Detecting Early Indicator Cars in an Automotive Database: A Multi-Strategy Approach

Ruediger Wirth & Thomas P. Reinartz

Daimler-Benz AG
Research & Technology F3S/E
PO Box 2360
89013 Ulm, Germany
e-mail: {wirth,reinartz}@dbag.ulm.DaimlerBenz.COM

Abstract

No company so far achieved the ultimate goal of zero faults in manufacturing. Even high-quality products occasionally show problems that must be handled as warranty cases. In this paper, we report work done during the development of an early warning system for a large quality information database in the automotive industry. We present a multi-strategy approach to flexible prediction of upcoming quality problems. We used existing techniques and combined them in a novel way to solve a concrete application problem. The basic idea is to identify sub populations that, at an early point in time, behave like the whole population at a later time. Such sub populations act as early indicators for future developments. We present our method in the context of a concrete application and present experimental results. At the end of the paper, we outline how this method can be generalised and transferred to other KDD application problems.

Introduction

No company so far achieved the ultimate goal of zero faults in manufacturing. Even very high-quality products, like Mercedes-Benz vehicles, occasionally show problems that must be handled as warranty cases. At Mercedes-Benz, such cases are recorded in a large database, which contains information on all Mercedes vehicles. For each vehicle, information on its technical configuration and its repair history is stored. This information includes faults and their repairs, the mileage at the repair time, costs associated with the repair, and the area where the repair was performed, for example. Currently, this quality information database contains about 7 million vehicles; the net size of the database is about 25 Giga bytes.

Domain experts access this database for various tasks including

- observation of product quality in the field,
- early detection of important faults and their causes,
- prediction of warranty costs,
- initiation of actions for product improvement.

Most database analyses are currently limited to standardised SQL queries and descriptive statistics. Nevertheless, users expect a lot of useful information hidden in these databases that is not accessible by conventional methods. *Knowledge Discovery in Databases* (KDD) (cf. Piatetsky-Shapiro & Frawley 1991; Frawley, Piatetsky-Shapiro, & Matheus 1992; Fayyad *et al.* 1996) has the potential to disclose this hidden information and also to improve the analysis capabilities by facilitating the tasks of the users.

A successful application of KDD techniques in the quality information domain at Mercedes-Benz will have high impact on the business. Warranty costs belong to the best-kept secrets of a company. Typically, a lot of money and prestige are involved. As an indication of the magnitude of this problem area, consider publicized recall actions of car manufacturers. Such actions can easily cost dozens of millions of dollars, not counting the negative impact on the image of the company.

The presented application domain has been tackled following the task model proposed by Reinartz & Wirth (1995) and Wirth & Reinartz (1995). In a thorough application analysis we first elicited the users' requirements, expectations and prior knowledge. Various application goals were identified in this phase. One of these goals is the development of an early warning system for quality problems. In this paper, we present a particular multi-strategy approach to prediction that is very useful for an early warning system. This approach is very flexible, not tied to particular techniques, and can be generalised to other KDD problems such as focusing.

Scientifically, this application domain serves to evaluate and improve KDD techniques. In particular, we develop and refine a methodology of KDD to guide a user through the process. This methodological approach is task oriented and not driven by techniques. It is based on a systematic refinement of tasks that will be finally mapped to techniques. This task refinement provides us with a framework to set up trials with different techniques until the best combination of techniques is found. Thus, the iterative and interactive nature of the KDD process (Fayyad *et al.*, 1996, Brachman & Anand 1994) is supported. Here, we focus on the application point of view.

As argued in Wirth & Reinartz (1995), the main bottleneck for KDD applications is not the lack of techniques. The challenge is to exploit and combine existing algorithms in the most profitable way in the context of real applications. Most successful KDD applications (e.g., SKICAT (Fayyad, Weir, & Djorgovski 1993), OpportunityExplorer (Anand & Kahn 1993)) did not rely on sophisticated new techniques. The reason for their success was the intelligent combination and adaptation of known techniques w.r.t. the constraints of the application problem. Once developed, these methods (i.e., combinations of techniques) can be transferred to other similar application problems. The application reported in this paper follows the same spirit. We rely on existing (although slightly modified) techniques, which we combined in a novel way to solve a concrete application problem. The resulting method can now be transferred to similar KDD application problems.

