DATA MINING ON THE INSTALLED BASE INFORMATION
Possibilities and Implementations

Rashid Bakirov
Center for Sensor Systems (ZESS), University of Siegen, Siegen, Germany

Christian Stich
ABB Corporate Research Germany, Ladenburg, Germany

Keywords: Data mining, Installed base, Proposals to customers, Failure prediction.

Abstract: Managing the installed base at customer sites is a key for customer satisfaction. Hereby installed base comprises installed systems and products at customer sites which are currently being serviced by the producer company. The purpose of the present study is developing use cases for data mining on the installed base information of a large manufacturing company and specifically ABB, and constructing data mining models for their implementation. The aim is to use the available information to enhance customer-tailored sales and proactive service. This includes recommendations to customers and failure prediction. The developed models employ association rules mining, classification and regression, realized with the help of data mining tools Oracle Data Mining and Weka. Results have been evaluated using statistical means, as well as discussed with the experts at the company. These results suggest that with the reasonable amount of data, installed base information is a potential source for data mining models useful for business intelligence.

1 INTRODUCTION

At the present time, with the improvement in data storage software and hardware systems, the amount of available data about each individual or organization is rapidly increasing. Today's intense global competition, cost pressure, unstable markets and empowered customers are the reasons why industrial enterprises have to fully utilize this knowledge. An important aspect of knowledge discovery for customer-specific offering and services is using historical data to find hidden relations using advanced data analysis techniques known as "data mining".

To enable customer tailored services, it is important for companies to keep track of their installed base, in other words, the products they produce and need to maintain. An installed base system could be used as a basis for data mining, identifying possible relations between products, discovering patterns of service jobs, predicting interest of customers to products and in many other various applications. Results will lead to improved customer sensitive offerings and service which would ultimately increase the sales of products, lower costs for the customers and predict service needs for the providing company.

In this work theoretical applications of data mining on the installed base are analyzed, and implementations are shown. They are implemented on an internal data ware house of ABB, which is called ServIS, the **Serv**ice **I**nformation **S**ystem. ABB is a multinational corporation headquartered in Zurich, Switzerland. ABB's core businesses are power and automation technologies and it holds market-leading positions in most key product areas. ABB employs more than 116,000 people and operates in approximately 100 countries (ABB Group, 2010). ServIS is a comprehensive information system, that encompasses all of the products and systems, customers' services and a range of additional data from all five divisions of ABB. It was developed by ABB Corporate Research Center Germany and provides a tool for outlining and maintaining ABB's installed base information. Goal of ServIS is to operatively provide ABB's field service and technicians with information about customers' sites and equipment installed there.

The purpose of this paper is to present use cases for data mining which target real business problems, and show their realization using available technolo-

Bakirov R. and Stich C..
DATA MINING ON THE INSTALLED BASE INFORMATION - Possibilities and Implementations.
DOI: 10.5220/0003186506490654
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 649-654
ISBN: 978-989-8425-40-9
Copyright © 2011 SCITEPRESS (Science and Technology Publications, Lda.)

gies. The first part of the work gives information about similar research and demonstrates possible usages of data mining on a large production company's installed base data. The second part describes the application the aforementioned use cases to the real data warehouse of ABB, evaluates the results and identifies possible shortcomings of these approaches. The final part concludes the paper, giving a summary of the conducted work and outlines perspectives for the future works this area. Paper provides references for used algorithms.

2 DATA MINING ON THE INSTALLED BASE INFORMATION

2.1 Review of Similar Works

Data mining on installed base presents many different interesting opportunities for companies but not much research work has been done in this field. Most of related work deals only with subsets of installed base.

DaimlerChrysler, now Daimler, has successfully employed data mining on their installed base. In the paper "Forecasting the Fault Rate Behavior for Cars" (Lindner and Studer, 1999), authors forecast the number of complaints for separate details of the cars that will be produced in the future using neural networks and decision trees. The implementation of the system proposed by the authors was successfully tested and has been transferred to the product controlling department.

In whitepaper "Data Mining in Equipment Maintenance and Repair: Augmenting PM with AM" (Exclusive Ore Inc., 2003), the possibility of anticipatory maintenance based on previous services data is discussed. A database of a large locomotive manufacturer was examined as a case study. Authors concluded that, data mining of equipment maintenance and repair data can help discover anticipatory maintenance (AM) procedures, improper repairs or other maintenance, ways to improve repairs, undocumented repair methods, as well as warn about likely failures in advance. This all would result in less future failures, and lower downtime by failures.(Exclusive Ore Inc., 2003).

