Computers and Electrical Engineering 000 (2017) 1–10

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Concept drift detection for data stream learning based on angle optimized global embedding and principal component analysis in sensor networks

Shenglan Liu[a,b], Lin Feng[b,*], Jun Wu[a,b], Gang Hou[c], Guangjie Han[d]

[a] School of Control Science and Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China
[b] School of Innovation Experiment, Dalian University of Technology,Dalian 116024, China
[c] School of Software, Dalian University of Technology, Dalian 116620, China
[d] Department of Internet of Things Engineering, Hohai University, Changzhou 213022, China

ARTICLE INFO

Article history:
Received 16 April 2016
Revised 5 September 2016
Accepted 5 September 2016
Available online xxx

Keywords:
Industrial Internet of Things (IIoT)
Sensor networks
Data stream
PCA
AOGE

ABSTRACT

As the significant component in Industrial Internet of Things (IIoT), sensor networks have been applied widely in many fields. However, concept drift in data stream produced in sensor networks always brings great difficulty for the robustness of data processing. To solve the problem, we propose a novel concept drift detection method based on angle optimized global embedding (AOGE) and principal component analysis (PCA) for data stream learning in sensors networks. AOGE and PCA analyze the principal components through the projection variance and the projection angle in the subspace, respectively. And then the occurrence of concept drift is determined by observing the change of subspace for each data stream patch. The experiments in synthetic datasets and Intel Lab data demonstrate witness the effectiveness of our method.

1. Introduction

With the rapid development of Industrial Internet of Things (IIoT) in recent years, devices are interconnected with each other to monitor and control effectively and efficiently. And data model and management in the process of manufacture affect the quality of IIoT significantly. As the core components, sensor networks play a significant role in the intelligent monitoring and managements in IIoT. Due to their microminiaturization and intelligence, sensor networks have been widely applied in many practical fields [1], e.g. Internet of Vehicles, mobile phones, health and military etc. In [2], multi-path self-organizing strategy have been comprehensive analyzed and illustrated in IIoT. In monitoring area, Mainwaring et al. [3] proposed a system architecture based on sensor networks for real-world habitat monitoring. Besides these aspects, wireless multimedia sensor networks is surveyed on algorithms, protocols and hardware[4].

Sensor network is an important part in IIoT. Currently there are also many related works proposed concerning improving the performance of sensor networks in data processing. Baker and Ephremides [5] proposed Linked Cluster Algorithm (LCA) to handle the mobility of high population of nodes. In [6], Algorithm for Cluster Establishment (ACE) as an emergent algorithm was presented utilizing the self-organizing properties of three rounds of feedback between nodes. Abbasi and

* Corresponding author.
E-mail address: 824513174@qq.com (L. Feng).

http://dx.doi.org/10.1016/j.compeleceng.2016.09.006
0045-7906/

Younis [7] analyzed and compared different clustering algorithms for wireless sensor networks. In sensor networks, a large quantity of data including different information are collected in the way of timestamped topology. Apparently, these data are dynamic in real-time and overwhelming volume because of the applicability and real-time of sensors. Therefore, it is in fact aiming at dealing with data stream obtained from real-time sensors. In [8], Wang et al. proposed an online approach to address the inhomogeneous sparseness for real-time road traffic monitoring through utilizing real-time data rather than mining historical data. Pereral et al. [9] proposed DAM4GSN architecture to capture sensor data from mobile phones though combining an open source sensor data stream engine with the Android platform.

Besides, as the domains of data mining, many great works referred to data stream has been presented in the past years [10–15]. Comparing with traditional static data, data stream has some special properties: the data in data stream is dynamic and the labels and distribution of data may change in trends over time known as concept drift. OUYANG et al. [16] presented the state-of-the-art algorithms in the field of data stream and then gave a critical judgment on these existing methods. Meanwhile, they point out that the concept drift detection methods should compare the similarities and differences directly between concepts in different time, while most of existing drift detection methods take a "circuitous" route with analyzing the reasons of concept drift and predicting the possible results of concept drift. In [17], Kifer et al. introduced a novel method for detection and estimation of changes in data stream based on the assumption that the points in the stream are generated independently. Lazarescu and Venkatesh [18] used selective memory to track concept drifting, and thus detect drifting more accurately and filter the data noise more effectively.

