AENSI Journals

Australian Journal of Basic and Applied Sciences

ISSN:1991-8178

Journal home page: www.ajbasweb.com

Feature Selection Technique Using Principal Component Analysis For Improving Fuzzy C-Mean Internet Traffic Classification

Adi Suryaputra Paramita

Information Technology Department, Faculty of Entrepreneurial Creative Industry, University of Ciputra, UC Town Citraland, Surabaya, Indonesia

ARTICLE INFO

Article history:
Received 25 June 2014
Received in revised form
8 July 2014
Accepted 10 August May 2014
Available online 30 September 2014

Keywords:
Traffic, Feature, Classification, PCA, Accuracy

ABSTRACT

Background: Feature selection for within datasets has an important role in the process of internet traffic classification , feature selection in the presence of more precise data would make the internet traffic classification more accurate and can capable of providing more precise information for optimization of internet bandwidth. One of the important things in the feature selection technique is how to choose the discriminant feature whichcould in turn deliver better results during the classification process. **Objective:** Choose the discriminant features in internet traffic dataset using Principal Component Analysis (PCA) technique to improve classification accuracy **Results:** PCA technique improving the accuracy of Internet traffic classification using Fuzzy C – Mean,in this research the accuracyobtain 88.49 %and better than another feature selection technique **Conclusion:** PCAtechnique can be one of the solution for pre processing feature selection before internet traffic classification, because PCA can choose discriminant and principal feature in internet traffic classification dataset

To Cite This Article: Adi Suryaputra Paramita , Feature Selection Technique Using Principal Component Analysis For Improving Fuzzy C-Mean Internet Traffic Classification. ***Aust. J. Basic & Appl. Sci.,*** *8(14): 13-18, 2014*

INTRODUCTION

In general, internet traffic classification methods can be classified into port-based method, payload-based, or heuristic protocol behavior analysis-Based classification method and classification based on statistical data. However due to the development of applications that using flexible port and the number of applications that are trying to avoid port-based and payload-basedclassification method, a more intelligent method to identify the type of applications that use the available bandwidth. Is necessary to use Machine Learning algorithms (Abuagla&Mohd Nor, 2009). Previous research on the internet classification is a classification by using The algorithm Self Organizing Map (SOM) is performed at Monash University . Wanga *et al*. introduce clustering mechanism based on the volume of internet bandwidth usage(Wanga, Abrahamb, and KaSmitha, 2005), while a few art using naïve bayes algorithm. The feature selection is done in order to classify the data generated which could have potentially members that have the same features . At the conclusion state that the method of features selection can produce a good performance for the detection of the use of the Internet traffic is still modest complexity. The use of the Internet traffic such as databases , games , and attacks such as worms and viruses are not taken into account (Parka, Tyanb, and Kuoa, 2006).Internet traffic classification research by taking the data usage overall Internet traffic carried by Chengjie GU ,Shunyi ZHANG , and Xiaozhen XUE , in April 2011. At these research internet traffic classification done by using Fuzzy K Mean Algorithm with making changes to the kernel algorithm . In that study it was found that the accuracy of the classification algorithm from Kernel Fuzzy K Mean Algorithm has been increased compared to Fuzzy K Mean usual . But in that research state that the Fuzzy K Mean Algorithm can't perform the optimization characteristics of the data being entered and also on Fuzzy K Mean all the features of the data are considered to have the same contribution to the cluster that will be generate. This is what causes the accuracy of the classification is less accurate and its accuracy still needs to be improved . This happens because the Kernel Fuzzy K Mean Algorithm , the number of clusters that are formed have been determined from the outset that as many as K. At the conclusion of this study says that still needs to be done a study to find the features that are suitable and appropriate to improve the accuracy of classification internet traffic. From the last internet traffic classification result.

Based on these previous research, there is an opportunity to study Internet traffic classification using machine learning algorithms .In this research will using Fuzzy C Mean algorithm. One advantage of this

Corresponding Author: Adi Suryaputra Paramita, Senior Lecturer, Business Information Systems Programme, Department of Information Technology, University of Ciputra, Surabaya, Indonesia.
E-mail: adi.suryaputra@ciputra.ac.id

Australian Journal of Basic and Applied Sciences, 8(14) Special 2014, Pages: 13-18

algorithm is the number of classno 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 . In previous research, the use of feature extraction using Sequential Feature Selection (SFS) and Correlation Feature Selection (CFS) . Both of this feature selectionalgorithm will only find a collection of the best sub - sets of data from the existing data and not look for features that are discriminant and principal of a body dataset. Based on this problem this research will using Principal Component Analysis technique in order to find discriminant and principal feature for internet traffic classification.

