Anomaly Detection in Spot Welding Machines in the Automotive Industry for Maintenance Prioritization

Laislla C. P. Brandão [1]
lcpb@ecomp.poli.br

Aldonso Martins-Jr [1]
amoj2@ecomp.poli.br

Gabriel A. Kopte [1]
gak@ecomp.poli.br

Edson Filho [1]
jeaf@ecomp.poli.br

Alexandre M. A. Maciel [1]
amam@ecomp.poli.br

[1] *Department of Computer Engineering*
School of Engineering, Universidade de Pernambuco
Recife, Brazil

Abstract

Based on the need of prioritization of maintenance activities in a BOSCH Spot Welding process in the automotive industry, this work aims to develop anomalous equipment selection methodologies for assisting it. The first one is proposed based on data exploration by checking every possible set of alarms of the machines. A second one is created using multiple data clustering models in order to identify machines that behave differently from the others for certain time periods. Bayesian networks were also applied to assist the identification of cause-and-effect relationships between the warning and error logs. The clustering method proved effective in identifying anomalies, which were later inspected on the shop floor.

Keywords: data mining, maintenance, spot welding, automotive industry, anomalies.

1 Introduction

The Automotive Industry are in constant process of transformation and, in order to thrive in the era of Industry 4.0, automakers need to quickly adapt in order to continuously achieve their goals and overcome challenges, increasing the lifespan of their assets and productivity through the use of technologies such as Big Data Analytics, enabling intelligent manufacturing [1].

Having in mind that maintenance activities must make the best possible use of scheduled downtime and resources to minimize its losses, it is always necessary to select which activities should be performed. In many cases, traditional maintenance policies can be satisfactory, but when maintenance and failure costs are high [2], managing maintenance using data analysis becomes a better choice.

By analyzing the data generated by the welding machines and using data mining techniques, this study plans to develop methodologies that assist in the detection of anomalous behavior patterns, to support the reduction of unscheduled stops, maintenance time and repair costs and provide a rise in the efficiency and quality of the welding process.

2 Background

2.1 Spot Welding

The industrial process of manufacturing an automobile has four major workshops. The one of interest for this work is the Body-in-White Shop, responsible for the construction of the car bodies by joining its metal parts.

A common process to all body-in-white shops globally is the joining of metal sheets using a technique known as Spot Welding. The equipment used, called welding gun, uses two metallic copper electrodes to apply a force of union between metallic plates creating welding spots by passing high level electrical current through them and the metal worksheets in between [3]. The heat generated by the passage of high electric current through the small section of the electrodes melts the metals of the two plates, providing the union of the parts when the material solidifies again.

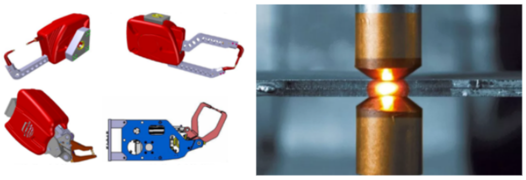

Figure 1: Example of welding guns (left) The union of two pieces of metal through spot weld (right)

A body-in-white shop contains thousands of welding guns and every machine stores information about the weld application, configuration parameters and measurements of each spot weld in a standard SQL Server database. It is available on the local manufacturing machine on the same industrial network as other welding equipment and robots.

2.2 Anomaly Detection

Cluster analysis is one of the most important research fields in data mining. Clustering belongs to the category of unsupervised learning as it does not depend on training samples, this is why they are the straightforward technique for

DOI reference number: 10.18293/SEKE2022-118

anomaly detection. For these algorithms, given a number of clusters k as an initial parameter, the set of data objects is divided into k categories, or groups.

An example of this type of algorithm is K-Means. One can use several sets of different starting centers for various iterative calculations and choose the best one as the final result, but one cannot guarantee that this result is the optimal solution, while several iterations consume a lot of time, a lot of uncertainty, so it is very important to select the ones suitable starting group centers [4].

Another algorithm, DBSCAN (Density-Based Spatial Clustering), is a pioneering density-based algorithm. It can discover clusters of any arbitrary shape and size in databases that contain even noise and outliers, although it has some problems, such as being subject to dilemmas when deciding meaningful clusters from datasets with varying densities [5].

When the algorithm is able to minimize an error function, it is often called C-Means where c is the number of clusters, and if the classes used are using the Fuzzy technique, then it is known as Fuzzy C-Means (FCM) [6]. Its benefit is the formation of new clusters from data points that have membership values close to existing classes. Fuzzy C-Means has the advantage of being very good for problems of many dimensions.

