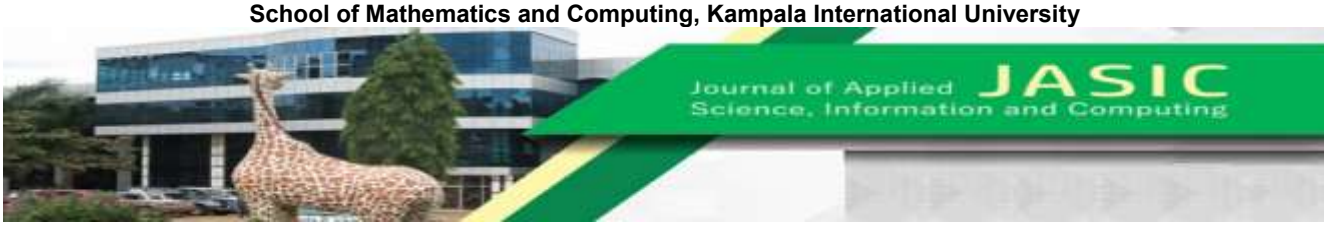
Vol. 1, No. 1 (June 2020)
Journal of Applied Sciences, Information, and Computing
School of Mathematics and Computing, Kampala International University



ISSN: 1813-3509 https://jasic.kiu.ac.ug

The Effects of Dimensionality Reduction in the Classification of Network Traffic Datasets Via Clustering

***Lawal Abbas, Maryam[1], Ajiboye, Adeleke Raheem[2]**

Department of Computer Science, Faculty of Communication and Information Sciences,
University of Ilorin, Ilorin, Nigeria.
marabblawone@gmail.com
ajibraheem@unilorin.edu.ng

Abstract
Unsupervised learning has emerged as an alternative meta-learning approach that is capable of accurately classifying the massive amount of data generated by modern-day applications. It is useful for active monitoring and provision of improved service quality by the network administrators. Extracting the optimal and most essential features with high discriminative power remains one of the critical challenges in unsupervised learning due to the absence of the class labels. The main objective of this research is to determine the effects of Dimensionality Reduction in Feature Selection via the clustering of internet traffic data sets. To achieve this overall goal, internet traffic data sets were retrieved, analyzed and clustered into application classes. A reduced form of these datasets was obtained and clustered using feature selection techniques. The results of the original and reduced data sets were compared and evaluated. The effects of two feature reduction techniques; Correlation-based Feature Selection (CFS) and Information Gain Attribute Evaluator were examined in K-means, Expectation Maximization and the Farthest-first clustering algorithms. The effectiveness of the candidate clustering algorithms was determined and the evaluation was based on overall accuracy, precision, recall, and Receiver Operating Characteristic (ROC) area metrics. Results revealed that both CFS and information gain significantly increase the performance of the three algorithms.

Keywords: Dimensionality reduction, clustering techniques, feature selection, traffic datasets, algorithm.

1. Introduction

Advancement in information and communication technology has transformed our lives such that we can rarely do anything without the internet. Most of our communications and activities are carried out online ranges from chatting, instant messaging, gaming, video and voice calls, and webinars to online transactions. It is evident that various issues and problems in the network such as security threats and bandwidth requirements will arise due to the various traffic flows from different terminals and thereby making network management cumbersome. Network operators need to be aware of the type of flows in their network to respond promptly to various issues that may arise and to be able to deliver good Quality of Service (QoS) to their clients. Network traffic classification follow a procedure that involves the packet or data in an internet traffic flow which can be categorized into various application classes such as web, multimedia, games, mail, etc (Vidyapeetham, Rajasundari, & Vidyapeetham, 2017).

The network traffic classification that involves a classical approach, assign applications to some already established port numbers. This approach has been advantageous in the past, but has been proven to be inaccurate in recent years, as many applications now adopt communication techniques with dynamic port numbers, and also using some common protocol port numbers in the application layer of the OSI model such as HTTP and FTP to avoid being detected. Although, port-based traffic classifications are simple, easy and a relatively fast, several studies have shown that they yield less accuracy in flow classification. (Dreger, Feldmann, Mai, Paxson, 2006); (Moore & Papagiannaki, 2005).

The need to resolve the problems of classification that is based on port has led to the proposal on payload-based classification. Payload-based analysis has a precise Deep Packet Inspection (DPI) mechanism which is used to know whether a packet contain some basic features of known applications. Research has shown that these approaches work so well for the current Internet traffic, including the point to point traffic (Choi, Kim, Yoon, Park, Lee, Kim, 2004). However, this approach has several shortcomings among which are: computational complexities, pre-knowledge of target application protocol and massive overheads in developing signature databases. These restrictions seem to be very crucial as the software engineers increasingly adopt data security transmission techniques, most especially the traffic encryption algorithms, for disallowing the management devices from detecting and shaping traffic accurately (Shi, Li, Zhang, Cheng, & Wu, 2017).

Despite the challenges recorded in both the port-based and payload-based approaches, previous studies have focused on achieving a much better classification approaches. Prominent among the direction taken is using the techniques of machine learning, which is based on classifying traffic flows on the use of transport layer statistic, (Anthony, Mark, Perry, & James, 2004). Statistical-based traffic classification is established on an underlying assumption that the traffic at the layer 3 of the OSI model has certain statistical features such as the packet inter-arrival time and others which include: distribution of flow duration, flow idle time, and packet lengths which are normal for differentiating some classes of internet applications from another (Nguyen & Armitage, 2008a).

One of the key challenges of traffic classification is the extraction of the optimal and relevant features for mining using the appropriate algorithm. What appears to be irrelevant and redundant features in a dataset introduce additional overheads of storage requirements and increase the time required for execution during the implementation of data mining algorithms and, thereby hinder the effectiveness of the algorithms. Selecting and identifying the relevant features for unsupervised classifications, such as clustering is generally more difficult. This is because external validation criteria, such as class labels, are not available for feature selection unlike supervised learning approach that used to be presented with class labels to serve as a guide for the feature extraction process. The features describing the dataset have an enormous impact on identifying these cluster labels, thus making the feature selection for unsupervised learning technique relatively difficult.

The dimensionality reduction is a significant step in the creation of Machine Learning model as it act as a preprocessing step, in which the redundant variables and attributes of very low relevance are removed for better quality results. A given data space is mapped onto a new, smaller dimensionality space and the original examples are then represented in the new space. There used to be a direct proportion as the dimensionality of the domain expands. This gives a corresponding rise in the number of features. The dimensionality reduction in machine learning is beneficial in so many ways. These include an improvement in classification performance, improvement in learning efficiency and to abate the complexity of the results already learned for the sake of understanding the underlying process. (Mladeni, 2006).

