Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

An angle-based subspace anomaly detection approach to high-dimensional data: With an application to industrial fault detection

Liangwei Zhang*, Jing Lin, Ramin Karim

Division of Operation and Maintenance Engineering, Luleå University of Technology, 97187 Luleå, Sweden

ARTICLE INFO

Article history:
Received 11 February 2015
Received in revised form
17 April 2015
Accepted 30 May 2015
Available online 11 June 2015

Keywords:
Big data analytics
Anomaly detection
High-dimensional data
Fault detection

ABSTRACT

The accuracy of traditional anomaly detection techniques implemented on full-dimensional spaces degrades significantly as dimensionality increases, thereby hampering many real-world applications. This work proposes an approach to selecting meaningful feature subspace and conducting anomaly detection in the corresponding subspace projection. The aim is to maintain the detection accuracy in high-dimensional circumstances. The suggested approach assesses the angle between all pairs of two lines for one specific anomaly candidate: the first line is connected by the relevant data point and the center of its adjacent points; the other line is one of the axis-parallel lines. Those dimensions which have a relatively small angle with the first line are then chosen to constitute the axis-parallel subspace for the candidate. Next, a normalized Mahalanobis distance is introduced to measure the local outlier-ness of an object in the subspace projection. To comprehensively compare the proposed algorithm with several existing anomaly detection techniques, we constructed artificial datasets with various high-dimensional settings and found the algorithm displayed superior accuracy. A further experiment on an industrial dataset demonstrated the applicability of the proposed algorithm in fault detection tasks and highlighted another of its merits, namely, to provide preliminary interpretation of abnormality through feature ordering in relevant subspaces.

1. Introduction

Increasing attention is being devoted to Big Data Analytics and its attempt to extract information, knowledge and wisdom from Big Data. In the literature, the concept of Big Data is mainly characterized by the three “Vs” (Volume, Velocity and Variety) [1] together with “c” to denote “complexity” [2]. High dimensionality, one measure of the volume of data (the other measure being instance size) [3], presents a challenge to Big Data Analytics in industry. For example, high dimensionality has been recognized as the distinguishing feature of modern field reliability data (incl. System Operating/Environmental data, or SOE data), i.e. periodically generated large vectors of dynamic covariate values [4]. Due to the “curse of dimensionality”, it has also been regarded as the primary complexity of multivariate analysis and covariate-response analysis in reliability applications [5,6].

Anomaly detection, also called outlier detection, aims to detect observations which deviate so much from others that they are suspected of being generated by different mechanisms [7]. Efficient detection of such outliers can help, in a timely way, to rectify faulty behavior of a system and, consequently, to avoid losses. In view of this, anomaly detection techniques have been applied to various fields, including industrial fault detection, network intrusion detection and so forth [8–10]. High dimensionality complicates anomaly detection tasks because the degree of data abnormality in relevant dimensions can be obscured or even masked by irrelevant dimensions [5,11,12]. For instance, in an industrial case (see Section 4.2), when detecting the fault “cavitation” in a hydro-turbine, many irrelevant dimensions (e.g. “hydraulic oil level” and “output power”) can easily conceal signals relevant to this anomaly (e.g. “substructure vibration”) and impede the discovery of the fault. Moreover, outliers are very similar to normal objects in high-dimensional spaces from the perspective of both probability and distance [5]. The use of

Abbreviations: ABOD, Angle-Based Outlier Detection; ABSAD, Angle-Based Subspace Anomaly Detection; ANN, Artificial Neuron Network; AUC, Area Under Curve; CMMS, Computerized Maintenance Management System; FCM, Fuzzy C-Means; FPR, False Positive Rate; ICA, Independent Component Analysis; LOF, Local Outlier Factor; MSPC, Multivariate Statistical Process Control; PCA, Principal Component Analysis; ROC, Receiver Operating Characteristic; SNN, Shared Nearest Neighbors; SOD, Subspace Outlier Detection; SOE, System Operating/Environmental (data); SPE, Squared Prediction Error; SVM, Support Vector Machine; TPR, True Positive Rate

* Corresponding author.
E-mail address: liangwei.zhang@ltu.se (L. Zhang).

http://dx.doi.org/10.1016/j.ress.2015.05.025
0951-8320/© 2015 Elsevier Ltd. All rights reserved.

