labor code.

Figure 7 provides examples of association rules between mileage-at-repair, vehicle features, and labor code. Rule 9 in Figure 7 indicates that approximately one third ( $P=32.39\%$ ) of the vehicles develop a problem described by the labor code Y7640 after reaching 58,000 miles of usage. However,  $Q=5.69\%$  in Rule 9 indicates that majority of the vehicles (94.31%) developed Y7640 type problems at lower mileage. Rules 10, 11, and 12 develop useful relationships between a vehicle feature (transmission T05), mileage-at-repair, and a labor code (P9945). Figure 8 illustrates Q value distribution for the aforementioned type of rules for fourteen mileage-at-repair intervals defined by the clustering algorithm. Results in Figure 8 reveal that majority of the vehicles with the transmission T05 developed a problem described by the labor code P9945 when the vehicle usage mileage was between 18,000 to 40,000 miles. This also indicates that approximately 75% of these types of problems occur within the three-year or 36,000-mile warranty period, an important threshold for most automotive manufacturers.

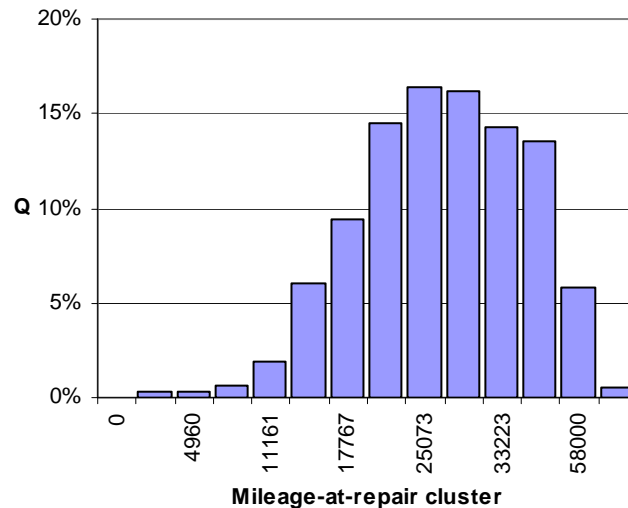


Figure 8. Q value trend of rules that relate transmission T05 to each of the fourteen mileage-at-repair intervals identified by the clustering algorithm.

Finally, the computation analysis presented next illustrates the importance of a database “projection” operation (see Section 3) in the association rule generation algorithm. Summaries of computer processing

times required to generate association rules from various-size datasets with and without a database “projection” operation are shown in Table 7. From Table 7, one may see that the database “projection” operation significantly reduces the computation time of the association rule generation algorithm when evaluating association rules from large datasets. For example, for a dataset with 50,000 objects, the association rule generation algorithm that utilizes “projection” operation generates all the possible rules within a second on a 2 GHz PC. Without database operation, it takes more than 17 minutes to produce the same results. From Table 7 one may see that as the size of the dataset increases, the effectiveness of the “projection” operation presented in Section 3 becomes more apparent and critical. To obtain a solution for a database with 300,000 objects, the algorithm that uses a “projection” operation produces the desired list of association rules within 42 seconds, while without this procedure it takes more than ten hours to produce the same results.

Table 7. Analysis of computer processing time of the association rule generation algorithm.

Number of objects	Computation Time (min:ss)	
	Association rule generation algorithm	
	with database “projection” operation	without database “projection” operation
50,000	00:01	17:19
100,000	00:03	71:46
200,000	00:06	285:31
250,000	00:39	440:48
300,000	00:42	> 600:00

## 5. Conclusions

In this research a new association rule generation algorithm for the analysis of automotive warranty data was presented. The algorithm combined elementary set concept and database operation functions to extract useful knowledge from warranty data. This knowledge was represented in the form of IF-THEN association rules, where the IF portion of the rule contained product feature attributes and THEN portion included problem-related labor code. Once association rules are generated, the algorithm applies statistical analysis techniques to evaluate the significance of each rule. The rules that pass the significance test are reported in the solution. Application of the association rule generation algorithm was presented with a data mining case study from the automotive industry. The knowledge (rules) extracted from the automotive warranty data were used to identify root causes of a particular warranty problem and to develop useful conclusions. Detailed discussion on the source and characteristics of warranty data were also presented. Experience with the automotive example showed that the application of data mining techniques to real world problems is difficult. Unless the information technology systems are designed to capture the desired information and able to easily piece together the disparate pieces of data, the quality of the knowledge extracted from the data may be uncertain. Therefore, future research in warranty data mining should concentrate on the development of techniques that allow for the fast and easy integration of data from different sources (i.e., assembly plants, dealerships, and repair shops) into one unified warranty data warehouse. Development of effective text mining algorithms that can extract knowledge from textual warranty input may significantly help this process. Finally, integrating association rule