The focus of this paper is on data preprocessing steps, it also investigates the impact of dimensionality reduction in the classification of network traffic datasets using an unsupervised learning approach. The study further evaluates the effects of Correlation-based Feature Selection and information gain attribute evaluator on three clustering algorithms, namely, the K-means, the Expectation Maximization (EM) algorithm and Farthest first. The result of implementing each algorithm is also presented; other sections in this article are structured as follows: the various internet traffic classification techniques that use unsupervised learning approaches are reviewed in Section 2. Section 3 comprised of the theory, material used and methods employed by the clustering algorithms; the feature reduction technique is also presented. Section 4 illustrate and briefly discuss the results, while in Section 5, the conclusion and research for further studies is presented.

2. Review of Related Works

The classification of network traffic datasets using machine learning approaches has attracted a remarkable high number of researchers over the years. A learning machine is capable of learning automatically from a given experience, restructure and improve its knowledge-base. ML are employed in various sectors of life which includes medical diagnosis and treatments, banking and financial guide, search engines, pattern recognition, security and image screening, weather forecasting, sales and market diagnosis, and lots of others. The concept of machine learning for internet flow classification came into beam light in 1994 where it was utilized for intrusion detection (Frank, 1994) which served as a milestone for further works employing machine learning in Internet traffic classification.

2.1 Unsupervised learning approach

One of the earliest work on traffic classification employing unsupervised machine learning technique was carried out by Anthony et al. (2004). The study proposed the possibility of internet traffic classification using packet header statistics, attributes of less importance in the classification process are identified and removed before clustering. The study by Zander et al. (2005) proposed a traffic classification and application identification using Autoclass algorithm. The approach used an approximation based on the EM algorithm, it helps to find the global maximum. The study was based on the actual performance that employed a feature selection technique of the learning algorithm.

A technique that allows for early identification of the application associated with a traffic flow by using the information available in the header of the first few packets was proposed by Bernaille, Teixeira, Akodkenou, Soule, & Salamatian (2006). Erman & Arlitt (2006) built a descriptive model that classifies traffic data using the concept of clustering algorithms; the study explored the use of three clustering algorithms namely, k-means, DBSCAN and Auto class algorithms in classifying network traffic. Also, Liu, Li, & Li (2007), in a related research, investigates the different levels in network traffic-analysis using k-means to obtain better classification accuracies. In the process, Log transformation was used to transform the flow discriminators.

Also, in a related research carried out by Nguyen & Armitage (2008b), the study demonstrated an automated method that was based on the use of machine learning techniques for choosing the right and a sub-flows that is

representative, through which a well trained ML-based traffic classifiers model may be achieved.

2.2 Feature Selection

Feature or Attribute selection can be described as the process of choosing a subset of attributes from the original set. The number of features in data space has a direct proportion to the dimension of the data space. In other words, as the features in a domain increases, the dimensionality of that domain also expands. Feature selection techniques typically consist of four significant steps: subset formation, subset assessment or evaluation, the stopping criterion satisfaction, and validation of results (Liu & Yu, 2005).

The techniques of selecting feature is usually designed based on different evaluation criteria, and as such, can be categorized into three, namely: filter model, wrapper model, and hybrid model. The filter approach uses some inherent property of the data for selecting and evaluating features without using the underlying classification or descriptive mining techniques that will eventually be applied. The wrapper approach wraps the feature search around a predetermined learning algorithm that will ultimately be used. It utilizes the learned results as the evaluation criteria to select the features. The wrapper methods aimed at improving the data mining performance, but comparing to filter model, it appears to be more computationally costly. The hybrid model takes the advantage of the other two models such that it exploits their different evaluation metrics in several search stages. (Dy, 2008)

Researches on dimensionality reduction techniques have been on for decades (Blum & Langley, 1997). Several studies conducted around 1997 explored only a few domains with not more than 40 features. These amounts have increased substantially in the past years as most articles now explore domains with several thousands of attributes or features (Guyon & Elisseeff, 2003). One of the main reasons why dimension reduction is widely studied is the "curse of dimensionality" (Hastie, Tibshirani, 2001). An increase in the number of dimensions makes a fix data sample to become exponentially sparse. Dimension reduction involves mapping the feature space of the original onto a new, reduced dimensional space. This can be achieved either by using feature selection method which involve choosing a subset from the original feature space or by coming up with new dimensions from the original features or by combining the two processes (Mladeni, 2006).

One of the earliest work reported in the literature on dimension reduction was conducted in 2000, where a feature selection technique known as Visual Feature Subset Selection using EM clustering (Visual-FSSEM) was introduced (Dy & Brodley, 2000). This method incorporates the use of visualization techniques, clustering, and user interaction that guides the feature subset search; this also gives a better understanding of the data. Mladeni (2006) investigated the effect of using different feature subset selection methods and in the process, the real-world data was explored to show the performance of several feature selection methods on document categorization problems. The study proposed by Zhang, Lu, Qassrawi, Zhang, & Yu (2012), addressed the imbalanced class problems.

In one of the recent research conducted on feature selection, the hybrid wrapper feature selection technique was based on membership degrees which was proposed in (Georgiev, Gueorguieva, Chiappa, & Krauza, 2016). The approach outlined two major criteria for cluster evaluation and selection of an optimal clustering scheme; these are compactness and separation. Robustness of traffic classification performance are highly challenged by zero-day applications, which are not previously known in a system for classifying traffic. To address this issue, Zhang, Chen, Xiang, Zhou, & Wu (2015) combined supervised and unsupervised machine learning techniques to form a Robust Statistical Traffic Classification (RSTC) scheme. This scheme can identify the traffic of zero-day applications and capable of accurately discriminating pre-defined application classes. This study reported in (Shi et al., 2017) proposed an efficient and robust feature extraction and selection technique for traffic classification. The fractal theory was employed to express the complex nonlinear behavior of the traffic and self-similarity were intensively analyzed.

3. Material and Methods

The clustering algorithms evaluated in this research include: K-means, Expectation Maximization (EM) and Farthest First (FF) algorithms. These algorithms can easily be implemented and they produce clustering models that can be more easily interpreted. They also support the clustering of new instances.

The experiments carried out in the course of this study were conducted in the environment of WEKA (Waikato Environment for Knowledge Analysis) using version 3.8.2 software suite. It is one of the widely used tools that implement a number of algorithms for data mining and it is commonly used in the Machine Learning community. The tool was developed by the University of Waikato, New Zealand (Bouckaert et al., 2017). The experiments carried out in this study was conducted on an Acer Aspire R 14 notebook with Windows operating system (Windows 10), Intel(R) Core (TM) i5 processor (2.4GHz) of 8GB RAM.

