Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 35 (2019) 623–628

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

2nd International Conference on Sustainable Materials Processing and Manufacturing (SMPM 2019)

Performance Evaluation of Data Mining Techniques in Steel Manufacturing Industry

Thembinkosi Nkonyana*[a], Yanxia Sun[a], Bhekisipho Twala[c], Eustace Dogo[a]

[a]*Department of Electrical and Electronic Engineering Science, University of Johannesburg, 2006, South Africa*
[c]*Department of Electrical and Mining Engineering, University of south Africa, Florida 1710, South Africa*

Abstract

Industry 4.0 has evolved and created a huge interest in automation and data analytics in manufacturing technologies. Internet of Things (IoT) and Cyber Physical System (CPS) are some of the recent topics of interest in the manufacturing sector. Steel manufacturing process relies on monitoring strategies such as fault detection to reduce number of errors which can lead to huge losses. Proper fault diagnosis can assist in accurate decision-making. We use in this study predictive analysis to help solve the complex challenges faced in industrial data. Random Forest, Artificial Neural Networks and Support Vector Machines are used to train and test our industrial data. We evaluate how ensemble methods compare to classical machine learning algorithms. Finally we evaluate our models' performance and significance. Random Forest outperformed other ML methods in our study.

Keywords: Machine Learning, Manufacturing, Fault Diagnostics

1. Introduction

Industry 4.0 has evolved and created a huge interest in automation and data analytics in manufacturing technologies.

* Corresponding author. Tel.: +2711 559 2110.
tnnkonyana@uj.ac.za

2351-9789 © 2019 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the organizing committee of SMPM 2019.
10.1016/j.promfg.2019.06.004

Internet of Things (IoT) and Cyber Physical System (CPS) are some of the recent topics of interest in the manufacturing sector. We have seen an adaptation of sensors in manufacturing plants, tools and intelligent devices connected together with sensors. The sensors are however collecting various amounts of data from intelligent devices connected together in the manufacturing environment. The manufacturing industry is however as part of its objectives expected to produce and deliver high quality products which are safe and must be efficiently cost effective. Industrial manufacturing can suffer and experience problems due to failure to deliver on this objectives. Steel products can however endure deviation which arise from factors such as shape, appearance, structure, dimensions and many more which don't match the specification for given standards. Aspects such as industrial plant monitoring aim to achieve reduced manufacturing costs, improved system performance, increased product quality and early detection of defects in products [[1]; [2]; [3]]. Fault detection diagnostic can however help quickly identify defects in products produced from manufacturing industry. The process of identifying faults can be done with the use of visual inspection, which is not recommended as it is time consuming and may produce inaccurate decisions and the other way is by using instruments and equipment that can be able to capture faults. If this faults in steel are not found early in the manufacturing process, there can be unfavorable results such as product failure, non-availability of products, and materials which cannot be used. There is however a need to find patterns in the generated data for accurate decision making.

Machine Learning (ML) is a branch of AI. Many industries in different domains make use and apply ML techniques with the ability to make the system learn from previous knowledge in order to develop or make predictions for future events and current [8]. ML techniques can be used to identify patterns and detect faults as a way of assisting in process monitoring, and efficient processes in manufacturing [4]. Motivation: the purpose of this study is to identify and analyze techniques which can be used to perform fault detection and diagnostics. Fault diagnostic seeks to find defective states and certain poor conditions within various manufacturing product and systems. Therefore measurements of a particular product are used to monitor a bad state of that product. The organization of the paper is as follows: Section one is introduction and machine learning techniques. Review of related work to fault detection and diagnosis in manufacturing in section two and background. Section three covers the experimental setup with explanation and discussion. Section four covers the discussion of the experimental results. Lastly, we conclude of the findings of the paper by use of discussions and future research possibilities.

1.1. Support Vector Machine

Support Vector Machines (SVM) is a very popular machine learning algorithm developed by Vapnik [[5]; [6]]. SVM's fall have the benefit of applying various kernel functions. Kernel functions in SVM's work in a way that maps the input feature space, in cases where data turn to be linearly separable [7]. This particular method is very popular for handling large data, and is made up of two categories which is for purposes of classification and regression. It used a hyperplane that makes use of a maximum margin by separating two boundaries [8].