The most popular model of association patterns between groups of items uses item set frequencies to quantify association level, but there are also the Bayesian, that use Bayes' theorem to represent knowledge. In simple cases, the structure of Bayesian networks can be defined by an expert and used to make inferences about a given problem. In other more complex applications the structure and parameters of the network can be learned [7] [8].

2.3 Related Work

Many works found propose a prioritization approach from a maintenance perspective, with data-oriented approaches and aiming at the preventive diagnosis of problems. Several methods have been developed to identify performance bottlenecks related to maintenance activities in production systems [9]. The underlying logic behind all of them lies in analyzing the machines' event log data [10].

Data mining techniques are used to detect bottlenecks in production systems, although several methods proposed focus on analytical logs referring to the process, these data do not provide diagnostic information explaining what the root causes of incidents are [11].

In one of the approaches, Bayesian networks and attribute relevance analysis are used to process a dataset of failure records of industrial machinery components, with the purpose of using the conditional probabilities generated by the networks, as well as the relevance of the rankings of criteria for creating a decision-making model [12].

At the same time, some works specifically related to spot welding technology and case studies applied to BOSCH were also identified. Due to the great variety and diversity of data collected in the welding process and which are relevant for monitoring product quality, the work of Svetashova et al. [13] reports that they find significant challenges for the modeling of machine learning algorithms and these challenges are presented in conjunction with the predictive quality monitoring model.

3 Materials and Methods

3.1 Database Description

The CRISP-DM methodology was used for providing a framework for carrying out data mining projects, regardless of the industry sector and the technology used.

As said, this database automatically stores spot welding information in specific datasets for each purpose. The *ExtError_RDS_V* dataset contains records of important events that occurs in one of the welding controllers in the process, which may have warning codes or errors. This dataset has 13 columns and the existing fields are shown in Table 1.

For this work, the focus was given to one single production line, the bottleneck line, thus the most important one, and all its welding machines. The data considered was the gathering of a month of production.

Table 1: Dataset *ExtError_RDS_V* Data Description

COLUMN	DESCRIPTION
date	Event registration date
line	Record of the production line where the robot is located
protRecord_ID	Single consecutive number (automatic value)
dateTime	Timestamp record of the moment of occurrence of the event
timerName	Unique identification of the registered event source welding machine
errorCode1	Warning or Error code event source
errorCode1_txt	Warning or Error event source - Text in selected language
errorCode2	Secondary error code
errorCode2_txt	Secondary error code - Text in selected language
isError	Flag to identify if the event is a Warning or Error
isError_txt	Literal text Warning or Error
servodynDVState	Servo status code
servodynDVState_txt	Servo Status - Text in selected language
tablename	Address of the folder on the computer where the ExtError tables are located

3.2 Data Preprocessing

The data was loaded into a NoSQL database in the Google Cloud Computing (GCP) cloud called BigQuery and the approach chosen was firstly the adaptation of the fact tables of the database in an OLAP (Online Analytical Processing) architecture, since it allows greater flexibility and performance in data analysis. Data was transformed from categorical columns into non-categorical columns from the beginning, the primary keys with ID for the dates and machines were stored separately in the Dimension tables and the new columns *id_date* and *id_machine* were used

DOI reference number: 10.18293/SEKE2022-118

instead of *date* and *timerName*.

White spaces at the end of the text were removed, especially in error description columns, and a mapping of the missing data in the database was carried out, these values being later filled with a symbolic value of (-1) not to impact the operation of subsequent algorithms.

In order to make possible a more in-depth analysis regarding each of the alarms present in the database, the columns *errorCode1_txt* and *errorCode2_txt* were concatenated creating a new column, *errorCode_txt*. This brings a second level of detail for each error.

After that, a new table was created, pivoting and grouping the data of *ExtError_RDS_V* by dates and machines and presenting the number of occurrences of each of the alarms of *errorCode_txt* arranged in separate columns, one for each alarm. The new table, called *ExtError_group*, has 40 columns, two of which are the date and machine ID and the others represent each of the possible alarms in the welding guns, encompassing both warnings and errors.

3.3 Exploratory Data Analysis

The analysis were initiated by looking at the distinct values for each column of the original *ExtError_RDS_V* database and creating histograms based on their categories. At first, a focus was given to the *errorCode1_txt* column in order to identify the unique descriptions of the existing alarms (errors and warnings), and after that, an analysis of occurrences of each possible alarm was performed in the new concatenated column *errorCode_txt* to verify the types of alarms most common in the whole dataset.