generation methods with sequential pattern analysis techniques, which concentrate on finding relationships between attributes of warranty problems that occur over time, should also be the focus of future research.

## 6. Impact

### **Educational**

Data mining models and algorithms developed in this research as well as documented automotive case studies have been incorporated into the new course IMSE 560 Data Warehousing and Mining offered by the Department of Industrial and Manufacturing Systems Engineering. One of the main tasks of the proposed project was the testing and validation of the data mining algorithm on real automotive warranty data and the study of the effectiveness (i.e., number of extracted rules, importance of extracted rules, insight gained) and efficiency of the algorithms (i.e., computational time and quality of solutions). All of these results are presented to IMSE 560 students, which significantly enhances the learning process. In addition, the approaches, algorithms, and results of this research project were also introduced into the course IMSE 533 Manufacturing Systems, taught by the Professor Zakarian, co-PI of this project, in the Fall 2004 and 2005 semesters. Practical data mining problems are introduced and studied in this course.

Finally, three different Graduate Student Research Associates were supported by this research grant. All three successfully completed their MS degree requirements in College of Engineering and Computer Science at the University of Michigan-Dearborn.

1. Pallavi Bhalerao, MS, Computer and Information Sciences, May 2006. Currently at Motorola, Ms. Bhalerao developed models, algorithms and software for sequential pattern mining in automotive warranty data.
2. Yuri Siradeghyan, MS, Computer and Information Sciences, May 2006. Currently at Microsoft Research, Mr. Siradeghyan worked on the development of association rule generation algorithms and software for the mining of automotive warranty data.
3. Adisorn Preutisranyanont, MS, Engineering Management, December 2005. Mr. Preutisranyanont worked on the development of association rule generation algorithms for automotive warranty data.

### **Industrial**

In this research, a new association rule generation algorithm for warranty data analysis was developed. Application of the algorithm was illustrated with an automotive warranty data analysis example. The knowledge extracted from the warranty data was represented by association rules, which relate product attributes to warranty problems, and were used to identify associations between product features and the occurrence of a particular warranty problem.

This research also developed sequential pattern mining algorithms for extracting hidden knowledge from a large amount of warranty data. The algorithm used elementary set concept and database manipulation techniques to search for patterns or relationships among occurrences of warranty claims over time. Significant patterns provide knowledge of one or more product failures that lead to future product fault(s).

Data mining algorithms developed in this research provide solid technical framework for analysis of large manufacturing data sets. Also, the industrial case study developed in this research clearly illustrates how the results obtained from the algorithm can be used in industrial settings for quality control purposes and for the identification of root causes of production problems.

## 7. Acknowledgments

The principal investigators would like to thank the General Motors Corporation and its Quality Warranty Data Group for their close collaboration on this project. We would like also to thank all three Graduate Students Research Assistants for their hard and tireless efforts on this project.

