A Density-Based Clustering Method for Machinery Anomaly Detection

Jing Tian
University of Maryland
jingtian@calce.umd.edu

Michael H. Azarian
University of Maryland
mazarian@calce.umd.edu

Michael Pecht
University of Maryland
pecht@calce.umd.edu

Abstract - Anomaly detection is a critical task in condition-based maintenance of machinery. In many applications clustering-based anomaly detection is preferred due to its ability to analyze data which may not follow a well studied distribution and are unlabeled. This paper introduces a density-based clustering method for machinery anomaly detection. This method assumes that the data from healthy states are located in regions with high densities and the data from faulty states are located in low density regions. By finding the boundaries of these regions, data from the anomalous states can be identified. The values of the densities for healthy machinery and faulty machinery are evaluated. The rate of change of the density from healthy to faulty is identified as a fault threshold. This method can be valuable for applications where faulty data are too difficult or costly to acquire.

Keywords - anomaly detection, clustering, machinery, feature extraction, machine learning

I. INTRODUCTION

Rotating machinery such as bearings and gears are widely used in electromechanical systems like computer cooling fans, wind turbines, and induction motors. Machinery failures have been a research focus due to their frequency and criticality. For example, in induction motors, bearing failures account for more than 40% of the system failures [1]. In wind turbine, where availability is a major concern, gearbox failure is the top contributor for system downtime [2]. Repairing or replacing failed machinery can require costly maintenance, not to mention the cost of the consequent system downtime.

A machinery fault is an abnormal condition that leads to the failure of the machinery, which is a state in which the machinery cannot perform its required function. Commonly observed faults in bearings include pits, indents, and wear. For gears, pits, root crack, wear and missing teeth are typical faults. In most cases, when a fault emerges, the machine can still perform its required function until the fault develops to a certain degree. Usually there is a time gap between the emergence of a fault and the failure. If faults can be detected at an incipient stage and if their development can be monitored, prognostics and health management (PHM) can be performed to reduce the failure frequency and severity as a result of optimized maintenance.

In-situ monitoring is preferable for PHM of machinery because it provides non-intrusive monitoring during the actual life cycle of the machinery. Widely used in-situ monitoring data of machinery include vibration acceleration signals and current signals of the motor that is linked with the machinery. For example, if a fault develops on a computer cooling fan bearing, the vibration profile of the cooling fan will be changed, which can be monitored by an accelerometer. If the fault leads to an increase of friction in the bearing, the current profile would be also changed. By analyzing vibration signals and current signals, the fault can be detected. To analyze the raw signal, fault features are extracted. Commonly extracted features are statistical characteristics of the signal, such as peak-to-peak, rms, and kurtosis [3] of the signal's amplitude in the time domain, characteristic frequency components in the frequency domain, and wavelet coefficients and empirical mode decomposition energy in time-frequency domain. Usually a single feature is not adequate to reflect the machinery health conditions, and multiple features are extracted.

These features reflect different aspects of the health conditions of the machine. They need to be analyzed together to determine whether some data points are anomalous. The assessment is achieved by anomaly detection.

Anomaly detection techniques include classification-based techniques, nearest neighbor-based techniques, statistical techniques, and clustering-based techniques [4]. Classification-based anomaly detection techniques construct classes of healthy and anomalous states from labeled data. An anomaly is detected if a test point is classified as belonging to an anomalous class. Representative methods include support vector machine [5] and hidden Markov model [6]. These techniques require training data from the faulty system, which are often unavailable. Nearest neighbor-based anomaly detection techniques assume that anomalies occur far from the nearest neighbors in the healthy reference data. A representative method is k-nearest neighbor [7]. Outliers in the healthy reference data may lead to false negative errors since they can be regarded as close neighbors by an anomaly. Also, these techniques do not consider the influence of the distribution of the data on anomaly detection. Statistical anomaly detection techniques assume that anomalies occur in the low probability regions of a stochastic model of the healthy data. These techniques rely on the assumption that the data follow certain distributions. However, real data may not follow these distributions. Representative work includes nonparametric statistical analysis [8]. Clustering-based techniques assume that normal data points and faulty data points belong to different clusters. Clustering techniques such as *k*-means algorithm [9] are applied to partition the data into clusters. Clusters for faulty data are identified by evaluating properties of the clusters. In this paper, the research focus is on clustering-based techniques because of the following merits. First, labeled data are not required, and thus these techniques address a practical challenge that healthy data and faulty data are often mixed without labels. Second, they are robust to outliers. Third, some clustering techniques do not require the data to follow particular statistical distributions.

