Jurnal Teknologi

Full paper

Feature Selection Technique Impact for Internet Traffic Classification Using Naïve Bayesian

Tony Antonio*, Adi Suryaputra Paramita

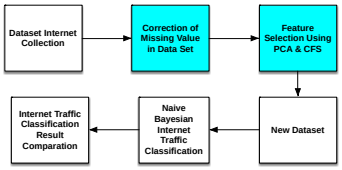
Information Technology Department, University of Ciputra, UC Town, Citraland, Surabaya, Indonesia

*Corresponding author: tonyantonio@ciputra.ac.id

Article history

Received : 15 August 2014
Received in revised form :
15 October 2014
Accepted :15 November 2014

Graphical abstract



Abstract

Feature selection technique has an important role for internet traffic classification. This technique will present more accurate data and more accurate internet traffic classification which will provide precise information for bandwidth optimization. One of the important considerations in the feature selection technique that should be looked into is how to choose the right features which can deliver better and more precise results for the classification process. This research will compare feature selection algorithms where the Internet traffic has the same correlation that could fit into the same class. Internet traffic dataset will be collected, formatted, classified and analyzed using Naïve Bayesian. Formerly, the Correlation Feature Selection (CFS) is used in the feature selection to find a collection of the best sub-sets data from the existing data but without the discriminant and principal of a body dataset. We plan to use Principal Component Analysis technique in order to find discriminant and principal feature for internet traffic classification. Moreover, this paper also studied the process to fit the features. The result also shows that the internet traffic classification using Naïve Bayesian and Correlation Feature Selection (CFS) have more than 90% accuracy while the classification accuracy reached 75% for feature selection using Principal Component Analysis (PCA).

Keywords: Feature; selection; classification; internet; traffic

1.0 INTRODUCTION

One of the objectives of internet traffic classification researches is to improve the internet traffic classification accuracy. In the past, internet traffic classification research method can be classified into port-based method, payload-based or heuristic protocol, behavior analysis-based and statistical data based methods. Due to the development of the application of flexible port, the research method has left the port-based and payload-based to focus on a more intelligent method in order to utilize the available bandwidth. Several researches can be mentioned such as Machine Learning (ML) Algorithms [1], Classification using The Algorithm Self Organizing Map Algorithm (SOM) developed by Monash University which introduced clustering mechanism based on the volume of internet bandwidth usage [2].

The feature selection is applied to classify the generated data. Part of the data members may have the same features. Although the feature selection method could give better performance for the detection of the use of the Internet but the traffic still has modest complexity. The use of the Internet traffic for database and games (which are vulnerable to worms and viruses) are not taken into account [11].

Significant research in feature selection for Machine Learning is done by Zhao Jing-jing, Huang Xiao-hong in 2008. The result clearly explains that feature selection is the most important step in ML. Good feature does not only improve the accuracy of algorithms but also improve the computational performance [3]. Gu reported that there still need more work to find the features that are suitable and appropriate to improve the accuracy of classification internet traffic. [7].

This paper will compare 2 feature selection algorithms if the Internet traffic has same correlations that could fit into the same class. Formerly, the Correlation Feature Selection (CFS) is used in the feature selection to find a collection of the best sub-sets data from the existing data but failed to find the discriminant and principal of a body dataset. We plan to use Principal Component Analysis technique in order to find discriminant and principal feature for internet traffic classification. Moreover, this paper also studied the process to fit the features.

2.0 RESEARCH METHOLODY

The purpose of this work is to investigate the feature selection technique effect for naïve Bayesian internet traffic classification. The block diagram of the research methodology is shown in Figure 1. The second and the third block (blue ones) indicate the main contribution of the work.

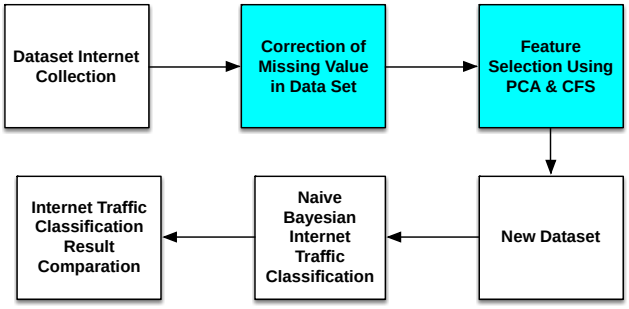


Figure 1 Research methodology

The first phase is to collect internet traffic dataset. The Moore set internet traffic dataset which has been used in previous research, is collected from http://www.cl. cam.ac.uk/research/srg/ netos/nprobe/data/papers/sigmetrics/. The next phase is to find the selected good features in the Internet traffic dataset. The Principal Component Analysis (PCA) and Correlation Feature Selection (CFS) are applied to find discriminant feature. When the good features in the dataset has been obtained, the next process is to classify the Internet traffic dataset using Naïve Bayesian. The result from internet traffic classification will be evaluated and monitored in the last phase (refer Block 6 in Figure 1).

