Some Recent KDD-Applications at DaimlerChrysler AG

E. Hotz
DaimlerChrysler AG, Global Service and Parts
edgar.hotz@daimlerchrysler.com

U. Grimmer G. Nakhaeizadeh
DaimlerChrysler AG, Research & Technology, Ulm
{udo.grimmer,rheza.nakhaeizadeh}@daimlerchrysler.com

Abstract: The mission of the Information Mining department, DaimlerChrysler Research and Technology, is exploring, exploiting, and enriching data mining and text mining methods to provide complex decision and product support systems. A key requirement for being able to provide a complex system lies in the different core technologies used, such as symbolic machine learning, statistical learning procedures, association learning, neural networks, distributed data mining, text mining, and model selection procedures. On the basis of these core technologies, we extract and analyze information from data collected in vehicles, from business data, financial data, and documents.

Awareness of the above-mentioned technologies together with know-how about other topics like optimization and case-based reasoning build only one part of the expertise of our research department. The complementary part consists of knowledge about different application domains such as computational marketing, computational finance, car market modeling, data cleaning, and knowledge from various technical domains. In this paper some topics of the second part are presented.

1 Prediction of warranty and goodwill costs

1.1 Problem Description

The warranty planning in the automobile industry is characterized by several problems:

- the need to calculate planning figures for cash reserves within the balance caused by legal or voluntary obligations in the context of warranty and goodwill
- a growing number of actually manufactured vehicles and a growing number of production plants all over the world, resulting into a more complex and multilayered production structure
- a different warranty and goodwill policy for different sales markets.

The system WAPS, that is the result of a joint project between DaimlerChrysler Research & Technology and the Sales and Services Department, should master the business needs

described above (see [HNPS99] for more details).

Up to now the warranty cost prediction has been realized by using a conventional planning method based on the amount of warranty and goodwill costs observed in the last budget year. This amount is modified by information available about the expected inflation, about the quality index of the vehicles and about the development of the sales figures for the different vehicle series. Some of this information is rather qualitative than quantitative requiring the additional expert knowledge to estimate the effects on the amount of the budget. This is a complicated procedure that takes about four weeks for each planning period and requires involving different experts from several departments. Regarding the fact that an enormous data stock stored about the vehicles and their behaviour in the context of warranty and goodwill costs is available, the question coming up was whether it is possible to use Data Mining technology and construct an automatic prediction tool based only on historical data which are available for the warranty and goodwill costs. This has led to the system WAPS that will be described in the next sections.

1.2 Structure of the available data

We discuss in this section the structure and source of the data relevant for the analysis and prediction of warranty and goodwill costs. The available historical data about the warranty costs is a part of the database QUIS (QUaltity and Information System) that can be considered as a kind of data warehouse containing information on produced vehicles and their damages. The fundament of QUIS is a normalized database model including more than fifty tables with more than 400 fields. The most extensive table contains more than 110 million rows. Figure 1 describes the process environment of data generation and data collection for QUIS. As figure 1 shows, data collection begins in the repair workshops.

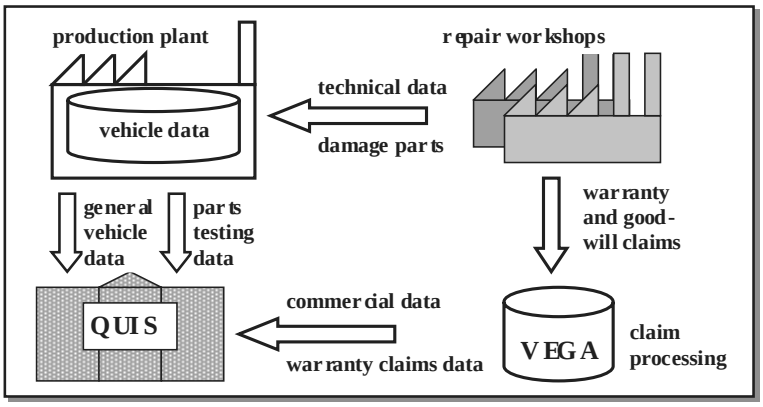

Figure 1: Process environment of data generation and data collection for QUIS

On one hand, if the repair is covered by warranty or goodwill agreement, the workshop

claims the refunding of the warranty costs. The data on such claims are stored and processed in a database called VEGA. If the claim is accepted the warranty costs data and the damage parts data are transferred to QUIS. On the other hand, some of the damage parts are sent to the production plants to analyze the damage causes. The analysis results will be transferred together with general vehicle data (vehicle ID number, date of production, motorizing version etc.) to QUIS as well.