The paper is organised as follows. First, we describe the application problem in more detail. Then, we outline our solution approach. We first identified the tasks that need to be solved and refined these tasks until they could be mapped to KDD techniques. We describe our concrete realisation and show experimental results. Finally, we critically assess the results, sketch potential benefits of our approach to KDD in general, and point out future work.

Application Problem

Currently, Mercedes-Benz is developing an early warning system on top of the quality information database mentioned in the previous section. The early warning system should discover upcoming quality problems as soon as possible. The earlier quality problems are detected the earlier product improvement actions can be initiated to save future warranty costs. Furthermore, an early warning system can help to avoid expensive recall actions. If an upcoming quality problem with a line of cars is discovered early enough, there may be time to overcome this problem by preventive maintenance actions at regular maintenance intervals, for example. Expensive follow-up faults can be prevented and the customer does not realise that there could have been a problem in the future.

The current system is based on conventional information technology and simple descriptive statistics. It is rigid, restricted in its capabilities, and consumes a lot of resources because it operates on the whole database. Using a KDD approach, we aim at making the system more flexible, less expensive, faster in terms of computing time, and capable to warn even earlier.

The solution approach described in this paper was inspired by the following observation. The experts look at sub populations of the cars if they want to get a first idea about the rate of certain faults. For instance, people in certain countries seem to be more concerned with the paint of a car. If the experts want to check the quality of the paint they look at the cars in this country first. Any problem with the paint is likely to be observed there first. Taxis are another sub population which is watched closely because they drive many miles in a short period.

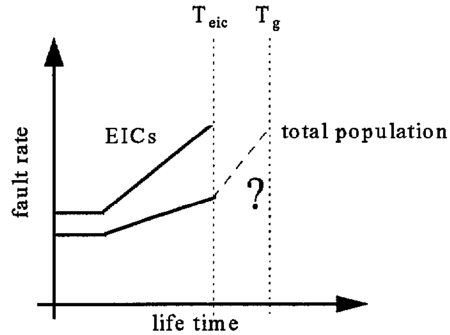


Figure 1:Predicting fault rates using early indicator cars

The question was, could this approach be generalised and automated. Do there exist sub populations of cars which, in certain aspects, behave like the whole population of cars at a later point in time? If yes, how can they be characterized by easily observable attributes? In the following, we call such sub populations early indicator cars (EICs).

Figure 1 illustrates the idea of EICs. Assume we observe a certain aspect of the fault profile, for instance a fault rate, over the life time of cars. The lower line in the diagram shows the accumulated fault rate for a certain fault for cars from the same production period. Assume that we are now at time T_{eic} and we want to predict the fault rate at time T_g. If we had EICs then we could compute the fault rate for the EICs at time T_{eic} and use this value as predicted value for the fault rate of the whole population at time T_g. In this example, the EICs would indicate an increasing fault rate although the total fault rate still looks fine at time T_{eic}.

The discussion above suggests the following high-level procedure. First, select a production period which is used for learning. It needs to be a period in the past such that the values of the relevant attributes at time T_g are already known. In the following, we will call this production period the learning production period. From this learning production period we derive a procedure for the identification of EICs which is independent from a particular production period. This identification procedure will then be applied to subsequent production periods. Each production period then has its own set of EICs which are used for prediction.[1]