3.1 K-means

K-means is a partitioned-based descriptive mining techniques that divides data objects into a number of specified partitions, where the individual partition denotes a cluster. Partitioned-based algorithms employ divisive techniques to cluster data by optimizing a certain criterion function that is either locally defined on a subset of the data objects or that is defined globally over all the data objects (Jain, Murty, & Flynn, 2000). K-means appears to be the simplest and most frequently used clustering algorithm which employs a squared error criterion. It partitions data

object set into k disjoint subset based on a similarity measure by minimizing the square-error. Computing the squared error involves taking the distance squared between each object and the average of its cluster. The square error can be computed by using the formula given in equation (1).

$$E = \sum_{i=1}^{K} \sum_{j=1}^{n} \left| dist(x_j, c_i) \right|^2 \quad (1)$$

Where x represents each data object, and c the respective center of each cluster. Clustering techniques follow a number of steps:

3.1.1 Place a point *k* into the space S
3.1.2 Each object should be assigned to the cluster that has the closest centroid
3.1.3 Calculate the positions of the *k* centroid again.
3.1.4 Repeat the steps taken in (ii) and (iii) until the centroids remain constant.

K-means begins with k random initial partitions called clusters, computes new centers for these clusters by minimizing the squared error and repartitions the data objects based on the new centers that are formed. This step is reiterated until the square error remains significantly the same or no data object is reassigned to a cluster. The *K-means* algorithm is very popular among the clustering algorithms, this is due to its ease of implementation, and its time complexity is *O*(n), where *n* denotes the number of data objects in the data space.

3.2 Expectation Maximization (EM)

EM clustering technique is a well-known probabilistic based method for grouping data into a number of clusters that is represented by model parameters (Jin & Han, 2016). One of the techniques employs in EM is the use of the finite Gaussian mixtures model and the technique repeatedly estimates a set of parameters, this process continues until the desired convergence value is obtained. The mixture is defined as a set of *k* probability distributions where the distribution individually corresponds to a cluster. A data object is assigned with a membership probability for each cluster. EM is an iterative and efficient procedure which can be used to calculate the Maximum Likelihood (ML) estimate even if such data contain missing or hidden values. In maximum likelihood estimation, the model parameter(s) for which the observed data are the most likely are estimated. EM algorithm clusters data based on a number of steps are as follows:

3.2.1 Initialize i to 0 and choose θ_i arbitrarily
3.2.2 (E-step): Compute $Q(\theta \mid \theta_i)$
3.2.3 (M-step): Choose θ_i+1 to maximize $Q(\theta|\theta_i)$

i. If θ_i != θ_i+1, then set i to i+1 and return to step ii

Where θ is an unknown hidden variable.

Each iteration of the EM clustering algorithm comprised of two main processes, these are: The E-step, and the M-step. The missing data are estimated in the E-step, which gives the observed data and thse present estimate of the model parameters. The likelihood function is maximized In the M-step, this is done bearing in mind that the missing data are known. Instead of using the actual missing data, the estimates of the data that are missing from the E-step are used. The convergence is very certain since the algorithm is guaranteed to ensure an increase in the likelihood at each iteration step.

3.3 Farthest first

Farthest first is one of the variants of K-means algorithm. Farthest first place the cluster center at the location that is further from the present cluster. This point is expected to be located within the data area. The farther apart points are considered to be clustered first. The characteristic features of the farthest first clustering algorithm hastening the clustering process and makes it to be relatively fast since less reassignment and adjustment of parameters is required (Revathi & Nalini, 2013). Farthest first works as a simple, fast approximate clusterer and can serve as an initializer for the simple K-means algorithm. Farthest-point heuristic-based method has the time complexity of O(nk), where n denotes the number of data objects in the dataset and k represent the number of desired clusters. The farthest-point clustering technique is particularly suitable and fast for data mining applications in large scale (Sharma, Bajpai, & Litoriya, 2012).

3.4 *Data collection*

The datasets clustered in the course of this research were retrieved from a research facility site that hosts about 1,000 users connected to the internet through a full-duplex Gigabit Ethernet link. The data which is available in real-time consists of 10 partitions and each captured at different periods of the day hours. Each of these datasets comprised of a number of objects described by a group of similar features. Each object within each dataset represents a single flow of TCP packets between client and server and is generated by different application sources. These application classes are: WWW, MAIL, P2P, FTP-CONTROL, FTP-PASV, ATTACK, DATABASE, FTP-DATA, SERVICES, INTERACTIVE, MULTIMEDIA, and GAMES. The dataset is of very high dimensions and consists of 249 attributes. The description and meaning of each of these attributes are available in a technical report (Moore, Zuev, & Crogan, 2005). Due to the large size of the datasets, only two of these datasets were used in this research. Table 1 describes some basic information about the two datasets.

Table 1. 1: Basic information about each dataset

DATASET	START-TIME	END-TIME	DURATION	FLOW
Dataset1	2003/8/20 00:34:21	2003/8/20 01:04:43	1821.8	24863
Dataset2	2003/8/20 02:45:19	2003/8/20 03:14:03	1724.1	22932

When examining the trends of flows that belong to each type of class, about 80% of the flows recorded in the datasets were WWW traffic. This means that the data sets are imbalanced and will make the machine algorithms to be biased towards the majority class. To obtain a more balanced dataset and deal with the problem of imbalance class, spread sub-sampling technique was used to sample out the dataset with a reduced number of flows. Table 1.2 summarizes the statistics of one of the experimental subsets.

Table 1. 2: Summary of Dataset1

Applications	Number of flows	% of flows
www	6000	47.4233323
Mail	4146	32.7695226
Ftp-control	149	1.17767942
Ftp-pasv	43	0.33986721
Attack	122	0.96427442
P2p	339	2.67941827
Database	238	1.88112551
Ftp-data	1319	10.4252292
Multimedia	87	0.68763832
Services	206	1.62820107
Interactive	3	0.02371167
Games	0	0
Total	12652	100%

3.5 Dimensionality Reduction Techniques

3.5.1 Correlation-based Feature Selection

The Correlation-based Feature Selection (CFS), is a multivariate filter which determines the actual worth of a subset of features. This can be achieved by taking into cognizance the predictive ability of each feature along with the level of redundancy that exist between them (Karegowda, Manjunath, & Jayaram, 2010). CFS uses correlation coefficients to estimate the relationship or correlation that exist between a subset of features and class, as well as the inter-correlations between the features. A group of features is said to be of a good feature subset when there exist some features that are highly correlated with the class or that are predictive of the class but are uncorrelated with each other (Hall, 1999). In other to determine the best feature subset for a given dataset, CFS is used in combination with search techniques, for instance: bi-directional search, backward elimination, forward selection, best-first search and genetic search.

