

Outliers Detection and Analyzing for Diabetic and Non-Diabetic Patients by Using Two-Phase Clustering

Hsu Mon Htwe, Aye Lelt Lelt Phyu

*University of Computer Studies, Yangon
corncracker1@gmail.com*

Abstract

Outliers are the set of objects that are considerably dissimilar from other values in a random sample from a population. Outliers are important because they can change the results of data analysis. In this paper, two-phase clustering approach is used to fulfil the detection of abnormal diabetic and non-diabetic patients. The patients' data is taking from one of datasets donated to UCI machine learning repository, Pima Indians diabetes dataset from National Institute of Diabetes and Digestive and Kidney Diseases. In the phase-1, patients' data are clustered according to their similar features. In the phase-2, a minimum spanning tree (MST) is constructed to detect clusters with outliers. The conclusion remarks on resulted abnormal patients are given by the expert researcher of diabetes disease and analyses on attributes' features are also presented in this paper. Thus, it is intended to detect abnormal patients of positive or negative in diabetes test result and to illustrate analyses on them.