Applied Soft Computing xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

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

Discernible visualization of high dimensional data using label information

Asef Pourmasoumi Hasan Kiyadeh, Amin Zamiri, Hadi Sadohgi Yazdi*, Hadi Ghaemi

Computer Department, Ferdowsi University of Mashhad, P.O. Box: 9177948974, Mashhad, Iran

ARTICLE INFO

Article history:
Received 12 August 2013
Received in revised form 23 July 2014
Accepted 18 September 2014
Available online xxx

Keywords:
Visualization
Star Coordinate
High dimensionality reduction
Fisher's discriminant form

ABSTRACT

Visualization methods could significantly improve the outcome of automated knowledge discovery systems by involving human judgment. Star coordinate is a visualization technique that maps k-dimensional data onto a circle using a set of axes sharing the same origin at the center of the circle. It provides the users with the ability to adjust this mapping, through scaling and rotating of the axes, until no mapped point-clouds (clusters) overlap one another. In this state, similar groups of data are easily detectable. However an effective adjustment could be a difficult or even an impossible task for the user in high dimensions. This is specially the case when the input space dimension is about 50 or more.

In this paper, we propose a novel method toward automatic axes adjustment for high dimensional data in Star Coordinate visualization method. This method finds the best two-dimensional view point that minimizes intra-cluster distances while keeping the inter-cluster distances as large as possible by using label information. We call this view point a discernible visualization, where clusters are easily detectable by human eye. The label information could be provided by the user or could be the result of performing a conventional clustering method over the input data. The proposed approach optimizes the Star Coordinate representation by formulating the problem as a maximization of a Fisher discriminant. Therefore the problem has a unique global solution and polynomial time complexity. We also prove that manipulating the scaling factor alone is effective enough for creating any given visualization mapping. Moreover it is showed that *k*-dimensional data visualization can be modeled as an eigenvalue problem. Using this approach, an optimal axes adjustment in the Star Coordinate method for high dimensional data can be achieved without any user intervention. The experimental results demonstrate the effectiveness of the proposed approach in terms of accuracy and performance.

1. Introduction

Our 3-dimensional perspective limits our conceptual experience of higher dimension space. Nevertheless our interaction with high dimension spaces is getting more and more inevitable. Increasing development of science and technology has led to substantial growth in data production beyond any human being conception capability. Billions of Web pages in cyberspace, huge geographical information, large amounts of biological data and numerous amounts of business databases are just small portions of the available data.

This has been the main motivation of obtaining geometric models (like graphs where there are 2 or 3 variables) of multivariate relationships arising in analyzing large sets of high dimensional data. Consequently numerous "visualization" approaches are proposed that try to achieve the best map from *k*-dimensions to 2 or 3 dimensions which is discernible for the human brain, effortlessly.

Unprecedented growth of data production and the limited ability of the human brain have made data visualization an interesting subject in computer science during recent years. As Card et al. described, visualization is "the use of computer-supported interactive, and visual representation of abstract data to amplify cognition." Visualization is considered as one of the most intuitive methods for cluster detection and validation, and especially is performing well for the representation of irregularly shaped clusters [27,32].

Other approaches of overcoming the problems of high dimensionality are dimension reduction [4,20] and feature selection [24]. Data sampling and data summarization could also help to cope with

* Corresponding author.
E-mail addresses: asef.pourmasoumi@stu-mail.um.ac.ir (A.P.H. Kiyadeh), amin.zamiri@stu.um.ac.ir (A. Zamiri), h-sadoghi@um.ac.ir (H.S. Yazdi), hadi.qaemi@stu-mail.um.ac.ir (H. Ghaemi).

http://dx.doi.org/10.1016/j.asoc.2014.09.026
1568-4946/

large amount of data records [17,28]. Scientists interested in these fields face a similar problem in exploratory analysis or visualization of multivariate data.

Star Coordinate is a visualization technique for mapping k-dimensional data into Cartesian coordinates, in which the coordinate axes are arranged on a circle of a two-dimensional plane with the origin at the center of the circle. It is proved that in this mapping technique, a cluster can always be preserved as a point-cloud (or cluster) in the visual space through linear mappings. But the main problem arises when these mapped point-clouds overlap one another, making their boundaries indistinguishable. Therefore the user is given the ability to push and pull or rotate the axes until the desired outcome is achieved. However, an advantageous adjustment is difficult or even impossible for the human agent to achieve, when visualizing high dimensional data. As a result, some researchers have proposed various dimension reduction methods, as pre-processing steps before applying the Star Coordinate visualization technique.

In this paper, we focus on the problem of automatic axes adjustment in Star Coordinate technique for improved visualization results. Our goal is to find the best projection possible that can represent the original data topology in k-dimensional data especially where k is greater than 50, effectively making manual axes adjustment impossible. The rest of the paper is organized as follows. Section 1.1 presents a discussion of related work. The main features of the Star Coordinate algorithm are briefly discussed in Section 1.2. Then, the proposed method is introduced in Section 2. In Section 3, we present the experimental results that validate the cost model. Section 4 presents a discussion of the experimental results. Finally, Section 5 concludes the presented approach.