If there is a known correlation between each feature in a dataset and the class, and the inter-correlation between each pair of feature is known, then, the correlation that exist between a composite feature that consist of addition of all the features and the class variable is predictable from the Pearson's correlation equation (Rodriguez-Galiano, Luque-Espinar, Chica-Olmo, & Mendes, 2018). This equation is given in equation 2.

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k+k(k-1)\overline{r_{ii}}}} \quad (2)$$

Where r_{zc} denotes the correlation between the additions of feature subsets and also the variable for the class, k represents the number of subset features, r_{zi} denotes the average values of the correlations between the subset features and the class variable, while r_{ii} is the mean value of the inter-correlation between subset features (Hall, 1999). The search techniques commonly used in Artificial intelligence such as Best First Search and Genetic Search were used as a search strategy with the correlation-based feature selection.

3.5.2 Information Gain Attribute Evaluator

Information gain algorithm is an algorithm used for ranking purposes, which evaluates an attribute to determine its worth, by measuring this feature with respect to the class. This measurement is done based on the entropy of a system (Alhaj, Siraj, Zainal, Elshoush, & Elhaj, 2016). Entropy is a widely used measuring approach in information theory. The technique unveils the level of purity of an arbitrary collection of examples. The technique is sometimes referred to as a measure of system's unpredictability. For instance, the entropy of Y can be represented as shown in (3).

$$H(Y) = -\sum_{y\in Y} p(y)\log_2(p(y)) \quad (3)$$

where p(y) denotes the marginal probability density function for the random variable Y.

If X is another feature and the observed values of Y in the training dataset S are partitioned, and this partitions are guided by the value of X, and the entropy of Y with respect to the partitions induced by X is less than the entropy of Y before partitioning, then there exist a relationship between features Y and X. Therefore, the entropy of Y after observing X is illustrated in (4).

$$H(Y/X) = -\sum_{x\in Y} p(x)\sum_{y\in Y} p(y/x)\log_2(p(y/x)) \quad (4)$$

Given the values of x, p (y |x) represents the conditional probability of y.

Also, given an entropy as a measure of impurity in a training dataset S, a measure can be defined in a way to reflect additional information about Y, which is provided by X that represents the amount by which the entropy of Y decreases. This measure is referred to as Information Gain and is illustrated in (5).

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y) \quad (5)$$

The information gained about Y after observing X gives the same information gained about X after observing Y. The feature ranking algorithms have lower computational complexities and do not overfit. They have been successfully implemented and proven to have worked well for certain datasets (Guyon & Elisseeff, 2003) and (Novakovic, 2009). Information gain has a major drawback; it is biased to features that has more values, even when they don't provide additional information (Novakovic, 2009; Rodriguez-Galiano et al., 2018). The ranker algorithm was utilized in conjunction with information gain to yield features that are most relevant.

3.6 Evaluation Metrics

The 10 fold cross-validation method was the validation approach used for testing of the effectiveness of the machine learning algorithms and the results were evaluated based on five standard performance metrics. These results show the overall accuracy, precision, recall, false positive rate and the ROC area.

4. RESULTS AND DISCUSSION

Figures 1, 2 and 3 shows the cluster assignments after the implementation of k-means, EM and Farthest-first respectively. Each distinct color in each figure represents a cluster, the square boxes represent errors or misclustered data points while correctly clustered data points are shown as x.