A variety of criteria have been developed to identify the clusters of anomalies. One criterion assumes healthy data are close to healthy clusters, while anomalous data are far from healthy clusters. For example, the *k*-means algorithm partitions the data into clusters according to the mutual distances between the data points. Close data points are grouped into the same cluster. If we know the healthy clusters, other clusters are anomalies. To apply this method, healthy clusters must be known in advance, and the number of clusters should be pre-determined.

It has been observed in experiments and field data that healthy data and faulty data have different densities. This observation is explored in this paper, and an anomaly detection method is developed based on a density-based clustering.

II. THEORETICAL BACKGROUND OF DENSITY-BASED CLUSTERING

In density-based clustering, for each object of a cluster, the neighborhood of a given radius has to contain a minimum number of data points (MinPts), and if this requirement is satisfied, a cluster is initiated [10], [11]. Following this idea, some popular density-based clustering methods have been developed, including density-based spatial clustering of applications with noise (DBSCAN) [12], and generalized density-based spatial clustering of applications with noise (GDBSCAN) [13]. Both methods require the user to input two parameters: MinPts and the radius of the neighborhood. These two algorithms face two challenges. First, there is no guideline to determine the radius of the neighborhood. Second, if clusters have a large difference in densities above a certain value, these methods may fail.

When a healthy machine becomes faulty, the data can exhibit a sharp decrease of density and therefore DBSCAN and GDBSCAN are not suitable to separate the healthy and faulty data. Ordering Points to Identify the Clustering Structure (OPTICS) [10] was developed as a generalization of DBSCAN. It does not need the radius of the neighborhood as an input and it can partition clusters with a large difference in densities.

The OPTICS algorithm works by ordering all data points in a sequence according to two distances, namely core distance and reachability distance. Given a data point *p*, if MinPts are found in its neighborhood within a radius of *ε*, *p* is called a core point. The minimum *ε* that enables *p* to be a core point is called the core distance. The reachability distance of point *q* to *p* is their Euclidean distance or the core distance of *p*. The reachability distance is the larger of the two distances.

The reachability distance can be regarded as a measure of density. A larger reachability distance means a smaller density. Clusters are usually separated by sparse regions, which result in high value of reachability distances, so the peaks of the reachability distance can be used to identify the boundaries of the clusters.

Identification of the clusters with different densities using the reachability distance is illustrated in Fig. 1.

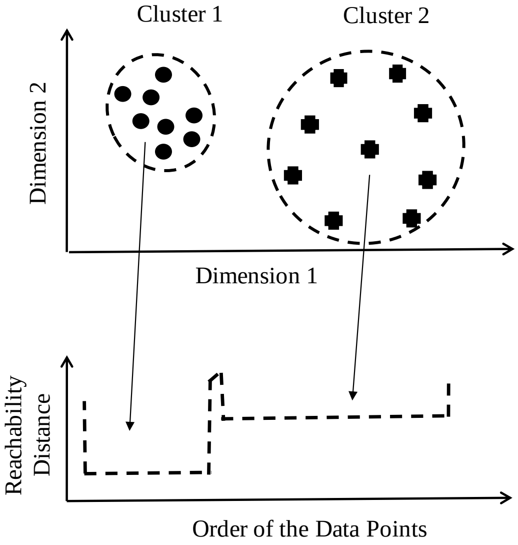

Fig. 1. Identification of clusters using reachability distance

.

