Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 45 (2017) 95–108

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Visual cognitive algorithms for high-dimensional data and super-intelligence challenges

Boris Kovalerchuk

Dept. of Computer Science, Central Washington University, USA

Received 4 February 2017; accepted 28 May 2017
Available online 6 June 2017

Abstract

In the long run the cognitive algorithms intend to make super-intelligent machines and super-intelligent humans. This paper presents a technical process to reach specific aspects of super-intelligence that are out of the current human cognitive abilities. These aspects are inabilities to discover patterns in large numeric multidimensional data with a naked eye. This is a long-standing problem in Data Science and Modeling in general. The major obstacle is in human inability to see n-D data by a naked eye and our needs in visualization means to represent n-D data in 2-D losslessly. While these means exist their number and abilities are limited. This paper expands the class of such lossless visual methods, by further developing a new concept of Generalized Shifted Paired Coordinates. It shows the advantages of proposed reversible lossless technique by representing real data and by proving mathematical properties.

Keywords: Cognitive algorithms; High-dimensional data; Visualization; Machine learning, generalized coordinates, super-intelligence

1. Introduction

The concept of *human-machine super-intelligence* is present in the literature for a long time (Hibbard, 2002). It includes prospects of both *super-intelligent machines* and *super-intelligent humans* that will far surpass the current human intelligence significantly lifting the human cognitive limitations.

The expected ways to achieve it range from progress in: (1) Artificial Intelligence (AI) and Computational Intelligence (CI), (2) new human abilities to evolve or directly modify their biology (Superintelligence), and (3) power of crowd interaction (Michelucci & Dickinson, 2015). A significant portion of publications in this area is the futuristic predictions of when super-intelligence can be achieved, and what the potential danger of expected achievements is. This paper is devoted to the different aspect, namely, a technical way to reach some specific aspects of super-intelligence that are beyond the current human cognitive abilities. It is to overcome inabilities to analyze a large amount of abstract numeric *high-dimensional* data and finding complex *patterns* in these data with a *naked eye*.

Human inability to discover patterns in n-D data using a naked eye is one of the major motivations for the emergence of visual analytics research area that is devoted to developing 2-D visual representations (visualizations) of n-D data. While multiple such representations have been developed, many of them are lossy, i.e., do not represent n-D data completely and do not allow restoring n-D data completely from their 2-D representation. Respectively our abilities to discover n-D data patterns from such incomplete 2-D representations are limited and potentially erroneous.

In contrast lossless visualizations of n-D data have no such limitations and have advantages as *cognitive enhancers*

E-mail address: borisk@cwu.edu

http://dx.doi.org/10.1016/j.cogsys.2017.05.007
1389-0417/

of the human cognitive abilities to discover n-D data patterns. Below we review the state of the art in this area, and outline the challenges that this paper addresses. Discovering patterns in big multidimensional data using visual means is a long-standing problem in Information Visualization, Visual Analytics, Visual Data Mining, and Data Science in general (Bertini, Tatu, & Keim, 2011; Grishin & Kovalerchuk, 2014; Hoffman & Grinstein, 2002; Inselberg, 2009; Kovalerchuk, 2014; Kovalerchuk & Grishin, 2014; Simov, Bohlen, & Mazeika, 2008; Tergan & Keller, 2005; Ward, Grinstein, & Keim, 2010; Wong & Bergeron, 1997). As we already outlined the major challenge is our *cognitive limitations*. We cannot see n-D data by a naked eye and need enhanced visualization tools ("n-D glasses") to represent n-D data in 2-D losslessly.

The number of available tools to overcome this cognitive limitation is quite limited. Principal Component Analysis (PCA) is a lossy n-D data representation when we use the first two main principal components to show n-D data in 2-D. Multidimensional scaling is also a lossy representation due to approximation of n-D distances. Simple tools such as heat maps, pie-and bar-graphs are applicable to relatively small datasets and dimensions. Parallel Coordinates (PC) and Radial (star) Coordinates (RC) today are the most known lossless n-D data visualization methods for relatively large data while suffering from occlusion.

There is a need to *extend* the class of lossless n-D data visual representations. A new class of such representations called the **General Line Coordinates (GLC)** and several of their specifications have been proposed in Grishin and Kovalerchuk (2014), Kovalerchuk (2014), and Kovalerchuk and Grishin (2014). These visualizations include Collocated Paired Coordinates (CPC) in orthogonal and radial forms. The benefits of these new visual representations and their advantages have been shown in Grishin and Kovalerchuk (2014), Kovalerchuk (2014), and Kovalerchuk and Grishin (2014) for analyzing data of the Challenger disaster, World Hunger, Semantic shift in humorous texts, and others.

This paper: (1) expands these new methods, (2) explores their *mathematical properties*, and (3) demonstrates advantages of these methods for *real-world data*. In exploration of the mathematical properties, we analyze how the methods represent known n-D data structures in 2-D. The importance to explore the mathematical properties of new methods in addition to comparing them with known methods on real-world data is in the ability to derive general properties that are common to all data of a given structure.

This paper is organized as follows. Section 2 presents the concept of lossless visualization of n-D data as cognitive enhancer for discovering n-D data patterns. Section 3 provides definitions of line coordinates. Section 4 provides algorithms and mathematical statements that demonstrate how n-D data representations in various general line coordinates simplify representation of n-D data in 2-D for better perceptual and cognitive abilities for visual pattern discovery. Section 5 shows advantages of different GLCs on real-world data. Section 6 demonstrates the technique to simplify the visual patterns in GLC. Section 7 relates super-intelligence issues to high-dimensional data. The paper is concluded with its summary and discussion of future studies.

2. Lossless visualization of n-D data as cognitive enhancer for discovering patterns

This paper described the proposed new algorithms for lossless visual representation of high-dimensional data and their connections with human super-intelligence challenges. We interpret these algorithms as cognitive algorithms that enhance human cognitive abilities to deal with modern Big data high-dimensional challenges. The paper focused on Generalized Shifted Paired Coordinates as a subset of General Line Coordinates. The advantages of these coordinates have been shown both mathematically and on the data. These advantages guide future studies to solve a major challenge. This challenge is finding conditions for a provable property of simpler and less overlapped lossless 2-D representation of the non-intersecting hyper-ellipses, hyper-rectangles, and other shapes in n-D.

The advantage of a wide class of General Line Coordinates is that it allows multiple different visualizations of the same data with the different perceptual and cognitive characteristics. This multiplicity increases the chances that humans will be able to reveal the hidden n-D patterns in these visualizations. It is not realistic to expect that a single visualization will do this for all possible data and all humans.

A full classification of the general line coordinates for the cognitively efficient n-D data visualization is a task for future research as well as the deeper links with Machine Learning to be able to build visually the learning algorithms using visual means in GLC such as Decision Trees. This is an area of future studies for the design of more complete processes and for expanding to other data mining/-machine learning methods.

Example. Assume that we have established that new data have the same mathematical structure that was explored before. Then we can use the derived matched structural properties. Consider n-D data with a mathematical structure where all n-D points of class C_1 are in the one hypercube and all n-D points of class C_2 are in another hypercube and the distance between these hypercubes is greater or equal to k lengths of these hypercubes.

Assume that it was established mathematically that for any n-D data with this structure a lossless visualization method V_1, produces visualizations of n-D vectors of classes C_1 and C_2 that do *not overlap* in 2-D. Next assume that this property was tested on new n-D data and was confirmed. In this case we can apply visualization method V_1 with confidence that it will produce desirable visualization without occlusion of two classes. Similarly if the structural property is negative to ability to visualize the pattern with-