Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776548

A ANOMALIES PREDICTED NOVEL METHOD FOR FAULT DETECTION IN BIG DATA ANALYTICS

Dr Akash Saxena, Pawan Agarwal
Compucom Institute of Technology and Mangement,Jaipur

Abstract
Numerous real-world applications are hampered by the fact that typical anomaly detection methods used in full-dimensional fields perform much worse as dimensionality rises. The method for choosing an important feature subspace and performing anomaly detection in the associated subspace projection is proposed in this study. Maintaining detection accuracy under high-dimensional conditions is the goal. The proposed method determines the angle between each pair of two lines for a particular anomaly candidate: the first line is joined by the relevant data point and the centers of its adjacent points, and the second line is one of the axis-parallel lines. The candidate's axis-parallel subspace is thus made up of those dimensions that have a comparatively modest angle with the initial line. An further experiment using an industrial dataset showed how the suggested approach could be used for fault detection jobs and emphasized one of its benefits, which is the capacity to provide a first interpretation of abnormality through feature ordering in pertinent subspaces.

Keywords: Big Data Anomalies, Fault Detection, Agile Methodologies

1. Introduction

Big data is already a common phenomenon thanks to the extensive usage of information and communication technology (ICT). Sensor-heavy Condition Monitoring Systems (CMS), enterprise asset SCADA (Supervisory Control and Data Acquisition) systems, and enterprise asset management (EAM) systems are just a few examples of the sources that are generating data for the industry at previously unheard-of rates and scales. For operation and maintenance research, they offer a fast growing resource, particularly as practitioners and scholars become more aware of the possibility for utilising latent value from these data. Manufacturing is one of the five key industries where big data analytics can have a disruptive impact, according to a recent McKinsey Institute analysis (Manyika et al. 2011). The field of eMaintenance, which is a part of electronic manufacturing, benefits from big data analytics (Figure 1.1 shows the integration of eMaintenance, e-Manufacturing, and e-Business systems). Supporting maintenance decision-making is one of eMaintenance's core goals.

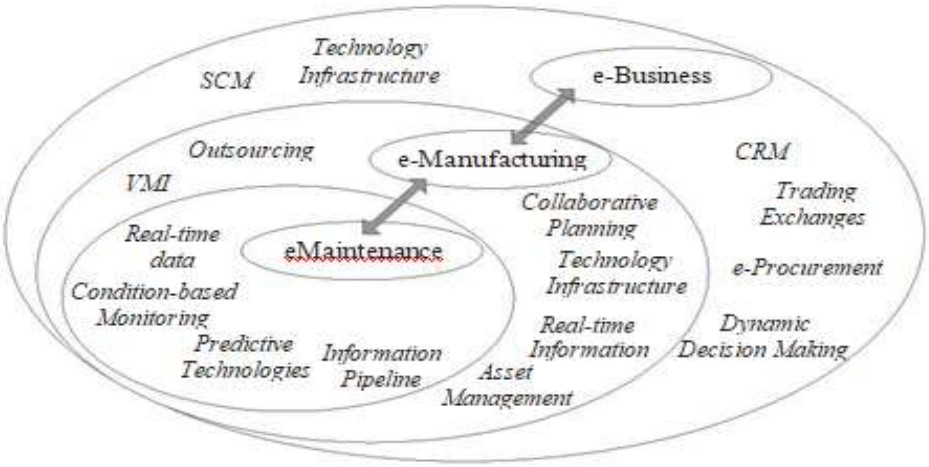

Figure 1 The integration of eMaintenance, e-Manufacturing, and e-Business systems.

The big data age has already begun. The rate at which data stored on servers in different domains increases far beyond expectations. Traditional database management systems have played an important role in storing, processing, managing, analyzing, and visualizing databases in the last decade [1]. Due to advances in the digital age, data collected from a variety of sources, such as sensors, scientific prediction, telecommunications, social media, bioinformatics, and healthcare, are in structured, semi-structured, and unstructured formats. It is necessary to research the knowledge of the volume of stored data [2]. The evolution of big data has attracted the attention of researchers, private and public organizations, as a result of the benefits of storing, processing, managing, analyzing and visualizing vast amounts of data.