From the new dataset, *ExtError_group*, it is possible to analyze separately the influence that each one of the alarms has on the behavior of the machines. For instance, the error *"welding error"* with the sub-description *"cancellation by pliers movement"* considering all machines from the perspective of each date is shown in Fig. 2.

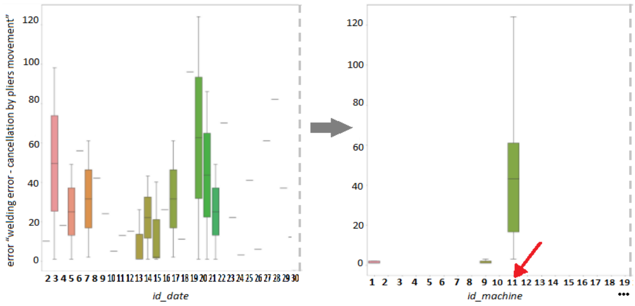

Figure 2: Variation of the error *"welding error - cancellation by pliers movement"* for each date id (left) and for each machine id (right).

When analyzing this same error from the perspective of each machines, a very important information is acquired. By looking at the boxplots, the machine id=11 stands out when compared to the others in terms of data variance, which shows that there may be an opportunity associated with this machine.

3.4 Modeling

Two approaches were used to prepare the data for the models, one using the original data from the *ExtError_group* dataset and the other using the normalized data through the StandardScaler method.

K-Means, DBSCAN and Fuzzy C-Means clustering algorithms were the techniques used for the anomaly detection. The K-Means and Fuzzy C-Means methods require the number of clusters as an initial parameter, while DBSCAN requires an agglutination radius and the minimum number of records to form a cluster. To determine these parameters, a search using the two approaches for each method and dataset was performed: Maximum Silhouette score and Elbow Curve (Inertia).

Therefore, a total of 12 models were analyzed in search of machines of interest: Three Methods (K-Means, DBSCAN and Fuzzy C-Means) × Two Datasets (Non-standardized and Standardized) × Two Approaches (Maximum Silhouette and Elbow Curve, using Silhouette for DBSCAN and Inertia for the others two methods.

Bayesian networks were used to find causality in the error and warning logs, trying to relate apparently non-relevant errors and warnings with errors related to critical failures for the machines of interest. The structure was trained using the Hill Climbing Search algorithm.

4 Evaluation and Results

4.1 Results

For the first of the 12 models, the result of the Inertia metric in the search for parameters (number of clusters) for the K-Means method and Non-standardized data is the point where a discontinuity occurs (the elbow). Here it was possible to see that the ideal number of groups that best represents the data from *ExtError_group* is four for the Elbow Curve method, showed in Fig. 3(a).

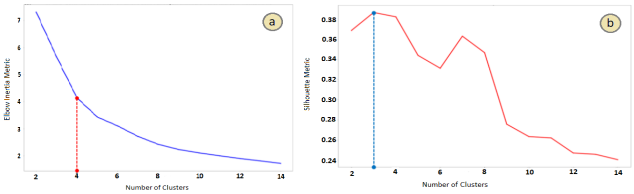

Figure 3: Result of Elbow Method Inertia vs. number of clusters (a) and Result of Silhouette Score vs. number of clusters (b).

This parameter is used to train the K-Means algorithm and then apply the model to the data to obtain an association

DOI reference number: 10.18293/SEKE2022-118

of each of the records in the database to one of the four clusters. This model had an overall Silhouette Score of 0.38 and an average Euclidean distance of 48.9.

Fig. 4(a) shows the number of day-machine pairs grouped in each cluster for the K-Means model with four clusters and the approach using Elbow Curve Inertia with Non-standardized data. It is observed that 28 events were isolated in group number 2, clearly different from the others grouped in large clusters. Minority groups tend to show rare and/or unusual events that may indicate good opportunities for preventive maintenance plans by characterizing machines that behaved anomalously in a small set of days.

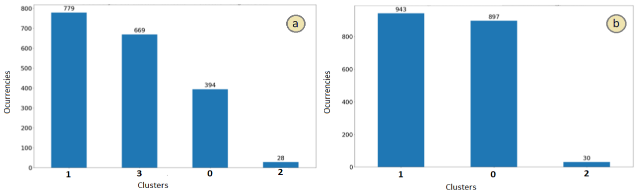

Figure 4: Distribution of clusters for K-Means model with four clusters (a) and three clusters (b) and Non-standardized data.

By looking at the records of each machine distributed among the four clusters, it was verified that the 28 events of interest occurred on the same machine, id=5. This machine was then declared a machine of interest, as it may be associated with anomalous functioning, and this information should be compared with the shop floor and the information held by the stakeholder, the welding specialist.