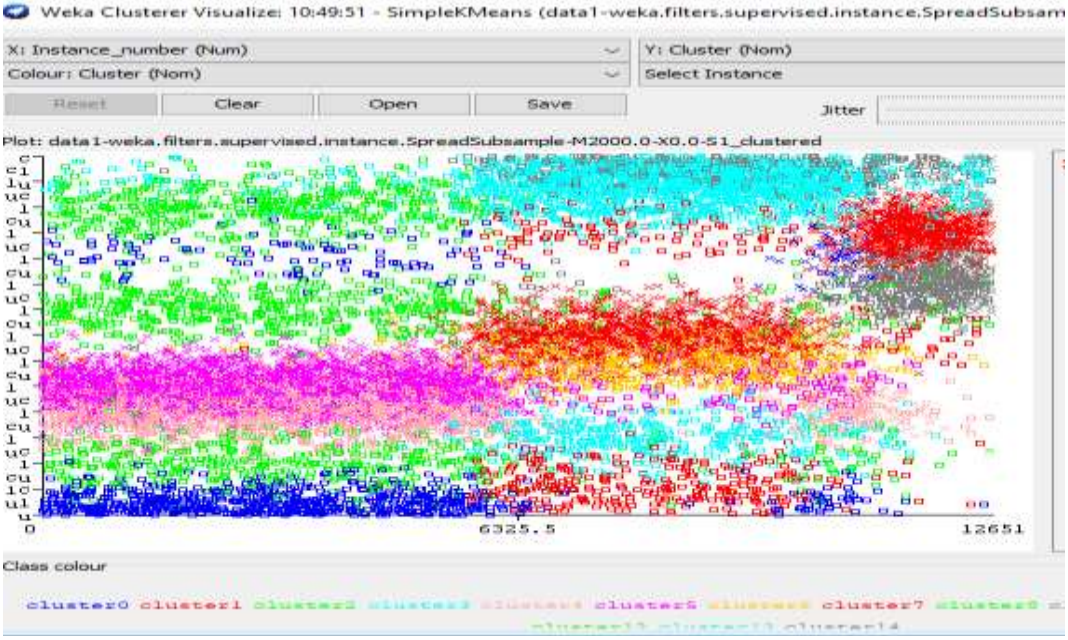


Figure 1: K-means luster assignments

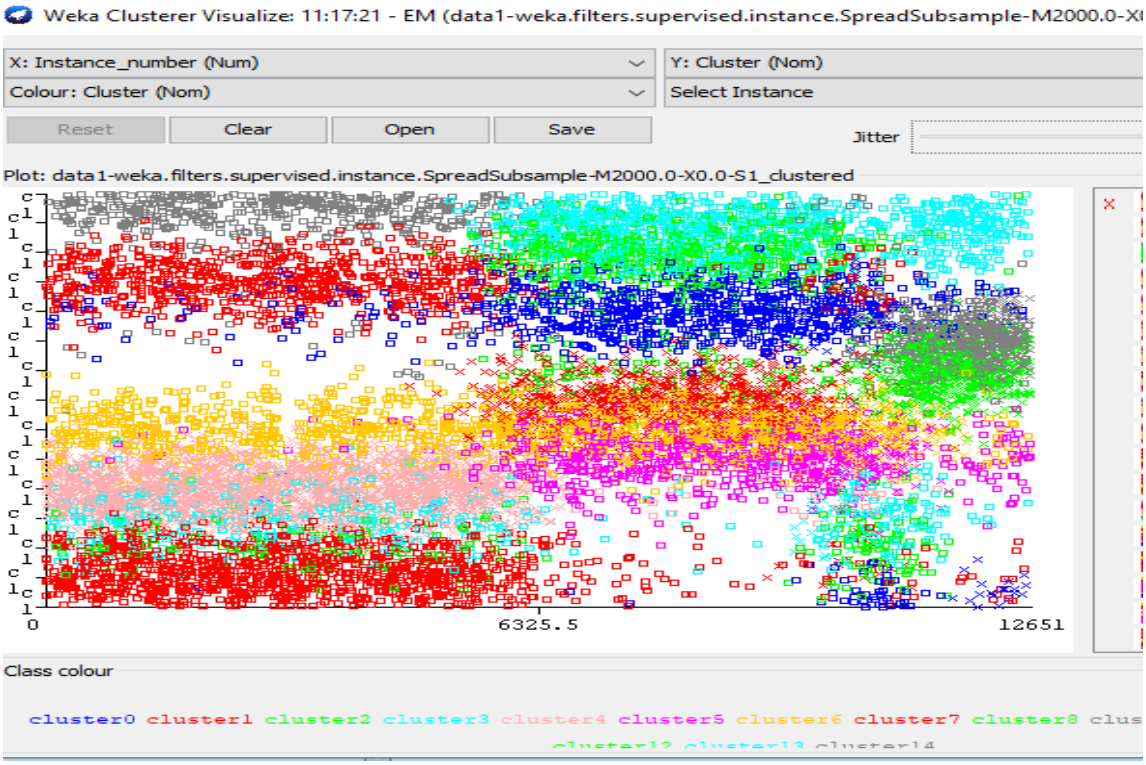


Figure 2: EM cluster assignments

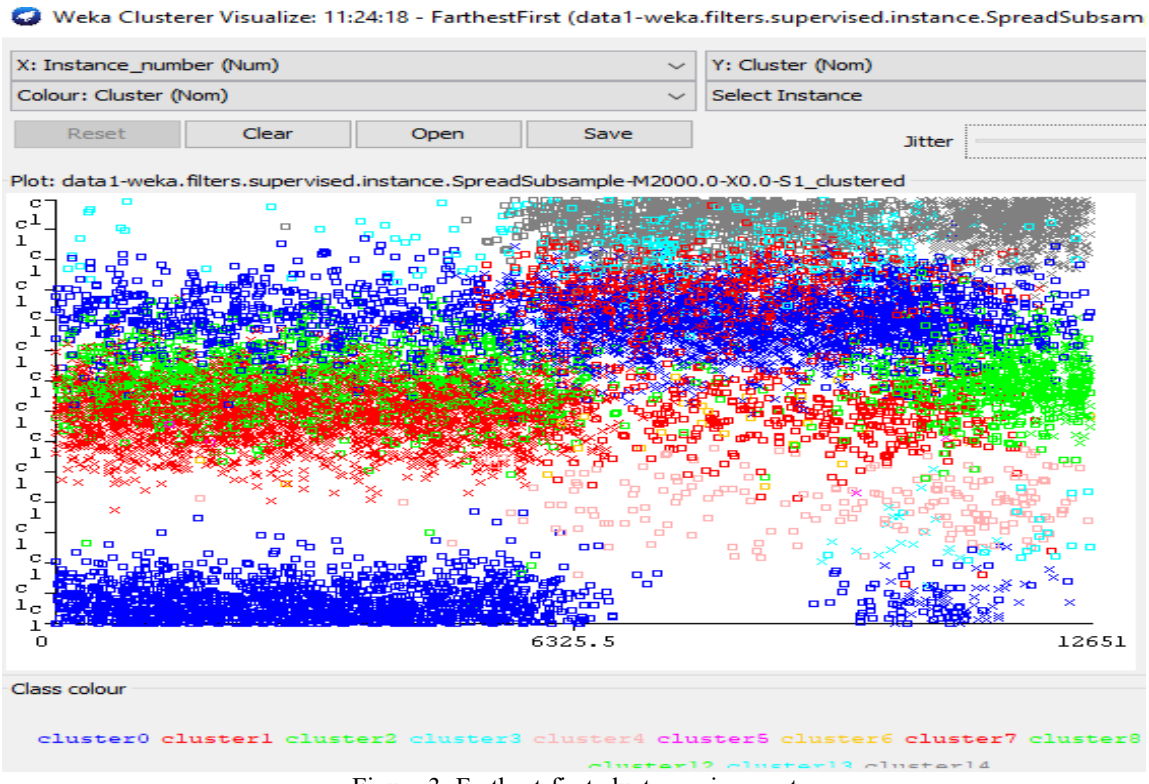


Figure 3: Farthest-first cluster assignment

K-means results: Figure 4 and Figure 5 compare the results of k-means clustering before and after dimension reduction using CFS and Information gain for the two datasets.

