A Data Mining Approach to the Joint Evaluation of Field and Manufacturing Data in Automotive Industry

Christian Manuel Strobel[1] and Tomas Hrycej[2]

[1] Universität Karlsruhe, Germany
mstrobel@statistik.uni-karlsruhe.de
[2] DaimlerChrysler AG, Research and Technology, 89081 Ulm, Germany
Tomas.Hrycej@daimlerchrysler.com

Abstract. The manufacturing quality can be evaluated only by considering the failure behavior of the product in the field. When relating manufacturing events to failure events, the main challenge is to master the huge number of combinations of both event types, of which each is only covered by a small number of occurrences. Additionally, this leads to the problem of selection of interesting findings - the appropriateness of the selection criterion for consequent decision making is a critical point. Another challenge is the necessity of mapping the process of manufacturing tests to a vector of variables characterizing the manufacturing process. The solution presented, focuses on correct rule generation and selection in the case of combinations with low coverage. Therefore statistical and decision theory approaches were used. The multiple hypothesis aspect of the rule set has also been considered. The application field was quality control of electronic units in automotive assembly, with thousands of variables observed.

1 Introduction

In the automotive assembly, testing of individual functions is frequently embedded in the assembly process. The focus is on testing the functionality of electronic units, whose number amounts, in the premium car segment, are up to several hundreds.

The ultimate criterion of manufacturing quality is the behavior of the product in the field. This is why it's important to learn as much as possible about the relationships between the assembly and testing procedure on the one hand, and the field quality measures on the other hand. In this case the task is to examine two different databases, each representing one part of the product life cycle: the assembly database, which stores the results of the electronic unit tests during the manufacturing process, and the field failure data collected in a company-wide service and warranty database.

The relationships between the assembly process and the field failure can be captured, in a the classical Data Mining manner, by association rules for which exist numerous algorithms. However, several fundamental characteristics of our application made some additional developments necessary:

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): PKDD 2006, LNAI 4213, pp. 625–632, 2006.

1. There is a considerable number of assembly process features and failure types. Even in large event data bases, each event combination is very scarce, which leads to the necessity to evaluate the statistical properties of rules based on small samples. Also it's not necessary to consider complex combinations of three and more variables, because the support can be expected to tend to zero.
2. The interestingness of results found has to be viewed in economic terms: interesting rules are those, that have a large economic potential.
3. The large number of process features and failure types typically leads to a large number of rules found. Even with rule selection criteria considering the statistical significance, a certain amount of the rules found are random. Therefore, to have an idea about the reliability of the results, a rough assessment of the proportion of such random rules is desirable.

These aspects are specific, but not restricted to this application. Many large scaled industrial applications exhibit similar characteristics.

2 Process Model of Manufacturing Tests

The testing process for electronic units in the automotive industry is rather complex. Each car is tested several times during the assembly process to ensure an error-free delivery to the costumer. During each test, a pre-defined, assembly-status based scale of electronic units and their individual and interrelated functionality are tested. Each test is either performed automatically or with the additional input of a assembly worker and has basically two outcomes: either positive (without errors but eventually with warnings), possibly accompanied by the list of warnings, or negative, accompanied by the list of errors. Therefore, if a particular test results in a warning or error message of the tested devices, several consequent actions may be taken i.e. ignoring the message, repeating the test, taking the car out of the assembly line (so called rework) and so on.

This variety of procedures implies, that the same test may be performed more than once, with different results, each of which has to be interpreted individually. In order to combine and compare the field results with the test results, it is necessary to transform the test information from the manufacturing process into a binary representation, holding relevant information from the whole testing process of a specific electronic unit in a car during assembly - the so called *history of a electronic unit*. It captures the information which error was reported[1], whether this error emerge in other tests or not, did this error lead to rework and did this error appear as a warning in another test. Additionally, the production weekday and month are logged to capture time fluctuations of the quality.

3 Rule Generation

To improve the assembly process and its testing procedure, the relationship between the attributes of the assembly and testing process on the one hand, and the

[1] In the binary coding: has a given error appeared or not.

field failures on the other hand has to be clarified and hence can mathematically be described by association rules.

