Clustering and Feature Selection Technique for Improving Internet Traffic Classification Using K-NN

Trianggoro Wiradinata and Adi Suryaputra Paramita

Abstract—**This research will use the algorithm K-Nearest Neighbour (K-NN) to classify internet data traffic, K-NN is suitable for large amounts of data and can produce a more accurate classification, K-NN algorithm has a weakness takes computing high because K-NN algorithm calculating the distance of all existing data. One solution to overcome these weaknesses is to do the clustering process before the classification process, because the clustering process does not require high computing time, clustering algorithm that can be used is Fuzzy C-Mean algorithm, the Fuzzy C-Mean algorithm does not need to be determined in first number of clusters to be formed, clusters that form on this algorithm will be formed naturally based datasets be entered, but the algorithm Fuzzy C-Mean has the disadvantage of clustering results obtained are often not the same even though the same input data, this is because the initial dataset that of the Fuzzy C-Mean is not optimal, to optimize initial datasets in this research using feature selection algorithm, after main feature of dataset selected the output from fuzzy C-Mean become consistent. Selection of the features is a method that is expected to provide an initial dataset that is optimum for the algorithm Fuzzy C-Means. Algorithms for feature selection in this study used are Principal Component Analysis (PCA). PCA reduced non significant attribute to created optimal dataset and can improve performance clustering and classification algorithm. Results in this study is an combining method of classification, clustering and feature extraction of data, these three methods successfully modeled to generate a data classification method of internet bandwidth usage that has high accuracy and have a fast performance.**

Index Terms—**Clustering, classification, feature, bandwidth.**

I. Introduction

One of previous classification research on Internet traffic by taking the data usage overall Internet traffic done by Chengjie Gu, Shunyi Zhang, and Xiaozhen Xue, in April 2011. This research flagship is improving the Kernel algorithms for Fuzzy K-Mean which is increasing the classification accuracy compared than traditional Fuzzy K-Mean. But in the study said that the algorithm Fuzzy C Mean unable to perform optimization characteristics of the data being entered and also on Fuzzy K Mean all the features of the data is considered to have the same contribution to the cluster that will be generated. This is why the level of accuracy of clustering produced less accurate and still needs to be improved accuracy [1].

Manuscript received August 11, 2015; revised January 23, 2016. This research was supported in part by Fundamental Research Grant from Indonesian Higher Education Directorate (DIKTI).

Trianggoro Wiradinata is with the Economic Management Faculty, University of Ciputra, Indonesia (e-mail: twiradinata@ciputra.ac.id).

Adi Suryaputra Paramita is with the Information Systems Programme, University of Ciputra, Indonesia (e-mail: adi.suryaputra@ciputra.ac.id).

This occurs because the algorithm Fuzzy C Mean Kernel, many clusters were formed has been determined from the outset that as many as K. At the conclusion of this study said that they need to discover what features are suitable and appropriate to improve the accuracy of classification Internet traffic.

This research will be conducted chosen a classification algorithm that others, namely the K-Nearest Neighbor (K-NN) for internet bandwidth usage classification process, in which the difference between the K-NN and Fuzzy K Mean is on a computational algorithm, in which the K-NN calculates all distances distribution of existing data, so the results classification was formed would be more accurate because it takes into account all the possibilities that exist, because the process of rigorous computational algorithms K-NN finally have a weakness in terms of performance is the slow process of classification.

In addressing the weakness of K-NN algorithm in this study will be conducted experiments by forming first ready-classified datasets, the dataset forming process is done by clustering beforehand. Clustering process is done so that the spread of the data occurs naturally based on similarity of existing data, as the data is scattered then carried out a process of classification, clustering process is expected to accelerate the performance of K-NN algorithm. This clustering algorithm is an algorithm that meets the Fuzzy C Mean. At Algorithm Fuzzy C Mean, number of clusters to be formed does not need to be determined in advance, so the number of clusters that formed later would show the grouping of data occurs. In a recent study in 2012 conducted by Xiaojun Lou, Junying Li, and Haitao Liu still stated that the Fuzzy C Mean generally have a weakness for the output partition / cluster for the same dataset [2].

