

Data Mining and Machine Learning Techniques (Mahout on Hadoop) For Automotive Industries: A Survey

Jaiganesh Madasamy

Research Scholar

Department of Computer Science,
School of Computing Sciences

Vels Institute of Science, Technology &
Advanced Studies, Chennai, India
jaiganesh.madasamy@gmail.com

Dr. R. Parameswari,

Associate Professor

Department of Computer Science,
School of Computing Sciences

Vels Institute of Science, Technology &
Advanced Studies, Chennai, India
dr.r.parameswari16@gmail.com

ABSTRACT

Data Mining and machine learning are the key technologies when it comes to the process and products with automatic learning and optimization to be used in the automotive industry of the future. In this paper, I have focused to compare a variety of techniques and approaches and its impact on the automotive industries and major sub processes in the automotive supply chain (Development, Procurement, Production, Logistics, Marketing, Sales and Service). There are huge data available within the Automotive Industries information system. However, there a lack of effective tools to discover hidden relationships and trends in data mining, advanced data mining and machine learning techniques via Mahout Application in Hadoop can help the automotive industry more efficient and enhance its customer focus throughout all its operations and activities.

Keywords: Data Mining, Machine Learning, Automotive Industries.

I. INTRODUCTION

1) The Data Mining Process

Data Mining (DM) is a most popular knowledge acquisition method for knowledge discovery. The traditional Cross-Industry Standard Process [1] for Data Mining (CRISP-DM) [2] only included 6 process steps. Instead, based on the business understanding, data understanding, data preparation, modelling, and evaluation sub-steps, CRISP proceeds directly to the deployment of results in business processes.

As per new proposed business process, an additional optimization step included that in turn comprises multi-criteria optimization and decision-making support. This approach is represented in Fig. 1.

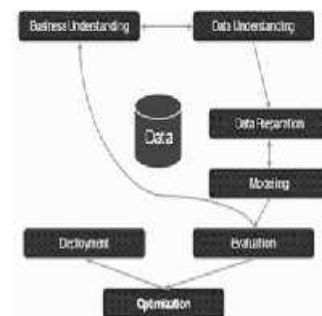


Fig.1: Traditional CRISP-DM process with an additional optimization step

2) Machine Learning Process

Machine learning (ML) is a type of artificial intelligence (AI) [3] that allows software applications to become more accurate in predicting outcomes without being explicitly programmed.

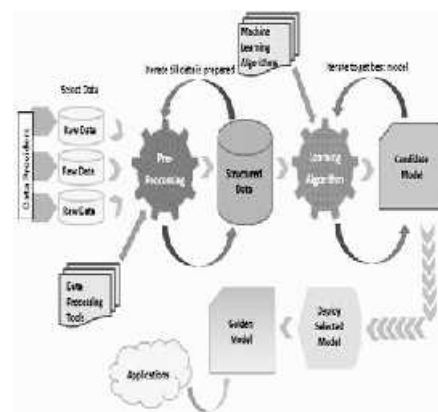


Fig. 2: Machine Learning Process

Machine learning is a branch of science that deals with programming the systems in such a way that they automatically learn and improve with experience. Here, learning means recognizing and understanding the input data and making wise decisions based on the supplied

data. The processes involved in machine learning are similar to that of data mining and predictive modelling. ML process is represented in Fig. 2

It is very difficult to cater to all the decisions based on all possible inputs. To tackle this problem, algorithms are developed. These algorithms build knowledge from specific data and past experience with the principles of statistics, probability theory, logic, combinatorial optimization, search, reinforcement learning, and control theory. [4]

3) DM and ML in the Automotive Industry

At a high level of abstraction, supply chain in the automotive industry can generally be represented with the following sub processes.

- Development
- Procurement
- Production
- Logistics
- Marketing
- Sales and Service

(i) Development

Vehicle development has become a largely virtual process that is now the accepted state of the art for all manufacturers. CAD models and simulation are used extensively in all stages of the development process. In terms of the analysis and further use of simulation results, data mining is already being used frequently to generate co-called "response surfaces". In this application, data mining methods (the entire spectrum, ranging from linear models to Gaussian processes, support vector machine and random forests) are used in order to learn a nonlinear regression model as an approximation of representation of the input vector for the simulation based on the relevant (numerical) simulation results.[5]

While simulation and the use of nonlinear regression models limited to individual applications have become the standard, the opportunities offered by optimizing analytics are rarely being exploited. Particularly with regard to such important issues as multidisciplinary (as well as cross-departmental) machine learning, learning based on historical data (in other words, learning from current development projects for future projects), and cross-model learning, there is an enormous and completely untapped potential for increasing efficiency.

(ii) Procurement

The procurement process uses a wide variety of data concerning suppliers, purchase prices, discounts, delivery reliability, hourly rates, raw material specifications, and other variables. Data mining methods allow the available data to be used, for example, to generate forecasts, to identify important supplier characteristics with the greatest impact on performance criteria, or to predict delivery reliability. In terms of optimizing analytics, the specific parameters that an automotive manufacturer can influence in order to achieve optimum conditions are also important. Overall, the finance business area is a very good field of

optimizing analytics, because the available data contains information about the company's main success factors.