Let $I = \{i_1, i_2, ..., i_m\}$ and $J = \{j_1, j_2, ..., j_m\}$ be sets of items or literals. Let D be a transaction database over $I \times J$ where each transaction $T = (T_a, T_b)$ is a subset of $I \times J$. Further let A be a set of items with $A \subseteq T_a$. The implication of the from $A \rightarrow B$ with $A \subseteq I$ and $B =\subseteq J$ is called *association rule*. The *support*[2] of a rule $A \rightarrow B$ can be defined as the numbers of transactions containing A and B ($n_{A \wedge B}$) beeing $n_{A \wedge B} := |\{T \in D : A \cup B \subseteq T\}|$, $n_A := |\{T \in D : A \subseteq T_a\}|$, $n_B := |\{T \in D : B \subseteq T_b\}|$ and $n := |D|$.

However, the causes of a failure may be more complex: Boolean functions of multiple assembly terms describe the influence. Sophisticated algorithms exist for solving this *frequent item set problem* [2,5]. Considering the fact that even the occurrences of simple terms are scarce, the support of simple rules of type $A \rightarrow B$ can be expected to be very low. This makes the probability of finding complex rules of three and more antecedent terms with a significant support to vanish. Instead, all pairs of significant (see the next section) simple rule pairs $A_i \rightarrow B, A_j \rightarrow B$ have been combined to rules of the form $f(A_i, A_j) \rightarrow B$ with f being some of 8 possible nontrivial Boolean functions of two arguments, having higher significance[3] than each simple rule.

4 Rule Significance

In machine learning community, it is usual to select interesting rules with help of measures such as support, confidence, lift or combinations of both [8]. This approach may be adequate for a support exceeding some substantial count such as 30. However, many applications, including the ours, lack this property as obvious from the right graph in Fig. (2). For example in quality monitoring, early warning is essential. The number of failures occurred grows with the time interval of observation. The same rules might have been found earlier, if the observation interval would have been shorter, and it's thus always preferable, to use the shorter interval. Consequently, interesting rules cannot, per definition, exhibit a high support - it is necessary to discover them with a lower support possible. This is why statistical measures of significance have to be used for application of this type.

A widely used measure is the χ^2-measure of mutual dependence. Unfortunately, this measure can answer only the question how certain it is, that there is *some* dependence between the variables, no matter how strong this dependence is. What is really needed, is a statistically founded statement about the strength of dependence, which would allow us to identify strong and significant influences on the failure rates. One possibility are interval estimates of interestingness measures. An example for this is what we have called ,,safe lift" - an interestingness

[2] Mostly the support of a rule is defined as fraction of transactions (see i.e. [6,1]). In our application we always refer to the number count when talking about support.

[3] Here significance refers to an appropriated interestingness measure as explained in the following sections.

measure based on the well known lift. It can be expressed with help of sample counts (see Section (3)) and substituting sample probabilities for real ones:

$$L(B|A) = \frac{P(B|A)}{P(B)} = \frac{\frac{n_{A\wedge B}}{n_A}}{\frac{n_B}{n}} \tag{1}$$

Obviously, neither $P(B|A)$ nor $P(B)$ can be determined exactly, since both are computed from sample counts ($n_{A\wedge B}, n_A$, n_Bandn respectively). Assuming that n_B and n are sufficiently large (which is frequently satisfied), we can confine ourselves to an interval estimate of $P(B|A)$, given sample counts $n_{A\wedge B}$ and n_A.

The confidence interval estimate for the posteriori conditional probability based on Gaussian approximation of binomial distribution (with sample count n_A, number of positive outcomes ($n_{A\wedge B}$)) is the Wilson confidence interval $[T^l_{n_A}(conf(A \to B)), T^u_{n_A}(conf(A \to B))]$ (see [4]):

$$T^l_{n_A} = \frac{1}{1+\frac{q^2_\alpha}{n_A}} \left(p + \frac{q^2_\alpha}{2n_A} - q_\alpha \sqrt{\frac{p(1-p)}{n_A} + \frac{q_\alpha}{4n^2_A}} \right) \tag{2}$$

with $q_\alpha := u_{1-\frac{\alpha}{2}}$ (the quantile of the standard normal distribution for given significance level α) and $p := \frac{n_{A\wedge B}}{n_A}$ being the empirical posterior conditional probability. For sufficiently large n_A, we can say $\frac{1}{1+\frac{q^2_\alpha}{n_A}} \to 1$, $\frac{q^2_\alpha}{2n_A} \to 0$, $\frac{q_\alpha}{4n^2_A} \to 0$ and therefore the lower interval bound (2) simplifies to $p - q_\alpha \sqrt{\frac{p(1-p)}{n_A}}$. The "safe lift" can thereby be defined as the ratio of the simplified lower interval bound and the posteriori probability of the consequence[4].