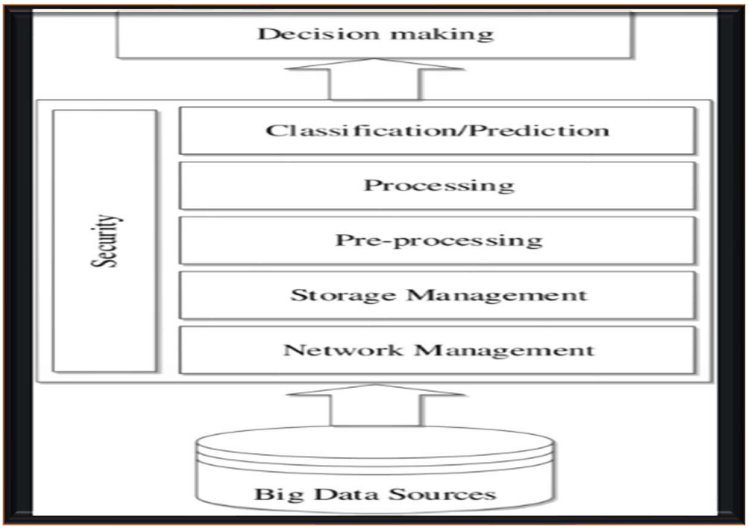

Figure 2 Big Data Management

2. Literature Review

Throughout an item's life cycle, The total of all technical, administrative, and management actions aimed to maintain or repair it so that it can perform the necessary role is maintenance [3]. In order to ensure that maintenance services are in line with the needs and business objectives of both customers and suppliers over the course of the product lifecycle,

eMaintenance is characterized as a multidisciplinary field based on the application of maintenance and information and communication technologies [4]. Additionally, an online resource monitoring and maintenance idea for maintenance management has been put out [5]. Another way to look at it is as a theory to help you move from reactive to proactive maintenance, or from corrective to proactive maintenance. Implement CBM to use e-monitoring to track the health of your system.

In CBM, maintenance procedures are suggested based on forecasts of impending failures, and Based on its condition monitoring system, the equipment's health operational circumstances. The frequency of needless scheduled maintenance visits can be greatly decreased with a properly designed CBM programme, lowering maintenance expenses.PM activities [6][9]. Improved equipment health management, cheaper costs, andthe cost of asset life cycle, the prevention of disastrous failure, etc [7][8].

Fault detection is a crucial part of a CBM system, according to both the OSA-CBM design and the ISO-13374 standard. It can provide instructional data for the procedures that come after, such as fault detection, prognosis, and action recommendations [10]. In industrial systems, subsystems, and components, fault detection aims to locate problematic states and circumstances. Early system fault detection may lessen the risk of unanticipated breakdowns and guarantee industrial systems are reliable and secure [11]. One of the most intriguing uses of Big Data and dependability is fault detection, which is a fundamental part of an Integrated Systems Health Management system [12].

3. Problem Definition

Finding defective conditions and states in industrial systems, subsystems, and components is the aim of fault detection. As was discussed in the section above, fault detection applications employ measurements as inputs to determine the health status of the thing being monitored. Given the growing number of sensors in industrial systems, such as thermometers, vibroscopes, and displacement State measures are often high-dimensional (e.g., flow metres, metres, etc.)[13]. In what are referred to be fast-flowing data streams, these high-dimensional metrics frequently enter firms quickly. The relationships between Observations may be quite nonlinear because nonlinearity is a feature that occurs naturally. Nonlinear modelling is regarded to be one of the toughest challenges. where reliability and big data collide [14].

3.1 Research Objectives

The major goal of this research is to look into, consider, and create strategies for eMaintenance solutions based on Big Data analytics to facilitate maintenance decision-making.