The Max Silhouette approach was used to select the number of clusters for training the second K-means algorithm, also applied to Non-standardized data, showed in Fig. 3(b). Three clusters were selected for this model and it was trained, obtaining the association between records and clusters. This model had an overall Silhouette Score of 0.38 and an average Euclidean distance of 41.4.

Fig. 4(b) shows the amounts of day-machine pairs grouped in each cluster of the second model, with K-Means with three clusters and the approach using Max Silhouette, and Non-standard data. It is seen that 30 events were isolated in group number 2, another notably minority group. As discussed earlier, this can characterize a set of machines with anomalous behavior.

Once again, observing the records of each machine distributed among the clusters, it was seen that the 30 events of interest in cluster number 2 are from the same machine id=5, reinforcing that this machine is a machine of interest for maintenance prioritization.

The Bayesian network was created to the machine with id=5 (the anomaly data from cluster number 2, the minority class). The resulting network was analyzed and filtered with the help of the stakeholder. Three relationships of interest were identified. Relationships 1 and 2 indicate a lack of current error related to a maximum lag warning. It refers to the phase shift of current with respect to voltage in the welding process. Relation 3, on the other hand, indicates a current oscillation problem. The main possible causes for the errors presented in 1, 2 and 3 are the same: abrasion of the welding electrode, measuring circuit or auxiliary cables; interference from other processes on the same network; and weld transformer problems (insufficient capacity).

Table 2 summarizes the result of the same methodology applied to these and the other 10 models created, and Table 3 summarizes the total occurrences of machines of interest identified by each of the models that were applied.

Table 2: Summary of algorithms, parameter selection methods, applied parameters, metrics (Silhouette Score - Distance) and the identified machines of interest.

ALGORITHM AND GROUPS	STANDARD DATA	METHOD SELECTION OF NUMBER OF GROUPS	S. SCORE AND DIST.	ID MACHINES OF INTEREST
K-Means-3	No	Max Silhouette score	0.38 - 48.9	5
K-Means-2	Yes	Max Silhouette score	0.43 - 13.0	Not identified
K-Means-4	No	Elbow Method (Inertia)	0.38 - 41.4	5
K-Means-4	Yes	Elbow Method (Inertia)	0.32 - 10.1	17, 18
DBSCAN-2[a]	No	Max Silhouette score	0.47 - none	2, 4, 17
DBSCAN-2[b]	Yes	Max Silhouette score	0.15 none	2, 3, 4, 13, 17, 18
DBSCAN-4[a]	No	Elbow Method (Silhouette)	0.33 - none	5, 7, 17, 18, 26
DBSCAN-4[c]	Yes	Elbow Method (Silhouette)	0.16 - none	2, 4, 13, 17, 33
FC-Means-2	No	Max Silhouette score	0.37 - 51.6	23, 33, 34
FC-Means-2	Yes	Max Silhouette score	0.43 - 13.0	Not identified
FC-Means-4	No	Elbow Method (Inertia)	0.31 - 39.2	Not identified
FC-Means-4	Yes	Elbow Method (Inertia)	0.32 - 10.1	17, 18

[a] Result of eps=19 and min_samp=10.

[b] Result of eps=18 and min_samp=7.

[c] Result of eps=10 and min_samp=3.

Table 3: Summary of the total occurrences of machines of interest identified by the various models.

ID MACHINES	17	18	2	4	5	33	13	7	3	34	23	26
OCCURRENCES	6	4	3	3	3	2	2	1	1	1	1	1

4.2 Discussion

For each cluster found, it is necessary to map the inherent characteristics that isolate one from the others. The implications of each group and latent opportunities in this analysis are under continuous discussion with the stakeholders. In general, it is expected that most clusters are associated with groupings of only warnings and a mix of warnings and alarms events, both cases in proportions that are common to the process. Minority classes are the ones likely to have anomalous events that must be analyzed with greater care.

Following Lima et al.[12], The Bayesian network analysis identified that there may be a failure in the process regarding the quality of energy in the weld on the machine

DOI reference number: 10.18293/SEKE2022-118

id=5. The root cause is believed to be associated with a failure in the milling process of the welding electrodes, or in the cables and measurement and power circuits of this machine. There is also the possibility that the welding controller is not able to supply the power and energy required for the application of the spot, either due to a wear event (gaps, contact faults), mechanical conditions of alignment and orthogonality or due to electrical defects in electronic components of the welding controller itself.