1.1. Related work

Numerous approaches have been proposed for the visualization of multi-dimensional datasets. Scatterplot matrix [9], parallel coordinates [21] and dimensional stacking [31] have been developed to address this issue. Parallel coordinates (PC) [21] is a well-known method in which features are represented by parallel vertical axes linearly scaled within their data range. Each sample is represented by a polygonal line that intersects each axis at its respective attribute data value. Parallel coordinates can be used to study the correlations among various attributes by spotting the locations of the intersection points [44]. Also, they are useful for detecting the data distributions and functional dependencies. The main challenge of parallel coordinate approach is the limited space available for each parallel axis. There are several extended method for parallel coordinate, such as Circular Parallel Coordinates [19] and Hierarchical Parallel Coordinates [13].

Ester et al. [10] proposed DBSCAN to discover arbitrarily shaped clusters. It may not handle data sets that contain clusters with different densities. The OPTICS method, derived from the DBSCAN algorithm, uses visualization for visual cluster analysis [1] and is useful for finding density-based clusters in spatial data. Like most of the clustering algorithms, OPTICS is a parametric approach. Yang et al. [39] proposed a visual hierarchical dimension reduction technique, which groups dimensions and visualizes data by using the subset of dimensions obtained from each group. In [2] and [36], some features that affect the quality of visualization have been introduced and some of the above systems are compared based on listed features.

Another famous approach for data visualization is Star Coordinate [25] and its extensions, such as VISTA [6]. The proposed method is based on the Star Coordinate technique. Star Coordinates arranges coordinate axes on a two-dimensional surface, where each axis shares the same origin point. It uses a linear mapping to avoid the cluster breaking after k-dimensional to 2D space mapping. (This has been proven in [8] mathematically). So far, several extensions for VISTA have been introduced. iVIBRATE [7] is a framework for visualizing large datasets using data sampling and the Star Coordinate model. In [37], an Enhanced VISTA is proposed which improves visualization and eases the human computer interaction. The experiments have shown that visual cluster rendering can improve the understanding of clusters, and validate and refine the algorithmic clustering result effectively [25].

VISTA is a very good interactive approach for visualization of k-dimensional data where $K < 50$, and its efficiency has been proven by various articles. The main shortcoming of this method is that the dimension must be less than 50. Since, according to each dimension of data, a coordinate axis is drawn, when the number of dimensions is more than 50, working with VISTA tools would be very exhausting for humans and, practically, its interactivity property would be useless. This problem becomes more serious when the number of dimensions is much greater than 50. However, there are many datasets with a large amount of features in the world, e.g., textual data, image data, bioinformatics data, etc.

In this paper we propose a novel semi-supervised visualization method for high dimensional data, where a fraction of the data is labeled. The visualization result achieved by applying this method is optimal in terms of discernibility by the user. This work extends Star Coordinates capabilities in working with high-dimensional datasets.

1.2. Star Coordination

Star Coordinates is a visualization technique for mapping high-dimensional data into two dimensions. In this technique a 2D plane is divided into k equal sectors (θ_i, the angle of the sectors, is set to $2\pi i/k$ by default). Therefore there are k coordinate axes, with each axis representing one dimension of data and all axes sharing their origins at the center of a circle on the 2D space (Fig. 1) having the same length [25]. Data points are scaled to the length of the axis, in way that the smallest is mapped to the origin and the largest to the other end of the axis. Then unit vectors on each coordinate axis are calculated accordingly to allow scaling of data values to the length of the coordinate axes.

The mapping of a point from k-dimensional space to a point in the two dimensional Cartesian coordinates is determined by the sum of all unit vectors ($\vec{u}_{xi}$, $\vec{u}_{yi}$), on each coordinate multiplied by the value of the data element for that coordinate, as shown in Formula (1):

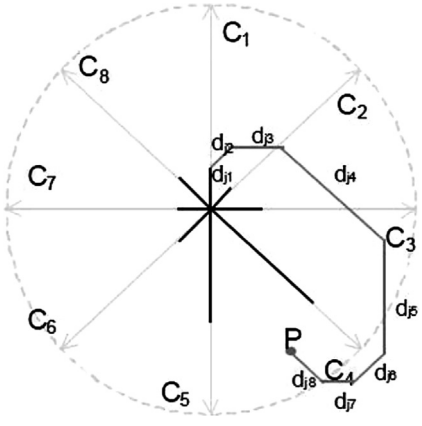

Fig. 1. The image of an 8-dimensional point in Cartesian coordinate [25].

Please cite this article in press as: A.P.H. Kiyadeh, et al., Discernible visualization of high dimensional data using label information, Appl. Soft Comput. J. (2014), http://dx.doi.org/10.1016/j.asoc.2014.09.026