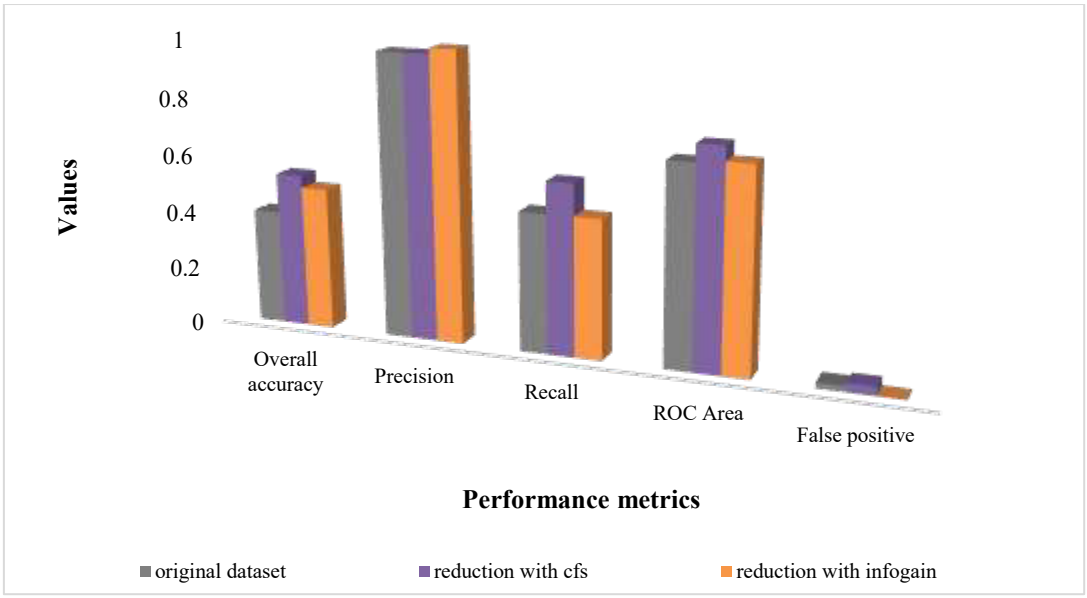


Figure 4: K-means results for dataset1

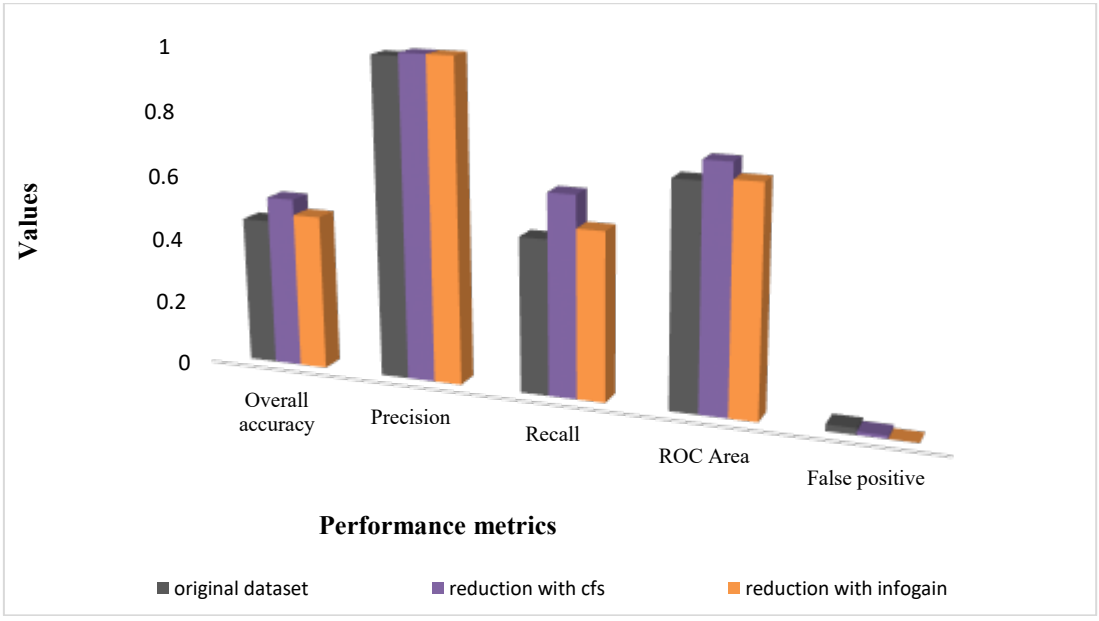


Figure 5: K-means results for dataset2

Figure 4 and Figure 5 show that a remarkable performance increment can be observed for K-means, having an overall accuracy increased by 34.5% and 16% in dataset1 and dataset2 respectively. The area under the ROC for K-means increased slightly by 9%, and around 30% recall increment was also observed. No significant change in false-positive was observed. Feature reduction with information gain yielded a considerable increase in overall accuracy of K-means by about 24% in dataset1, and with a slight increase of 6% dataset2, no significant increase was observed for precision, recall, false-positive and ROC-area in both datasets.

Expectation Maximization results

Figure 6 and Figure 7 compares the results of EM before and after dimension reduction with CFS and information gain for the two datasets.

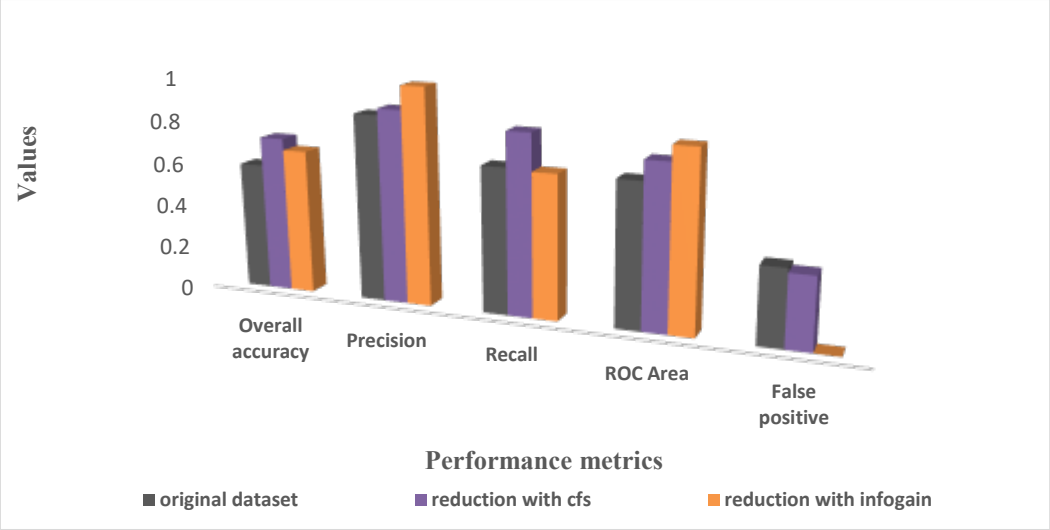


Figure 6: EM results for dataset1

Figure 7: EM results for dataset

Figure 6 and Figure 7 reveal that the EM overall accuracy increased significantly from approximately 23% in dataset1 and increased tremendously by 48% in dataset2, the recall value also increased substantially by 24% and 67.2% in dataset1 and dataset2 respectively. This performance increment is significant. Although, a slight decrease in precision and ROC-area was observed for dataset2, this was expected as the number of false-positives also increased greatly for dataset2. Results of feature reduction using information gain show that a slight increment of about 14% and 20% was observed for overall accuracy in dataset1 and dataset2 respectively. The area under the ROC increased significantly from 25% in both data sets, a slight decrease in the recall of about 1.5% was observed for dataset1, but a 16% increment was observed for dataset2. The precision values increased to 1 in both datasets which denotes a 6% and 16% increment in dataset1 and dataset2 respectively.