The research goals are more specifically:

- The use of Big Data analytics that may be applied to high-dimensional maintenance datasets, such as for defect detection.
- The big data analytics approach that can be used for high-dimensional maintenance data streams, such as online dynamic fault detection.
- The analytics framework for big data for nonlinear maintenance datasets, such as one that can be applied to defect finding in nonlinear systems.

4. Proposed Methodology

The expansion of knowledge is made possible through research, which adds new knowledge to the corpus [15]. Itis described as a "systematic approach consisting of enumerating the problem, establishing a hypothesis, gathering the facts or data, analysing the findings, and arriving at particular conclusions" in the technical sense. conclusions, whether they be generalisations or remedies to the specific issue at hand pertaining to some theoretical formulation [15].The three main categories of research methods are mixed, quantitative, and qualitative. While qualitative research is based on non-numerical data, quantitative research is based on quantifying quantity or amount. Mixed methods fall somewhere in the middle. additional two, in between. These strategies are fully explained in depth in [16]. Further, There are several types of research, including exploratory research, descriptive research, and research as well as analytical research. The first study to investigate a topic or gain fresh understanding of it is called exploratory research.

It aims to familiarise oneself with the phenomenon and lay the foundation for further investigation. A literature review, focus group interviews, or other techniques may be used in exploratory research, which frequently uses qualitative approaches. Through these techniques, researchers can explore novel occurrences and get a deeper knowledge of them. They can also suggest new routes for future research or make it easier to choose the techniques to apply in a later study[17].

The goal of descriptive research is to accurately describe a phenomenon's features. It may employ a hybrid, qualitative, or quantitative methodology. It frequently involves gathering facts about events, which are then put into order, tallied, depicted, and explained. In descriptive research, observational techniques, surveys, and case studies are widely utilised. The development of a hypothesis might result from descriptive research's ability to generate insightful findings.

Explanatory study, sometimes referred to as causal research, tries to test the theory that different factors have a cause-and-effect link that explains how a certain phenomenon got its characteristics. In explanatory research, quantitative methods are typically used. To reveal the causal connections inside a phenomenon, statistical techniques, particularly hypothesis testing, are used. Explicit study could research approach To draw conclusions about a phenomenon; it may also result in new concepts for additional exploratory research. The degree of ambiguity in the research problem can be used to distinguish between the three different forms of research. Exploratory research typically lacks key variables that have been predetermined, Explanatory research often includes key variables that are well defined, whereas descriptive research typically includes both key variables and important Before to the investigation, relationships were defined. Despite the fact that exploratory, descriptive, and explanatory research The three are not exclusive of one another but are normally done in order. Research investigations evolve throughout time, the study objectives may change over time, making it possible to pursue all three at once concurrently.

