**Bulletin of Electrical Engineering and Informatics**
**ISSN: 2302-9285**
Vol. 6, No. 2, June 2017, pp. 159~165, DOI: 10.11591/eei.v6i2.608 

# 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*

the data is scattered then carried out a process of classification, this spread data from clustering process is expected to accelerate the velocity performance of K-NN algorithm. In Clustering process which done by Fuzzy C Mean the number of clusters to be formed will assemble naturally, so the number of clusters that formed would show the grouping of data occurs. In previous study in 2012 conducted by LOU Xiaojun, LI Junying, and Haitao LIU stated that the Fuzzy C Mean generally had a disadvantage for the output partition or cluster for the same dataset [2].

Based on previous research, there is an opportunity to do deeper research about Internet traffic clasffication in machine learning area. In this research K-NN algorithm will use for classification and Fuzzy C Mean algorithm for clustering. One advantage of Fuzzy C-Mean is no need to set the specific number of class in the beginning of clustering process such as Fuzzy K Mean algorithm. The class will formed naturally and expected to represent real data. However Fuzzy C Mean require a feature selection for data to use that Internet traffic has the same correlation could fit into the same class. Another thing that could be the development in this research is the process of how to find the important and precise fit features in dataset.

This research used the same dataset with previous research, the internet traffic data is collected from http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmetrics/. After dataset collected the next phase in this research is to find the discriminant features in the Internet traffic dataset. Principal Component Analysis (PCA) is the algortihm to select the discriminant feature in this research. The procedures of discriminant feature selection using PCA will be seen in the picture below.

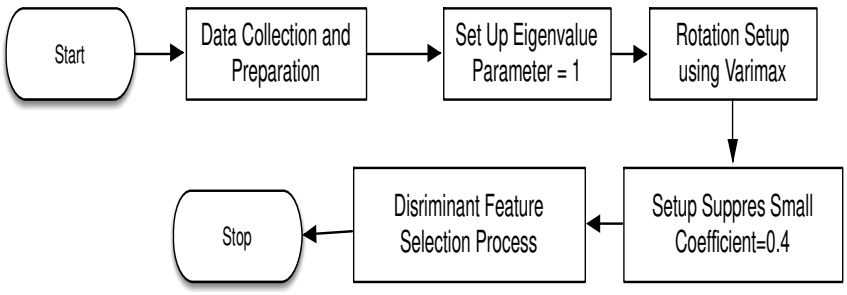


Figure 1. Discriminant feature selection procedures

Flowchart in Figure 1. illustrates the process of discriminant feature finding in some dataset using Principal Component Analysis. The discriminant feature selection process is to find correlated data in dataset and established a correlation matrix in the end of process, the eigenvalue in this process set to 1, the eigenvalue is the number of variants associated with factors used for eigenvalues is worth more of 1 which have an impact on the features only have eigenvalues greater than 1 will be retained, while the variance factor less than 1 will be reduced in accordance with the standards as written in the F.Wang article titled factor Analysis and Principal Component Analysis in 2009 [3], meanwhile varimax rotation used to maximize the amount of variance of the squared correlation between variables and factors . This is achieved if every variable that has given a high load on a single factor but near zero loads on the remaining factors and if the factors are given based on only a few variables with a very high load on this factor, while the remaining variables have the burden close to zero on this factor. In Figure 1. Seen that suppressed altogether Small Coefficients filled with 0.4, it will take a long time due to the features that have values below 0.4 will be ignored and not be forming new features, the use of coefficient 0.4 will yield significant results in the recommended by JP Stevens (1992) [4], this is related to the significant results quoted by Andy Field in his book Discovering Statistics Using SPSS(1992) [5].