(iii) Production

Every sub-step of the production process will benefit from the consistent use of data mining. It is therefore essential for all manufacturing process parameters to be continuously recorded and stored. Since the main goal of optimization is usually to improve quality or reduce the incidence of defects, data concerning the defects that occur and the type of defect are required, and it must be possible to clearly assign this data to the process parameters. This approach can be used to achieve significant improvements, particularly in new types of production process.

(iv) Logistics

In the field of logistics, a distinction can be made between procurement, logistics, production logistics, distribution logistics, and spare parts logistics.

Procurement logistics consider the process chain extending from the purchasing of goods through to shipment of the material to the receiving warehouse. When it comes to the purchasing of goods, a large amount of historical price information is available for data mining purposes, which can be used to generate price forecasts and, in combination with delivery reliability data, to analyse supplier performance. As for shipment, optimizing analytics can be used to identify and optimize the key cost factors.

A similar situation applies to production logistics, which deals with planning, controlling, and monitoring internal transportation, handling, and storage processes. Depending on the granularity of the available data, it is possible to identify bottlenecks, optimize stock levels, and minimize the time required.

Distribution logistics deal with all aspects involved in transporting products to customers. Since the primary considerations here are the relevant costs and delivery reliability, all the subcomponents of the multimodal supply chain need to be taken into account from rail to ship and truck transportation through to sub aspects such as the optimal combination of individual vehicles on a truck. In terms of used-vehicle logistics, optimizing analytics can be used to assign vehicles to individual distribution channels (e.g., auctions, Internet) on the basis of a suitable, vehicle-specific resale value forecast in order to maximize total sale proceeds.

In spare parts logistics, the provision of spare parts and their storage, data-driven forecasts of the number of spare parts needing to be held in stock depending on model, age and model (or sold volumes) are one important potential application area for data mining, because it can significantly decrease the storage costs.

(v) Marketing

The focus on marketing is to reach the end customer as efficiently as possible and to convince people either to become customers of the company or to remain customers. The success of marketing activities can be

measured in sales figures, whereby it is important to differentiate marketing effects from other effects, such as the general financial situation of the customers. Measuring the success of marketing activities can therefore be a complex endeavour, since multivariate influencing factors can be involved.

It would also be ideal if optimizing analytics could always be used in marketing, because optimization goals, such as maximizing return business from a marketing activity, maximizing sales figures while minimizing the budget employed, optimizing the marketing mix, and optimizing the order in which things are done, are all vital concerns. Forecast models, such as those for predicting additional sales figures over time as a result of a specific marketing campaign, are only one part of the required data mining results – multi criteria decision-making support also plays a decisive role in this context.

(vi) Sales and Service

Customer feedback, warranty repairs, and production are potentially inter-meshed as well, since customer satisfaction can be used to derive soft factors and warranty repairs can be used to derive hard factors, which can then be coupled with vehicle-specific production data and analyzed. In this way, factors that affect the occurrence of quality defects not present or foreseeable at the factory can be determined. This makes it possible to forecast such quality defects and use optimizing analytics to reduce their occurrence. Nevertheless, it is also necessary to combine data from completely different areas – production, warranty, and after-sales – in order to make it accessible to the analysis.

II. LITERATURE SURVEY

Zhiqiang Ge, Zhihuan Song, Steven X. Ding and Biao Huang (2017) [6], In this paper, methodologies of Data Mining and analytics in the process industry are introduced, with a systematically review through the viewpoint of machine learning. It is no doubt that machine learning plays a key role in data mining and analytics in the process industry. While both unsupervised and supervised machine learning methods have already been widely used in the process industry, which approximately accounts for 90% - 95% of all applications, the semi supervised machine learning has been introduced in recent years, thus its application will become more popular in the near future.

Kritika Yadav, Mahesh Parmar (2017) [7], analyse the performance of MGNREG scheme and various Data Mining Techniques are used. This research shows that data extracted from the bank and National Informatics Centre, Processed and analysed with Data Mining tool – Weka provide useful information to the government which could be for further improvement in the scheme. Using Decision tree algorithm (J 48) the payment of wages for the state of Rajasthan can be discovered. Horatiu Constantin Palade, Sergiu Stefan Nicolaescu, Claudiu Vasile Kifor (2016) [8], In this paper, Big Data methodologies in the Automotive Industries are introduced. Big Data will enhance the results and will

improve the Supply Chain and reduce the lean time of quality approaches like Six Sigma, Design for Six Sigma, Kaizen, Lean manufacturing; no more statistic results, clear data is provided.

S.Gunasekaran, C.Chandrasekaran (2011) [9], In this paper, methodologies of Data Mining and analytical in the Automobile Industries are used. The survey done in the present study on the data mining application techniques – Definitions, ID3 Algorithm (j48 in weka tool), k-means method for the automobile industries.