Based on these previous researches, there is an opportunity to study Internet traffic classification using machine learning algorithms. In this research will using K-NN algorithm and Fuzzy C Mean algorithm. One advantage of this algorithm is the number of class no need specified from the beginning such as Fuzzy K Mean algorithm. It is expected that the class is formed 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 of these studies is how the process of finding the features and precise fit.

II. Literature Review

A. Internet Traffic Classification

In this phase internet traffic classification is done by using

Fuzzy C-Mean algorithm, this phase perform calculation accuracy of classification that is generated by the Fuzzy C Mean in addition to this phase also calculate class recall and class Precision of classification that have been generated. The formula for the calculation of accuracy, Class Precision and Class Recall is as below.

B. Discriminant Feature Selection

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

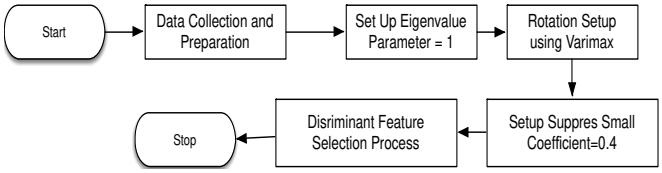

Fig. 1. Discriminant feature selection procedures.

In Fig. 1, shown that Principal Component Analysis used to find discriminant feature in some dataset. The selection method used to find correlated data and finally established a correlation matrix, while the eigenvalue set to 1 which the eigenvalue is the number of variants associated with factors used for eigenvalues is worth more of 1, 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 article titled F. Wang factor Analysis and Principal Component Analysis in 2009 [3], whereas 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 load 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 Fig. 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].

C. K-Nearest Neighbor (K-NN)

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 [6], [7].

III. Research Methodology

The purpose of this study is how to improving K-NN Classification accuracy and performance by using Fuzzy C-Mean Clustering and Principal Component Analysis (PCA). PCA first technique for analyzing internet traffic dataset and to find the discriminant feature. 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. The research methodology to achieve these research objectives, as shown in Fig. 2.

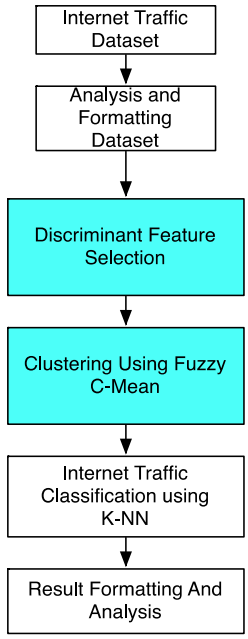

Fig. 2. Research methodology.

The scientific contribution in this research shown in the blue box on Fig. 2. The first phase of this research is collecting internet traffic dataset which used another internet traffic research, 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 next phase after data collected is find the best correlated dataset through discriminant features using Principal Component Analysis (PCA). The next process after the dataset already formed, is clustered the dataset using

Fuzzy C-Mean algorithm. Fuzzy C-Mean will grouping and disseminate dataset into a group that has the same data characteristics. After that K-NN will classified the dataset into classification class. The result from internet traffic classification will be evaluate and monitoring after K-NN classification done.

IV. Experimental Result

In this research used 3 dataset from mooreset (http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmetrics/) the dataset is dataset 2, dataset 3 and dataset 10. This dataset is used in previous research. The class and number of data from each dataset is present in Table I.

Table I shown that dataset 10 is at the most number of flow, the www class is the majority of class in all dataset, the dataset 2 is consist of 12 class, dataset 3 is consist of 10 class and dataset 10 is consist of 11 class. The class dataset 2 is the most varied. The experiment results in this classification model shown in Table II to Table VI.

TABLE I: 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 NUMBER OF FLOW	23801	22931	65036

TABLE II: Execution Time Result

Algorithm	Dataset		
	Dataset 2	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 III: Accuracy Result

Algorithm	Dataset		
	Dataset 2	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 IV: Classification Summary Result Dataset 2