1 From another point of view, the identification procedure for EICs could be viewed as a particular prediction model.

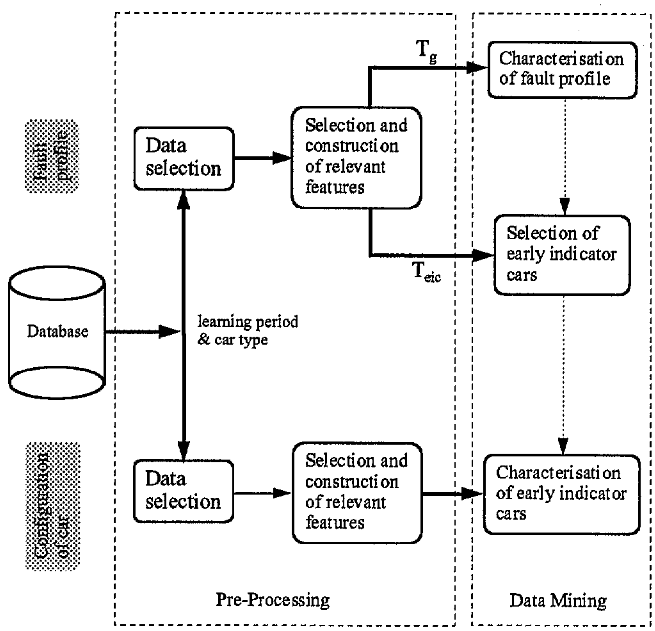


Figure 2: Main steps for detecting EICs

In the context of an early warning system, EICs are an attractive concept because they promise the following benefits:

- Significant changes in the fault profile will be detected earlier. Such changes include upcoming new quality problems, a change in the fault rate or average costs for certain faults.
- Analyses can be performed on smaller sets of vehicles. Thus, resources are saved which can be used for different or more detailed analyses. Some analyses, for instance the application of sophisticated learning algorithms, may not even be possible on the whole population.[2]
- The EICs can be used to predict various aspects of the fault profile, e.g., values, interesting events, and trends. Thus, prediction is much more flexible than with traditional statistical methods like time series analyses.

While EICs alone will not be sufficient for a comprehensive early warning system, they will certainly form a very powerful central part.

Solution Approach

In this section, we describe our approach to detection of EICs. We pursued a task-oriented approach. We first identified the tasks that needed to be solved, refined these tasks and made them more precise until they could be mapped to KDD techniques. Different techniques could have been applied to solve the tasks. In the next sections, we describe one particular choice of techniques and report experimental results.

Basic idea

Our solution approach consists of three major steps.
1. Characterize the fault profile for the whole population of cars at a certain time T_g.
2. Select EICs.
3. Generate an identification procedure for EICs.

In the first step, the fault profile of the whole population of cars at a certain time T_g is characterized. A fault profile is a vague term which could be characterized in various ways. In any case, it will be described in terms of fault relevant attributes.

In the second step, EICs need to be selected. For this purpose, we compute the values for the fault attributes at an earlier time T_{eic}. Those cars which fit the fault profile of the whole population at T_{eic} are considered to be EICs.

The identification procedure for EICs must take into account that EICs need to be identified at production time or shortly after. Therefore, we cannot use attributes that relate directly to the fault profile. EICs need to be characterized by easily observable attributes, like type and configuration of a car or areas where it was sold. This requires the use of two separate sets of attributes. One set is related to the fault profile and the other set contains attributes that can be observed at production time.
Figure 2 shows the procedure.

2 Therefore, the EIC approach is closely linked to the focusing problem in KDD.

Realisation

In this section, we describe techniques and set-up of experiments we conducted to assess the applicability of the approach outlined in the previous sections.

The main choice for the realisation concerned the fault profile. We decided to represent the fault profiles as a hierarchy of classes. Classes contain cars which are similar according to a set of attributes which domain experts judged to be relevant for the fault profile.

The characterization of a fault profile can then be mapped to a (conceptual) clustering task (e.g., Gennari, Langley, & Fisher 1989). For several reasons, we selected the hierarchical conceptual clustering algorithm ECOBWEB (Reich 1994) for this clustering task. ECOBWEB can handle both numeric and symbolic attributes and provides some other features, like the hierarchy correction to avoid ordering effects, which proved to be useful. Since we chose a clustering approach for the first step, the second step must be defined accordingly.