The machines shown in Table 3 gave rise to greater opportunities for maintenance intervention. Some were not even in the radar of prioritization and these results turned out to be of great importance. A great emphasis was given to the machines id=17 and id=18, as they appear in several models as machines of interest. The machines were inspected on the shop floor and issues such as the early abrasion of the welding electrodes were raised and treated, restoring their base conditions.

5 Conclusions

5.1 Conclusion

One of the main goals of this work was to explore the data in search of anomalies that could lead to latent opportunities for the priorization of activities in specific machines. Data was migrated from the on-premises SQL database to the cloud, where it was consumed for processing and analysis. A data preprocessing was carried out in order to prepare and model them to obtain the necessary information for the purpose of finding anomalous patterns in the data.

The types of alarms are very unbalanced due to the normal operation of the welding process. Most of the data is composed of Warnings, which do not necessarily imply losses in the production process. Some warnings may simply mean records of the normal functioning of the process, such as records of milled electrodes or signaling that these need to be milled, although in some cases, prealarms, they may be indicative of errors that may occur later.

From the results obtained with the work, the authors came with the definition of two methodologies to identify machines of interest. The first one consists of scanning all types of machine failures still in the data exploration stage. When finding a machine with a variance above the others, it is considered a machine of interest for inspections and close attention of the Maintenance team. The second methodology, associated with the result of the Unsupervised Clustering algorithm, consists of classifying the events organized by day and machine, in which the database columns are composed of each possible warning or error in the process, with the values being the amount of event occurrences for the machine-data pair. It was possible to apply several different models to this data and identify the records grouped into minority classes as machines of interest.

5.2 Future Work

Other datasets, such as the spot welding process parameters and measurements *ExtMeasuresProt_V*, are likely to be used in the future to improve the findings produced with this work, to improve analysis and associate failure modes in more detail. Also, the use of this dataset might help to identify new priorities that also benefit quality control, not exclusively the maintainability of the welding process.

Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

[1] R. Y. Zhong, X. Xu, E. Klotz, and S. T. Newman, "Intelligent manufacturing in the context of industry 4.0: A review," *Engineering*, vol. 3, no. 5, pp. 616–630, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095809917307130

[2] A. Parida and U. Kumar, "Maintenance productivity and performance measurement," in *Handbook of maintenance management and engineering*. Springer, 2009, pp. 17–41.

[3] B. Zhou, Y. Svetashova, S. Byeon, T. Pychynski, R. Mikut, and E. Kharlamov, "Predicting quality of automated welding with machine learning and semantics: a bosch case study," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2933–2940.

[4] H. Zou, "Clustering algorithm and its application in data mining," *Wireless Personal Communications*, vol. 110, no. 1, pp. 21–30, 2020.

[5] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "Dbscan: Past, present and future," in *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*. IEEE, 2014, pp. 232–238.

[6] J. Nayak, B. Naik, and H. Behera, "Fuzzy c-means (fcm) clustering algorithm: a decade review from 2000 to 2014," *Computational intelligence in data mining-volume 2*, pp. 133–149, 2015.

[7] K. B. Korb and A. E. Nicholson, *Bayesian artificial intelligence*. CRC press, 2010.

[8] J. Pearl, "Bayesian networks," 2011.

[9] M. Subramaniyan, A. Skoogh, A. S. Muhammad, J. Bokrantz, B. Johansson, and C. Roser, "A data-driven approach to diagnosing throughput bottlenecks from a maintenance perspective," *Computers & Industrial Engineering*, vol. 150, p. 106851, 2020.

[10] C. Roser, M. Nakano, and M. Tanaka, "Comparison of bottleneck detection methods for agv systems," in *Winter Simulation Conference*, vol. 2, 2003, pp. 1192–1198.

[11] C. Yu and A. Matta, "A statistical framework of data-driven bottleneck identification in manufacturing systems," *International Journal of Production Research*, vol. 54, no. 21, pp. 6317–6332, 2016.

[12] E. Lima, E. Gorski, E. F. Loures, E. A. P. Santos, and F. Deschamps, "Applying machine learning to ahp multicriteria decision making method to assets prioritization in the context of industrial maintenance 4.0," *IFAC-PapersOnLine*, vol. 52, no. 13, pp. 2152–2157, 2019.

[13] Y. Svetashova, B. Zhou, T. Pychynski, S. Schmidt, Y. Sure-Vetter, R. Mikut, and E. Kharlamov, "Ontology-enhanced machine learning: a bosch use case of welding quality monitoring," in *International Semantic Web Conference*. Springer, 2020, pp. 531–550.

DOI reference number: 10.18293/SEKE2022-118