Detecting concept drift in high dimensional data streams also plays an important part in learning data stream of sensor networks. Traditional machine learning methods are not appropriate to tackle with these data due to the curse of dimensionality. One of the classical detecting concept drifts methods is that an information-theoretic approach [19] is introduced through measuring the differences between two given distributions with Kullback-Leibler distance. However, there are some limitations utilizing KL distance: 1) it needs discretization to calculate probability density; 2) it can only handle concept drift between two classes and multiclass can only be decided based on results of two classes; 3) the process of bootstrap and discretization is time consuming. In [20], Dries and Rückert presented three novel test methods based on statistics for draft detection and evaluated the performance of several different methods applied for concept drift detection. Manifold learning theory has been applied successfully for high dimensional data stream [21,22]. These methods mainly focus on evaluating data manifold so that reveal interesting properties of data stream. Effectiveness of detection concept drifting should be considered by manifold learning approach.

In this paper, a novel data stream learning framework based on principal component analysis (PCA) and angle optimized global embedding (AOGE) for concept drift detection in sensor networks. The superior of this concept drift detection method based on subspace learning depends mainly on the performance of PCA and AOGE. As one of the classical subspace learning methods, PCA has achieved great success in many fields. Through analyzing the projection variance of sampled data, PCA describes the dispersion among the data and chooses the principal components based on the maximized projection variance. Different from PCA, AOGE analyzes the projection angle of sampled data for choosing principal components. And thus AOGE has more robust performance than PCA in real-world datasets with lots of noisy data. So both PCA and AOGE can detect the concept drift in data stream, and the combination between them can further improve the effectiveness of concept drift detection. After that, the subspace learning for data stream patches based on AOGE can improve the efficiency of data processing by reducing the dimension. Finally, the data classifications based on Extreme Learning Machine (ELM) and Support Vector Machine (SVM) validate the performance of our method in three synthetic datasets and Intel Lab data.

The rest paper is organized as follow: PCA and AOGE are introduced in Sections 2 and 3 respectively; the data stream learning framework is presented in Section 4; Section 5 gives the experimental results and analysis; finally, we draw the conclusion of this paper in Section 6.

2. Principal component analysis (PCA)

As a classical subspace learning method, PCA has demonstrated the superior performance in many practical fields. With the basic assumption that the principal components of high-dimensional data are distributed in low-dimensional orthogonal subspace, PCA analyzes the global distributed structure of these data based on the principle that the embedded data in this subspace have the largest variance.

Supposed that a set of data in data stream is $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{D\times n}$, and the corresponding low-dimensional project is denoted as $Y = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^{d\times n}$, the aim of subspace learning is to find a set of orthogonal basis $U \in \mathbb{R}^{D\times n}$ where $Y = U^T X$. Through $\hat{X} = X\left(I - \frac{1}{n}11^T\right)$ where$1 = [1, \cdots, 1]^T_{1\times n}$ and I is the identity matrix, we first centralize the sample data X. Based on PCA, the optimal problem can be expressed as follow:

$$\begin{aligned} \max \quad & E_{PCA} = \sum_{i=1}^{n} \left\| U^T (x_i - \bar{x}) \right\|^2 = tr\left(U^T \hat{X}\hat{X}^T U\right) \\ s.t. \quad & U^T U = I \end{aligned} \tag{1}$$

Where $\bar{x}$ corresponds to the center of all the data. The solution in Eq. (1) is equal to solve the following problem:

$$\hat{X}\hat{X}^T U = \lambda U \tag{2}$$

Please cite this article as: S. Liu et al., Concept drift detection for data stream learning based on angle optimized global embedding and principal component analysis in sensor networks, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2016.09.006