Research Methodology:

The purpose of this work is how to improving Fuzzy C-Mean accuracy by using Principal Component Analysis (PCA) Technique to find the best dataset and how to improving feature selection method by using Principal Component Analysis as the first technique for analyzing internet traffic dataset and to find the discriminant feature. to achieve these objectives, the research methodology as shown in Figure 1.

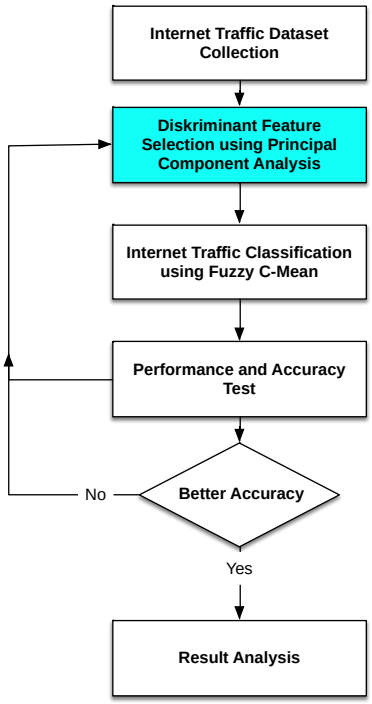

Fig.1: Research Methodology.

The main contribution in this research is shown in the blue box on figure1. The first phase of this research is collecting internet traffic dataset in this research, this research will using mooreset internet traffic dataset, this data is used in previous research, the internet traffic data is collected fromhttp://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. After the discriminant features in the dataset obtained, the next process is to classify the Internet traffic dataset using Fuzzy C - Mean algorithm. The result from internet traffic classification will be evaluate and monitoring after Fuzzy C-Mean classification done

Discriminant Feature Selection:

Dataset used in this study consists of 244 attributes and contains 65036 records, the selection of the discriminant features that will use the methods of Principal Component Analysis (PCA), discriminant feature selection procedures will be seen in the picture below

Australian Journal of Basic and Applied Sciences, 8(14) Special 2014, Pages: 13-18

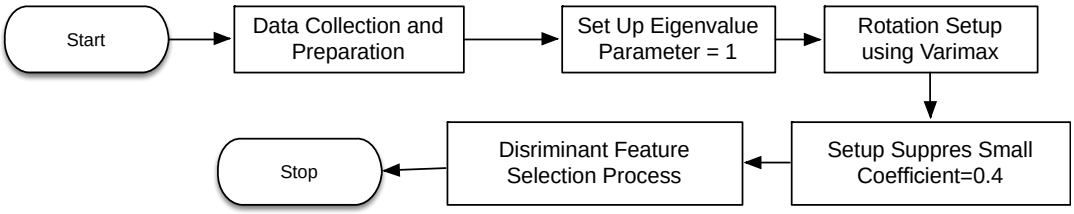

Fig. 2: Discriminant Feature Selection Process.

In Figure 2. Seen that the discriminant feature selection method using Principal Component Analysis use to find data that are correlated to finally established a correlation matrix , while the eigenvalue used is one in which the eigenvalue is the number of variants associated with factors used for eigenvalues is worth more of 1, it will have an impact on the features that 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 on in 2009, whereas varimax rotation was 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 Figure 4 . Seen that supressed 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) , this is related to the significant results quoted by Andy Field in his book Discovering Statistics Using SPSS(1992)

Discriminant feature selection result showed 244 of the data perform reduction to only 37 new features, a new feature is considered a main feature and principle. New features created by the PCA is the processing of the 244 original features, where the first new feature can be formed from the merger of some of the old features. In the feature extraction data using PCA there are some attributes / features in the dataset reduced by PCA directly and not be forming attribute / new features created by the PCA, the features are reduced by PCA algorithm is a feature no 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 75, 76, 77, 78, 91, 92, 97, 98, 103, 104, 105, 106, 143, 174, 181, 188, 193, 219, 229, 239. Feature names are reduced by PCA can be seen in table 1 and table 2 below

Table 1 : Summary Of Feature Reduced By PCA

No	Cause Of Feature Reduction	Number Of Feature
1	Unidentified Data	7
2	Percentage of data equal value 100%	11
3	Percentage of data equal value 95-99%	7
4	Percentage of data equal value 50-94%	6

Table 2: Features Not Involved In New Feature.