2.1 Principal Component Analysis

Esbensen, in (Esbensen and Rehearsal, 2009), explained that the main component analysis (PCA) is a multivariate data analysis method mostly used for exploratory analysis of data, outlier detection, rank (dimension) reduction, graphical clustering, classification, and regression. The proper understanding of PCA is a prerequisite for the controlling other latent variable methods, including Principal Component Analysis regression, multivariate calibration and classification. Current use of PCA is associated with the latent data structure visualization with a graphical plot. Since PCA allows interpretation based on all variables simultaneously, it will lead to a deeper understanding which one of individual variables is possible.

PCA mostly used as the first data analysis conducted on multivariate data sets, although further data analysis with other methods even more advanced one may be required. PCA is designed to model the data. This model is characterized by the correlation between some nontrivial. For scientific data sets at large, such as the natural sciences, industry, and technology all of the variables are involved.

One of the important types of data set that does not comply with these prerequisites is the orthogonal experimental design. The minimal notation and nomenclature which are introduced in this method is based on standard linear algebra notation and traditions that have evolved in chemometrics. It is considered as a starting value for a geometric interpretation of PCA as a method of projection, followed by a short mathematical background and some insights about the algorithms and their historical development. For example when a data set is didactical, a realistic size is used to describe a typical way that is applied in the PCA. The data for PCA should be collected in a two-way arrays or matrices, called X, where the column vector represents the 'variable' (eg attributes, wavelength, retention time, parameters of physical / chemical, toxicity values, and biological responses), and the row vector represents the measured variable component, which is often also referred to as a case, the sample, measurement, and so on. We have to choose the appropriate variables and objects for subsequent data analysis in accordance with the purpose of the research conducted.

The main formula of Principal Component Analysis and PCA Projection are shown in Figures 2 and 3.

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \cdots + \mathbf{t}_A\mathbf{p}_A^{\mathrm{T}} + \mathbf{E} \quad \text{Equation (1)}$$

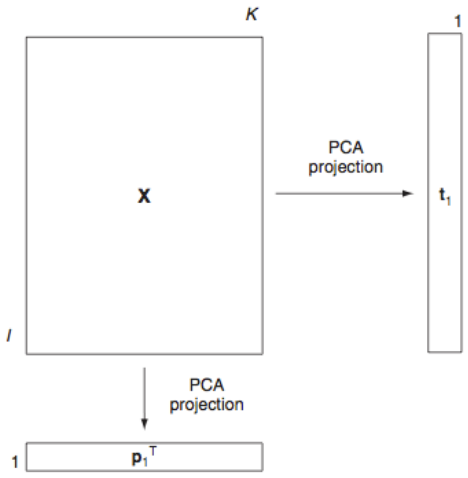


Figure 2 PCA projection from matrix X into 2 vector

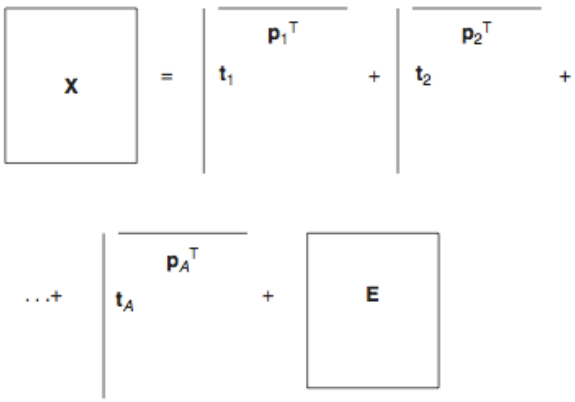


Figure 3 PCA new component development from matrix X

2.2 Correlation Feature Selection

Correlation-based Feature Selection (CFS) is a heuristic evaluation that takes into account the benefit of individual features for predicting the class along with the level of inter-correlation between them. CFS puts a high score as a subset of data that contains features that highly correlated with the inter-class but have a low correlation with each other (Zhao, J., Huang, X., Sun, Q., & Ma, Y. 2008). CFS evaluates a subset of attribute values by considering the individual predictive ability of each feature data and the level of redundancy. The correlation coefficient is used to estimate the relationship between a subset of attributes and classes, as well as the correlation between features. Relevance of a group of features increases due to the correlation between features and feature classes but on the other hand it will decrease due to the increasing of the inter-correlation. CFS is used to determine the best feature subset and is usually combined with search strategies such as forward selection, backward elimination, two-way search and genetic search. The formula of Correlation-based feature selection is shown in Equation 2 (Karegowda, A. G., Manjunath, A. S., Ratio, G., & Evaluation, C. F. 2010).