III. ANOMALY DETECTION METHOD USING DENSITY-BASED CLUSTERING

When a machine is healthy, it behaves in a consistent pattern where the features extracted from health monitoring signals are distributed with small variances. Therefore, the healthy data form dense clusters. When the machine begins to degrades, its behavior deviates from being normal, and the features extracted from the signals are distributed with shifted mean and larger variances. As a result, the anomalous data form sparse clusters. By estimating the density of the data, anomalies can be detected. However, in the available literature there is no density-based anomaly detection method for the situation where the healthy and anomalous data are mixed. In this research, OPTICS is applied to fill this research gap.

To apply OPTICS in anomaly detection, raw signals should be processed so that the features representing different aspects of the machinery healthy state are extracted and their correlations are reduced. OPTICS is then applied to combine the features to make an aggregated evaluation. This procedure is realized by the anomaly detection method of this paper.

The method consists of six steps. At first, health monitoring signals are selected according to their sensitivity to the fault and their availability for in- situ monitoring. For example, in computer cooling fan bearing monitoring [14], vibration acceleration signals and motor current signals are usually monitored because they are sensitive to bearing faults, and they can be monitored in-situ.

In the second step, statistical features such as rms and kurtosis of the vibration signals that represent different aspects of the health conditions are extracted. The statistical features have different scales, and they are correlated and may be high-dimensional. Therefore they are normalized in the third step using techniques such as Z-score. In the fourth step, dimensionality reduction technique such as principal component analysis (PCA) is applied to generate fault features that have reduced correlation and dimensionality. These three steps are regarded as a feature extraction module that transforms raw signals to a feature space within which clustering can be performed. In the fifth step, the OPTICS algorithm is applied to partition the data into dense and sparse clusters. In the final step, decisions about the health states are made based on the densities of the clusters.

The framework is illustrated in Fig. 2.

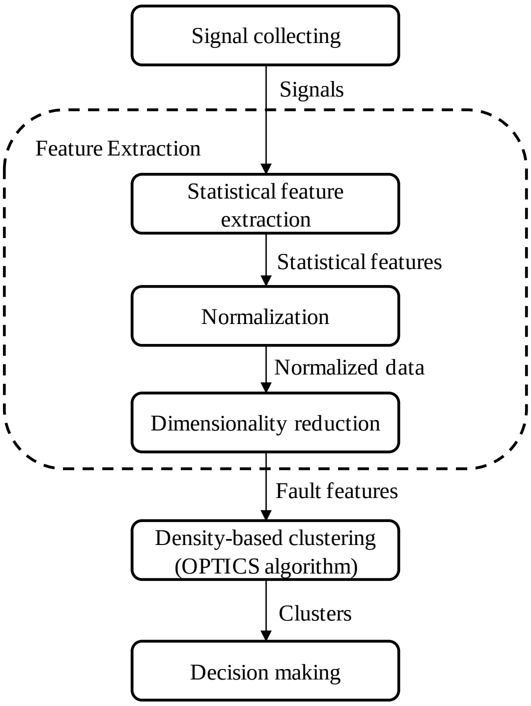

Fig. 2. Framework of the methodology

IV. EXPERIMENTAL STUDY

The density-based clustering anomaly detection method developed in this research was evaluated with the data from a cooling fan accelerated life experiment, which was described in [14, 15].

A. Experimental Setup

A new cooling fan with a ball bearing was tested. Normally, the ball bearing was lubricated by grease and oil. To accelerate the test, the bearing was only lubricated by oil. After an initial measurement, the cooling fan was run at its rated speed of 4,800 rpm in a chamber at the fan’s rated maximum operating temperature of 70℃.

Vibration acceleration signals and motor current signals were collected after the following time intervals: 0 hours, 8 hours, 16 hours, 24 hours, 48 hours, and 72 hours. For each measurement, the cooling fan was run at room temperature of about 20℃, and 10 seconds of signals were collected at a sampling rate of 102,400 Hz for both the vibration signal and the motor current signal. At the end of the test, there were 60 seconds of data consisting of 6144,000 data points. The collected signals formed a 6,144,000 by 2 matrix. Each row is an observation, and each column is a signal.