## 8. References

- Baird, P. *Robert Bosch Corporation Failure Modes Effects Analysis (FMEA)*. Read Associates: Farmington Hills, MI, 2000.
- Blischke, W.R., and D.N.P. Murthy. *Warranty Cost Analysis*. Marcel Dekker: New York, 1994.
- Blischke, W.R., and D.N.P. Murthy. *Product Warranty Handbook*. Marcel Dekker: New York, 1996.
- Breiman, L., and J.H. Friedman. *Classification and regression trees*. Wadsworth International: Belmont, CA, 1984.
- Cali, J. *TQM for Purchasing Management*. McGraw Hill: New York, 1993.
- Cantu-Paz, E., and C. Kamath. "Inducing oblique decision trees with evolutionary algorithms." *IEEE Transactions on Evolutionary Computation* 7 (2003): pp. 54-68.
- Chakrabarti, S., S. Roy, and M.V. Soundalgekar. "Fast and accurate text classification via multiple linear discriminant projections." *The VLDB Journal* 12 (2003): pp. 170-85.
- Cooke, T. "Two variations on Fisher's linear discriminant for pattern recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (2002): pp. 268-73.
- Feng, J., J. Wang, and J. Wang. "An Optimization Model for Concurrent Selection of Tolerances and Suppliers." *Computers and Industrial Engineering* 40 (2001): pp. 15-33.
- Garcia-Molina, H., J.D. Ullman, and J.D. Widom. *Database Systems: The Complete Book*. Prentice Hall: Upper Saddle River, NJ, 2001.
- Gehrke, J.E., R. Ramakrishnan, and V. Ganti. "RainForest - A framework for fast decision tree construction of large datasets." *Data Mining and Knowledge Discovery* 4 (2000): pp. 127-62.

- Han, J., and M. Kamber. *Data Mining: Concept and Techniques*. Morgan Kaufmann: San Francisco, 2001.
- Hotz, E., U. Grimmer, W. Heuser, G. Nakhaeizadeh, and M. Wieczorek. "REVI-MINER, a KDD-environment for deviation detection and analysis of warranty and goodwill cost statements in automotive industry." *Proceeding of the 7<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2001): pp. 432-37.
- Hotz, E., G. Nakhaeizadeh, B. Petzsche, H. Spiegelberger. "WAPS, a data mining support environment for the planning of warranty and goodwill costs in the automobile industry." *Proceeding of the 5<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (1999): pp. 417-19.
- Hu, X.J., and J.F. Lawless. "Estimation from truncated lifetime data with supplementary information on covariates and censoring times." *Biometrika* 83 (1996): pp. 747-61.
- Johnson, D.E. *Applied Multivariate Methods for Data Analysts*. Duxbury Press: Pacific Grove, CA, 1998.
- Johnson, R.A., and D.W. Wichern. *Applied Multivariate Statistical Analysis*. 4<sup>th</sup> ed. Prentice-Hall: Upper Saddle River, NJ, 1998.
- Kalbfleisch, J.D., J.F. Lawless, and J.A. Robinson. "Methods for the analysis and prediction of warranty claims." *Technometrics* 33 (1991): pp. 228-73.
- Karim, M.R., W. Yamamoto, and K. Suzuki. "Statistical analysis of marginal count failure data." *Lifetime Data Analysis* 7 (2001): pp. 173-86.
- Kusiak, A. "Decomposition in data mining: an industrial case study." *IEEE Transactions on Electronics Packaging Manufacturing* 23 (2000).
- Kusiak, A. "Rough set theory: A data mining tool for semiconductor manufacturing." *IEEE Transactions on Electronics Packaging Manufacturing* 24 (2001): pp. 44-50.
- Kusiak, A., J.A. Kern, K.H. Kernstine, and B.T. Tseng. "Autonomous decision-making: a data mining approach." *IEEE Transaction on Information Technology in Biomedicine* 4 (2000): pp. 274-85.
- Kusiak, A., and C. Kurasek. "Data mining of printed-circuit board defects." *IEEE Transaction on Robotics and Automation* 17 (2001): pp. 191-97.
- Kusiak, A., I.H. Law, and M. Dick. "The G-algorithm for extraction of robust decision rules - children's postoperative intra-atrial arrhythmia case study." *IEEE Transactions on Information Technology in Biomedicine* 5 (2001): pp. 234-55.
- Lawless, J.F. "Statistical analysis of product warranty data." *International Statistical Review* 66 (1998) pp. 41-60.
- Lawless, J.F., and J.D. Kalbfleisch. "Some issues in the collection and analysis of field reliability data." *Survival Analysis: State of the Art*. Ed. J.P. Klein and P.K. Goel. Amsterdam: Kluwer, 1992. pp. 141-52.
- Lu, M.W. "Automotive reliability prediction based on early field failure warranty data." *Quality and Reliability Engineering International* 14 (1998): pp. 103-8.