Nomenclature

$\boldsymbol{X}$	design matrix
m	number of data points (rows) in $\boldsymbol{X}$
n	number of dimensions (columns) in $\boldsymbol{X}$
N	the set of feature space $\{1,\ldots,n\}$
LOS	vector of local outlier scores
$\boldsymbol{S}$	matrix consists of the retained subspaces and local outlier score on each retained dimension
i	the ith data point (row) in $\boldsymbol{X}$
j	the jth element of a vector, or the jth dimension (column) of a matrix, or the retained subspace
v	vector representation of a point
p	a data point (outlier candidate)
RP	a set of reference points of a point
q	data point represents the geometric center of all the points in $RP(p)$
l	line connected by two points (e.g. p and q)
NN_k	k nearest neighbor list of a point
Sim_{SNN}	similarity value of two points derived by the SNN method
SNN_s	s nearest neighbor list of a point derived by the SNN method
$PCos\left(\vec{l},\vec{\mu}_n(j)\right)$	average absolute value of cosine between line l and the jth axis in all possible combinations of the two-dimensional spaces (j,j^-), where $j^- \in N\backslash\{j\}$
d	number of retained dimensions of a point
G	threshold for singling out large $PCos$ values

Greek symbols

α	acute angle between line l and x axis
β	acute angle between line l and y axis
γ	angle between a projected line and one of the axes in the retained subspace
σ	a row vector, containing the column-wise standard deviation of the design matrix
ε	a significantly small positive quantity
μ	an axis-parallel unit vector
θ	an input parameter for selecting relevant subspaces
$\boldsymbol{\Sigma}$	covariance matrix of a set of points

Accents

$\bar{\square}$	mean vector of a matrix
$\vec{\square}$	vector representation of a line

Superscripts

$\square^*$	a normalized matrix (e.g. $\boldsymbol{X}^*$)
$\square^T$	transpose of a vector or a matrix
$\square^{-1}$	inverse of a matrix
$\square^\#$	a non-zero scalar quantity obtained by zero-value replacement (e.g. $l_j^\# = 10^{-5}$, if $l_j = 0$)
$\square^-$	one of the remainder dimensions of the original feature space excluding a specific dimension (e.g. $j^- \in N\backslash\{j\}$)
$\square'$	projection of point, set of points or line on the retained subspace (e.g. $RP(p)'$)

The symbol □ denotes a placeholder.

traditional techniques to conduct anomaly detection in full-dimensional spaces is problematic, as anomalies normally appear in a small subset of all the dimensions.

Industrial fault detection aims to identify defective states of a process in complex industrial systems, subsystems and components. Early discovery of system faults may ensure the reliability and safety of industrial systems and reduce the risk of unplanned breakdown [13,14]. Fault detection is a vital component of an Integrated Systems Health Management system; it has been considered as one of the most promising applications wherein reliability meets Big Data [4]. From the data processing point of view, methods of fault detection can be classified into three categories: (i) model-based, online, data-driven methods; (ii) signal-based methods; and (iii) knowledge-based, history data-driven methods [13]. Given the complexity of modern systems, it is too complicated to explicitly represent the real process with models or to define the signal patterns of the system process. Thus, knowledge-based fault detection methods, which intend to acquire underlying knowledge from large amounts of empirical data, are more desirable than other methods [13]. Existing knowledge-based fault detection methods can be further divided into supervised and unsupervised ones, depending on whether the raw data have been labeled or not, i.e. indicating whether the states of the system process in historical data are normal or faulty. Generally, supervised learning methods like Support Vector Machine (SVM), Fuzzy C-Means (FCM), Artificial Neural Network (ANN), and several others can provide reasonably accurate results in detecting or even isolating the hidden faults [9,15]. However, when there is a lack of sufficient labeled data, often the case in reality, fault detection must resort to unsupervised methods. In unsupervised fault detection methods, normal operating conditions are modeled beforehand, and faults are detected as deviations from the normal behavior. A variety of unsupervised learning algorithms have been adopted for this purpose, such as Deep Belief Network, k Nearest Neighbors, and other clustering-based methods [16,17], but few have tackled the challenges of high-dimensional datasets.

Other types of Multivariate Statistical Process Control (MSPC) methods, including Principle Component Analysis (PCA) and Independent Component Analysis (ICA), have also been widely used in fault detection [18,19]. But PCA-based models assume multivariate normality of the in-control data, while ICA-based models assume latent variables are non-Gaussian distributed [20,21]. Both MSPC methods make strong assumptions about the specific data distributions, thereby limiting their performance in real-world applications [22]. Moreover, although PCA and ICA can reduce dimensions and extract information from high-dimensional datasets, their original purpose was not to detect anomalies. Further research has confirmed PCA-based models are not sensitive to faults occurring on the component level [23]. To improve this, several studies have integrated MSPC methods with assumption-free techniques, such as the density-based Local Outlier Factor (LOF) approach [22,24]. Though better accuracy has been reported, LOF still suffers from the "curse of dimensionality", i.e. the accuracy of LOF implemented on full-dimensional spaces degrades as dimensionality increases, as will be shown in Section 4.1.

Although in many industrial applications for fault detection, detecting anomalies from high-dimensional data remains relatively under-explored, several theoretical studies (see Section 2 for a review) have started to probe this issue, including, for example, subspace anomaly detection by random projection or heuristic searches over subspaces. These methods, however, are either arbitrary in selecting subspaces or computationally intensive.