Data mining was also used by Xerox to analyze the service need of their installed base including service of customer replaceable units (CRU). Aim of the project was "measurable cost reduction in services delivery, including field service and consumable supplies replenishment" (Minhas, 2003). Xerox used Oracle database for storing the information and Oracle Data Mining (ODM) to for clustering of customer usage based on various attributes. (Minhas, 2003)

2.2 Use Cases of Data Mining in Installed Base

The usage of data mining in the specific installed base database will naturally depend on the data, data model, structure and quality of data. This section describes its use cases. It is assumed that the database has full information about products, customers and services, as well as any necessary additional data. Databases like that need to be constructed from various datasets, tables and databases acquired from different divisions of company.

One use of data mining is cross-selling and making recommendations to customers. Cross-selling involves selling additional products to the same customer, or bundling products in packets and selling them together. Using data mining methods would help to identify which products would interest specific customers. It is possible not only to create bundles of products, but also to recommend interesting products to the customers. Two approaches to make a recommendation are possible. One is to show what did other customers, who bought a particular product, also purchase, as seen on Amazon Internet store (Linden et al., 2003). This requires relatively few information - only transaction ID, and ID's of products which appear on this transaction. Second approach is to recommend the products which are purchased by similar customers. The similarity criteria could be combined from different inputs, for example company location, size, industry type etc. These two described approaches can be used together. They can be applied to to identify suitable target audiences of marketing campaigns, directed both to attract new customers and make offers to the existing ones. Here, it is attempted to discover customers, who will be interested in offer, rather than offers, which might interest the customer. The customers could then be assigned with response scores for the particular campaign. An example of this use case can be the following. Suppose that two products, A and B are frequently bought together. Then, a company can offer a product B to the customer that purchases A. In the second approach, company can search for the customers, who have only A, to use them as marketing targets for B.

A potential area of data mining's application is failure prediction. It should be possible to use the repair history of the product, or of the other products on the same site to make predictions about future failures. This method is less expensive than real time condition monitoring using sensors and probes. Discovery of common repair sequences would also assist in improving the repair process itself. (Exclusive Ore Inc., 2003). Suppose that product B regularly fails after failure of product A. Then the servicing company can check the status of B each time after A fails. Also, if it is concluded that the reason of B's failure is the failure of A, while repairing B, the parts of it which has connection with A could have higher probability of malfunction and can be checked first.

Another usage of data mining, is prediction of customer leaving. There are two different models regarding this topic. One is to predict if the customer will leave in certain amount of time and another is to predict, for how long will customer stay with the company. Data mining can also be applied to estimate customers' prospective value, a revenue that customer will bring during his remaining lifetime. This can be used in identifying "good" and "bad" customers (Berry and Linoff, 2004). Both models of attrition prediction and other applications of data mining can assist in corporate planning. Predicting amount of failures for automobiles has successfully been applied at DaimlerChrysler (see section 2.1). Same methods could be applied on general installed base data, to predict failure rates of various products for financial, logistical and other planning.

Data quality control is another important task, particularly when managing a large installed base data warehouse with many various information sources. With the help of data mining, it is possible to improve automation of this task. Two possible approaches in this case are discovery of numerical outliers and identification of cases that do not fall into general patterns identified by other data mining techniques. Identification of cases that do not fall into general patterns can also lead to discovery of holes in data. Discovering and treating outliers is an important preprocessing task before applying other data mining algorithms.

3 DATA MINING ON ABB'S INSTALLED BASE

ABB's ServIS is based on a data warehouse which stores data in Oracle relational database. This database consists of more than 200 tables with installed base and administrative data. Information on the ABB products, materials, customers, services etc. which is included in the system, can be used as a basis for data mining. Amount of available data was quite large, encompassing about 30,000 products or 700,000 equipment units. Nevertheless, its quality was not always on desired level for the data mining purposes. These are missing data such as NULLs, "other" or "unknown" entries, as well as scarcity of historical data and lack of information that could be useful for data mining purposes.

