Information Sciences 423 (2018) 80–95

Contents lists available at ScienceDirect

Information Sciences

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

Self-Organised direction aware data partitioning algorithm

Xiaowei Gu[a], Plamen Angelov[a,b,*], Dmitry Kangin[a], Jose Principe[c]

[a] *School of Computing and Communications, Lancaster University, Lancaster, LA1 4WA, UK*
[b] *Honorary Professor, Technical University, Sofia 1000, Bulgaria*
[c] *Computational NeuroEngineering Laboratory, Department of Electrical and Computer Engineering, University of Florida, USA*

ARTICLE INFO

Article history:
Received 17 November 2016
Revised 22 August 2017
Accepted 9 September 2017
Available online 19 September 2017

Keywords:
Autonomous learning
Nonparametric
Clustering
Empirical Data Analytics (EDA)
Cosine similarity
Traditional distance metric

ABSTRACT

In this paper, a novel fully data-driven algorithm, named *Self-Organised Direction Aware* (*SODA*) data partitioning and forming data clouds is proposed. The proposed *SODA* algorithm employs an extra cosine similarity-based directional component to work together with a traditional distance metric, thus, takes the advantages of both the spatial and angular divergences. Using the nonparametric Empirical Data Analytics (EDA) operators, the proposed algorithm automatically identifies the main modes of the data pattern from the empirically observed data samples and uses them as focal points to form data clouds. A streaming data processing extension of the *SODA* algorithm is also proposed. This extension of the *SODA* algorithm is able to self-adjust the data clouds structure and parameters to follow the possibly changing data patterns and processes. Numerical examples provided as a proof of the concept illustrate the proposed algorithm as an autonomous algorithm and demonstrate its high clustering performance and computational efficiency.

1. Introduction

Tremendous increase in the volume and complexity of the data (streams) combined with rapid development of computing hardware capabilities requires a fundamental change of the existing data processing methods. Developing advanced data processing methods that have elements of autonomy and deal with streaming data is now becoming increasingly important for industry and data scientists alike [4,11].

Data partitioning and clustering techniques have been widely used in different areas of the economy and society [3,16,35]. However, despite being considered to be an unsupervised form of machine learning, traditional clustering techniques require *prior* knowledge and handcrafting to operate. Users need to define a number of parameters and make assumptions in advance, i.e. bandwidth [16], number of clusters [18,25,40], radii [15,21,28,41], grid size [31], type of the distance metric [18,26,40], kernel type [10,16,42], etc. Moreover, the parameters and thresholds that are required to be pre-defined are often problem- and sometimes user-specific, which inevitably leads to the subjective results; this is usually ignored and neglected portraying clustering and related data partitioning techniques as unsupervised.

Generally, clustering algorithms may use miscellaneous distances to measure the separation between data samples. However, the well-known Euclidean and the Mahalanobis [27,33] distance metrics are the most frequently used ones. In some fields of study such as natural language processing (NLP), for example, the derivatives of the cosine (dis)similarity,

[*] Corresponding author.
E-mail addresses: x.gu3@lancaster.ac.uk (X. Gu), p.angelov@lancaster.ac.uk (P. Angelov), dkangin@gmail.com (D. Kangin), principe@cnel.ufl.edu (J. Principe).

http://dx.doi.org/10.1016/j.ins.2017.09.025
0020-0255/© 2017 Elsevier Inc. All rights reserved.

which is a pseudo metric, are also used in the machine learning algorithms for clustering purpose [3,38,39]. Nevertheless, once a decision is made, only one type of distance/dissimilarity can be employed by the clustering algorithms.

Empirical Data Analytics (EDA) [5–7] is a recently introduced nonparametric, assumption free, fully data-driven methodological framework. Unlike the traditional probability theory or statistic learning approaches [9], EDA is conducted entirely based on the empirical observation of the data alone without the need of any *prior* assumptions and parameters. It has to be stressed that the concept of "nonparametric" means our algorithms is free from user- or problem- specific parameters and presumed models imposed for the data generation, but this does not mean that our algorithms do not have meta-parameters to achieve data processing.