B. Feature Extraction

At first, observations of the signals were segmented sequentially. Each segment has 20,480 observations, equal to 0.2 seconds of measurement. Altogether there were 300 segments. Vibration features and motor current features were extracted from each segment. Five commonly used time domain statistics were used as features, as listed in Table I.

Table I. STATISTICAL FEATURES

Signals	Vibration	Current
Features	rms	rms
	Kurtosis	Standard deviation
	Peak-to-peak	-

A 300 by 5 matrix of statistical features was extracted, where there were 300 observations for each of the 5 statistical features. The first 10% of the observations (30 observations) were used as reference data to set up a baseline. The mean and standard deviation of the reference data were calculated, and the whole 300 observations were normalized by calculating Z-scores referring to the mean and standard deviation of the reference data.

An analysis of the Pearson's correlation coefficients of the reference data shows that some of the statistical features are highly correlated, and PCA was applied to reduce the correlation. The first three PCs account for 98.8% of the total variance. The remaining two PCs account for 1.2% of the total variance, so discarding them would not result in any significant loss of information. The result is shown in Fig. 3. The reference data were concentrated in a dense region within the circle in Fig. 3.

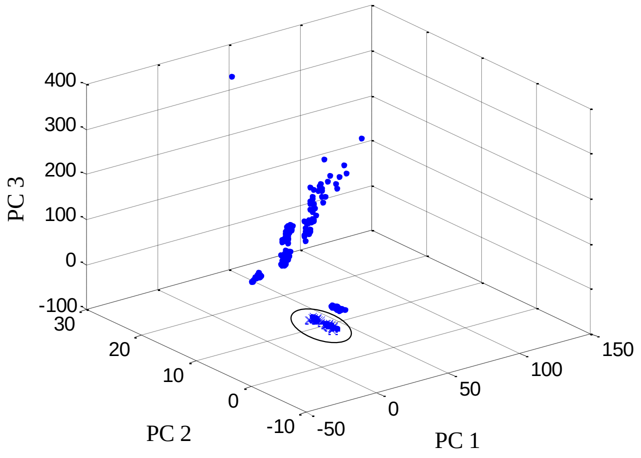

Fig. 3. Scatter of the extracted fault features

C. *Anomaly Detection Using Density-Based Clustering*

The density-based clustering algorithm OPTICS was applied to the extracted 300 by 3 matrix in the feature space.

In the initial measurement at 0 hours, 50 observations were collected in the features space. We assume at least more than half of the observations, such as 30 observations, can be regarded as healthy, and these healthy observations can form a cluster, so at least 10% of all the 300 observations should be able to form a cluster. Therefore, we chose 10% of the data size as MinPts.

After the analysis of OPTICS, the data were rearranged that the points connected by the similar reachability distance were ordered together. The order-reachability distance plot is shown in Fig. 4. The *y* axis is the value of the reachability distance, which is the reciprocal of the density. Each valley is a cluster, and the peaks are boundaries between the clusters.

From Fig. 4, we can identify at least 5 clusters. The first cluster contains the reference data, so this cluster was used to represent the healthy state. The second and third clusters have smaller reachability distances. In other words, they have higher densities, so they are not regarded as faulty. The last two clusters have much larger reachability distances. These are two sparse clusters, and they are likely to be faulty.

Using the reachability distance as a health indicator, an anomaly detection threshold can be defined based on the empirical distribution of the first cluster, which represents the healthy state. Using 99th percentile, the threshold was calculated to be 9.83. After the 151st observation, reachability distances of all the observations are larger than this threshold. Therefore, cluster 4 and cluster 5 in Fig. 4 are the two anomalous clusters. Cluster 4 corresponds to the data collected after 24 hours of test, and cluster 5 corresponds to the data collected after 48 and 72 hours of test.