As the ServIS is based on Oracle 10g database, the first choice of software for data mining was Oracle Data Mining (ODM) (Oracle Corporation, 2010), which is an built-in option of the database, with Oracle Data Miner GUI. Additionally free Weka (Witten and Frank, 1999) software was chosen as an alternative, to estimate results from ODM. Following sections present information on tested data mining models on ABB's installed base information.

3.1 Association Rules on ABB Products

This model aims to discover ABB products that tend to be on the same site - plant, factory, etc. An implementation of association rules discovery is done using Apriori algorithm (Agrawal and Srikant, 1994). Similar method is also used by Amazon to find suggestions for the customers (Linden et al., 2003) This algorithm is available in most data mining packages including ODM.

After constructing a data table, algorithm was executed with different generalization levels of products. To find optimal level, as suggested by Berry and Linoff (Berry and Linoff, 2004), we used a dynamic approach, considering the item of the most specific generalization level and comparing amount of item's occurrences to heuristically identified threshold value. If this amount is less than threshold, we considered the next less specific generalization level, continuing the process until the generalization level which satisfies the threshold condition is found. Experimental threshold values of 50, 100, 200, 400 were used, based on the number of occurrences in the middle generalization level, with the maximum of 506 and average of 23. 10% confidence and 1% support thresholds were used for models.

We got at most 14 rules as a result of a single model. This was dynamic generalization model with 200 occurrences threshold. This performance can be explained by a specific nature of ABB. Even while the number of products is quite large, 22,000 compared with an average of 50,000 products in large supermarket (Nestle, 2002), ABB has lesser diversity of products and fewer amount of customers when compared to a supermarket. Customers of ABB that come from the same industries tend to purchase similar mix of products. Combining results from all models, we get 9 rules with confidence levels higher than 75% with various support levels ranging from 1% to 3%.

This method is easily reproducible and reusable.

It can be used in combination with other data mining or statistical methods, such as prediction of customer interest described in the next section, for better final outcome. Association rules on products can be applied to provide aid in cross-selling and make recommendations to customers. It is possible to realize automatic recommendation system on company's Internet portal, where customers reviewing a particular product are notified of products, that were purchased together with this one.

3.2 Prediction of Customer Interest

This model tries to determine customers with possible interest in a given product. The idea is to use all the data about customers that might be relevant in decision making and a binary attribute, showing whether customer bought particular product. The approach is first to build a classification model, predicting that binary variable on sample data, then apply the result model to all available data and make predictions about each single case. The used data included information about technical site (location, industry, number of products and materials on the site), information about the company which owns the site (location, number of sites the company owns, relation to ABB). Additionally, binary variables that show whether the popular products in the database exist on the given site were included. This greatly improved the accuracy of models.

Models were built to predict the existence of some of ABB's more widespread products on the site. The chosen decision trees classification technique produces a whitebox model - in other words, the possibility to see resulting rules and determine on which attributes does the final result depend the most. ODM implements a variation of CART decision tree algorithm (Breiman et al., 1984). As there were not any missing values in input table and decision tree is resistant to outliers and does not require normalization, no other pre-processing methods except from gathering data from different tables were used. To build models, we used stratified sample, which keeps original distribution of data, of around 10,000 cases for each model. 60% of sample were used in tree building and 40% in its testing. An example of sample build data for prediction for product A with distribution of class (goal) attribute can be seen in figure 1.

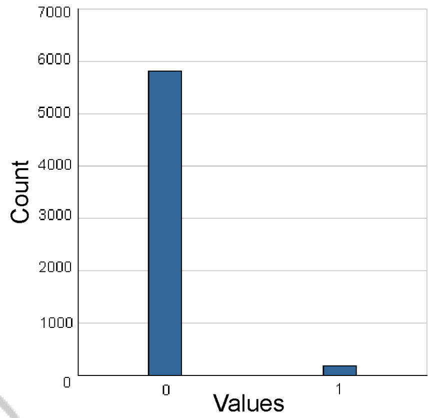

Figure 1: Sample build data for product A prediction.

Models were built and tested for several most frequent products in ABB's installed base to make statement more significant. We will review results on an example of A's prediction. After the decision tree is built, it is applied to test data. With the default probability threshold value of 0.5, model correctly predicts 99.19% of negative and 36.73% of positive cases. In some cases a user might want to bias the prediction in the direction of making more correctly identified positive cases at the cost of less correctly identified negatives, or in the reverse direction. This can be implemented by altering the threshold value. Optimal value for thresholds - the one which maximizes true positive rate and minimizes the false positive rate, may be found using ODM ROC-chart (Receiver Operator Characteristic) (figure 2). In our case it's 0.11. If we use this threshold, model correctly predicts 91.71% of negative and 92.86% of positive cases.