In this paper, we introduce a new autonomous algorithm named Self-Organised Direction Aware (*SODA*) data partitioning. In contrast to clustering, a data partitioning algorithm firstly identifies the data distribution peaks/modes and uses them as focal points [7] to associate other points with them to form data clouds [8] that resembles Voronoi tessellation [34]. Data clouds [8] can be generalized as a special type of clusters but with many distinctive differences. They are nonparametric and their shape is not pre-defined and pre-determined by the type of the distance metric used (e.g. in traditional clustering the shape of clusters derived using the Euclidean distance is always hyper-spherical; clusters formed using Mahalanobis distance are always hyper-ellipsoidal, etc.). Data clouds directly represent the local ensemble properties of the observed data samples.

The *SODA* partitioning algorithm employs both a traditional distance metric and a cosine similarity based angular component. The widely used traditional distance metrics, including Euclidean, Mahalanobis, Minkowski distances, mainly measure the magnitude difference between vectors. The cosine similarity, instead, focuses on the directional similarity. The proposed algorithm that takes into consideration both the spatial and the angular divergences results in a deeper understanding of the ensemble properties of the data.

Using EDA operators [6,7] the *SODA* algorithm autonomously identifies the focal points (local peaks of the typicality, thus, the most representative points locally) from the observed data based on both, the spatial and angular divergences and, based on them, discloses the ensemble properties and mutual distribution of the data. The possibility to calculate the EDA quantities incrementally enables us to propose computationally efficient algorithms.

Furthermore, a version of the *SODA* algorithm for streaming data is also proposed, which is capable of continuously processing data streams based on the offline processing of an initial dataset. This version enables the *SODA* algorithm to follow the changing data pattern in an agile manner once primed/initialised with a seed dataset. The numerical examples in this paper demonstrate that the proposed autonomous algorithm constantly outperforms the state-of-the-art methods by producing high quality clustering results and has high computational efficiency.

The remainder of this paper is organised as follows. Section 2 introduces the theoretical basis of the proposed methodology and approach. Section 3 presents the main procedure of the proposed *SODA* partitioning algorithm. The streaming data processing extension is described in Section 4. Numerical examples and performance evaluations are given in Section 5. This paper is concluded by Section 6.

2. Theoretical basis

Firstly, let us consider the real data space $\mathbf{R}^m$ and assume a data set/stream as $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\cdots\}$, where $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}]^T \in \mathbf{R}^m$ is a m dimensional vector, $i = 1, 2, 3, \cdots$; m is the dimensionality; subscript i $(i = 1, 2, 3, \cdots)$ indicate the time instances at which the i^{th} data sample arrives. In real situations, data samples observed at different time instances may not be exactly the same, however, with a given granularity of measurement, one can always expect that the values of some data samples repeat more than ones. Therefore, within the observed data set/stream at the n^{th} time instance denoted by $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, we also consider the set of sorted unique values of data samples $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{n_u}\}$ $(\boldsymbol{u}_i = [u_{i,1}, u_{i,2}, \ldots, u_{i,m}]^T \in \mathbf{R}^m)$ from $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and the corresponding normalised numbers of repeats $\{f_1, f_2, \ldots, f_{n_u}\}$, where n_u $(1 < n_u \le n)$ is the number of unique data samples and $\sum_{i=1}^{n_u} f_i = 1$. The following derivations are conducted at the n^{th} time instance as a default unless there is a specific declaration.

2.1. Distance/Dissimilarity components in SODA

As it was described in Section 1, the *SODA* approach employs:

i) a magnitude component based on a traditional distance metric;
ii) a directional/angular component based on the cosine similarity;

and, thus, it is able to take advantage of the information extracted within a metric space and within a pseudo-metric, similarity oriented one, namely, the spatial and directional divergences.

The magnitude component can be, but is not limited to, the well-known Euclidean or Mahalanobis distances as well as other known full metric types of distances. For the clarity of the derivation, the most widely used Euclidean distance metric will be used in this paper as the magnitude component, and thus, the magnitude component is expressed as:

$$d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\| = \sqrt{\sum_{l=1}^{m} \left(x_{i,l} - x_{j,l}\right)^2};\ i, j = 1, 2, \ldots, n \tag{1}$$