Pattern Recognition Letters 106 (2018) 7–12

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Blind spots in Star Coordinate Visualization: Analysis and correction

Swee Chuan Tan*, Jeksen Tan

Singapore University of Social Sciences, Singapore

ARTICLE INFO

Article history:
Received 24 April 2017
Available online 13 February 2018

Keywords:
Star Coordinate Plot
Data visualizations
Data analytics

ABSTRACT

Star Coordinate Plot is a simple and efficient technique for visualizing multidimensional data. Since the proposal of this method in early 2000, several researchers have attempted to address its weakness of tending to project data points toward the origin of the star coordinate space. But so far no one has provided a critical analysis of the issue in the literature. As a result, the weakness of Star Coordinate Plot is still not well understood. In this paper, we first provide an explanation of its weakness by pointing out two design constraints in the original Star Coordinate Plot. We show how these constraints result in three categories of data points that are lost in the process of translating from n-dimensional space to a two-dimensional star coordinate space. We then propose the Enhanced Star Coordinate data visualization method to address these constraints. Our experimental results show that the proposed method is superior to the original Star Coordinate Plot on several datasets used for evaluations.

1. Introduction

Visual analytics is an effective and intuitive approach for uncovering interesting and useful patterns in data [1,2]. A key component of visual analytics is data visualization [3], which helps analysts gain insights into complex multi-dimensional data [4,5]. When it comes to understanding and analyzing real-world data, visualization can be used to complement the more sophisticated data mining methods, such as support vector machines [6] and neural networks [7].

This paper examines the Star Coordinate (SC) Plot [8], a data visualization technique that represents an n-dimensional data space using a two-dimensional radial plot. Such a plot has n axes, namely A_1, A_2, ..., A_n. Let there be an imaginary horizontal reference line R with an arbitrary point as the origin. All the axes here are radial lines starting from the origin. An axis, A_i, is to incline at an angle of $2(i-1)\pi/n$ with R. This arrangement makes the axes of an SC Plot to be equally spaced by an angle of $2\pi/n$. Each axis has a range of [0, 1], with value zero at the origin, and value one at the other end of the axis.

SC Plot has two strengths. Firstly, it is a simple method that can easily be implemented and extended for more complex data visualization tasks. One example of extension is the three-dimensional SC Plot [9]. Secondly, SC Plot requires only constant time to render each data point, giving it a nice property of linear time complexity with respect to data size. This makes it a good candidate for real-time, data-intensive visualization applications.

Fig. 1 shows how a four-dimensional data point V is mapped to a point V^* in a two-dimensional SC Space, and this is achieved by adding up the four normalized components of V (which are shown as green arrows) within the SC Space.

However, we will show that there are conditions under which non-zero data vectors can be mapped to the *origin* of SC space. This problem results in a loss of visual information when displaying data pattern on the visualization output.

In a typical Cartesian Coordinate system, every axis is a number line that shares the same origin. For each number line, one side of its origin represents positive numbers, and the other side represents negative numbers. However, in SC Plot, every dimension is represented by a single axis scaled to a range of [0, 1] using min–max normalization. In other words, every component of a vector, including those with negative values, will be normalized to a range of [0, 1]. This means that additions of normalized vector components within the SC space can become less meaningful. This is the first design constraint of SC Plot.

The second design constraint is concerned with the placement of two (statistically) independent dimensions in exact opposite directions (e.g., refer to dimensions A_2 and A_4 in Fig. 1). Such a placement tacitly assumes that one dimension (e.g., A_2) is the exact opposite of the other (e.g., A_4), which may not necessarily be the case. If A_2 and A_4 are two independent attributes, then they must not offset one another when summating the vector components.

To address the above constraints, we first introduce a sign-preserving data normalization technique that scales the data to a

* Corresponding author.
E-mail address: jamestansc@suss.edu.sg (S.C. Tan).

https://doi.org/10.1016/j.patrec.2018.02.005
0167-8655/

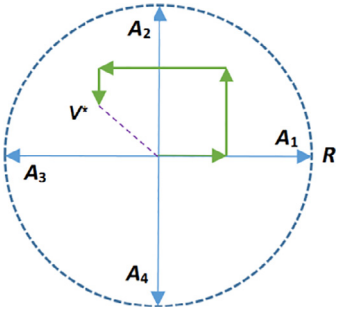

Fig. 1. Star Coordinate projection of a four-dimensional data point to a two-dimensional point V^* within SC Space. (This diagram has been adapted from [9]).(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

bounded range of [−1, 1]. We then include a negative region on the plot, which allows vectors to be displayed meaningfully. At the same time, this arrangement does away with the need to place independent axes in exact opposite directions.

The rest of this paper is organized as follows. Section 2 provides a systematic analysis of the weakness of the original SC Plot. Section 3 describes the proposed method. Section 4 presents experimental setup and results. Section 5 discusses related work and Section 6 concludes this paper.

2. Blind spots of the original star coordinate space

Recall that we have discussed two constraints in the original SC Plot. The first constraint is concerned with the fact that the signs of negative data values are not preserved in min–max normalization. The normalization process causes each axis to have a range of [0, 1]. As a result, vector algebra cannot be meaningfully applied to any normalized vector component that is originally negative.

The second constraint is concerned with the placement of two independent dimensional axes in exact opposite directions, which can result in less precise presentation of data patterns on an SC Plot. Rightfully, two independent dimensions shall be placed on two orthogonal axes. Although it is expedient to relax this rule in exchange for more dimensions to be represented on an SC Plot, we should still follow this rule as much as possible.