Fuzzy C-Means clustering is a technique to clustering of each data point in dataset which determined by the degree of membership. This algorithm was first introduced by Jim Bezdek in 1981. First step of Fuzzy C-Means is to determine cluster centers, which marked the average location for each cluster. In the initial condition, the center of the cluster is still not

accurate. Each data point has a degree of membership for each cluster. By improving the cluster centers and the degree of membership of each data point repeatedly, it will be seen that the center cluster will move towards the right location. This loop is based on the minimize of an objective function that describes the distance from the given data point to the center of the cluster that is weighted by the degrees of membership of data points. Output of Fuzzy C-Means is a row of cluster centers and some degree of membership for each data point. First of all, the method provides membership values, which can be useful for assessing the validity of the cluster structure obtained. Second, the method has a simple and efficient algorithm which makes it applicable in a broad class of situations [6].

Algorithm k-nearest neighbor (k-NN or KNN) is an algorithm used for the classification of the object based on the distance between the objects. The data used for the classification process in the K-NN projected into multiple dimensions, where each dimension represents the features of the data. The space is divided into sections based on the classification of data that are classified. A point in this space marked class c if class C is the most common classification of the k nearest neighbors of the dot. Near or far neighbors Euclidean.Pada usually calculated based on the distance learning phase, the algorithm is simply to store the vectors of features and classification of learning data. In the classification phase, the same features are calculated for test data (which classification is not known). The distance of this new vector of all learning data vector is calculated, and the number k closest retrieved. K-NN algorithm accuracy is greatly influenced by the presence or absence of features that are not relevant, or if the weight of such features is not equivalent to its relevance to the classification. Research on these algorithms largely discusses how to choose and give weight to the feature, in order to become a better classification performance [7-8].

## 2. Research Method

The purpose of this research is improving K-NN Classification accuracy and performance through Fuzzy C-Mean Clustering and Principal Component Analysis (PCA). PCA first technique for analyzing internet traffic dataset and to find the discriminant feature [9]. Fuzzy C-Mean is a technique for improving the K-NN performance, Fuzzy C-Mean is the solution to help K-NN in data clustering, Fuzzy C-Mean will make the distribution and grouping of data so as to make the K-NN does not need to perform the calculation of all distances between existing data [10]. The research methodology to achieve these research objectives, as shown in Figure 2.

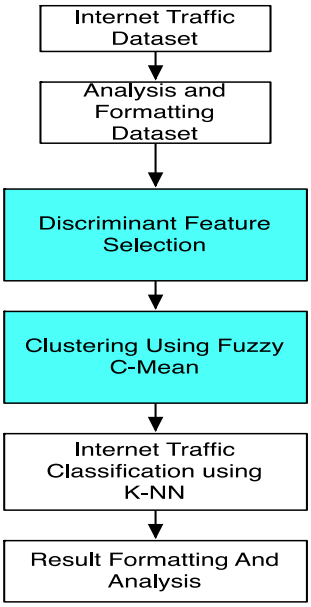


Figure 2. Research methodology

The scientific contribution of this research shown in the blue box on Figure 2. The first stage of this research is gathering internet traffic dataset, the internet traffic dataset used in this research are mooreset dataset, this data is collected from http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmetrics/, the dataset used in this research are dataset2, dataset3 and dataset10. The next stage after data collected is discover the best correlated dataset thru discriminant features selection using Principal Component Analysis (PCA). The next process after the dataset already shaped, is clustering the dataset using Fuzzy C-Mean algorithm. Fuzzy C-Mean will grouping and disseminate dataset into a group that has the same data characteristics. Afterwards the K-NN will classified the dataset into classification class. The result from internet traffic classification will be assess and observing after K-NN classification done.

## 3. Results and Analysis

Overall PCA had significant contribution to K-NN algorithm, when PCA applied into this classification model, the execution time of K-NN algorithm increase sharply, meanwhile Fuzzy C-Mean contribution regarding execution time only seen in dataset10 which is have largest amount of data. On the other hand accuracy in dataset10 is decrease when PCA and Fuzzy C-Mean applied in classification model, meanwhile the accuracy in dataset 2 and dataset 3 increase mode than 1% when PCA and Fuzzy C-Mean applied. The detail of all dataset present in the Table 1. below, Table 2. until Table 6. give information about the detail of experimental result.