Results of Farthest-first

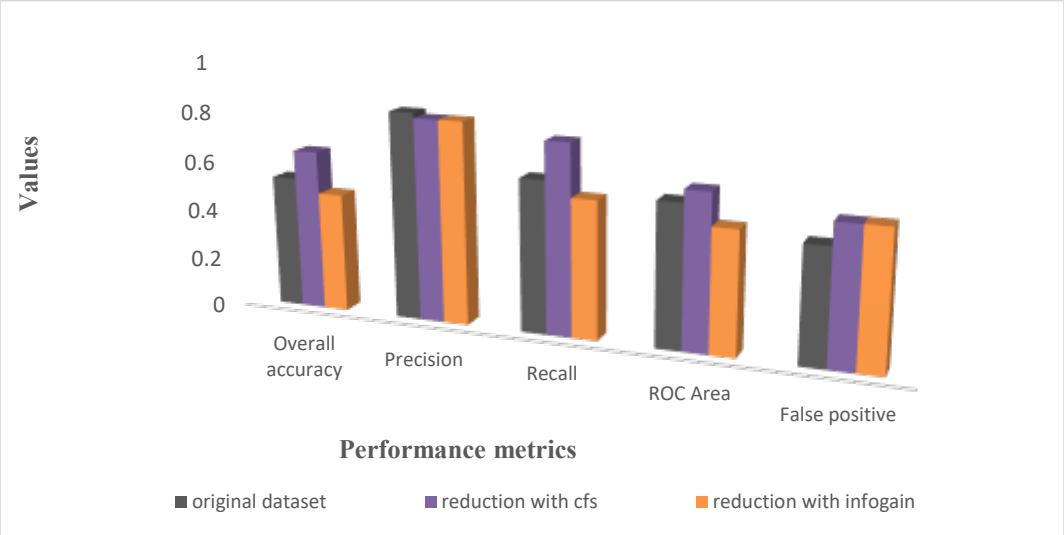


Figure 8: Results of Farthest-first for dataset1

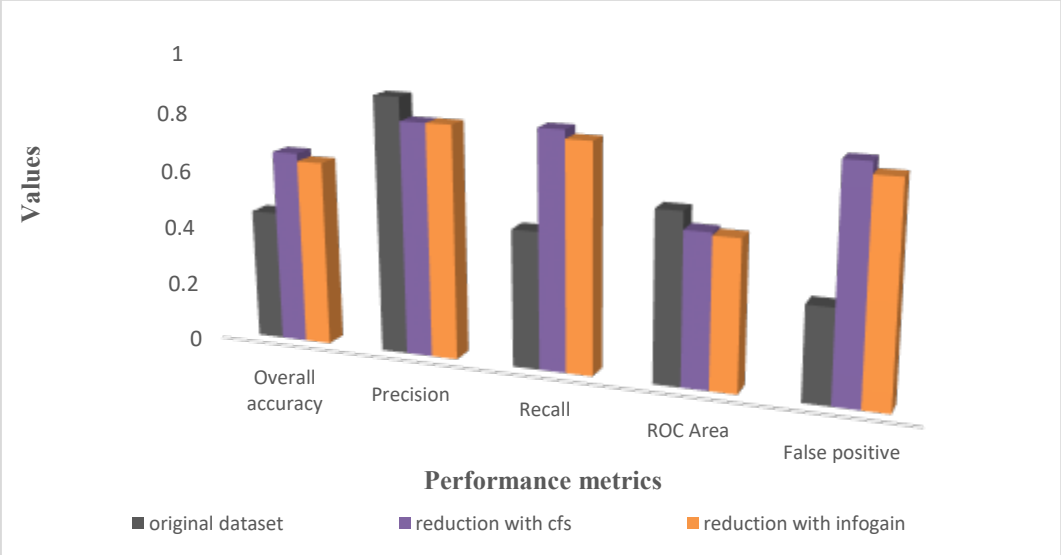


Figure 9: Results of Farthest-first for dataset

Figure 8 and Figure 9 reveal that feature reduction with CFS yields a remarkable increase in overall accuracy by 22% and 48% in dataset1 and dataset2 respectively, the recall values increased tremendously by 72.3% in dataset2 and 18% in dataset1. An increased false-positive of about 140% was observed for dataset2 and 20% for dataset1. The precision values reduced slightly by around 2% in dataset1 and approximately 9% in dataset2. A slight increase in ROC area of about 9% was also observed for dataset1; this value reduces by almost 10% for dataset2. On the other hand, the results of feature reduction with information gain revealed a slight reduction in overall accuracy of around 10% for dataset1 but a remarkable 42% increase for dataset2. A slight reduction of less than 8% in dataset2 and 2% in dataset1 was also observed or precision. The values of false positive increased tremendously by more than 100% and around 20% in dataset1 and dataset2 respectively. The area under the ROC curve also reduces by approximately 14% in dataset1 and 12% in dataset2. Also, a remarkably high increase in recall of about 66% was observed for dataset2, although, this measure reduces by about 10% in dataset1.

5. CONCLUSION AND FURTHER STUDIES

In this paper, Correlation-based Feature Selection and Information Gain Attribute Evaluator were used for dimensionality reduction through the process of filtering of irrelevant and redundant features. Both CFS and Information Gain significantly improve the performance of three clustering algorithms; K-means, EM, and Farthest first in the classification of internet traffic data sets. Although, the overall performance of these clustering algorithms seems to be generally poor, most especially when compared to the clustering results obtained in

previous studies. This difference in result may be due to the differences in the selection of the number of *k* used for initialization. Some previous studies also choose the value of k to be 150 for a 10-class application classification. This number seems to be outrageous for a 10-class clustering even though selecting an optimal number of clusters is an ill-posed problem of critical relevance in clustering analysis.

Further research should be directed towards the classification of UDP flows and the application of other dimensionality reduction techniques and clustering algorithms to further reveal some interesting findings in internet traffic classification problems.

REFERENCES

Alhaj, T. A., Siraj, M. M., Zainal, A., Elshoush, H. T., & Elhaj, F. (2016). Feature selection using information gain for improved structural-based alert correlation. *PLoS ONE*, *11*(11), 1–18. https://doi.org/10.1371/journal.pone.0166017

Anthony, M., Mark, H., Perry, L., & James, B. (2004). flow clustering using machine learning techniques. In *PAM*.