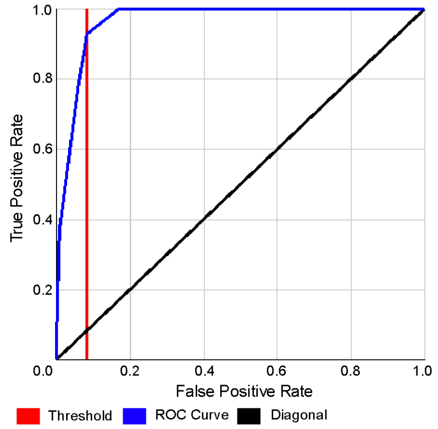

Figure 2: Prediction ROC of A. Area under curve is 0,96.

When we test the model with all of the available data, which also includes training data, with the same threshold, we get similar results - 91.88% (35,703 of 38,859) true negatives and 90% (909 of 1009) true positives. Ideally, training data has to be excluded from test set. In our case, since training data is only roughly 16% of all the available data and the results are similar with results of the model bulidng test, this

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

slight error in evaluation is neglected.

Achieved results show that the models perform well when predicting product type that appear fairly often in installed base, as well as on those which appear less often. In our case, and the most of related cases true positive predictions is more important than true negatives. We achieve more true positives by altering prediction threshold. Of course, more frequent does product appear in dataset; more significant are model's results. With products that rarely appear, a very heavy cost matrix biasing has to be used for a model to make positive predictions. This greatly decreases true negatives accuracy rate. Industry branch of the site was identified as the most significant attribute, being the first splitting criterion in 3 out of 4 models for different products. Other significant attributes were location of the site and amount of product types on the site.

A drawback of this approach is that a separate model has to be constructed for every product, on which predictions are made. This approach can be combined with the association rules mining to achieve even more accurate predictions. Results of customer interest prediction models can be used for recommendations to customers for marketing purposes, as well as data quality control.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

3.3 Prediction of Future Repair Jobs

This model attempts to determine amount of repair jobs for various materials in installed base. It also aims to identify attributes that have the largest effect on amount of repairs. Different variations of this model were tested; regression - predicting exact amount of repair jobs and classification, splitting goal value into none/low/high categories (ternary classification) or making binary "yes"/"no" categories (binary classification) predictions. This approach is similar to the method of Lindner and Studer (section 2.1). Both approaches use amount of repairs that happened before to forecast repairs that will happen in the future. The difference is the lack of available historical data for us to use. Instead, we use additional information about site and material which can be found in ServIS. General information about site of location (see previous section) as well as information about material (type, age) and repairs history from previous years were used as predictors for the model. Because of data quality problems, dataset used for models in this section consisted of only 778 cases. This results in lower significance of achieved results.

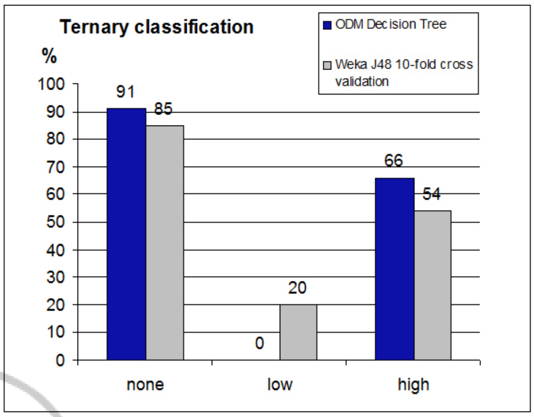

Figure 3: Ternary classification results.

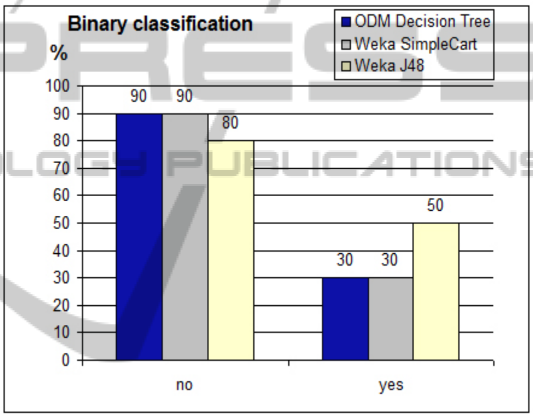

Figure 4: Binary classification results.

