Journal of King Saud University – Computer and Information Sciences (2017) 29, 2–18

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Detection and visualization of non-linear structures in large datasets using Exploratory Projection Pursuit Laboratory (EPP-Lab) software

Souad Larabi Marie-Sainte

Information Technology Department, College of Computer and Information Sciences, King Saud University, Saudi Arabia

Received 5 June 2015; revised 28 March 2016; accepted 18 April 2016
Available online 11 May 2016

KEYWORDS

Exploratory Projection Pursuit;
Genetic Algorithm;
Particle Swarm Optimization;
Tribes;
Clustering;
Outliers

Abstract This article consists of using biologically inspired algorithms in order to detect potentially interesting structures in large and multidimensional data sets. Data exploration and the detection of interesting structures are based on the use of Projection Pursuit that involves the definition and the optimization of an index associated with each direction or projection. The optimization of a projection index should provide a set of multiple optima that is expected to correspond to interesting graphical representations in low dimensional space. The implementation of the bio-inspired algorithms along with the projection pursuit develops a new software called EPP-Lab. Projection pursuit is widely used in different scientific domains (biology, pharmacy, bioinformatics, biometry, etc) but not widely present in the well-known softwares. EPP-Lab is dedicated to recognize and visualize clusters and outlying observations on one dimension from high dimensional and multivariate data sets. It includes different statistical techniques for results analysis. It provides several features and gives the user the option to adjust the parameters of the selected bio-inspired methods or to use defaults values. EPP-Lab is a unique software for detection, visualization and analysis of non-linear structures. The performance of this tool has been validated by testing different real and simulated data sets.

1. Introduction

In computer science and especially Artificial Intelligence, metaheuristic is a technique intended to find an approximate solution for hard and combinatorial optimization problems in a reasonable time. Metaheuristic methods include nature-inspired by social behavior (Particle Swarm Optimization, Ants Colony Optimization, Bee Colony, Firefly, etc) or by Darwin evolutionary biology (Genetic Algorithms, Genetic programming, Evolution strategies, etc). Most of them solve hard and combinatorial problems (Sevkli et al., 2014; Goswami and Mandal, 2013; Upadhyay et al., 2014) by handling a set of solutions and maintaining a balance between diversification (exploration of the solution space) and intensification (exploitation of the accumulated knowledge). This research work concerns the

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2016.04.003
1319-1578 © 2016 The Author. Production and hosting by Elsevier B.V. on behalf of King Saud University.

use of biologically inspired algorithms to find out potential interesting structures in large and multidimensional data sets. Extracting hidden information from data of large dimensions involves employing exploratory data analysis methods. A Principal Component Analysis is one known method of statistical analysis that is based on projecting the data on dimensions that maximize the dispersion of the observations. However, maximizing the dispersion does not constantly lead to the detection of interesting structures. In this study, Projection Pursuit (PP), one of Exploratory Projection Pursuit methods, is used. It consists of finding interesting low dimensional projections of high dimensional multivariate data (Jones and Sibson, 1987; Huber, 1985). Based on human visualization, low dimension means (one- two- three) dimensions. PP focuses on the definition and the optimization of an index associated with each direction or projection space. To optimize the projection indices, exact optimization methods (Newton, Steepest ascent, etc) were applied. These methods require the properties of regularity that most of the projection indices do not provide. On the other hand, the main characteristic required for PP is to find multiple optima corresponding to different interesting structures (Friedman, 1987; Morton, 1989; Sun, 1993). However, these exact methods cannot provide multiple optima. Hence, the use of bio-inspired algorithms is encouraging. They can, not only find a global optimum (approximate solution) but also several local optima (using several runs) corresponding to different potential interesting projections. Among the different bio-inspired algorithms, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and a hybrid Particle Swarm Optimization method called Tribes are employed. The performance of these selected methods combined with PP has been validated in Berro et al. (2010) and Mari-Sainte et al. (2010).

This study is focused on the detection of clusters and outlying observations. Clustering is one of the main tasks of data mining and mainly importuned in different domains (Alghamdi et al., 2014; Aljumah et al., 2013). Outlier detection involves removing anomalous observations from data (Hogge and Austin, 2004). Outliers occur due to mechanical faults, fraudulent behavior, human error, device fault or natural deviations in populations. Their detection can eliminate contaminating effect on the data set.

Although PP has been used for this purpose in different domains (biology, bioinformatics, image processing, biometry, etc), its implementation is not satisfactory (Caussinus and Ruiz-Gazen, 2009).

Therefore, having a software including PP has become indispensable in many scientific disciplines that use this statistical analysis method. In this article, one-dimensional projection pursuit method, including five projection indices, is implemented along with the selected bio-inspired methods in order to obtain a powerful tool called Exploratory Projection Pursuit Laboratory (EPP-Lab).

EPP-Lab is dedicated to look for hidden non-linear structures in high dimension data sets, particularly clusters and outlying observation. This software is designed with collaboration of statisticians. It gives the user the option to adjust the parameters of the bio-inspired methods implemented or to use the defaults values. In addition, it provides new ways for the analysis of the results. EPP-Lab is a unique tool for clusters-outliers detection, visualization and analysis.

To validate the performance of this software, several real and simulated data sets are treated, and some data sets are large and with complex structures.

The rest of this paper is organized as follows. Section 2 introduces the principle notion and definition of PP. Section 3 presents the related works. Section 4 addresses the methodology of the implemented techniques. Section 5 gives a global presentation of EPP-Lab and its main features. Section 6 illustrates the EPP-Lab application and results using several real and simulated data sets to determine clusters and detect outliers, in addition to a small part of comparison studies along with the convergence study. Section 7 tackles the computation time and the limitation of EPP-Lab. Finally Section 8 concludes this research work.

2. Projection Pursuit

PP method seeks to look for low- (one-, two-, three-) dimensional projections that provide potential interesting structures hidden in multidimensional and large data sets. The notion of "interesting" structures is defined by a suitable projection index function $I(a)$, depending on a normalized projection vector a. This index tries to find the degree of nonlinear structure present in the projected distribution. Let denote by $X : N \times P$ the data set matrix of N cases and P variables. Let X_i is the ith column vector in $\Re^p$ associated with the ith observation. This study focuses on one-dimensional projection. The projection vector can be defined from $\Re^p$ to $\Re$ as $z = Xa, a$ is a P-dimensional vector defining the linear transformation, and z is a N-dimensional vector corresponding to the coordinates of the projected observations. So, determining a projection is equivalent to determining a. In other words, this matter is equivalent to optimize a selected projection index.

The present work focuses on four particular one-dimensional indices. The Friedman–Tukey index (Friedman and Tukey, 1974) is the first index proposed in the context of EPP and it is interesting for the detection of outliers. The Friedman's index (Friedman, 1987) belongs to the family of the polynomial-based indices, it performs particularly well in detecting separations or clusters compared with other indices from the same family (Sun, 1993). The kurtosis index is based on the fourth moment and has been studied in Peña and Prieto (2001) and Achard et al. (2004). We also consider a new proposal suited to the detection of clusters called "discriminant index" (Berro et al., 2010). All these indices are well defined and applied for the detection of clusters and/or outliers in Berro et al. (2010) and Mari-Sainte et al. (2010).

One of the most important characteristic of PP method is sphering the data. The sphering step is generally applied to data before the PP step. It consists of eliminating scale and correlation structure in the data sets in order to find other aspects of the data. It also ensures the difference between any structures found by PP and those found by Principal Components Analysis. This characteristic is implemented in EPP-Lab.

PP is less widely used compared with PCA but is more powerful than PCA in many cases because PCA only considers the second order moment and may miss interesting hidden structures that can be easily discovered by another EPP technique (Jones and Sibson, 1987). Indeed, important aspects of the data structure are likely to appear in none of the principal subspaces as this may be seen in Fig. 1. Suppose that the whole dimension is larger than 2 and the straight lines are parallel to subspaces of dimension 2. The first principal plane of PCA, roughly the horizontal axis in this example, is clearly unable