The second step is the selection of EICs. We compute values of the fault attributes for a selected earlier time T_{eic} for all cars included in the first step. The resulting instances are then examined how similar they are to classes from the first step. If an instance is similar to an existing class, then we assume that this instance fits the fault profile at time T_g and could thus serve as an early indicator car. If the instance is not similar to any existing class, then it cannot be an early indicator.

This second step can be reduced to a test whether the instances of time T_{eic} can be assigned to existing classes of the hierarchy. For this test, ECOBWEB was slightly modified. ECOBWEB is an incremental hill climbing system. Each instance is presented incrementally and ECOBWEB tries to incorporate this instance. In the simplest case, the instance can be added to an existing class. The description of the class will then be adapted but the structure of the hierarchy remains unchanged. Other operators change the structure of the hierarchy by splitting classes, merging classes, or adding a new class. All of these operators are tentatively applied, their result is evaluated, and the best operator is permanently executed.

We changed ECOBWEB such that the operators are evaluated but not executed. If one of the structure changing operators wins, then the instance does not fit into the existing hierarchy. In other words, it does not fit into the fault profile of time T_g. If the structure preserving operator wins, then the instance fits and is considered to be an early indicator car.

Using this test, we selected a sub population of cars that are considered as early indicator cars. If we were only interested in the present production period we could simply take this set for future prediction tasks. Since we are really interested in identifying early indicator cars for future production periods, we need to characterize the early indicator cars independent of the production period. This characterization could then be used to identify the set of EICs for each subsequent production period.

Therefore, in the third step we try to discriminate EICs and NonEICs using attributes that can be easily observed at production time or shortly after. We selected attributes which described the configuration of the car (e.g., engine type, transmission type, special equipment), the sales area, and the average mileage, which can be easily calculated from the information collected at the first maintenance visit to the garage.

This task can be mapped to a standard classification task with two classes, EIC and NonEIC.[3] Both classes are described using the configuration attributes and the average mileage per day. The latter attribute is computed from the mileage that is observed at the first maintenance in the garage. The resulting instances were then assigned to two classes according to the result of the second step. Finally, we employed two standard rule learning programs, CN2 (Clark & Niblett 1988) and C4.5 (Quinlan 1993), to learn descriptions for these two classes.

These descriptions are independent of the fault profile and a particular production period. The identification procedure for EICs for a production period now is simply to select all cars from this period which fit the description.

Experimental Results

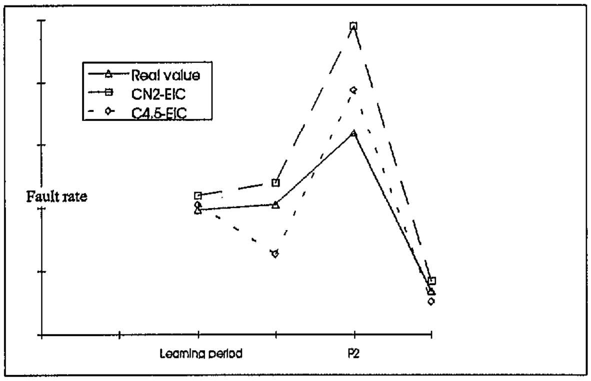
In this section, we present results of our experiments. Numbers in diagrams have been modified because the actual numbers are confidential. The interpretation of the diagrams remains valid.

For our experiments, we chose a learning production period of three months in 1991. The times T_g and T_{eic} were set to 12 and 6 months, respectively. Faults were restricted to faults in the injection system. In cooperation with domain experts we selected 8 attributes which are relevant for the fault profile. Among others, these attributes described average mileage between two faults, average time between two faults, total number of faults in the period, and costs per faults. For each car produced in the learning production period, values for these attributes were computed for life times T_g and T_{eic}. There were 2628 cars considered.

First, ECOBWEB was used to cluster the resulting instances and to construct a class hierarchy based on the attribute values for T_g. Then, attribute values at time T_{eic} were used for selection of EICs as described in the previous section. About one third of the cars (i.e., 862 cars) were classified as EICs.