The above constraints result in three undesirable situations where certain non-zero data points in the original data space are erroneously projected to a zero vector (i.e., the origin) in the star coordinate space. We name such data points as *blind spots*.

To facilitate the following discussions, we define some notations. Let $\boldsymbol{D}$ be an n-dimensional data space. Let $V=(v_1, v_2, \ldots, v_n)$ in $\boldsymbol{D}$ be a non-zero data vector to be mapped to a point V^* in the star coordinate space $\boldsymbol{S}$. And the origin of $\boldsymbol{S}$ is denoted as $\boldsymbol{0}$. Let min_i and max_i be the minimum and maximum values in dimension i, respectively. That is, for m data points in $\boldsymbol{D}$, $min_i = \min\{v_{i1}, v_{i2}, \ldots, v_{im}\}$, and $max_i = \max\{v_{i1}, v_{i2}, \ldots, v_{im}\}$. Finally, let the angle $\theta = 2\pi/n$.

The first type of blind spots arises from the application of min–max normalization that scales each dimension to a range of [0, 1], as shown in Proposition 1.

Proposition 1. *If all the component values of V are the minimum values of their respective dimensions (i.e., $v_i = min_i\ \forall i \in \{1, 2, \ldots, n\}$), then $V^* = \boldsymbol{0}$ in $\boldsymbol{S}$.*

Proof. The min–max normalization of any component value v_i of a vector is defined as:

$$v_i^* = (v_i - \min_i)/(\max_i - \min_i).$$

And the x and y coordinates of V^* are respectively defined as:

$$x = \sum_{i=-1}^{n} v_i^* \cos(i\theta), \quad \text{and}$$
$$y = \sum_{i=1}^{n} v_i^* \sin(i\theta).$$

If $v_i = min_i$ then $v_i^* = 0 v_i = 0$, $\forall i \in \{1, 2, \ldots, n\}$.
Hence $x = y = 0$, and $V^* = \boldsymbol{0}$ in $\boldsymbol{S}$ □

The second type of blind spots are original data vectors that contain all the same component value. To prove such blind spots, we use the following known results on the finite sums of sine and cosine functions [10].

Lemma 1. *For any integer $n \geq 1$, the following sine-series and cosine-series hold.*

$$(1)\ \sum_{i=1}^{n} \sin i\theta = (\cos\tfrac{\theta}{2} - \cos(n + \tfrac{1}{2})\theta)/2\sin\tfrac{\theta}{2}$$
$$(2)\ \sum_{i=1}^{n} \cos i\theta = (\sin(n + \tfrac{1}{2})\theta - \sin\tfrac{\theta}{2})/2\sin\tfrac{\theta}{2}$$

Proposition 2. *If V contains the same value for all its normalized components (i.e., $v_1^* = v_2^* = \ldots = v_n^* = c$, where c is some constant), then $V^* = \boldsymbol{0}$ in $\boldsymbol{S}$.*

Proof. The x coordinate of V^* is:

$x = c\sum_{i=1}^{n} \cos(i\theta)$. Since $\theta = 2\pi/n$ and using (2) from Lemma 1, we get:

$$\begin{aligned} 2\sin\left(\frac{\pi}{n}\right)x &= c\left[\sin\left(n+\frac{1}{2}\right)\frac{2\pi}{n} - \sin\frac{\pi}{n}\right] \\ &= c\left[\sin\left(2\pi+\frac{\pi}{n}\right) - \sin\frac{\pi}{n}\right] \\ &= c\left[\sin 2\pi\cdot\cos\frac{\pi}{n} + \cos 2\pi\cdot\sin\frac{\pi}{n} - \sin\frac{\pi}{n}\right] \\ &= c\left[\sin\frac{\pi}{n} - \sin\frac{\pi}{n}\right] \\ &= 0. \end{aligned}$$

Similarly, the y coordinate of V^* is:

$y = c\sum_{i=1}^{n} \sin i\theta$. Using (1) from Lemma 1, we get:

$$\begin{aligned} 2\sin\left(\frac{\pi}{n}\right)y &= c\left[\cos\frac{\pi}{n} - \cos\left(n+\frac{1}{2}\right)\frac{2\pi}{n}\right] \\ &= c\left[\cos\frac{\pi}{n} - \cos\left(2\pi+\frac{\pi}{n}\right)\right] \\ &= c\left[\cos\frac{\pi}{n} - \cos 2\pi\cdot\cos\frac{\pi}{n} + \sin 2\pi\cdot\sin\frac{\pi}{n}\right] \\ &= c\left[\cos\frac{\pi}{n} - \cos\frac{\pi}{n}\right] \\ &= 0. \end{aligned}$$

Since $x = y = 0$, and $V^* = \boldsymbol{0}$ in $\boldsymbol{S}$. □

The third type of blind spots are original data vectors that contain the same component value in each pair of dimensions where axes are placed directly opposite in the star coordinate space.

Proposition 3. *Let $\boldsymbol{D}$ be an n-dimensional data space where n is even. Let any two dimensions in $\boldsymbol{D}$ that are positioned directly opposite of one another in $\boldsymbol{S}$ be A_i and $A_{i+n/2}$, we would then expect $\angle(A_i, A_{i+n/2}) = \pi$. If V contains identical normalized component values in each possible pair of A_i and $A_{i+n/2}$ (i.e., $v_i^* = v_{i+n/2}^*$), then $V^* = \boldsymbol{0}$ in $\boldsymbol{S}$.*

Proof. The x and y coordinates of V^* are:

$$x = \sum_{i=1}^{n} v_i^* \cos(i\theta) = \sum_{i=1}^{n/2} v_i^*[\cos(i\theta) + \cos(i\theta + \pi)],$$