Feature Number	Feature Name	Modus(Mode)	The Amount of Data	Data Percentage
1	source_port	80	53767	82.67%
102	missed_data_b a	0	64702	99.49%
209	Time_since_last_connection	Unidentified		
210	No_Transitions_bulk/trans	0	43865	67.45%

In Table 2 it appears that there are four features that although not reduced but not be forming for new features created by the PCA because the value of its contribution under 0.4 and does not qualify as disclosed by JP Stevens, thus indirectly also features reduced . The features available in table 2 when further analyzed the features of existing data in table 2 is a feature that is information alone is not an Internet bandwidth usage data so that by the PCA algorithm considered these features do not contribute significantly to form components the main of this dataset

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 and Class Precision Recall is as below

$$\boldsymbol{Accuracy} = \frac{\sum_{k=1}^{n} TP}{\sum_{k=1}^{n} TP + \sum_{k=1}^{n} FP} \text{x } 100\%$$

$$\boldsymbol{Precision} = \frac{TP}{TP+FP} \text{x } 100\%$$

Australian Journal of Basic and Applied Sciences, 8(14) Special 2014, Pages: 13-18

$$\boldsymbol{Recall} = \frac{TP}{TP+FN} \text{x } 100\%$$

True Positive (TP) is the number of unclassified data in the correct class. False Positive (FP) is the amount of data that is considered to be in the wrong class by the application when the data should already be in the correct class. False negative is the amount of data that was in the wrong class. The results of the implementation of the Fuzzy C-Mean algorithm was formed 11 class the class are:

1. WWW
2. P2P
3. MAIL
4. SERVICE
5. FTP-PASSIVE
6. ATTACK
7. IOTERACTIVE
8. DATABASE
9. FTP-CONTROL
10. FTP-DATA
11. GAMES

Experimental Result:

Based on the experimental results and the Selection of the right main features using Principal Component Analysis (PCA) will contribute significantly to the accuracy of clustering , it is seen that there are differences in accuracy of 5.23 % of accuracy generated by the extraction of features with the first election of the major features and no prior major feature selection , feature extraction while the right will also improve the accuracy of a clustering . When compared with previous studies an increase in the accuracy of the method proposed in this study there was an increase of 0.43 % . Significant increases in this study is the method proposed in this study could increase the number of clusters occupied by the data correctly , in this study the number of clusters assigned the correct data is 7 clusters of 11 clusters are formed , while the methods used in previous research only able to put the data on a cluster that is true as much as 5 clusters. Disadvantage of the method in this research compared to the previous research method is the execution time, the execution time in this research method is slower 62 second compared to the previous method. The experimental result is shown on table 4 until table 8 below

Table 4: Execution Time ResultComparation.

Algorithm	Execution Time (Second)
Fuzzy C-Mean Pure	166
Fuzzy C-Mean With PCA Discriminant Feature Selection	211
Fuzzy Kernel K-Means (Previous Research)	149

Table 5: Internet Traffic Classification Accuracy ResultComparation.

Algorithm	Accuracy
Fuzzy Kernel K-Means (Previous Research)	88.06%
Fuzzy C-Mean Pure	83.73%
Fuzzy C-Mean With PCA Discriminant Feature Selection	88.49%

Tabel 6: Class Recall Result Comparation.

Class	Fuzzy C-Mean Pure	Fuzzy C-Mean With PCA Discriminant Feature Selection	Fuzzy Kernel K-Means (Previous Research)
WWW	99.82 %	97.51 %	95.84 %
P2P	18.58 %	5.45 %	1.12 %
MAIL	0 %	65.23 %	72.22 %
SERVICE	0 %	28.30 %	0 %
FTP-PASSIVE	0 %	13.62 %	0 %
ATTACK	0 %	0 %	0 %
IOTERACTIVE	0 %	0 %	0 %
DATABASE	0 %	0 %	0 %
FTP-CONTROL	0 %	43.21 %	32.10 %
FTP-DATA	0 %	0.51 %	51.69 %
GAMES	0 %	0 %	0 %

Table 7: Class Precision Result Comparation.

Class	Fuzzy C-Mean Pure	Fuzzy C-Mean With PCA Discriminant Feature Selection	Fuzzy Kernel K-Means (Previous Research)
WWW	83.92%	90.43 %	91.13%
P2P	41.13%	40.96 %	38.89%
MAIL	0 %	71.68 %	73.04%
SERVICE	0 %	74.07 %	0 %
FTP-PASSIVE	0 %	83.33 %	0 %
ATTACK	0 %	0 %	0 %
IOTERACTIVE	0 %	0 %	0 %
DATABASE	0 %	0 %	0 %
FTP-CONTROL	0 %	38.89 %	4.53 %
FTP-DATA	0 %	7.14 %	45.27 %
GAMES	0 %	0 %	0 %

Tabel 8: Summary Internet Traffic Classification Method Comparison.