Considering that they are based on a very limited dataset, models in this section have fairly high accuracy rates. For instance, ODM regression model that is based on Support Vector Machines (Vapnik, 1995) has 27.48% predictive confidence. Predictive confidence shows how much better given model is than a "naive" model, a model that predicts average value for the dataset. Predictive confidence is the percentage increase in accuracy of prediction gained by the tested model over a naive model. Binary and ternary classification results can be seen on the figures 3 and 4. We can see that in both cases, Weka J48 (implementation of C4.5 decision tree algorithm) provides better results. The most important predictor for decision tree models was average of repair jobs on material in previous years. Other significant attributes for both classification algorithms were type of material, amount of materials, industry branch as important.

Models described in this section can help in financial and logistical planning. For example, the budget of company can be adjusted to include repair costs and the spare parts could be prepared beforehand. This knowledge can also be a factor for choosing profitable service contract terms. Another usage of the models is optimization of proactive service (section

2.2). Choice of a particular model will depend on a use case, for example when assessing the behavior of a product, ternary classification could be enough, but planning service budget or preparing spare parts for product is better done with numerical values. Other factors in choosing model are its accuracy and importance of obtaining rules, as for example ODM regression does not provide any. Another interesting data mining approach to service jobs history is discovering common sequences of failures on the site and using these results to make prognoses about future repairs. This was not implemented in our work due to scarce data.

4 CONCLUSIONS

In this paper we have analyzed the possibilities of application of data mining techniques on the installed base in form of use cases. Data mining helps to discover relationships between products, customers, service jobs etc. This information can be helpful in many areas, among them cross-selling, marketing, proactive service, contracts management, data quality control, corporate planning. We implemented the use cases using both mining tools. After tests and assessment we have reached a conclusion that results of preliminary tests were good and that it was worth continuing research in this direction. Prediction of customer interest in products have provided the best results. These results have proven that data mining methods could be successfully implemented on installed base data.

With the increased amount of data and improvement of its quality, more data mining models described in section can be applied to ABB's installed base system. Then the obtained results are expected to be more accurate and significant. Some refinement and restructuring of data is necessary in order to apply certain models. After development and tests of various models, the ones that provide useful results for the outlined use cases can be implemented in real business applications. Model training and application processes can be extracted from ODM in form of PL/SQL packages, that may be used in respective software or web-interfaces to show data mining results. For instance, results of association rules mining can be used to make automated suggestions to customers in company's web portal. PL/SQL packages can be scheduled to run automatically to keep models and results up to date with new data.

Successful business applications based on data mining on the installed base information would create a win-win situation both for a business and a customer, offering appropriate products and systems, identifying cross-selling opportunities and predicting the maintenance needs. This would result in easier product selection and decrease of downtime costs for the customer and increased profits for the business.

REFERENCES

ABB Group (2010). ABB Group Annual Report 2009.

Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499. Morgan Kaufmann Publishers Inc.

Berry, M. J. A. and Linoff, G. S. (2004). *Data mining techniques: For marketing, sales, and customer relationship management*. Wiley, Indianapolis, Ind., 2. edition.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth International Group, Belmont Calif.

Exclusive Ore Inc. (2003). Data Mining in Equipment Maintenance and Repair: Augmenting PM with AM.

Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations item-to-item collaborative filtering. In *Internet Computing, IEEE*. IEEE Computer Society.

Lindner, G. and Studer, R. (1999). Forecasting the Fault Rate Behavior for Cars. In *Proceedings of Workshop "From Machine Learning to Knowledge Discovery in Databases" at the ICML 1999, Bled, Slowenia, June, 26-31, 1999*.

Minhas, R. (2003). Towards Intelligent Machines: A Xerox Initiative for Customer Assisted Self Help.

Nestle, M. (2002). The soft sell: how the food industry shapes our diets.

Oracle Corporation (2010).

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer, New York.

Witten, I. H. and Frank, E. (1999). *Data mining: Practical machine learning tools and techniques with Java implementations*. The Morgan Kaufmann series in data management systems. Morgan Kaufmann, San Francisco, Calif.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS