Sains Malaysiana 52(8)(2023): 2431-2451
http://doi.org/10.17576/jsm-2023-5208-19

A New Single Linkage Robust Clustering Outlier Detection Procedures for Multivariate Data

(Suatu Prosedur Baharu Pengesanan Data Terpencil Berasaskan Pengelompokan Rangkaian Tunggal Teguh bagi Data Multivariat)

SHARIFAH SAKINAH SYED ABD MUTALIB[1,2], SITI ZANARIAH SATARI[1,*] & WAN NUR SYAHIDAH WAN YUSOFF[1]

[1]Centre for Mathematical Sciences, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia
[2]Faculty of Computer, Media and Technology Management, University College TATI, Jalan Panchur, Telok Kalong, 24000 Kemaman, Terengganu, Malaysia

Received: 3 January 2023/Accepted: 1 August 2023

ABSTRACT

Outliers are abnormal data, and the detection of outliers in multivariate data has always been of interest. Unlike univariate data, outlier detection for multivariate data is insufficient with a visual inspection. In this study, we developed a new single linkage robust clustering outlier detection procedure for multivariate data. A robust estimator, Test on Covariance (TOC) is used to robustified the similarity distance measure, producing robust single linkage clustering. The performance of the new single linkage robust clustering outlier detection procedure is investigated via a simulation study using three outlier scenarios and historical multivariate datasets as illustrative examples. Three performance measures are used, which are *pout*, *pmask,* and *pswamp*. The performance of the new single linkage robust clustering procedure also compared with single linkage clustering using Euclidean and Mahalanobis distances as similarity distance measures as well as TOC. It is found that the new single linkage robust clustering procedure performs well in Outlier Scenario 3 when the mean and covariance matrix are shifted. The new procedure also performs well by successfully detecting all outliers, does not have masking effects in two out of five datasets and does not have swamping effect in all datasets. In conclusion, the new single linkage robust clustering outlier detection procedure is a practical and promising approach and good for simultaneously identifying multiple outliers in multivariate data.

Keywords: Multivariate data; outliers; single linkage clustering; Test on Covariance; robust clustering

ABSTRAK

Data terpencil ialah data tidak normal dan pengesanan data terpencil untuk data multivariat sentiasa menjana minat. Tidak seperti data univariat, pengesanan data terpencil untuk data multivariat tidak mencukupi dengan pemeriksaan visual. Dalam kajian ini, kami membangunkan satu prosedur baru pengesanan data terpencil berasaskan pengelompokan rangkaian tunggal teguh bagi data multivariat. Penganggar teguh, *Test on Covariance* (TOC) digunakan untuk meneguhkan ukuran jarak persamaan, menghasilkan pengelompokan rangkaian tunggal teguh. Prestasi prosedur baru pengesanan data terpencil berasaskan pengelompokan rangkaian tunggal teguh disiasat melalui kajian simulasi menggunakan tiga senario data terpencil dan set data sedia ada multivariat sebagai contoh ilustrasi. Tiga ukuran prestasi digunakan, iaitu *pout*, *pmask* dan *pswamp*. Prestasi prosedur baru pengesanan data terpencil berasaskan pengelompokan rangkaian tunggal teguh juga dibandingkan dengan pengelompokan rangkaian tunggal menggunakan jarak *Euclidean* dan *Mahalanobis* sebagai ukuran jarak persamaan beserta *TOC*. Didapati bahawa prosedur baru pengesanan data terpencil berasaskan pengelompokan rangkaian tunggal teguh berprestasi baik dalam Senario Data Terpencil 3 apabila min dan matriks kovarians dianjakkan. Prosedur baru juga berfungsi dengan baik apabila berjaya mengesan semua data terpencil dan tidak mempunyai kesan *masking* dalam 2 daripada 5 set data dan tidak mempunyai kesan *swamping* dalam semua set data. Kesimpulannya, prosedur baru pengesanan data terpencil berasaskan pengelompokan rangkaian tunggal teguh ialah pendekatan yang praktikal dan menjanjikan, serta bagus untuk mengesan data terpencil yang berkelompok secara serentak dalam data multivariat.

Kata kunci: Data multivariat; data terpencil; pengelompokan rangkaian tunggal; pengelompokan teguh; *Test on Covariance*