- Majeske, K.D., T.C. Lynch, and G.D. Herrin. "Evaluating product and process design changes with warranty data." *International Journal of Production Economics* 50 (1997): pp. 79-89.
- Majeske, K.D., and G.D. Herrin. "Assessing mixture model goodness-of-fit with an application to automobile warranty data." *Proceedings of the Annual Reliability and Maintainability Symposium* (1995): pp. 378-83.
- Mehta, M., Agrawal, R., and Risanen, J., "SLIQ: A fast scalable classifier for data mining," *Proceeding of the 5<sup>th</sup> International Conference on Extending Database Technology (EDBT)* (1996): pp. 18-32.
- Murthy, S.K. "Automatic construction of decision trees from data: a multidisciplinary survey." *Data Mining and Knowledge Discovery* 2 (1998): pp. 345-89.
- Murthy, D.N.P., I. Djameludin, and R.J. Wilson. "A consumer incentive warranty policy and servicing strategy for products with uncertain quality." *Quality and Reliability Engineering International* 11 (1995): pp. 155-63.
- Ohrn, A., L. Ohno-Machado, and T. Rowland. "Building manageable rough set classifiers." *Proceedings of AMIA Symposium* (1998): pp. 543-7.
- Pawitan, Y. *In All Likelihood: Statistical Modeling and Inference Using Likelihood*. Oxford University Press: London, 2001.
- Pawlak, Z. "Rough sets." *International Journal of Information and Computer Sciences* 11 (1982): pp. 341-56.
- Pawlak, Z. *Rough Sets – Theoretical Aspects of Reasoning about Data*. Kluwer Academic: Boston, 1991.
- Pawlak, Z. "Rough set approach to knowledge-based decision support." *European Journal of Operational Research* 99 (1997): pp. 48-57.
- Polatoglu, H., and I. Sahin. "Probability distributions of cost, revenue and profit over a warranty cycle." *European Journal of Operational Research* 108 (1998): pp. 170-83.
- Quinlan, J.R. "Induction of decision trees." *Machine Learning* 1 (1996): pp. 81-106.
- Quinlan, J.R. *C4.5: Programs for machine learning*. Morgan Kaufmann: San Mateo, CA, 1993.
- Robinson, J.A., and G.C. McDonald. *Issues related to Field Reliability Data, Data Quality Control: Theory and Pragmatics*. Marcel Dekker: New York, 1991.
- Ruggieri, S. "Efficient C 4.5." *IEEE Transactions on Knowledge and Data Engineering* 14 (2002): pp. 438-44.
- Sahin, I., and H. Polatoglu. *Quality, Warranty and Preventive Maintenance*. Kluwer Academic: Boston, 1998.
- Shafer, J., R. Agrawal, and M. Mehta. "SPRINT: A scalable parallel classifier for data mining," *Proceeding of International Conference on Very Large Databases* (1996): pp. 544-55.
- Singpurwalla, N.D., and S.P. Wilson. "Failure models indexed by two scales," *Advances in Applied Probability* 30 (1998): pp. 1058-72.

- Suzuki, K., W. Yamamoto, M.R. Karim, and L. Wang. "Data analysis based on warranty database." *Recent Advances in Reliability Theory: Methodology, Practice, and Inference*. Ed. N. Limnios and M. Nikulin. Birkhauser: Boston, 2000.
- Suzuki, K., M.R. Karim, and W. Yamamoto. "Statistical analysis of reliability warranty data." *Advances in Reliability*. Ed. N. Balakrishnan and C.R. Rao. Elsevier: London, 2001.
- Tickle, A.B., R. Andrews, M. Golea, and J. Diederich. "The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks." *IEEE Transactions on Neural Networks* 9 (1998): pp. 1057-68.
- Tsukimoto, H. "Extracting rules from trained neural networks." *IEEE Transactions on Neural Networks* 11 (2000): pp. 377-89.
- Wang, L., and K. Suzuki. "Nonparametric estimation of lifetime distribution from warranty data without monthly unit sales information." *Journal of the Reliability Engineering Association of Japan* 23 (2001): pp. 145-54.
- Zhang, G.P. "Neural networks for classification: A survey." *IEEE Transactions on Systems, Man, And Cybernetics – Part C: Applications and Reviews* 30 (2004): pp. 451-62.
- Zhang, D., and L. Zhou. "Discovering golden nuggets: Data mining in financial application." *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews* 34 (2004): pp. 513-22.