Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Support high-order tensor data description for outlier detection in high-dimensional big sensor data

Xiaowu Deng [a,b,c], Peng Jiang [a,*], Xiaoning Peng [b,c], Chunqiao Mi [b,c]

[a] College of Automation, Hangzhou Dianzi University, Hangzhou 310018, China
[b] School of Computer Science and Engineering, Huaihua University, Huaihua 418000, China
[c] Hunan Provincial Key Laboratory of Ecological Agriculture Intelligent Control Technology, Huaihua 418000, China

HIGHLIGHTS

- We construct a third-order tensor representation model for big sensory data.
- For outlier detection in big sensory data, we propose KSTDD and a tensorial kernel.
- The proposed method can improve the accuracy and efficiency of anomaly detection.

ARTICLE INFO

Article history:
Received 24 April 2017
Received in revised form 22 September 2017
Accepted 8 October 2017
Available online xxxx

Keywords:
Big sensor data
High-dimensional data
Outlier detection
CP factorization
KSTDD

ABSTRACT

The various high-dimensional sensor data can be collected by wireless sensor networks, video monitoring systems and multimedia sensor networks, while High-dimensional sensor data is inherently large-scale because each sensor node has spatial attributes and may also be associated with large amounts of measurement data evolving over time. Detecting outlier in high-dimensional big sensor data is a challenging task. Most of existing outlier detection methods is based on vector representation. However, high-dimensional sensor data is naturally described by tensor representations. The vector-based methods can lead to destroy original structural information and correlation for high-dimensional sensors data, result in the problem of curse of dimensionality, and some outliers cannot be detected. To solve this problem, support high-order tensor data description (STDD) and kernel support high-order tensor data description (KSTDD) are proposed to detect outliers for tensor data. STDD and KSTDD extend support vector data description from vector space to tensor space. KSTDD maintains the structural information of data, avoids the problem caused by the vectorization of tensor data, and improves the performance of outlier detection. Experiments on four sensor datasets show that the proposed method is superior to the traditional vectorized data analysis method.

1. Introduction

With the increasing advances in sensor technology for data collection, and advances in software technology (databases) for data organization, the size of collected data is emerging the explosive growth in various complex forms. The sensor data can be generated by wireless sensor networks, video monitoring systems and multimedia sensor networks. There are more and more all kinds of data being collected from these sources, which is the high-dimensional complex data in a sensor networks such as bio-sensors, video sensors and image sensors. These sensors can detect certain features through analyzing complex high-dimensional (also known as multivariate in statistics) data for multivariate numeric data such as bio-signals, video and images rather than only a univariate value. Let us consider the whole wireless sensor networks or neighborhoods of sensors [1,2] where there are numbers of sensors measuring various distinct properties at different locations. As an example, the traffic sensor network detects various properties including speed and volume of vehicles on a large road system. Under these circumstances, the temporal data collected by such a sensor network could become to be complex and high order. High-dimensional sensor data is inherently large-scale because each sensor node has spatial attributes and may also be associated with large amounts of measurement data evolving over time [3]. That is to say high-order sensor data is considered as big sensor data. In such massive and high-dimensional data detecting outliers can be a challenge because of the large-scale data. As a result, one interesting and rapidly growing area where outlier detection is prevalent is to analyze big sensor data.

[*] Corresponding author.
E-mail address: pjiang@hdu.edu.cn (P. Jiang).

https://doi.org/10.1016/j.future.2017.10.013
0167-739X/© 2017 Elsevier B.V. All rights reserved.

Outliers, according to the famous definition of Hawkins [4], is “an outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”. The process of detecting outliers is called outlier detection. Outlier detection is a popular research topic in data mining. Thus, many scholars have conducted in-depth research in this area. The introduction of various basic data theories and intelligent computing methods has resulted in many efficient and accurate outlier detection algorithms. Outlier detection plays an important role in identifying abnormal patterns and can be applied in different areas, including process control [5], environmental monitoring [6,7], video surveillance systems [8–10], network security [11–13], social networks [14–17], traffic monitoring [18–20], and medical diagnosis [21,22]. According to the classification of detection methods and techniques, outlier detection includes statistics-based [23], depth-based [24], deviation-based [25], distance-based [26], and density-based methods [27] as well as spectral decomposition method [28]. These methods are mostly employed for vector data. However, in many real applications, data are naturally described by high-order tensor representations (i.e., tensor data). For example, in network monitoring, the source host, target host, and port constitute third-order tensors (*source IP* × *destination IP* × *port*) [12]. Adding the time dimension results in a fourth-order tensor (*source IP* × *destination IP* ×*port* × *time*) [11]. In a sensor network, sensors collect various sensing data. The traditional data analysis method presents tensor data in a vectorized form and analyzes them. However, tensors can effectively express the structural information of and correlation among data, so sensing data collected by a sensor network are represented as tensors. Sensing data include the position of the sensor, the measurement value type (such as humidity, temperature, and light intensity), and the sampling time to form third-order tensors (*type* × *position* × *time*). Most outlier detection algorithms cannot deal with tensor data directly and need to present the tensor data in a vector form. Support vector data description (SVDD) and one-class support vector machine are examples of this type of algorithm.