Table 1. The Number of Data Flow

| Algorithm | Dataset | | |
|---|---|---|---|
| | Dataset 2 | Dataset 3 | Dataset 10 |
| WWW | 18559 | 18065 | 54436 |
| MAIL | 2726 | 1448 | 6592 |
| FTP-CONTROL (FC) | 100 | 1861 | 81 |
| FTP-PASV (FP) | 344 | 125 | 257 |
| ATTACK | 19 | 41 | 446 |
| P2P | 94 | 100 | 624 |
| DATABASE (DB) | 329 | 206 | 1773 |
| FTP-DATA (FD) | 1257 | 750 | 592 |
| MULTIMEDIA(MM) | 150 | 136 | 0 |
| SERVICES(SRV) | 220 | 200 | 212 |
| INTERACTIVE (INT) | 2 | 0 | 22 |
| GAMES (GM) | 1 | 0 | 1 |
| Total Data Flow | 23801 | 22932 | 65036 |

Table 1. shown that dataset 10 have 65036 data which is the highest amount of data flow, meanwhile dataset2 and dataset3 almost have same amount of dataset around 22000-23000 data. The most class in all dataset is WWW.

Table 2. Execution Time Result

| Algorithm | Dataset 2 | Dataset Dataset 3 | Dataset 10 |
|---|---|---|---|
| Traditional K-NN | 258 Seconds | 237 seconds | 1232 seconds |
| Traditional K-NN + Fuzzy C-Mean | 261 seconds | 249 seconds | 839 seconds |
| Traditional K-NN + Fuzzy C-Mean+ PCA | 66 seconds | 69 seconds | 249 seconds |
| Number of data | 23801 | 22932 | 65036 |

Table 2. present that PCA have significant contribution to accelarate the velocity of K-NN algorithm, the execution time increase around fourfold to five fold. On the other side when Fuzzy C-Mean applied in K-NN algorithm using dataset2 and dataset 3 the execution time become little bit slower while the execution time still faster when Fuzzy C-Mean applied in dataset 10.

Table 3. Accuray Result

| Algorithm | Dataset 2 | Dataset Dataset 3 | Dataset 10 |
|---|---|---|---|
| Traditional K-NN | 97.96% | 97.57% | 98.41 % |
| Traditional K-NN + Fuzzy C-Mean | 97.96% | 97.57% | 96.70% |
| Traditional K-NN + Fuzzy C-Mean+ PCA | 98.37% | 98.63% | 98.06% |
| Number of data | 23801 | 22932 | 65036 |

Table 3. shown while classification using dataset 2 and dataset 3 and apply PCA dan Fuzzy C-Mean the accuracy improve slightly, meanwhile if the classification using dataset10 and apply Fuzzy C-Mean as clustering algorithm and PCA as the discriminant feature selection the accuracy decrease slightly.

Table 4. Classification Summary Result Dataset 2

| Algorithm | Traditional K-NN | Dataset Traditional K-NN + Fuzzy C-Mean | Traditional K-NN + Fuzzy C-Mean+ PCA |
|---|---|---|---|
| Max Precision Value | 99.38% | 99.38% | 99.27 % |
| Min Precision Value | 0% | 0% | 0% |
| Number of Class in Dataset | 12 | 12 | 12 |
| Number of Class figure out in classification | 9 | 9 | 9 |
| Number of data | 23801 | 23801 | 23801 |

Table 5. Classification Summary Result Dataset 3

| Algorithm | Dataset | | |
|---|---|---|---|
| | Traditional K-NN | Traditional K-NN + Fuzzy C-Mean | Traditional K-NN + Fuzzy C-Mean+ PCA |
| Max Precision Value | 98.38% | 98.38% | 99.33 % |
| Min Precision Value | 0% | 0% | 0% |
| Number of Class in Dataset | 10 | 10 | 10 |
| Number of Class figure out in classification | 9 | 9 | 9 |
| Number of data | 22932 | 22932 | 22932 |