$$r_{sk} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad \text{Equation (2)}$$

$$FS = \max_{Sk}\left[\frac{r_{cf1}+r_{cf2}+\cdots+r_{cfk}}{\sqrt{k+2(r_{f1f2}+\cdots+r_{fiji}+\ldots+r_{fkf1})}}\right] \quad \text{Equation (3)}$$

3.0 EXPERIMENTAL

In this phase internet traffic classification is processed by Naïve Bayesian algorithm. Following the calculation accuracy of classification that is generated by the Naïve Bayesian, the algorithm will 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 shown in Equation 4.

$$\text{Accuracy} = \frac{\sum_{k=1}^{n} TP}{\sum_{k=1}^{n} TP + \sum_{k=1}^{n} FP} \text{ x } 100\% \quad \text{Equation (4)}$$

$$\text{Precision} = TP/(TP+FP)\text{x } 100\% \quad \text{Equation (5)}$$

$$\text{Recall} = TP/(TP+FN)\text{x } 100\% \quad \text{Equation (6)}$$

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 implementation of the Naïve Bayesian algorithm resulted in the formation of the following 11 classes:

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

The internet dataset in this research consists of 248 attributes and 65036 records. The number of flows of this internet traffic dataset is shown in the Table 1. The feature selection process will reduce the attribute of the internet dataset. The result of the experimental of feature selection is shown in Table 2.

Table 1 Number of flows in data sets

Flow Classes	Numbers
WWW	54436
MAIL	6592
FTP-CONTROL	81
FTP-PASSIVE	257
ATTACK	446
P2P	624
DATABASE	1773
FTP-DATA	592
SERVICES	212
IOTERACTIVE	22
GAMES	1

Table 2 Experimental result tables

Feature Selection Methods	Number of Attributes Retained/Selected	Number of Attributes Reduced
Principal Component Analysis (PCA)	68	180
Correlation Feature Selection	7	241

All the feature selection method is done by Weka as computational tools. The results show that PCA has reduced the feature to 68. The PCA feature reduction methodology is shown in Figure 4.

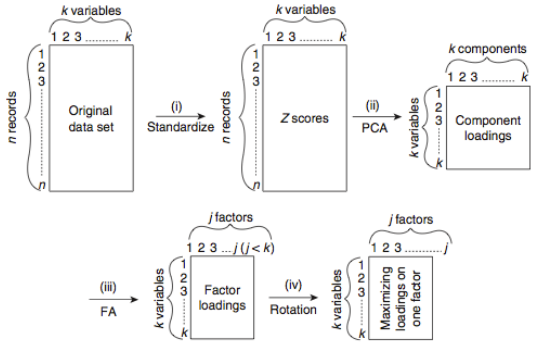


Figure 4 PCA dimensional reduction method

CFS only chose 7 from the 248 features and selected the best subset of the feature. Therefore, there are only 7 features with good correlation and that's the best subset. The Correlation Feature Selection (CFS) Methodology shown in Equation 2 and Equation 3.

4.0 RESULTS AND DISCUSSION

The results of the experiment are shown in Tables 3 to 7. These results show that the selection of the good features using Principal Component Analysis (PCA) and Correlation Feature Selection (CFS) will significantly increase the accuracy of classification. The differences is more than 20% while the CFS method gives the best impact for naïve Bayesian internet traffic classification.