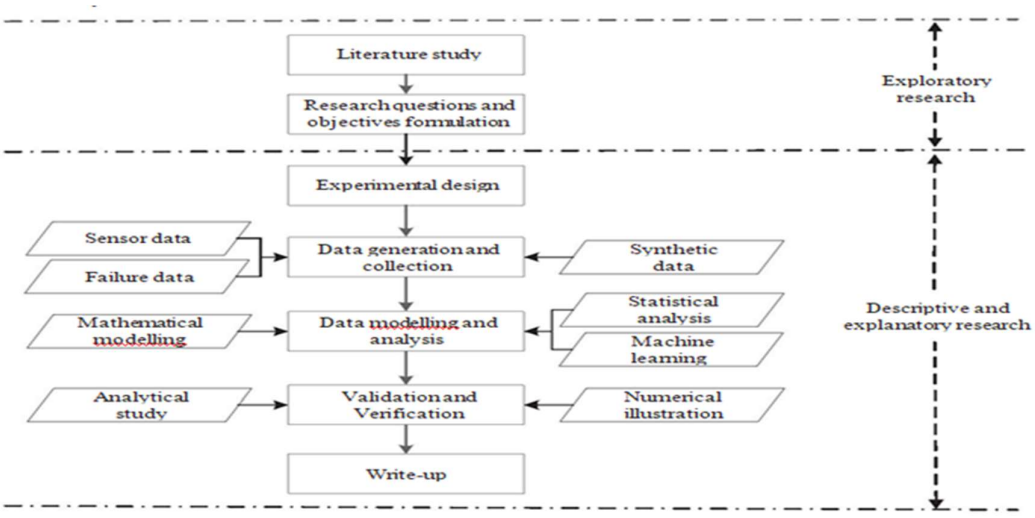

Figure 3 Research design metho

5. Result Analysis

We created an Angle-based Subspace Anomaly Detection method to address this query. The objective was to preserve detection accuracy while detecting anomalies in high-dimensional datasets. The strategy draws two lines for a single anomaly candidate based on the angle between all pairings of pertinent subspaces in full-dimensional space: the first line is linked by the issue region, and the second line is connected by the relevant subspace. second line is The second line is an axis-parallel line, and the first line is the centre of its surrounding points. The "pairwise cosine" *(pCos)* metric is used to compute the angle. The *pCos* is the cosine's average absolute value between projections of the two lines in all two-dimensional spaces that are conceivable. They are all merely two dimensional. The feature spans both one of the additional feature dimensions and the relevant axis dimension. space. The targeted subspace is chosen to consist of the dimensions with the largest *pCos* values. In order to determine the anomaly candidate's local outlierness normalised The Mahalanobis distance unit is used in its subspace projection. Both synthetic data and a dataset from the real world were used to evaluate the proposed technique, and the individual results are displayed below.

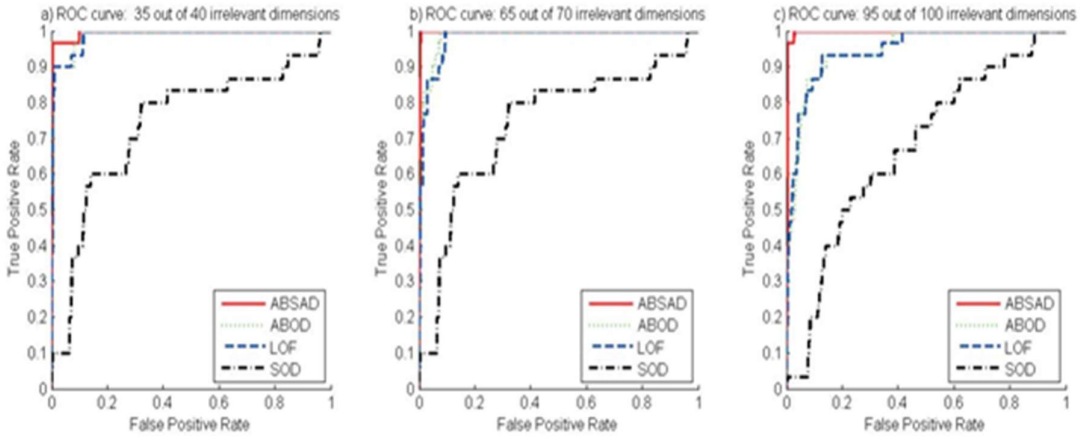

Figure 4 Comparison of ROC curves for various dimensionality settings

The algorithm not only has greater accuracy, but it can also recognise with accuracy the proportions that make anomalies stand out from their surrounding places. the final 30 rows of the matrix, where each entry in the algorithm's output represents the degree to which a sample deviates on a given dimension. The measurements connected to the 30 anomalies are listed in Graphic 4.2; the graphic illustrates Various groups of rows (471 to 480,respectively) and vertical lines serve as separators between various producing processes. There are horizontal lines dividing the numbers (481 to 490 and 491 to 500).A zero entry in the matrix indicates that the relevant dimension for the given data point is not preserved in the subspace since it has a low valueof *pCos*. An item that is not zero not only indicates that the dimension is a however, it also depicts the specific subspace's degree of deviance on this one dimension observation. The retained subspaces perfectly match the dimensions, as seen in Figure 4.2. the location of the faulty data (see Subsection 3.2.1). Additionally, the non-zero ranka foundational grasp of the extent to which each ingredient contributes to abnormality byvarious dimensions that were preserved.

