Evolving Systems (2016) 7:15–27
DOI 10.1007/s12530-015-9131-7

ORIGINAL PAPER

Online feature extraction based on accelerated kernel principal component analysis for data stream

Annie Anak Joseph[1,2] · **Takaomi Tokumoto**[1] · **Seiichi Ozawa**[1]

Received: 1 March 2014 / Accepted: 21 February 2015 / Published online: 22 March 2015

Abstract Kernel principal component analysis (KPCA) is known as a nonlinear feature extraction method. Takeuchi et al. have proposed an incremental type of KPCA (IKPCA) that can update an eigen-space incrementally for a sequence of data. However, in IKPCA, the eigenvalue decomposition should be carried out for every single data, even though a chunk of data is given at one time. To reduce the computational costs in learning chunk data, this paper proposes an extended IKPCA called Chunk IKPCA (CIKPCA) where a chunk of multiple data is learned with single eigenvalue decomposition. For a large data chunk, to reduce further computation time and memory usage, it is first divided into several smaller chunks, and only useful data are selected based on the accumulation ratio. In the proposed CIKPCA, a small set of independent data are first selected from a reduced set of data so that eigenvectors in a high-dimensional feature space can be represented as a linear combination of such independent data. Then, the eigenvectors are incrementally updated by keeping only an eigenspace model that consists of the sextuplet such as independent data, coefficients, eigenvalues, and mean information. The proposed CIKPCA can augment an eigen-feature space based on the accumulation ratio that can also be updated without keeping all the past data, and the eigen-feature space is rotated by solving an eigenvalue problem once for each data chunk. The experiment results show that the learning time of the proposed CIKPCA is greatly reduced as compared with KPCA and IKPCA without sacrificing recognition accuracy.

Keywords Online learning · Incremental learning · Feature extraction · Kernel principal component analysis

✉ Annie Anak Joseph
jannie@feng.unimas.my

Takaomi Tokumoto
takaomi.tokumoto@gmail.com

Seiichi Ozawa
ozawasei@kobe-u.ac.jp

[1] Graduate School of Engineering, Kobe University, Rokko-dai, Nada, Kobe 657-8501, Japan

[2] Faculty of Engineering, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

1 Introduction

With the fast development of the internet and computer technologies, data are continuously generated from real-world applications (e.g., stock price prediction) and such data are known as *data streams*. Examples of data streams include video images, computer network traffic, and so forth. In this paper, the term "data stream learning" is used under a situation where data are continuously generated over time and should be learned on an on-going basis without keeping past data (i.e., *online* or *incremental learning*). Applications of incremental learning are rapidly increasing and the effectiveness has been proved in various practical problems (Domingos and Hulten 2001; Babcock et al. 2002; Case et al. 1999). Actually, various kinds of online learning algorithms have been developed so far (Joseph et al. 2012; Aoki et al. 2013; Jang et al. 2011; Minku et al. 2009; Elwell and Polikar 2011; Zhao et al. 2006).

To date, incremental learning has given at least two contributions. The first contribution is that it can solve an issue where all data are unavailable at the beginning (Weng et al. 2000). In real situations, it is likely that sufficient training data have not been provided at one time and can only be provided consecutively. For example, in face recognition problems (Yang 2002), human faces could be changing over time due to various conditions such as aging, health