Table 3 Internet traffic classification results

Dataset	No Feature Selection	PCA Accuracy	CFS Accuracy	PCA+CFS Accuracy
Dataset 10	56.1074%	75.6212%	93.8357%	83.5629%

Table 4 Detail internet traffic classification results

Results	No Feature Selection	PCA	CFS	PCA+CFS
True Positive (TP) Rate	56,1%	75,6%	93,8%	83,6%
False Positive (FP) Rate	2,6%	11,6%	25%	25,4%
Avg. Precision	97,1%	92,8%	92,7%	90,1%
Avg. Recall	56,1%	75,6%	93,8%	83,6%
F-Measure	70%	82,5%	92,9%	85,2%
ROC Area	94,4%	88,7%	92%	88%

Table 5 Internet traffic classification class precision

Class	No Feature Selection	PCA	CFS	PCA+CFS
WWW	99%	96.8%	94.5%	94%
MAIL	99.5%	79.9%	95.4%	77.3%
FTP-CONTROL	3.6%	3%	0%	0%
FTP-PASSIVE	25.1%	13.9%	77%	17.6%
ATTACK	1.6%	16.9%	1%	16.4%
P2P	20.3%	11.7%	41.5%	13.6%
DATABASE	100%	100%	98.6%	100%
FTP-DATA	97.3%	71.4%	62.4%	57.6%
SERVICES	68.9%	1.8%	0%	4.4%
IOTERACTIVE	5.4%	2.9%	0%	2.5%
GAMES	0%	0%	0%	0%

Table 6 Internet traffic classification class recall

Class	No Feature Selection	PCA	CFS	PCA+CFS
WWW	54.8%	80.5%	99.5%	91.6%
MAIL	48.8%	36.1%	71.4%	24.8%
FTP-CONTROL	91.4%	4.9%	0%	0%
FTP-PASSIVE	76.7%	27.2%	33.9%	25.7%
ATTACK	90.1%	69.7%	0.7%	69.7%
P2P	36.5%	33%	7.1%	22.6%
DATABASE	98.4%	98.3%	93.4%	98.3%
FTP-DATA	97.1%	74.2%	63.7%	73.5%
SERVICES	91%	85.4%	0%	80.2%
IOTERACTIVE	68.2%	13.6%	0%	4.5%
GAMES	0%	0%	0%	0%

Table 7 Internet traffic classification method summary

Class	No Feature Selection	PCA	CFS	PCA+CFS
Number of class figure out	10	10	7	10
Minimum recall value	0%	0%	0%	0%
Minimum precision value	0%	0%	0%	0%
Maximum recall value	97.1%	98.3%	99.5%	98.3%
Maximum precision value	100%	100%	98.6%	100%
Accuracy	56.1074%	75.6212%	93.8357%	83.5629%

The result shows that the discriminant feature selection using Principal Component Analysis (PCA) improves the number of class figure out and class precisions as well. By using PCA there are 10 class figures out, the maximum recall value is 98.3% and maximum precision value is 100%. Meanwhile CFS gives the main contribution for improving the total accuracy. Naïve Bayesian internet traffic classification with CFS as a feature selection gives 93.8357% accuracy. Another result is the possibility of the combination between PCA and CFS as an alternative method for feature selection method. Table 7 shows that n the combination between PCA and CFS give 83.5629% accuracy and have better precision value compare to CFS alone. Based on the experimental result we propose a new model for Internet Traffic Classification which is shown in Figure 5.

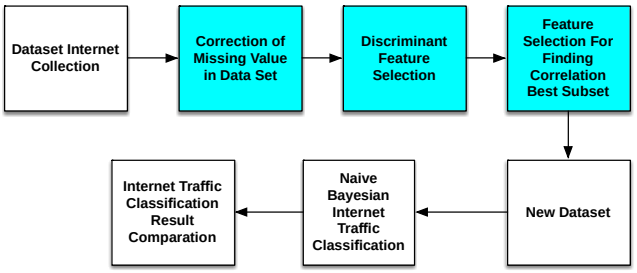


Figure 5 Internet traffic classification method purpose

Acknowledgement

Thanks for DIKTI research grant and University of Ciputra support

5.0 CONCLUSION

Feature selection technique using Principal Component Analysis (PCA) and Correlation Feature Selection (CFS) in this research has shown significant impact for improving internet traffic classification using Naïve Bayesian. The most significant result compare to the one without feature selection is in the classification accuracy. PCA improves the classification accuracy to 75.6212%. Meanwhile CFS improves the classification accuracy to 93.8357%. Since PCA is the best alternative to filter data set in the first phase, we can conclude that PCA is the best solution for discriminant feature selection. It will increase the number of class figure as the output. As for CFS, it is one of the best alternatives for figuring out the subset in dataset. Both PCA and CFS can be combined for feature selection methods. Future works suggested is to improve the feature selection method and combine it with another dimensional reduction algorithm.

References

[1] Mohd, A. B. 2009. Towards a Flow-based Internet Traffic Classification for Bandwidth Optimization. *International Journal of Computer Science and Security*. 3(2): 146–153.

