Improving K-NN Internet Traffic Classification Using Clustering and Principle Component Analysis

Adi Suryaputra Paramita*
Ciputra University
UC Town, Citraland, Surabaya 60291, Indonesia
*Corresponding author, e-mail: adi.suryaputra@ciputra.ac.id

Abstract

K-Nearest Neighbour (K-NN) is one of the popular classification algorithm, in this research K-NN use to classify internet traffic, the K-NN is appropriate for huge amounts of data and have more accurate classification, K-NN algorithm has a disadvantages in computation process because K-NN algorithm calculate the distance of all existing data in dataset. Clustering is one of the solution to conquer the K-NN weaknesses, clustering process should be done before the K-NN classification process, the clustering process does not need high computing time to conqest the data which have same characteristic, Fuzzy C-Mean is the clustering algorithm used in this research. The Fuzzy C-Mean algorithm no need to determine the first number of clusters to be formed, clusters that form on this algorithm will be formed naturally based datasets be entered. The Fuzzy C-Mean has weakness in clustering results obtained are frequently not same even though the input of dataset was same because the initial dataset that of the Fuzzy C-Mean is less optimal, to optimize the initial datasets needs feature selection algorithm. Feature selection is a method to produce an optimum initial dataset Fuzzy C-Means. Feature selection algorithm in this research is Principal Component Analysis (PCA). PCA can reduce non significant attribute or feature to create optimal dataset and can improve performance for clustering and classification algorithm. The resultsof this research is the combination method of classification, clustering and feature selection of internet traffic dataset was successfully modeled internet traffic classification method that higher accuracy and faster performance.

***Keywords**: classification, clustering, internet, bandwidth, PCA*

1. Introduction

Previous internet traffic classification research using the internet traffic data usage done by Chengjie GU, Shunyi ZHANG, and Xiaozhen XUE, in April 2011. This research main contribution is increasing the classification accuracy by improving the Kernel algorithms for Fuzzy K-Mean. But in that research said that the algorithm Fuzzy K-Mean insufficient to optimize the characteristics of the data in dataset and all the features of the data in dataset is consider to have the same contribution to the class that will be generated. It cause the level of accuracy in classification produced less accurate and still needs to improved [1]. This occurs because the algorithm Fuzzy K-Mean Kernel, many class were formed has been define from the outset that as many as K. At the conclusion of this research said that still need future works to discover what features are suitable and appropriate to improve internet traffic classification accuracy.

K-Nearest Neighbor (K-NN) algorithm will chosen for internet traffic classification algorithm in this research, the difference between K-NN and Fuzzy K Mean is on a computational algorithm dan process, in K-NN all distances distribution of existing data will calculate in computational process, it cause the results of accuracy in internet traffic classification would be more accurate, because K-NN process compute all the possibilities that exist. On the other hand, the process of conscientious computational of algorithms K-NN have a weakness in terms of performance, the process of classification become slower than another algorithm. In addressing the weakness of K-NN algorithm in this research will conduct the experiments by forming the classified datasets into cluster at the first phase, the classified dataset forming process is done by clustering algorithm. When clustering process is done, the spread of the data in dataset developed naturally based on similarity of characteristic data, as

Received February 25, 2017; Revised April 26, 2017; Accepted May 10, 2017