1.3 Prediction models

As mentioned in section 1.1 the Data Mining task in WAPS is a prediction task. Besides the statistical approaches like regression analysis, regression and model trees, SMART, Naive Bayes etc. one can use also alternative approaches based on neural networks. After some initial trials we have selected the regression analysis as a representative for the statistical approaches and a Backpropagation network as a representative for neural approaches.

The next step was to design an evaluation plan to compare the both alternatives and to select the best one. The evaluation in the testing time has shown that there is no significant difference between the prediction performance of two alternative approaches if we use STDV and Theil's U as comparison criteria. Using of MAPE, however, has led to the fact that almost for all vehicle series the regression analysis is the favorite model. Concerning these results and the fact that regression analysis is easier to implement in comparison to a neural network approach we have decided to select the regression model for further consideration. This model was however refined in further steps. Especially, we have tried to capture the dynamic aspects and structural changes of the input data by refining the regression model. This can be considered as an innovative approach in the implementation of WAPS.

WAPS can test now if a structural change has occurred in the past. If it is the case, instead to fit a unique regression model over the whole time, the system switches to a piecewise regression approach taking into account the structural change by dividing the whole time space into several subspaces.

1.4 Deployment of WAPS

During its development, WAPS has been used besides the conventional planning method in order to test its practicability in prediction of the total annual budget for warranty and goodwill in the passenger car section. The results were very promising. The needed planning time is 75% shorter than the time using the conventional method. During its development, we could convince the users that WAPS is not only able to support the planning process for warranty and goodwill costs and to provide additional security for the calculated planning figures, but it is able also to support the calculation of the amount of balance cash reserves for warranty and goodwill obligations in considerable shorter time. In this connection, the final version of WAPS which is implemented in Visual Basic for

PC, is in use in the Sales and Services Department of DaimlerChrysler in Stuttgart.

2 REVI-MINER, a KDD-Environment for Deviation Detection and Analysis of Warranty and Goodwill Cost Statements

2.1 General Remarks

In order to get the refund of vehicle repair costs, workshops of DaimlerChrysler AG worldwide regularly submit the warranty and goodwill cost statements to the central warranty department in Germany. These statements should be examined for validity and correctness, which is a very complex task for the warranty cost controllers. REVI-MINER is a KDD-environment which supports the detection and analysis of deviations in warranty and goodwill cost statements. The system was developed within a cooperation between DaimlerChrysler Research & Technology and the direction Global Service and Parts. We have implemented different approaches based on Machine Learning and Statistics that can be used for data cleaning in the preprocessing phase. The applied Data Mining models were developed by using a statistical deviation detection approach. The tool supports the controller within his task to audit the authorized workshops (see [HHG+01] for more details).

2.2 Data Cleaning Approaches

To check the quality of data the following approaches are developed that can be used for data cleaning in the preprocessing phase:

- Descriptive statistic approach: Stored (historic) data has been described by descriptive statistics. The descriptions have been compared to values known from the documentation, or other sources.
- Development of a statistical prototype based on normal distribution assumption. The basic idea behind this approach is to build models for time series from historic data under the assumption of statistical normal distribution. Once models have been build, the corresponding values for new data will be derived, and tests statistics are being applied. If statistically significant deviations are discovered, suspicious data sets are being tagged and warning messages are generated
- Application of GritBot (see [Qui01]). The third approach applied the commercial tool GritBot to each single QUIS database table, as well as several joined tables. Although GritBot is quite scalable, it also can process flat files only. For any analysis, GritBot needs two files: one file containing all the data, and another names file containing a short description of the names and formats of the data file. Both files need to be prepared manually.