Tensor (multi-dimensional array) is a type of data representation of multivariate relations and multidimensional characteristics in the real world. If tensor data are vectored, the original data structural information or correlation will be destroyed [29]. This destruction can result in the “curse of dimensionality” problem and overfitting when the number of training samples is small [30]. Some outliers cannot be detected [31]. Meanwhile, tensor representations can retain high-order correlations in the modes of data. For example, analysis of original tensor data can achieve good accuracy and can provide an explanation of the results [32]. Transforming high-order tensor data into a vector would destroy data correlation and structural information [33]. In video data of wireless multimedia sensor networks, tensor representations retain the relationship among the three modes corresponding to the horizontal and vertical dimensions of the video frame as well as time. Outlier detection methods based on tensors can detect abnormal video objects of the three modes. However, in outlier detection methods based on vectors, each frame is represented as a vector, and the mode of time is lost. If each frame is analyzed independent of the other frames, minimal noise can cause a frame to be considered an outlier. Therefore, maintaining the tensor data structure and designing an outlier detection algorithm that can be applied to original tensor data rather than vectorized data are necessary. The main challenge in tensorial data learning is constructing a model that preserves the structure of the data.

Anomalies occur occasionally in practical applications. Thus, recording abnormal data samples, such as the fault feature in fault diagnosis or non-health data in the diagnosis of diseases, is difficult. If a classifier for abnormal detection cannot obtain sufficient knowledge to learn, classification accuracy decreases and the detection error increases. However, substantial normal data can be obtained. The use of one-class classification (OCC) in such situations has elicited much attention. SVDD is one of the widely used OCC methods. The basic idea in SVDD is to construct a super sphere that allows all or most of the normal data points to be contained while simultaneously minimizing the volume of the hyper sphere. The data in the hyper sphere are considered normal, whereas the data outside the hyper sphere are regarded as outliers. However, SVDD [34,35] and its mutation algorithm [36–40] are oriented toward vector data and cannot handle tensor data directly. To the best of the authors’ knowledge, SVDD is extended from vector space to tensor space for the first time in the present study. We propose support tensor data description (STDD) and kernel support tensor data description (KSTDD) to deal directly with tensor data and retain structural information.

Table 1
Summary of specific terms used in the paper.

Symbol	Definition
X, Y, A, X_i, X_j	Tensors are represented by uppercase letters
x, y	Scalar
$vec\,(\cdot)$	Column vectorization of a tensor
R	Radius of the hyper sphere
$\circ$	Outer product
$\otimes$	Kronecker product
$\langle \cdot, \cdot \rangle$	Inner product
$\lVert \cdot \rVert_F$	Frobenius norm of a tensor

We propose our novel method for outlier detection in high dimensional big sensor data using KSTDD as well as handling the scalability problem when it comes to manipulating large datasets. By taking advantage of SVDD, tensor representation and tensor factorization, we present a class of algorithms to discover outliers in high-dimensional sensor data. We discuss detailed experimental results of the algorithms proposed. Specifically, in this paper we make the following contributions:

(1) We construct a third-order tensor representation model for high-dimensional sensor data.
(2) For anomaly detection in high-dimensional sensor data, we propose STDD and KSTDD which are the tensor version of SVDD. The CP decomposition is utilized to decompose a multi-channel dataset and map it to the tensor feature space to establish a structure-preserving tensorial kernel function.
(3) We provide extensive experimental results for identify outliers and comparative results. Experiments show that the proposed method can improve the accuracy and efficiency of anomaly detection while ensuring that the internal structure of tensor data is not destroyed.

The rest of this paper is organized as follows. In Section 2, several useful concepts and related knowledge are briefly presented. The proposed STDD is discussed in Section 3. KSTDD is applied to outlier detection in Section 4. The experimental evaluation is presented in Section 5. Section 6 provides the conclusions.

2. Tensor concepts and related knowledge

This section introduces tensor concepts and related knowledge. The mathematical definitions and related operations of tensors are provided, and the meanings of the symbols used in the paper are presented (Table 1). This presentation helps in understanding the current work and tensors.

Definition 1 (*Tensor*). A tensor $X \in \Re^{I_1 \times I_2 \times \ldots \times I_N}$ of order N is an N-way array where elements $x_{i_1 i_2 \ldots i_n}$ are indexed by $i_n \in 1, 2, \ldots, I_n$, $1 \leq n \leq N$. Tensors are evidently generalizations of vectors and matrixes. A zero-order tensor is a scalar, a first-order tensor is a vector, a second-order tensor is a matrix, and tensors of order three and higher are called high-order tensors.

Please cite this article in press as: X. Deng, et al., Support high-order tensor data description for outlier detection in high-dimensional big sensor data, Future Generation Computer Systems (2017), https://doi.org/10.1016/j.future.2017.10.013.