Algorithm	Dataset		
	Traditional K-NN	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 V: 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 VI: 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 I shown that dataset 10 has a 65036, dataset has 23801, dataset 3 has 22932. The data is majority in WWW class. Fuzzy C-Mean performed cluster formation process before conducted by K-NN classification, clustering by Fuzzy C-Mean is expected to improve the performance and execution time of the algorithm K-NN. PCA algorithm transform the dataset into new dataset, the new dataset is create after PCA made dimensional reduction for 3 dataset. Table II shown that Fuzzy C-Mean gave significant impact for K-NN Classification in dataset 10, the execution time decreases almost 400 seconds. But unfortunately K-NN not able to improve the computation time in dataset 2 and dataset 3, it can be conclude that K-NN will have significant improvement in computation while number of flow in dataset more than 30.000 data The feature reduction which done by PCA shows the most significant impact, Table II shown execution time improvement is more less 70%. K-NN accuracy is also improve when PCA applied as feature selection for K-NN, Table III shown that accuracy for dataset

2 and dataset 3 increase more than 1%. Table IV to Table VI shown that class precision have improvement when PCA applied as feature selection. Summary of this experimental result is Fuzzy C-Mean has significant improvement in execution time for dataset which have large number of data flow, the most significant improvement for execution time is when PCA applied as feature selection, the dimensional reduction from PCA removed non discriminant feature of dataset which give significant improvement for execution time, class precision and accuracy [8]-[11].

V. Conclusion

K-NN is one of the best algorithms for internet traffic classification, it shows in the accuracy result. Unfortunately K-NN have high execution time, To improve the performance of K-NN algorithms needed to carry out the reduction features PCA to removed non discriminant feature and Fuzzy C-Mean algorithm to form a cluster prior to the classification process, with the combination of two algorithm, K-NN algorithm would have a shorter execution time and accuracy can be increased.

VI. Future Works

The future works of this research is how to improving the class figure out in dataset, probably it could be done by using another feature selection algorithm such as Correlation Feature Selection or another algorithm.

Acknowledgment

We would like to thank to Indonesian Higher Education and Research for this opportunity and research grant, and also for University Of Ciputra for research facility.

References

[1] C. Gu, S. Zhang, and X. Xue, "Internet traffic classification based on fuzzy kernel K-means clustering," *International Journal of Advancements in Computing Technology*, vol. 3, no. 3, pp. 199-209, 2001.

[2] X. Lou, J. Li, and H. Liu, "Improved fuzzy C-means clustering algorithm based on cluster density related work," *Journal of Computational Information Systems*, pp. 727–737, January 2012.

[3] F. Wang, *Factor Analysis and Principal-Component Analysis*, Elsevier, 2009.

[4] J. P. Stevens, *Applied Multivariate Statistics for the Social Sciences*, 2nd ed. Hillsdale, NJ: Erlbaum, 1992.

[5] A. Field, *Discovering Statistic Using SPSS*, 3rd ed. Hillsdale, NJ: Erlbaum, 1992.

[6] L. Zhang, Q. Liu, W. Yang, N. Wei, and D. Dong, "An improved K-nearest neighbor model for short-term traffic flow prediction," *Proceeding of Social and Behavioral Sciences*, vol. 96, pp. 653–662, 2013.

[7] Y. H. Lee, C. P. Wei, T. H. Cheng, and C. T. Yang, "Nearest-neighbor-based approach to time-series classification," *Decision Support Systems*, vol. 53, no. 1, pp. 207–217, 2012.

[8] A. S. Paramita, "Feature selection technique using principal component analysis for improving fuzzy C-mean internet traffic classification," *Australian Journal of Basic and Applied Sciences*, vol. 8, no. 14, pp. 13–18, 2014.

[9] T. Antonio and A. S. Paramita, "Full paper feature selection technique impact for internet traffic classification using naïve Bayesian," *Jurnal Teknologi*, vol. 20, pp. 85–88, 2014.

[10] A. Fahad, Z. Tari, I. Khalil, I. Habib, and H. Alnuweiri, "Toward an efficient and scalable feature selection approach for internet traffic classification," *Computer Networks*, vol. 57, no. 9, pp. 2040–2057, 2013.

[11] T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Communications Surveys & Tutorials*, vol. 10, no. 4, pp. 56–76, 2008.

Trianggoro Wiradinata is the vice dean of the Management and Business School, University of Ciputra Surabaya, his main research is in IT adoption and IT evaluation in small medium enterprise in Indonesia.

Adi Suryaputra Paramita is the head of the Management Information Systems Department, University of Ciputra Surabaya, his main research include internet traffic classification, scientific algorithm for clusterint and classification.