5 Decision Oriented Rule Selection

To select the *best* rules within the founded rules, the goal of the application has to be considered. One possibility is the attempt, to quantify the benefit from knowing the rule.

Each rule concerning a field failure gives a hint to it's causes and can be used to take an action ro reduce the failure rate. In an idealized setting, the benefit of such an action can be quantified i.e. in monetary terms. Usually, these effects can only be quantified partially. The effectiveness of the action taken cannot be predicted exactly, but a reasonable assumption may be, that the action may reduce the conditional failure rate to the level of the unconditional rate. For example, if the failure rate of reworked cars is higher than an average one, the reworking process may be scrutinized to reach at least the average performance. Then, the total savings through the failure rate reduction would amount to:

$$n_A \left(P(B|A) - P(B)\right) c_B = \left(n_{A\wedge B} - \frac{n_A n_B}{n} \right) c_B \tag{3}$$

with c_B being the unit costs of the failure B - or so called potential.

[4] In [7,10] a similar interesting measure was introduced as a kind of correlation factor.

The quantification of the action costs will scarcely be feasible in advance, since the actions themselves will be mostly designed only after viewing the rules. This is why the action costs will not be able to be included in the decision measure, and so we did in this application.

6 Multiple Hypothesis Problem

For an application as the present one, it is characteristic that the total number of significant findings can be very high. Therefore, a specific problem becomes particularly serious: the multiple hypothesis problem.

Rules received from a sample data set are a result of a random process. This randomness is quantified by significance measures. Accepting a rule on a certain significance level means that it is highly improbable that the law described by the rule does not exist. Nevertheless, the rule still *might* be a result of pure chance - only the probability of this is low. If our goal was to accept or reject all rules together, the procedure of testing individual rules against a given individual significance level was completely faulty. If each rule was a result of a rejection of a null hypothesis on a significance level π, the probability, that at least one rule is still a random product (while its null hypothesis is valid) is $\pi = 1 - (1 - \alpha)^n$. Vice versa, to ensure that the totality of rules are valid on the given level π, the significance level of individual rules would be $\alpha = 1 - (1 - \pi)^{\frac{1}{n}}$, a drastically low figure[5] for n of the order or magnitude of hundreds or thousands. Fortunately, we do not have to be so ambitious. A good rule set is such, that we can trust at least a significant portion of it. But it is still desirable to assess how large the portion of trustworthy rules really is - otherwise we would risk there are none.

The qualitative relationships within this problem can be illustrated with help of hypotheses about Gaussian distribution. Suppose the null hypothesis is, that a sample is drawn from the distribution $N(0, 1)$ while the alternative hypothesis postulates $N(m, 1)$. For a given significance level α, if the sample value is greater than the quantile q_α (such that $1 - F(q_\alpha) = \alpha$), the probability of this sample value if the sample is drawn from the null-hypothesis distribution $N(0, 1)$ is α while the corresponding probability if the sample is drawn from $N(m, 1)$ is $\gamma = 1 - N(q_\alpha - m, 1)$. The values of γ in dependence from α for some mean values of the alternative hypothesis m are given in left graph of Fig. (1) and the the ratio of γ/α in the right graph respectively. Obviously, the ratio grows with diminishing α. This is exactly what we expect: for higher significance level, the separation between correctly identified alternative-hypothesis samples, and wrongly identified null-hypothesis samples improves.

Going back to the case of discovering rules, suppose there are N rules whereof M are valid. Consequently $N - M$ rules are invalid, corresponding to the null hypothesis of no dependence between the antecedent and consequent. With a significance level of α, we can expect, in average, the testing algorithm to deliver $E_i = (N - M)\alpha$ invalid, and $E_v = M\gamma$ valid rules. In reality, neither M nor γ is known in advance. However, if the algorithm proposes K rules, we know that

[5] In literature this value is often referred to as Bonferroni correction [3,9].