1.2. Random Forest

Random Forests (RF) belong to the family of decision trees and are very popular in the various aspects of machine learning application which can be categories for tasks such as regression and classification for supervised learning[8]. Decision trees are constructed based on a tree like graph or hierarchical decision structure which is made up of leaves that represent class labels and branches that represent divisions of features mapping to the class labels [[5]; [9]]. Random forest (RF) under the decision trees belong to the family of ensemble learning algorithms. Other types of decision trees are ID3, C4.5, C5.0, CART (Classification and Regression Tree), CHAID (CHI-squared Automatic Interaction Detector. C4.5 was actually a successor to ID3, and later came C5.0. Advantages of decision trees are as follows: they are easy to understand, can handle large dataset, do not require much data preparation, and have a built in feature selection. Disadvantages of decision trees are as follows: they are prone to overfitting, during prediction and can be complex in implementation.

2. Background and literature review

Faults in the manufacturing industry exist and it is vital to develop techniques to understand and diagnose faults which can yield good results in cost reduction and increased quality control and many more benefits. Methods of fault detection can categorized in types namely model based, knowledge based, and signal base [10]. The model based fault detection has to do with for a mathematical perspective with comparing the actual and the expected behavior, while secondly the knowledge based fault detection has to do with identifying and mapping of faults by using qualitative models which are associated with heuristic symptoms for purpose of reasoning to causes of faults, and lastly the signal based fault detection method are due to the nature of spectral analysis do however not include any model [10].

In the study of mining a metallurgy industry domain, a Random forest feature extraction was applied to test fault diagnostic scheme which was tested on a simple nonlinear system, and two faults of benchmark Tennessee Eastman process [4]. The authors indicate that random forest has the ability to be robust in fault diagnosis to process monitoring. In order to detect sensor faults in heating, ventilation and air conditioning (HVAC) systems for past performance data, the authors in [2] proposed a combination of rough set approach and artificial neural network. Their model proved to be very successful as they used rough set to reduce unnecessary features. A support vector machine is compared with a ML technique for anomaly detection in rotating components, and the objective was to perform classification between defects using fault seeded bearing tests. The result indicated that anomaly detection techniques were considered to perform better than SVM [3].

Moreover [11], compared in this particular study using optimized data from non-optimized sensor set solution between linear support vector machines, distance-weighted k-nearest neighbor (WKNN) and decision tree. With the aim of categorizing faults differing form various degrees for severity, they tested algorithms using past known data to predict unknown outcome. Techniques such as feature selection are most popular in machine learning, and a study in [7] they proposed a nonlinear SVM feature selection technique for the purpose of managing process monitoring and fault detection. As part of this process an evaluation of ranking features in order to assist in the process of fault diagnosis. Their feature selection method was successful in improving accuracy for fault detection and diagnostics.

Further studies experimented a proposed a novel technique for fault detection and diagnosis which is based on One-Class SVM. This method makes use of applying an SVM-recursive feature elimination for feature selection method. However this approach was investigated to compare conventional techniques such as Principal Component Analysis and Dynamic Component Analysis for measuring performance with metrics such as false alarm rates, detection latency and fault detection rates [12]. Their method however was more robust than PCA and DPCA for detecting faults and diagnostics. In another study [9] applied Random Forest with the aim of using similarity distance measurement as a metric for anomaly detection in a semiconductor manufacturing process. Making use of industrial data for experimentation, their method accurately detected faulty wafers, and they further suggest that RF method applied is best suited for anomaly detection in Big Data cases.

3. Experimental Setup

The dataset, which was used for the purpose of this study, was obtained from UCI Machine Learning Repository[13] . The name of the dataset is called “Steel Plates Faults Data Set”. The number of instances is 1941 and with 27 attributes. The attribute characteristics are presented in a form of integers and numbers

Table 1. List of classes and number of samples

Class	Fault Type	No of Samples
1	Pastry	158
2	Z_Scratch	190
3	K_Scatch	391
4	Stains	72
5	Dirtiness	55
6	Bumps	402
7	Other_Faults	673