	Fuzzy C-Mean Pure	Fuzzy C-Mean With PCA Discriminant Feature Selection	Fuzzy Kernel K-Means (Previous Research)
Number of class figure out	2	7	5
Minimum recall value	18.58 %	0.51 %	1.12 %
Minimum precision value	41.13 %	7.14 %	4.53 %
Maximum recall value	83.92 %	97.51 %	95.48 %
Maximum precision value	99.82 %	90.43 %	91.13 %
Accuracy	83.26 %	88.49 %	88.06 %

The most significance result in this research is the discriminant feature selection using Principal Component Analysis (PCA) and internet traffic classification using Fuzzy C-Mean is have significance impact for improving number of class figure out and improving classification accuracy. The accuracy in this research is increase 0.43% than previous research, and the most significance is 7 class is figure out, in the previous research method only figure out 5 class. By the experimental result the method in this research is have advantage in figure out more class number and give the higher accuracy

Conclusion:

Feature selection technique using Principal Component Analysis (PCA) in this research is shown have significance contribution for improving internet traffic classification using fuzzy c-mean, the most significance result compare than the method in previous research is the number of class figure out is increase from 5 class to 7 class, the another improvement is the accuracy in this method is increasing 0.43% from previous method. By this research is shown that Fuzzy C-Mean can represent the real data of internet traffic classification dataset, and PCA is one the best solution for discriminant feature selection or PCA is the best alternative for 1st phase filtering dataset.The opportunities for future work is improving the execution time for this method and to find out PCA contribution for another classification algorithm

REFERENCES

Abuagla, M., S. Mohd Nor, 2009. Towards a Flow-based Internet Traffic Classification for Bandwidth Optimization . International Journal of Computer Science and Security, 3(2).

Berget, I., B.H. Mevik, T. Næs, 2008. New modifications and applications of fuzzy -means methodology. Computational Statistics & Data Analysis, 52(5): 2403–2418. doi:10.1016/j.csda.2007.10.020

Budayan, C., I. Dikmen, M.T. Birgonul, 2009. Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping. Expert Systems with Applications, 36(9): 11772–11781.

Erman, J., A. Mahanti, M. Arlitt, I. Cohen, C. Williamson, 2007. Offline/realtime traffic classification using semi-supervised learning. Performance Evaluation, 64(9-12): 1194–1213.

Esbensen, K.H., 2009. Principal Component Analysis : Concept , Geometrical Interpretation , Mathematical Background , Algorithms , History , Practice. Elsevier.

Gu, C., S. Zhang, X. Xue, 2011. Internet Traffic Classification based on Fuzzy Kernel K-means Clustering. Internet Traffic Classification based on Fuzzy Kernel K-means Clustering. International Journal of Advancements in Computing Technology, 3(3): 199–209.

Karegowda, A.G., A.S. Manjunath, 2010. Comparative Study Of Attribute Selection Using Gain Ratio And Correlation Based Feature Selection. International Journal of Information Technology and Knowledge Management, 2(2): 271–277.

Lou, X., J. Li, H. Liu, 2012. Improved Fuzzy C-means Clustering Algorithm Based on Cluster Density Related Work. Journal of Computational Information Systems, 2(January), pp: 727–737.

Mingoti, S.A., J.O. Lima, 2006. Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. European Journal of Operational Research, 174(3): 1742–1759.

Parka, J., H. Tyanb, C. YaKuoa, 2006. Internet Traffic Classification For Scalable Qos Provision. IEEE Multimedia Conference and Expo.

Sun, M., J. Chen, 2011.Research of the traffic characteristics for the real time online traffic classification. The Journal of China Universities of Posts and Telecommunications, 18(3).

Wang, F., B. Rouge, 2009. Factor Analysis and Principal-Components Analysis. Elsevier, pp: 1–7.

Zhang, H., Lu, G., Qassrawi,

, Zhang, M.T., X. Yu, 2012. Feature selection for optimizing traffic classification. Computer Communications, 35(12): 1457–1471.

Wanga, X., A. Abrahamb, K. Smitha, 2005. Intelligent web traffic mining and analysis. Journal of Network and Computer Applications, 28: 147-165.

Wang, X., Y. Wang, L. Wang, 2004. Improving fuzzy c-means clustering based on feature-weight learning. Pattern Recognition Letters, 25(10): 1123–1132

Zhao, J., X. Huang, Q. Sun, Y. Ma, 2008. Real-time feature selection in traffic classification. The Journal of China Universities of Posts and Telecommunications, 15(S): 68–72.