Bernaille, L., Teixeira, R., Akodkenou, I., Soule, A., & Salamatian, K. (2006). Traffic classification on the fly. *ACM SIGCOMM Computer Communication Review*, *36*(2), 23. https://doi.org/10.1145/1129582.1129589

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*(1–2), 245–271. https://doi.org/10.1016/S0004-3702(97)00063-5

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2017). *WEKA Manual for Version 3-8-2*. Hamilton, New Zealand. Retrieved from papers3://publication/uuid/24E005A2-AA1B-4614-BAF5-4D92C4F37413

Dy, J. G. (2008). Unsupervised Feature Selection. In H. Liu & H. Motoda (Eds.), *Computational Methods of Feature Selection* (1st ed., pp. 19–35). Boca Raton, FL: Taylor & Francis Group, LLC. https://doi.org/10.1177/058310248902100419

Dy, J. G., & Brodley, C. E. (2000). Visualization and interactive feature selection for unsupervised data. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00* (pp. 360–364). Boston, MA USA: ACM. https://doi.org/10.1145/347090.347168

Erman, J., & Arlitt, M. (2006). *Traffic classification using clustering algorithms*. *SIGCOMM*. Pisa, Italy. https://doi.org/http://doi.acm.org/10.1145/1162678.1162679

Frank, J. (1994). Machine learning and intrusion detection: Current and future directions. In *Proc. National 17th Computer Security Conference*. Washington,D.C.

Georgiev, G., Gueorguieva, N., Chiappa, M., & Krauza, A. (2016). Feature selection using Gustafson-Kessel fuzzy algorithm in high dimension data clustering. In *2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015* (pp. 1–6). https://doi.org/10.1109/ICMLA.2015.57

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, *3*(3), 1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

H. Dreger, A. Feldmann, M. Mai, V. Paxson, and R. R. S. (2006). Dynamic application-layer protocol analysis for network intrusion detection. In *USENIX Security Symposium,*.

Hall, M. (1999). *Correlation-based Feature Selection for Machine Learning*. The University of Waikato Hamilton, NewZealand. https://doi.org/10.1.1.149.3848

Jain, A. K., Murty, M. N., & Flynn, P. J. (2000). Data Clustering : A Review. *ACM, Inc*, 60.

Jin, X., & Han, J. (2016). Expectation Maximization Clustering. *Encyclopedia of Machine Learning and Data Mining*, 1–2. https://doi.org/10.1007/978-1-4899-7502-7 344-1

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative Study of Attribute Selection using Gain Ratio and Correlation based Feature Selection. *International Jornal of Information Technology and Knowledge Management*, *2*(2), 271–277.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions On*, *17*(4), 491–502. https://doi.org/10.1109/TKDE.2005.66

Liu, Y., Li, W., & Li, Y.-C. (2007). Network Traffic Classification Using K-means Clustering. In *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)* (pp. 360–365). https://doi.org/10.1109/IMSCCS.2007.52

Mladeni, D. (2006). Feature Selection for Dimensionality Reduction. *Subspace, Latent Structure and Feature Selection*, *3940*, 84–102. https://doi.org/10.1007/11752790_5

Moore, A. W., & Papagiannaki, K. (2005). Toward the

accurate identification of network applications. *Passive and Active Network Measurement*, *3431*, 41–54. https://doi.org/10.1007/978-3-540-31966-5_4

Moore, A., Zuev, D., & Crogan, M. (2005). *Discriminators for use in flow-based classification*. *Queen Mary University of London, Department of Computer Science*. https://doi.org/10.1.1.101.7450

Nguyen, T. T. T., & Armitage, G. (2008a). A survey of techniques for internet traffic classification using machine learning. *Communications Surveys & Tutorials, IEEE*, *10*(4), 56–76. https://doi.org/10.1109/SURV.2008.080406

Nguyen, T. T. T., & Armitage, G. (2008b). Clustering to Assist Supervised Machine Learning for Real-Time IP Traf c Classi cation. *Communications Society*, 5857–5862.

Novakovic, J. (2009). Using Information Gain Attribute Evaluation to Classify Sonar Targets. In *17th Telecommunications Forum* (pp. 1351–1354). Serbia, Belgrade.

R. Tibshirani, and J. F. T. H. (2001). The Elements of Statistical Learning. *Springer*.

Revathi, S., & Nalini, T. (2013). Performance Comparison of Various Clustering Algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(2), 67–72. Retrieved from https://pdfs.semanticscholar.org/34dc/c12822ed9b0c0995afcfe306e56cefec1bc6.pdf

Rodriguez-Galiano, V. F., Luque-Espinar, J. A., Chica-Olmo, M., & Mendes, M. P. (2018). Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. *Science of the Total Environment*, *624*, 661–672. https://doi.org/10.1016/j.scitotenv.2017.12.152

Sharma, N., Bajpai, A., & Litoriya, R. (2012). Comparison the various clustering algorithms of weka tools. *International Journal of Emerging Technology and Advanced Engineering*, *2*(5), 73–80.

Shi, H., Li, H., Zhang, D., Cheng, C., & Wu, W. (2017). Efficient and robust feature extraction and selection for traffic classification. *Computer Networks*, *119*, 1–16. https://doi.org/10.1016/j.comnet.2017.03.011

T. Choi, C. Kim, S. Yoon, J. Park, B. Lee, H. Kim, and H. C. (2004). Content-aware internet application traffic measurement and analysis. In *IEEE/IFIP NOMS*. Seoul, South Korea, South Korea: IEEE. https://doi.org/https://doi.org/10.1109/NOMS.2004.1317737

Vidyapeetham, A. V., Rajasundari, T., & Vidyapeetham, A. V. (2017). A Comparative Performance Analysis on Network Traffic classification using Supervised learning algorithms. In *Advanced Computing and Communication Systems (ICACCS), 2017 4th International Conference on* (pp. 1–5). coimbatore, India: IEEE. https://doi.org/10.1109/ICACCS.2017.8014634

Zhang, H., Lu, G., Qassrawi, M. T., Zhang, Y., & Yu, X. (2012). Feature selection for optimizing traffic classification. *Computer Communications*, *35*(12), 1457–1471. https://doi.org/10.1016/j.comcom.2012.04.012

Zhang, J., Chen, X., Xiang, Y., Zhou, W., & Wu, J. (2015). Robust Network Traffic Classification. In *IEEE/ACM Transactions on Networking* (pp. 1–14). https://doi.org/10.1109/TNET.2014.2320577