Table 2. List of Attributes in the Steel dataset

Number	Feature Attributes	Number	Feature Attributes	Number	Feature Attributes
Attribute 1	X_Minimum	Attribute 11	Length_of_Conveyer	Attribute 21	Outside_Global_Index
Attribute 2	X_Maximum	Attribute 12	TypeOfSteel_A300	Attribute 22	LogOfAreas
Attribute 3	Y_Minimum	Attribute 13	TypeOfSteel_A400	Attribute 23	Log_X_Index
Attribute 4	Y_Maximum	Attribute 14	Steel_Plate_Thickness	Attribute 24	Log_Y_Index
Attribute 5	Pixels_Areas	Attribute 15	Edges_Index	Attribute 25	Orientation_Index
Attribute 6	X_Perimeter	Attribute 16	Empty_Index	Attribute 26	Luminosity_Index
Attribute 7	Y_Perimeter	Attribute 17	Square_Index	Attribute 27	SigmoidOfAreas
Attribute 8	Sum_of_Luminosity	Attribute 18	Outside_X_Index		
Attribute 9	Minimum_of_Luminosity	Attribute 19	Edges_X_Index		
Attribute 10	Maximum_of_Luminosity	Attribute 20	Edges_Y_Index		

The simulation is done using a Dell Latitude core i7, memory size is 8G ram. The code is written in Python environment. Three algorithms, two performance measures, gridsearch and k-Fold validation is utilized.

4. Results

The results show that Random forest is robust to other algorithms for the task of classifying faults. Random Forest achieve the highest accuracy of (0.778) while SVM followed by (0.736), and ANN followed with (0.696). We however wanted to further test the impact of parameter tuning and hyper parameter. For this task we left out ANN and only focused on SVM and RF. We made use of Grid Search as a tool to perform tuning to the algorithms and applied also K-fold validation to our results. The results we obtained from our simulation was that SVM performed better as we did a full algorithm tuning, and for RF we reduced some of the parameters. The results with grid search show that the best parameters results = {'C': 1, 'gamma': 0.1, 'kernel': 'rbf'} for SVM, and best accuracy = 0.778. RF however suffered as we looked at the least of its parameters and best parameters were {'bootstrap': False, 'criterion': 'entropy', 'max_depth': 20, 'min_samples_leaf': 10, 'n_estimators': 600} with Best accuracy = 0.755. The other ways to compare if this is important we can compare the results with Random Search method, but for the purpose of this study we will not perform that exercise. Recall takes in to account the number of true positives divided by the number of true positives plus the number of false negatives. Precision however is the number of true positives divided by the number of true positives plus number of false positives. In this case of study it would mean faults correctly classified divided by faults correctly classified plus faults incorrectly classified as faults.