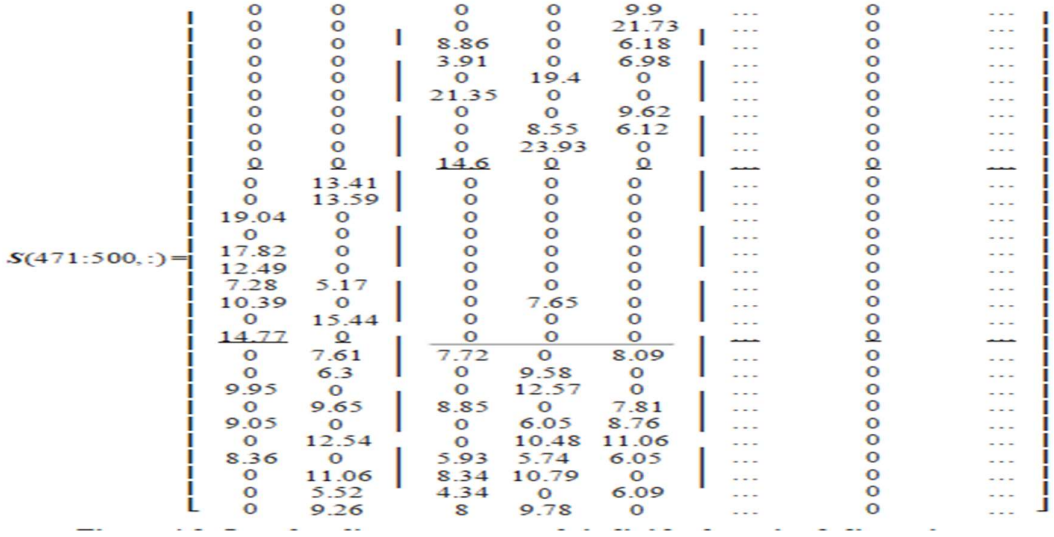

Figure 5: Local outlier scores for each maintained dimension individually

We used the ABSAD formula in a practical fault detection application to validate it. The data,

in particular, came from assessments of a hydro-generator unit's state of health at a hydropower facility in Sweden. We thought about the challenge of identifying errors in the case without losing generality.

6. Conclusion

The research is concluded in this chapter, which also provides a summary of the contributions and research directions. Based on the results of this study, the three research questions (RQs) presented in Chapter 1 have the following answers. The original high-dimensional space can be used to choose relevant subspaces using the proposed ABSAD technique. In other words, it can have dimensions that exhibit a significant difference between a point and its surrounding points. When The analytical analysis shows that the metric "pairwise cosine" is a restricted metric and that it becomes asymptotically stable as dimensionality increases when used to measure vectorial angles in high-dimensional spaces. The experiments on high-dimensional spaces show that the recommended approach can detect anomalies efficiently and has superior accuracy when compared to the suggested alternatives. synthetic datasets with different dimensionality settings.