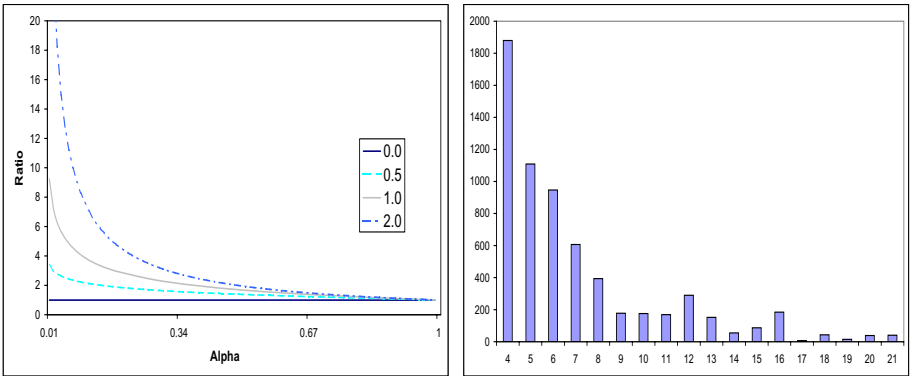

Fig. 1. Left: Ratio γ/α for different mean values of the alternative hypothesis m. Right: Histogram of the support for simple rules applying the number counts from Table (1).

this corresponds to the sum $K = (N - M)\alpha + M\gamma = N\alpha + M(\gamma - \alpha)$. Since $N\alpha > (N - M)\alpha$, we can expect a minimum of $K - N\alpha$ rules to be valid. Thus the ratio $\frac{K}{N\alpha}$ gives a good idea of proportion of usable rules in the rule set.

More precisely, we have estimated the number of invalid, randomly generated rules with help of its expected value $(N-M)\alpha$. $N\alpha$ was the upper bound of this estimate. For large values of $(N - M)\alpha$, the sample fluctuations around the expected value are small. For small $(N-M)\alpha$, an interval estimate is more reliable. The number of invalid rules generated by the algorithm is binomial distributed with sample size $N - M$ and probability α. The standard deviation of this number is $\sqrt{(N - M)\alpha(1 - \alpha)} \approx \sqrt{(N - M)\alpha}$. So the upper interval boundary is approximately $(N - M)\alpha + q\sqrt{(N - M)\alpha}$, with q being some quantile. Since we are ignorant of the value of M, the upper bound to be compared with the total number of rules found K is $L(\alpha) := N\alpha + q\sqrt{N\alpha}$.

7 Computing Results and Conclusion

There was a considerable number of variables in the described automotive test application with 3789 field failure (goal variables) and 3310 process attributes (influence variables). The aim of our application was to find relevant rules within this domain that are useful for the quality improvement process. Therefore there were 15,520,749 simple rules to examine and following the description of the previous section and applying the constraints from Table (1) we obtained a set of 7441 rules of which 6910 were simple and 525 complex rules.

The low support of complex rules is striking (see left graph of Fig. (2)), which seems to substantiate the commitment to a maximum of two antecedent variables (three or more influence variables would hardly produce a non-trivial support). Also the distribution of lift based on the contingence table (see right graph of Fig. (2)) shows, that there is a considerable number of seemingly valid rules,

Table 1. Rule pre-selection constraints used in the field feedback application

Constrains	Value	Constrains	Value	Constrains	Value
$min_{n_{A\wedge B}}$	4	Factor for std. deviation	1	α for safe lift	0.2
$min_{n_A} = min_{n_B}$	4	Factor for composed rules	1.5	minimum safe lift	1.5