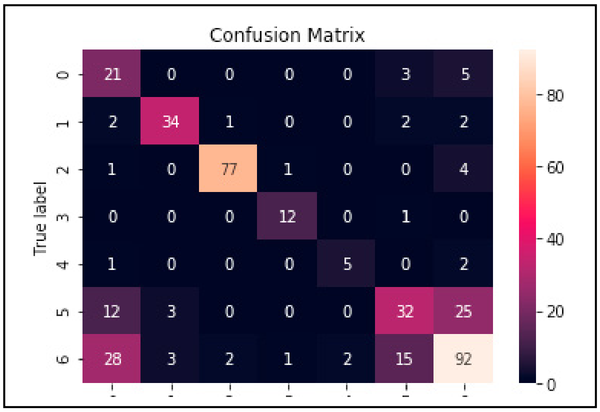

Fig. 1. ANN Confusion Metrix

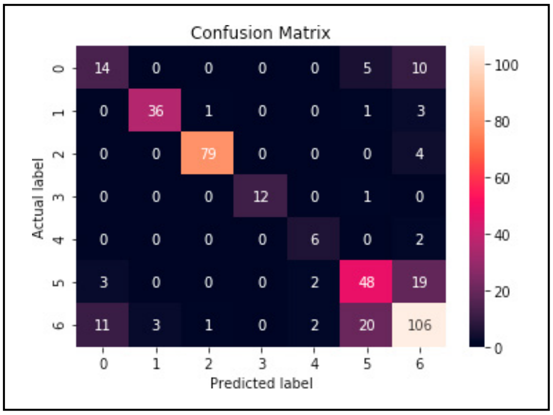

Fig. 2. RF Confusion Metrix

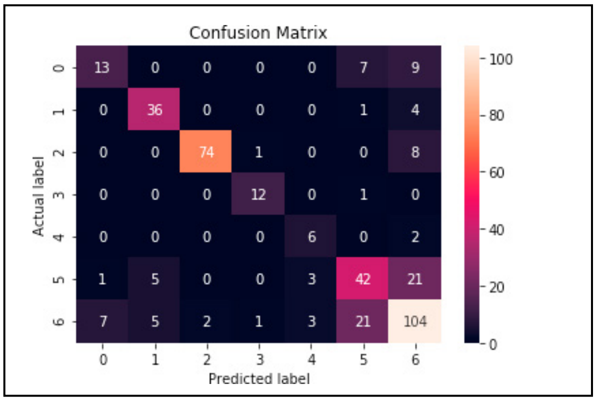

Fig. 3. SVM Confusion Metrix

5. Conclusion

Machine learning techniques can be used and applied to steel manufacturing process which relies on monitoring strategies such as fault detection to reduce number of errors which can lead to huge losses. Investment in understanding how to better ML algorithms can be applied in order to help in fault diagnosis which can assist in accurate decision-making. Future research is to (i) evaluate techniques for fault diagnostics in real time using predictive maintenance. The research is very essential as sensors record huge amounts of data that need Big Data Analytics (ii) to help with analysis for better decision making. Real time manufacturing process compromised due to lack of proper monitoring techniques for identifying faults.

Acknowledgements

This research is supported partially by South African National Research Foundation Grants (112142), and South African National Research Foundation Incentive Grant (No. 95687), Eskom Tertiary Education Support Programme (Y. Sun), Research grant from URC of University of Johannesburg and the Institute of Intelligence of University of Johannesburg.

References

[1] A. Sánchez-Fernández, F. J. Baldán, G. I. Sainz-Palmero, J. M. Benítez, and M. J. Fuente, “Fault detection based on time series modeling and multivariate statistical process control,” *Chemom. Intell. Lab. Syst.*, vol. 182, no. July, pp. 57–69, 2018.

[2] Z. Hou, Z. Lian, Y. Yao, and X. Yuan, “Data mining based sensor fault diagnosis and validation for building air conditioning system,” *Energy Convers. Manag,* vol. 47, no. 15–16, pp. 2479–2490, 2006.

[3] A. Purarjomandlangrudi, A. H. Ghapanchi, and M. Esmalifalak, “A data mining approach for fault diagnosis: An application of anomaly detection algorithm,” *Meas. J. Int. Meas. Confed.*, vol. 55, pp. 343–352, 2014.

[4] C. Aldrich and L. Auret, *Fault detection and diagnosis with random forest feature extraction and variable importance methods*, vol. 43, no. 9 PART 1. IFAC, 2010.

[5] R. A. Ariyaluran Habeeb, F. Nasaruddin, A. Gani, I. A. Targio Hashem, E. Ahmed, and M. Imran, “Real-time big data processing for anomaly detection: A Survey,” *Int. J. Inf. Manage.*, no. August, pp. 1–19, 2018.

[6] A. Widodo and B. S. Yang, “Support vector machine in machine condition monitoring and fault diagnosis,” *Mech. Syst. Signal Process.* vol. 21, no. 6, pp. 2560–2574, 2007.

[7] M. Onel, C. A. Kieslich, Y. A. Guzman, and E. N. Pistikopoulos, *Simultaneous Fault Detection and Identification in Continuous Processes via nonlinear Support Vector Machine based Feature Selection*, vol. 44. Elsevier Masson SAS, 2018.

[8] T. Nkonyana and B. Twala, *Artificial Intelligence and Evolutionary Computations in Engineering Systems*, vol. 668. Springer Singapore, 2018.

[9] L. Puggini, J. Doyle, and S. McLoone, “Fault detection using random forest similarity distance,” *IFAC-PapersOnLine*, vol. 28, no. 21, pp. 583–588, 2015.

[10] L. A. M. Riascos, L. A. Moscato, and P. E. Miyagi, “Detection and treatment of faults in manufacturing systems based on Petri Nets,” *J. Brazilian Soc. Mech. Sci. Eng.*, vol. 26, no. 3, pp. 280–289, 2004.

[11] M. Jung, O. Niculita, and Z. Skaf, “Comparison of Different Classification Algorithms for Fault Detection and Fault Isolation in Complex Systems,” *Procedia Manuf.*, vol. 19, no. 2017, pp. 111–118, 2018.

[12] S. Mahadevan and S. L. Shah, “Fault detection and diagnosis in process data using one-class support vector machines,” *J. Process Control*, vol. 19, no. 10, pp. 1627–1639, 2009.

[13] Semeion, Research Center of Sciences of Communication, Via Sersale 117, 00128, Rome, Italy