2.3 Description of Applied statistical-based Deviation Analysis Criteria

Discussions with the end users showed that the needed criteria to identify and analyze deviations in warranty and goodwill data should cover the main cost types (damage types)

- total cost (total number of repairs)
- labor cost (number of working hours)
- cost for repair material (number of repairs with deployment of repair material)
- cost for exchange of vehicle aggregates, e.g. gear unit, air conditioner unit, motor unit (number of repairs with deployment of aggregates)
- incidental cost.

The nature of the created criteria allows different analysis approaches for the detection of suspicious workshops. We will now describe one of the seven applied approaches.

2.4 Deviation Analysis for a single Workshop by Comparison to predefined Workshop Clusters

Workshop clusters can be built by assignment of workshops according to some chosen criteria, e.g.

- repair turnover during a specified period of time
- affiliation of the workshop to given workshop subgroups, as there are branch offices, trade partners, representatives, general representatives.

Comparison of workshop average total (labor, material, aggregate, incidental) costs per damage code with the respective values of the workshop cluster delivers first hints if there could be irregularities at the transaction of warranty and goodwill cases. We can measure absolute and relative deviations of workshop averages from averages of the respective workshop cluster and weigh the absolute deviation with the number of underlying repair cases.

$\bar{x}_{i,j}$: average costs of workshop i for damage code j

$\bar{x}_{k,l,j}$: average costs of workshop cluster with turnover group k and subgroup l for damage code j

$n_{i,j}$:] number of repair cases with damage code j for workshop i

$\frac{\bar{x}_{i,j}}{\bar{x}_{k,l,j}} - 1$: relative deviation of average costs for damage code j

$(\bar{x}_{i,j} - \bar{x}_{k,l,j}) \cdot n_{i,j}$: weighted absolute deviation of average costs for damage code j

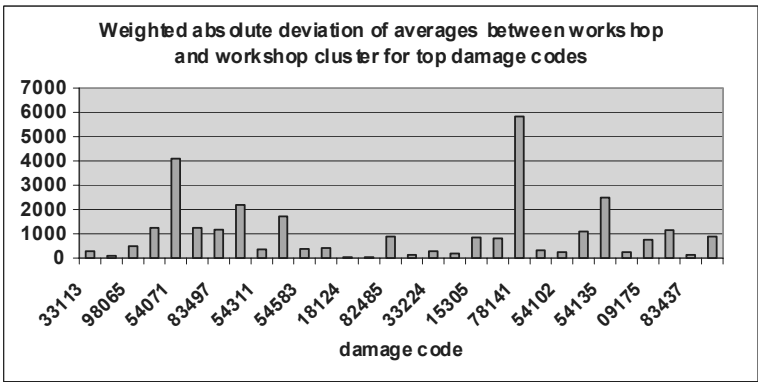

Figure 2: Weighted absolute deviation of averages between workshop and workshop cluster

2.5 Deployment of REVI-MINER

The Data Mining tool REVI-MINER supports controlling efforts to detect and avoid fraudulent activities within the workshop organization of DaimlerChrysler AG. Its functionality covers the essential phases of a Data Mining process and provides a user interface with easily manageable menus based upon VISUAL BASIC forms. REVI-MINER provides the methods for a fast, efficient and meaningful analysis of the warranty and goodwill data for workshops thus giving the experts of the revision department crucial hints upon possibly fraudulent activities.

Literaturverzeichnis

[HHG+01] E. Hotz, W Heuser, U Grimmer, G Nakhaeizadeh, and M. Wieczorek. REVI-MINER, a KDD-Environment for Deviation Detection and Analysis of Warranty and Goodwill Cost Statements in Automotive Industry. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 432–437, 2001.

[HNPS99] E. Hotz, G. Nakhaeizadeh, B. Petzsche, and H. Spiegelberger. WAPS, a Data Mining Support Environment for the Planning of Warranty and Goodwill Costs in the Automobile Industry. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 417–419, 1999.

[Qui01] R. Quinlan. GritBot - An informal tutorial, 2001.