Then, we selected 28 attributes that can be easily observed early in the life time of the car. We applied both CN2 and C4.5 with different parameter settings until the learning result seemed to be reasonable. It turned out that EICs and NonEICs were not easy to separate. In retrospect, this is not surprising since the vast majority of cars had no quality problems during the first 12 months. For a large portion of these cars their assignment to one of the two classes was almost random.

3 Later, we will see that results became better after we introduced a third class.



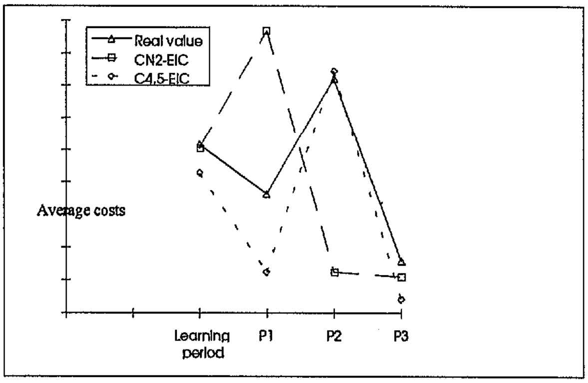


Figure 3: Evaluation results

After this observation, we tried to bias the algorithms towards cars that actually had a quality problem. We introduced a third class containing the EICs which had at least one fault. Now, both CN2 and C4.5 could better discriminate NonEICs from the other two classes. After some iterations with different parameter settings two rule sets were produced.

The resulting rules could be used in two ways to extract EICs. Rules for EICs could be applied directly to identify EICs or rules for NonEICs could be applied to exclude Non-EICs. The latter alternative turned out to be better.

This identification procedure was applied to cars from the learning production period and three subsequent production periods. This resulted in eight sets of EICs. Each set contained about one third of the cars produced during the respective period. The goal of the experiments was to predict different fault rates, average costs per car, and a hot list of faults. The values of the EICs six months after the production date were used to predict the corresponding twelve months values of the whole set.

Figure 3 shows the results for the prediction of the fault rate (total number of faults divided by the number of produced cars) at the left and the average cost per car at the right, respectively. Other tests, for instance predicting fault rates for individual parts produced similar results.

CN2-EICs overestimated fault rates but picked up trends accurately.[4] On costs they performed badly. The predictions of C4.5-EICs were always pretty close to the actual values, both for fault rates and costs. In many cases

In addition to predicting values, we also used EICs to predict hot lists of faults. It turned out that for most cases, ranking and relative frequency of individual faults were predicted accurately. In some cases, however, there were significant deviations from the predicted values. We suspect that these deviations are interesting in their own right. At least some of them may indicate unexpected developments that deserve attention by domain experts.

In summary, the results exceeded our expectations. They show that EICs exist and can be identified automatically. The feasibility of this approach has been proved.

Conclusions

In this paper, we presented a multi-strategy approach to flexible prediction and demonstrated its feasibility. Our experiments showed that the EIC approach can be used to predict future developments. It will now be further refined and incorporated in an early warning system for quality problems.

The experiments are an example for a typical KDD process. After the identification of the prediction learning task we went through several steps of pre-processing, applied various data mining techniques to the pre-processed data, and evaluated and revised the results in a post-processing stage of the KDD process. Methodological considerations helped us to structure this process and to generalise the results.

Applicability of the EIC approach

Based on our experience with transferring the EIC approach[5], we expect that its applicability and its benefits depend on the domain. Probably, EICs neither exist in all application domains nor can they address all aspects of an application problem. Nevertheless, if EICs exist, as in our quality information domain, they can be use for different purposes.

EICs provide a flexible way of prediction. They are not fixed to prediction of values of just one attribute. They can predict different attributes or events, depending on the application domain. They can also predict different trends, which is very important for any decision support system.

Furthermore, EICs are an intelligent, innovative approach to sampling. Analyses can be performed on EICs instead of the whole population. If the method is applied as described in this paper, then the time aspect has

4 In this application domain, trend prediction is sufficient. Specific quantitative values are not necessary.

5 If we generalise the approach described in this paper, it would be more appropriate to talk about early indicator instances instead of early indicator cars. Nevertheless, we use the acronym EIC throughout this paper.

to be considered. EICs are a sample for the whole population at a later point in time. However, the approach can easily be generalised such that the EICs are selected according to different criteria. For instance, in the quality information domain it is also interesting to consider regional effects. Identify a set of regions such that the cars sold in these regions behave like the whole population of cars.