V. Conclusions

A density-based anomaly detection method was developed in this paper. With appropriate fault feature extraction, the clustering technique is able to partition the data into clusters according to the density of the data, and a density measure named reachability distance is extracted as a health indicator. By examining this health indicator, clusters of anomalous data can be identified. This study evaluated the feasibility of using the cluster density of the health monitoring data to detect machinery anomalies. The density-based anomaly detection method developed in this research provides a novel approach to detect anomalies. Unlike distance-based anomaly detection method, where only the distance between the test data and the healthy reference data is used as a measure of health, the density-based method also analyzes the relationship inside the test data for anomaly detection. Therefore, this method is more robust. The method is unsupervised, so labeled training data are not required. It is suitable for application where faulty data are unavailable. Future work includes optimizing the minimum number of data points for OPTICS in anomaly detection.

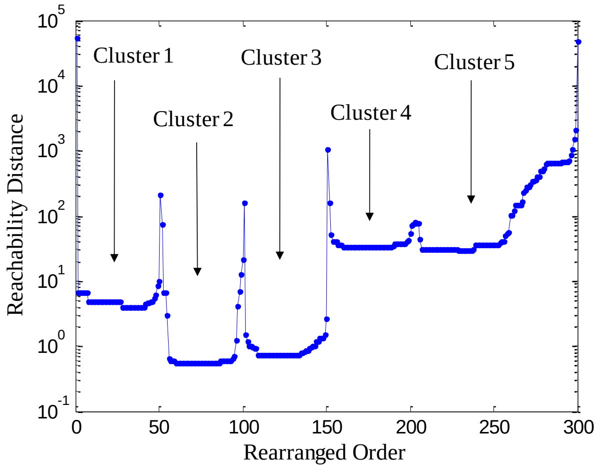

Fig. 4. Clustering Based on Reachability Distances

Acknowledgment

The authors would like to thank the more than 100 companies and organizations that support research activities at the Center for Advanced Life Cycle Engineering (CALCE) at the University of Maryland annually. Also special thanks go to the members of the Prognostics and Health Management Consortium at CALCE for their support of this work.

References

[1] C. Bianchini, F. Immovilli, M. Cocconcelli, R. Rubini, and A. Bellini, “Fault detection of linear bearings in brushless AC Linear motors by vibration analysis,” IEEE Tranactions on Industrial Electronics., vol. 58, no. 5, pp.1684-1694, 2011.

[2] H. Link, W. LaCava, J. van Dam, B. McNiff, S. Sheng, R. Wallen, M. McDade, S. Lambert, S. Butterfield, and F. Oyague,“Gearbox reliability collaborative project report: findings from phase 1 and phase 2 testing”, NREL Report, No. TP-5000-51885, 2011.

[3] D. Siegel, C. Ly, and J. Lee, “Methodology and Framework for predicting helicopter rolling element bearing failure”, IEEE Transactions on Reliability, vol. 61, no. 4, pp. 846-857, 2011.

[4] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey," ACM Computing Surveys (CSUR), vol. 41, pp. 15, 2009.

[5] M. Hejazi and Y. P. Singh, “One-class support vector machines approach to anomaly detection,” Applied Artificial Intelligence, vol. 27, pp. 351-366, 2013.

[6] G. Georgoulas, M. O. Mustafa, I. P. Tsoumas, J. A. Antonino-Daviu, V. Climente-Alarcon, C. D. Stylios, et al., "Principal Component Analysis of the start-up transient and Hidden Markov Modeling for broken rotor bar fault diagnosis in asynchronous machines," Expert Systems with Applications, vol. 40, pp. 7024-7033, 2013.

[7] M. Xie, J. K. Hu, S. Han, and H. H. Chen, "Scalable hypergrid k-NN-based online anomaly detection in wireless sensor networks," IEEE Transactions on Parallel and Distributed Systems, vol. 24, pp. 1661-1670, 2013.