References

1. X. Zhou, Y. Hu, W. Liang, J. Ma and Q. Jin, "Variational LSTM Enhanced Anomaly Detection for Industrial Big Data," in IEEE Transactions on Industrial Informatics, vol. 17, no. 5, pp. 3469-3477, May 2021, doi: 10.1109/TII.2020.3022432.
2. L. Huang et al., "Hybrid-Order Anomaly Detection on Attributed Networks," in IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2021.3117842.
3. W. Liang, L. Xiao, K. Zhang, M. Tang, D. He and K. -C. Li, "Data Fusion Approach for Collaborative Anomaly Intrusion Detection in Blockchain-based Systems," in IEEE Internet of Things Journal, doi: 10.1109/JIOT.2021.3053842.
4. S. Han et al., "Log-Based Anomaly Detection With Robust Feature Extraction and Online Learning," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 2300-2311, 2021, doi: 10.1109/TIFS.2021.3053371.
5. J. Zhang et al., "Viral Pneumonia Screening on Chest X-Rays Using Confidence-Aware Anomaly Detection," in IEEE Transactions on Medical Imaging, vol. 40, no. 3, pp. 879-890, March 2021, doi: 10.1109/TMI.2020.3040950.
6. P. Rathore, D. Kumar, J. C. Bezdek, S. Rajasegarar and M. Palaniswami, "Visual Structural Assessment and Anomaly Detection for High-Velocity Data Streams," in IEEE Transactions on Cybernetics, vol. 51, no. 12, pp. 5979-5992, Dec. 2021, doi: 10.1109/TCYB.2020.2973137.
7. O. Abdelrahman and P. Keikhosrokiani, "Assembly Line Anomaly Detection and Root Cause Analysis Using Machine Learning," in IEEE Access, vol. 8, pp. 189661-189672, 2020, doi: 10.1109/ACCESS.2020.3029826.
8. A. Alnafessah and G. Casale, "TRACK-Plus: Optimizing Artificial Neural Networks for Hybrid Anomaly Detection in Data Streaming Systems," in IEEE Access, vol. 8, pp. 146613-146626, 2020, doi: 10.1109/ACCESS.2020.3015346.
9. D. Luo, J. Lu and G. Guo, "Road Anomaly Detection Through Deep Learning Approaches," in IEEE Access, vol. 8, pp. 117390-117404, 2020, doi: 10.1109/ACCESS.2020.3004590.

10. Y. Lu et al., "Semi-Supervised Machine Learning Aided Anomaly Detection Method in Cellular Networks," in IEEE Transactions on Vehicular Technology, vol. 69, no. 8, pp. 8459-8467, Aug. 2020, doi: 10.1109/TVT.2020.2995160.
11. A. Libri, A. Bartolini and L. Benini, "pAElla: Edge AI-Based Real-Time Malware Detection in Data Centers," in IEEE Internet of Things Journal, vol. 7, no. 10, pp. 9589-9599, Oct. 2020, doi: 10.1109/JIOT.2020.2986702.
12. T. Sui et al., "A Real-Time Hidden Anomaly Detection of Correlated Data in Wireless Networks," in IEEE Access, vol. 8, pp. 60990-60999, 2020, doi: 10.1109/ACCESS.2020.2984276.
13. M. A. Elsayed and M. Zulkernine, "PredictDeep: Security Analytics as a Service for Anomaly Detection and Prediction," in IEEE Access, vol. 8, pp. 45184-45197, 2020, doi: 10.1109/ACCESS.2020.2977325.
14. R. Zhu et al., "KNN-Based Approximate Outlier Detection Algorithm Over IoT Streaming Data," in IEEE Access, vol. 8, pp. 42749-42759, 2020, doi: 10.1109/ACCESS.2020.2977114.
15. Q. Zhu and L. Sun, "Big Data Driven Anomaly Detection for Cellular Networks," in IEEE Access, vol. 8, pp. 31398-31408, 2020, doi: 10.1109/ACCESS.2020.2973214.
16. W. Yu, H. Bai, J. Chen and X. Yan, "Anomaly Detection of Passenger OD on Nanjing Metro Based on Smart Card Big Data," in IEEE Access, vol. 7, pp. 138624-138636, 2019, doi: 10.1109/ACCESS.2019.2943598.
17. B. Hussain, Q. Du, S. Zhang, A. Imran and M. A. Imran, "Mobile Edge Computing-Based Data-Driven Deep Learning Framework for Anomaly Detection," in IEEE Access, vol. 7, pp. 137656-137667, 2019, doi: 10.1109/ACCESS.2019.2942485.