Finally, EICs can also be used for detecting deviations. If EICs are believed to be good predictors then any deviation in the prediction deserves attention of domain experts.

Future work

Although the experimental results with the chosen techniques were very good, there is much room for refinement of the EIC approach. For the experiments reported here some more or less arbitrary choices were made which need to be investigated more systematically. The task refinement process made these choices and parameters explicit and provides a framework for further experiments. Further experiments will address both domain specific and methodological refinements of the EIC approach.

The domain specific refinements concern selection of attributes and data. For instance, the length of the production period was taken to be three months. Other values, perhaps taking seasonalities into account, may lead to better results. Also, times T_g and T_{eic} were arbitrarily set to twelve and six months, respectively. Again, other values may lead to better results. These choices might even depend on the type of fault. Furthermore, attributes describing the fault profile were highly aggregated. In subsequent experiments we will also investigate more fine-grained attributes.

Methodological refinements address the usage of different techniques for the various tasks of the EIC approach and the extension to other application domains. For instance, in the quality information domain we are performing similar experiments with Kohonen networks (Kohonen, 1988) for the clustering sub task. Additionally, we are exploring completely different methods for the representation of a fault profile. At the end, we will have a variety of combinations of techniques which realise the EIC approach. We will then apply the EIC approach to other domains. We expect that the choice of the best realisation of the EIC approach depends on both the domain and the data. Our methodological framework allows us to explore the alternatives systematically and to select the most appropriate one.

Acknowledgments: We thank our domain experts at Mercedes-Benz for their support and access to the quality information database. Lothar K. Becker partly implemented and evaluated our approach.

References

Anand, T., & Kahn, G. (1993). Opportunity explorer: Navigating large databases using knowledge discovery templates. *Proceedings of 1993 AAAI Workshop on Knowledge Discovery in Databases*, pp. 45-51.

Brachman, R.J., & Anand, T. (1994). The process of knowledge discovery in databases: A first sketch. *Proceedings 1994 AAAI Workshop on Knowledge Discovery in Databases*, pp. 1-12.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning 3*, pp. 261-283.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., & Uthurasamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: MIT Press.

Fayyad, U.M., Weir, N., & Djorgovski, S. (1993). Automated cataloging and analysis of sky survey image databases: The SKICAT system. *Proc. CIKM*, pp. 527-536.

Frawley, W.J., Piatetsky-Shapiro, G., & Matheus, C.J. (1992). Knowledge discovery in databases: An overview. *AI Magazine*, Fall, pp. 57-70.

Gennari, J.H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence*, 40, pp. 11-61.

Kohonen, T. (1988). *Self-organisation and associative memory*. Berlin: Springer.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Piatetsky-Shapiro, G., & Frawley, W.J. (eds.) (1991). *Knowledge discovery in databases*. Menlo Park, CA: AAAI Press / MIT Press.

Reich, Y. (1994). Macro and micro perspectives of multi-strategy learning. in: Michalski, R.S., & Tecuci, G. (eds.) *Machine Learning: A Multi-strategy Approach*, Vol. IV, San Francisco, CA: Morgan Kaufmann, pp. 379-401.

Reinartz, T.P., & Wirth, R. (1995). The need for a task model for knowledge discovery in databases. in: Kodratoff, Y., Nakhaiezadeh, G., & Taylor, C. (eds.) Workshop notes *Statistics, Machine Learning, and Knowledge Discovery in Databases*. MLNet Familiarisation Workshop, Heraklion, Crete, pp. 19-24.

Wirth, R., & Reinartz, T.P. (1995). *Towards a task model for Knowledge Discovery in Databases*. Internal Report, Daimler-Benz Research & Technology, Ulm, Germany.