[8] R. F. Luo, M. Misra, S. J. Qin, R. Barton, and D. M. Himmelblau, "Sensor fault detection via multiscale analysis and nonparametric statistical inference," Industrial & Engineering Chemistry Research, vol. 37, pp. 1024-1032, 1998.

[9] J. Zhao, K. Liu, W. Wang, and Y. Liu, "Adaptive fuzzy clustering based anomaly data detection in energy system of steel industry," Information Sciences, vol. 259, pp. 335-345, 2014.

[10] M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander, "OPTICS: ordering points to identify the clustering structure," ACM SIGMOD Record, vol. 28, no. 2, pp. 49-60, 1999.

[11] M. Daszykowski, B. Walczak, D. L. Massart, "Looking for natural patterns in analytical data. Part 2. Tracing local density with OPTICS," Journal of Chemical Information and Computer Sciences, vol. 42 pp. 500-507, 2002.

[12] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," Proceedings of KDD 96, Portland OR, pp. 226-231, August 2-4, 1996.

[13] J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Density-based clustering in spatial databases: The algorithm gdbscan and its applications," Data Mining and Knowledge Discovery, vol. 2 no. 2, pp. 169-194, 1998.

[14] H. Oh, M. Azarian, and M. Pecht, "Estimation of fan bearing degradation using acoustic emission analysis and Mahalonabis distance," MFPT: The Applied Systems Health Management Conference 2011, Virginia Beach, Virginia, May 10-12, 2011.

[15] Q. Miao, M. H. Azarian, and M. Pecht, "Cooling fan bearing fault identification using vibration measurement," IEEE: International Prognostics and Health Management Conference, Denver CO, June 21-23, 2011.

AUTHOR BIOGRAPHY

Jing Tian received the B. Eng degree in Machinery Design and Manufacturing and Automation from the University of Electronic Science and Technology of China. He is a Research Assistant doing Ph.D. research at the Center for Advanced Life Cycle Engineering (CALCE), University of Maryland, College Park. His sponsored research projects include drive train data analysis for condition-based maintenance, prognostics and health management (PHM) algorithm development, and machinery anomaly detection. Prior to joining CALCE in 2010, he had worked as an engineer and researcher for 7 years, where he performed data analysis and design work on several rugged computer products, which have been well accepted by the market. His research focuses on machine learning and its application in PHM.

Michael H. Azarian received the B.S.E. degree in chemical engineering from Princeton University and the M.E. and Ph.D. degrees in materials science and engineering from Carnegie Mellon University.

He is a Research Scientist with the Center for Advanced Life Cycle Engineering (CALCE), University of Maryland, College Park. Prior to joining CALCE he spent over 13 years in industry. His research focuses on the analysis, detection, prediction, and prevention of failures in electronic and electromechanical products. He is the holder of five U.S. patents.

Dr. Azarian is co-chair of the Miscellaneous Techniques subcommittee of the SAE G-19A standards committee on detection of counterfeit parts. He has previously held leadership roles in various IEEE reliability standards committees and co-chaired iNEMI's Technology Working Group on Sensor Technology Roadmapping. He is on the Editorial Advisory Board of Soldering & Surface Mount Technology.

Michael Pecht received the M.S. degree in electrical engineering and the M.S. and Ph.D. degrees in engineering mechanics from the University of Wisconsin, Madison.

He is the Founder of the Center for Advanced Life Cycle Engineering, University of Maryland, College Park, where he is also a George Dieter Chair Professor in mechanical engineering and a Professor in applied mathematics. He has consulted for over 100 major international electronics companies. He has written more than 20 books on electronic-product development, use, and supply chain management and over 400 technical articles.

Dr. Pecht is a Professional Engineer and a fellow of ASME and IMAPS. He is the editor-in-chief of IEEE Access. He was the recipient of the IEEE Reliability Society's Lifetime Achievement Award, the European Micro and Nano-Reliability Award, the 3M Research Award for electronics packaging, and the IMAPS William D. Ashman Memorial Achievement Award for his contributions in electronics reliability analysis.