Rudolf Kruse, Matthias Steinbrecher, Christian Moewes (2010) [10], in this paper, presented method for pattern discovery tasks for the development stage, the manufacturing and planning stage as well as for maintenance and aftercare for Automotive Industries.

Association rule – Bayesian network used to find interesting patterns and Markov networks used to the planning task when assembling a vehicle and recurring patterns in time series used for adjusting simulation parameters.

III. AN APPROACH ON HANDLING MACHINE LEARNING APPLICATION (MAHOUT) IN AUTOMOTIVE

In the current automotive industry, data mining algorithms used for analyzing bulk data to identify trends and draw conclusions. However, no data-mining algorithm can be well-organized enough to process very large data sets and provide results in quick time. Due to the heterogeneity of data and the volume of data, the issue of “data mining in simulation data” is extremely complex. We now have new frameworks that allow us to break down a task into multiple segments and run each segment on a different machine. Mahout is type of data mining framework that normally runs joined with the Hadoop infrastructure at its background to manage huge volume of data.[4]

Mahout is an open source project that is mainly used for creating scalable machine learning algorithms. At the most general level, machine learning (ML) algorithms can be subdivided into two categories: supervised and unsupervised learning algorithms.

a. Supervised Learning algorithms:

Supervised Learning is used to predict numerical values (regression) and for classification purposes (predicting the appropriate class), and the corresponding data is not limited to a specific format – ML algorithm are more than capable of processing images, audio files, videos, numerical data, and text. Classification examples include object recognition (traffic signs, objects in front of a vehicle), face recognition, credit risk assessment and voice recognition. Regression examples include determining continuous numerical values on the basis of multiple input variables, such as a self-driving car calculating its ideal speed on the basis of road and ambient conditions, determining a financial indicator such as gross domestic product based on a changing number of input variables (use of population education

levels, industrial production etc.) and determining potential market shares with the introduction of new models.

b. Unsupervised Learning algorithms:

Unsupervised learning algorithms do not focus on individual target variables, but instead have the goal of characterizing a data set in general. Unsupervised ML algorithms are often used to group (cluster) data sets to identify relationships between individual data points and group them into clusters.

Apache Mahout:

Apache Mahout is a highly scalable machine learning library that enables developers to use optimized algorithms. Mahout implements popular machine learning techniques such as recommendation, classification, and clustering. [4]

• Recommendation

The Recommendation is a popular machine learning technique that provides close recommendations based on past actions and user information such as previous purchases and ratings.

• Classification

Classification, also known as categorization, is a machine learning technique that uses known data to determine how the new data should be classified into a set of existing categories. Classification is a form of supervised learning.

• Clustering

Clustering is used to form groups or clusters of similar data based on common characteristics. Clustering is a form of unsupervised learning.

IV. CONCLUSION

The survey done in the present study of the data mining application and Machine Learning techniques for the automotive industry will surely help them to overcome the existing issues on the DM algorithms and using new framework that allow them to process large datasets and provide outcomes in quick time. Machine Learning has already found its way into our daily lives. At present, machine learning is used primarily in the following

areas:

Analytical data processing

Domains in which qualified decisions need to be made quickly on the basis of a large amount of (often heterogeneous) data

In the field of analytical data processing, the next few years will see us transition from exclusive use of decision-support systems to additional use of systems that make decisions on our behalf.

REFERENCES

- [1] <https://data-science-blog.com/blog/2017/05/06/artificial-intelligence-and-data-science-in-the-automotive-industry/#4>
- [2] https://en.wikipedia.org/wiki/Crossindustry_standard_process_for_data_mining
- [3] https://en.wikipedia.org/wiki/Machine_learning
- [4] <https://www.tutorialspoint.com/mahout>
- [5] Th. Bäck, C. Foussette, P. Krause: Automatische Metamodellierung von CAE-Simulationsmodellen (Automatic Meta- Modeling of CAE Simulation Models), ATZ – Automobiltechnische Zeitschrift 117(5), 64-69, 2015
- [6] Zhiqiang Ge, Zhihuan Song, Steven X. Ding and Biao Huang, "Data Mining and Analytics in the process industry: The Role of Machine Learning", IEEE Access, 2017
- [7] Kritika Yadav, Mahesh Parmar, "Analysis of Mahatma Gandhi National Rural Employment Guarantee Act Using Data Mining Technique", International Journal of Computational Intelligence Research, ISSN 0973 – Volume 13, Number 9 (2017), pp. 2221 – 2235
- [8] Horatiu Constantin Palade, Sergiu Stefan Nicolaescu, Claudiu Vasile Kifor, "Model of Handling Big Data and Knowledge Management in Automotive Industry", International Conference 2016.
- [9] S.Gunasekaran, C.Chandrasekaran, "A Survey of Automobile Industries Using Data Mining Techniques", International Journal of Science and Advanced Technologies (ISSN 2221 – 8386), 2011.
- [10] Rudolf Kruse, Matthias Steinbrecher, Christian Moewes, "Data Mining Application in the Automotive Industry", 4th International Workshop on Reliable Engineering Computing, 2010.