[2] Wang, X., Abraham, A., & Smith, K. 2005. Intelligent web traffic mining and analysis. *Journal of Network and Computer Applications*. 28(2): 147–165. doi:10.1016/j.jnca.2004.01.006.

[3] Zhao, J., Huang, X., Sun, Q., & Ma, Y. 2008. Real-time feature Selection in Traffic Classification. *The Journal of China Universities of Posts and Telecommunications*. 15(S): 68–72. doi:10.1016/S1005-8885(08)60158-2.

[4] Moore, A., Zuev, D., & Crogan, M. 2005. *Discriminators for Use in Flow-based Classification*. Queen Mary, University of London.

[5] Budayan, C., Dikmen, I., & Birgonul, M. T. 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. doi:10.1016/j.eswa.2009.04.022.

[6] Erman, J., Mahanti, A., Arlitt, M., Cohen, I., & Williamson, C. 2007. Offline/realtime Traffic Classification Using Semi-supervised Learning. *Performance Evaluation*. 64(9–12): 1194–1213. doi:10.1016/j.peva.2007.06.014.

[7] Gu, C., Zhang, S., & Xue, X. 2011. Internet Traffic Classification based on Fuzzy Kernel K-means Clustering 3. Internet Traffic Classification based on Fuzzy Kernel K-means Clustering. *International Journal of Advancements in Computing Technology*. *3*(3): 199–209. doi:10.4156/ijact.vol3.

[8] Fahad, A., Tari, Z., Khalil, I., Habib, I., & Alnuweiri, H. 2013. Toward an Efficient and Scalable Feature Selection Approach for Internet Traffic Classification. *Computer Networks*. 57(9): 2040–2057. doi:10.1016/j.comnet.2013.04.005.

[9] Lee, Y. H., Wei, C. P., Cheng, T. H., & Yang, C. T. 2012. Nearest-neighbor-based Approach to Time-series Classification. *Decision Support Systems*. 53(1): 207–217. doi:10.1016/j.dss.2011.12.014.

[10] Lin, G., Xin, Y., Niu, X., & Jiang, H. 2010. Network Traffic Classification Based on Semi-supervised Clustering. *The Journal of China Universities of Posts and Telecommunications*. 17(December): 84–88. doi:10.1016/S1005-8885(09)60577-X.

[11] Nguyen, T., & Armitage, G. 2008. A Survey of Techniques for Internet Traffic Classification Using Machine Learning. *IEEE Communications Surveys & Tutorials*. 10(4): 56–76. doi:10.1109/SURV.2008.080406.

[12] Park, J., Tyan, H., & Kuo, C. 2006. Internet Traffic Classification for Scalable QOS Provision. *2006 IEEE International Conference on Multimedia and Expo*. 1221–1224. doi:10.1109/ICME.2006.262757.

[13] Sun, M., & Chen, J. 2011). Research of the Traffic Characteristics for the Real Time Online Traffic Classification. *The Journal of China Universities of Posts and Telecommunications*. 18(3): 92–98. doi:10.1016/S1005-8885(10)60069-6.

[14] Sun, M., Chen, J., Zhang, Y., & Shi, S. 2012. A New Method of Feature Selection for Flow Classification. *Physics Procedia*. 24: 1729–1736. doi:10.1016/j.phpro.2012.02.255.

[15] Vieira, S. M., Sousa, J. M. C., & Kaymak, U. 2012. Fuzzy Criteria for Feature Selection. *Fuzzy Sets and Systems*. 189(1): 1–18. doi:10.1016/j.fss.2011.09.009.

[16] Wang, H., & Fei, B. 2009. A Modified Fuzzy C-means Classification Method Using a Multiscale Diffusion Filtering Scheme. *Medical Image Analysis*. 13(2): 193–202. doi:10.1016/j.media.2008.06.014.

[17] Wang, Y., Xiang, Y., Zhang, J., Zhou, W., & Xie, B. 2014. Internet traffic Clustering with Side Information. *Journal of Computer and System Sciences*. 80(5): 1021–1036. doi:10.1016/j.jcss.2014.02.008.

[18] Zhang, H., Lu, G., Qassrawi, M. T., Zhang, Y., & Yu, X. 2012. Feature Selection for Optimizing Traffic Classification. *Computer Communications*, *35*(12), 1457–1471. doi:10.1016/j.comcom.2012.04.01.

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

[20] Karegowda, A. G., Manjunath, A. S., Ratio, G., & Evaluation, C. F. 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.