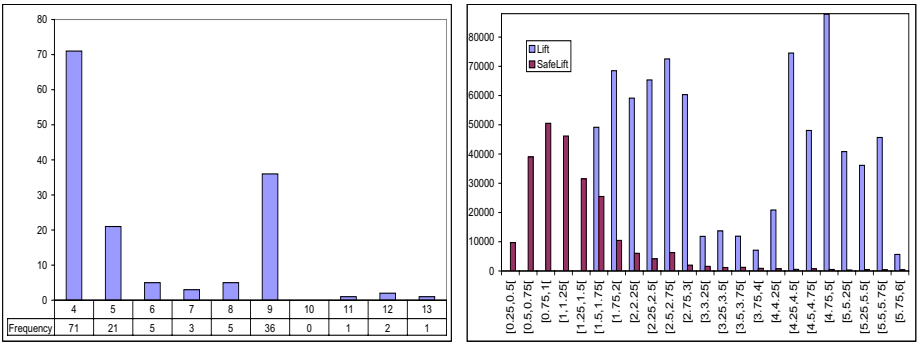

Fig. 2. Left: Histogram for $n_{A\wedge B}$ for significant complex rules Right: Histogram of lifts and safe lifts for significant rules

having a lift ≥ 1.5. However, the histogram of safe lift (Fig. (2)) exhibits, that the bulk of these rules are hardly statistically significant as only 80% of these rules have a safe lift ≥ 1.0. This shows, how important it is, to consider the statistical properties and justifies the pre-selection criterion of rules based on the safe lift.

The search for economically useful rules makes a trade-off between the extensive (considering many rules) and intensive (only the most important rules) approach necessary. This can be reached by the potential measure defined in Section 5. Ordering and accumulating the rules by their potential, results into a marginal utility function, which allows to set an economically meaningful cut-off point. The slope of this function can be compared with the estimated average costs of considering the rule and taking a corrective action. The cut-off point would have to be set at the potential value for which the slope and the costs are equal.

The multiple hypothesis problem can be illustrated by a selected subset of rules for a particular electronic unit. From the total number of 218363 possible rules of this subset, only 2900 were considered as candidates[6]. Using the constraints listed in Table (1) we found 61 rules. These rules were significant on different levels. For a given significance level, we can compute the expected mean number, and the upper bound of ,,random" rules, as discussed in the previous section. For example, there were 35 rules with $\alpha > 0.005$. From these rules, at

[6] Applying $n_A > 10$ and $n_B > 4$ reduced the overall amount of rules to the candidate rules.

most 17 can be expected to be random, using the last formula of Section 6. This is a considerable number, but still an acceptable proportion.

Concluding, in this work we have presented a framework for the analysis of quality data, with the goal to find relationships between the assembly and testing process and the failures in the field. The basic method was the search for association rules. However, several characteristics of our application lead to extensions of mainstream methods:

1. It was necessary to map the assembly process data to a feature vector, with help of expert knowledge.
2. The considerable number of variables involved lead to high number of value combinations and thus to low supports of individual candidate rules. This made considering the statistical properties of rule selection criteria necessary, to avoid the risk of obtaining random rules.
3. Due to the industrial context, the rules selection criteria had to consider the economic impact of the rules.
4. The large number of rules obtained by the sample, justifies the question how many of them are really valid and which are random. We used a simple framework to determine these proportions and showed that a large part of the rules founded are valid.

Since these characteristics are shared by many industrial applications, the solutions presented may be useful for a broad scope of Data Mining problems.

References

1. C. C. Aggrawal and P. Yu. Online generation of association rules. *ICDE*, 1998.
2. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. *Proceedings 20th Int. Conference on Very Large Data Bases*, pages 487–499, 1994.
3. Y. Benjamini and H. Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.*, B:289–300, 1995.
4. T. T. D.-A. Brown, L. D. Cai. *Statistical Science, 16(2)*, pages 101–133, 2001.
5. B. Goethals. Survey on frequent pattern mining. HIIT Basic Research Unit Department of Computer Science University of Helsinki, Finnland.
6. M. K. Jiawei Han. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.
7. G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*, pages 229–248, 1991.
8. J. D. U. S. Brin, R. Motwani and S. Tsur. Dynamic itemset counting and implication rules for market basket data. *AIn Proc. of the ACM SIGMOD Intl Conf. on Management of Data (ACM SIGMOD 1997)*, pages 265–276, 1998.
9. B. Walsh. Multiple comparisons: Bonferroni corrections and false discovery rates. Lecture Notes for EEB 581, May 2004.
10. C.-Y. W. Wei-Chou Chen, Shian-Shyong Tseng. A novel manufacturing defect detection method using data mining approach. *Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, 2004.