Table 6. Classification Summary Result Dataset 10

| Algorithm | Dataset | | |
|---|---|---|---|
| | Traditional K-NN | Traditional K-NN + Fuzzy C-Mean | Traditional K-NN+Fuzzy C-Mean+PCA |
| Max Precision Value | 99.77% | 99.09% | 99.60 % |
| Min Precision Value | 0% | 0% | 0% |
| Number of Class in Dataset | 11 | 11 | 11 |
| Number of Class figure out in classification | 10 | 10 | 9 |
| Number of data | 65036 | 65036 | 65036 |

Table 4. to Table 6. presents that PCA is have main contribution to increase the precision value except in classification using dataset 10. In Classification model which used dataset10 as a dataset and apply PCA dan Fuzzy C-Mean, the number of class figure out also decrease, but the number of class figure out in this case still higher than in classification which used another dataset.

## 4. Conclusion

K-NN is one of the best solution to classified the internet traffic, it presents in the accuracy result. Unfortunately K-NN have high execution time, To resolve the K-NN weakness in the performance PCA needed to carry out the reduction features and Fuzzy C-Mean algorithm to assemble a cluster prior to the classification process. The combination between PCA and Fuzzy C-Mean algorithm, K-NN algorithm would have a shorter execution time and accuracy could be increase. The future works of this research is how to improving the class identification in dataset, probably it could be done by using another feature selection algorithm such as Correlation Feature Selection or another algorithm

## References

[1] Gu C, Zhang S, Xue X. Internet Traffic Classification based on Fuzzy Kernel K-means Clustering. *International Journal of Advancements in Computing Technology (IJACT)*. 2011; 3(3): 199-209.
[2] Lou X, Li J, Liu H. Improved Fuzzy C-means Clustering Algorithm Based on Cluster Density Related Work. *Journal of Computational Information Systems*. January 2012; 8(2): 727–737.
[3] F Wang. Factor Analysis and Principal-Component Analysis. *Elsevier*. 2009.
[4] Stevens JP. Applied Multivariate Statistics for the Social Sciences. 2nd Edition. Hillsdale. NJ: Erlbaum. 1992.
[5] Andy Field. Discovering Statistic Using SPSS, 3rd Edition, Hillsdale. NJ: Erlbaum. 1992

[6] Berget I, Mevik BH, Næs T. New Modifications and Applications of Fuzzy-means Methodology. Comput. Stat. Data Anal. 2008; 52(5): 2403–2418. doi:10.1016/j.csda.2007. 10.020
[7] Zhang L, Liu Q, Yang W, Wei N, Dong D. An Improved K-nearest Neighbor Model for Short-term Traffic Flow Prediction. *Procedia-Social and Behavioral Sciences*. 2013; 96: 653–662. doi:10.1016/j.sbspro.2013.08.076
[8] Lee YH, Wei CP, Cheng TH, Yang CT. Nearest-neighbor-based Approach to Time-series Classification. *Decision Support Systems*. 2012; 53(1): 207–217. doi:10.1016/j.dss.2011.12.014
[9] Antonio T, Paramita AS. Full paper Feature Selection Technique Impact for Internet Traffic Classification Using Naïve Bayesian. *Jurnal Teknologi*. 2014; 20: 85–88.
[10] Paramita AS. Feature Selection Technique Using Principal Component Analysis for Improving Fuzzy C-Mean Internet Traffic Classification. *Australian Journal of Basic and Applied Sciences*. 2014; 8(14): 13–18.
[11] Fahad A, Tari Z, Khalil I, Habib I, Alnuweiri H. Toward an Efficient and Scalable Feature Selection Approach for Internet Traffic Classification. *Computer Networks*. 2013; 57(9): 2040–2057. doi:10.1016/j.comnet.2013.04.005
[12] Nguyen T, Armitage G. A survey of Techniques for Internet Traffic Classification Using Machine Learning. *IEEE Communications Surveys & Tutorials*. 2008; 10(4): 56–76. doi:10.1109/SURV.2008.080406