

Enhancing Intrusion Detection System Efficiency Using Feature Selection and K-Means Clustering

Pratik Jain¹, Neha Shrivastava², Tanvee Nema³, Dr. Bharat Batham⁴,
Mariyam Ejaz Maniyar⁵, Jayshree Pargee⁶

¹Department of Computer Science & Engineering, IPS Academy, Institute of Engineering & Science,
Pratikjain@ipsacademy.org

²Department of Computer Science & Engineering, IIST, Neha.shrivastava0903@gmail.com

³Department of Computer Science & Engineering, IPS Academy, Institute of Engineering & Science,
Tanvinema91@gmail.com

⁴Department of Computer Science & Application, Atal Bihari Vajpayee Hindi University, Bhopal,
Dr.bharat@abvhv.edu.in

⁵Department of MCA, K. K. Wagh Institute of Engineering Education & Research, Nashik,
Mariyamem@gmail.com

⁶Department of Artificial Intelligence & Data Science, Shri Vaishnav Vidhyapeeth Vishwavidyalaya, Indore,
Jayshreepargee@svvv.edu.in

HOW TO CITE:

Pratik Jain, Neha Shrivastava,
Tanvee Nema, Bharat Batham,
Mariyam Ejaz Maniyar, Jayshree
Pargee (2026). Enhancing
Intrusion Detection System
Efficiency Using Feature Selection
and K-Means Clustering.
International Journal of Special
Education, 41(16s), 1138-1150

ABSTRACT:

Intrusion Detection Systems (IDSs) also help protect the computer systems from intruders and their attacks. With the growing amount of traffic in networks, dealing with high-dimensional data has become one of the most challenging tasks for intrusion detection models. The KDD Cup 1999 dataset, the normal benchmark for evaluation of IDS technique, contains 41 network traffic features and among these some of the features during clustering may produce redundant or less useful information in forming the clusters. In this work, a novel feature selection-based framework is proposed to enhance the efficiency of detection system in the elimination of redundant/distracting features prior to the execution of K-Means clustering algorithm. Proposed method concentrates on low complexity representation of the data within the relevant subspace where normal and anomalous network traffic patterns are now linearly separable. The methodology consists of a set of procedures such as data pre-processing, feature selection, normalization, K-Means clustering along with performance evaluation based on a set of widely used clustering and classification measures. It is anticipated that the proposed scheme could achieve better running time and cluster quality and would make the IDS process easier and simpler without dramatically degrading the detection performance. This work presents a systematic approach to advance unsupervised intrusion detection through feature selection techniques on KDD Cup 1999 dataset.

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the
terms and conditions
of the Creative Commons Attribution
(CC BY)
license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords: Intrusion Detection System, KDD Cup 1999, Feature Selection, K-Means Clustering, Network Security, Data Mining, Machine Learning, Cyber Security.

1. Introduction

The expanding availability of computer networks and internet applications has greatly magnified the array of options open to cyber-attacks. PARTICULARLY at risk from these problems are the organizations in all sectors whose digital infrastructure have been breached and who use it to communicate, perform financial transactions, access cloud computing, keep data, and so on. Traditional security technology like firewalls & encryption are no longer enough to ensure full protection with attackers always coming up with new ways to exploit systems vulnerabilities. Intrusion Detection Systems (IDS) have become an essential component of our defense, actively monitoring and alerting on all unusual, suspicious

events, and potential unauthorized network access. Rather than trying to prevent attacks, IDS attempts to find hostile behavior by monitoring the patterns of traffic in a network and sending warnings when it sees unexpected activity. Detection techniques are also crucial in IDSs, and they can be broadly divided into three categories: signature-based detection, anomaly-based detection, and hybrid detection methodologies. Signature-based systems are reasonably accurate against known attacks but are of limited use against zero-day attacks. On the other hand, anomaly-based systems consider anything deviating from the normal behavior as an attack (and hence are able to detect zero-day attacks) but are considered to be less reliable.

Signature-based NIDS

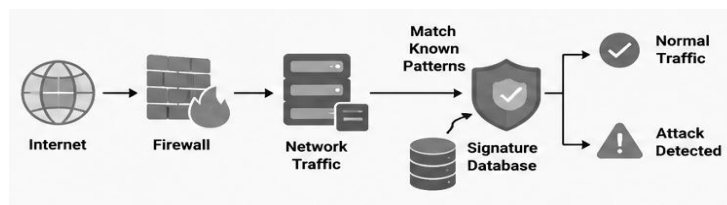


Figure 1.1 Signature – Based NIDS

Machine learning is becoming more attractive in intrusion detection, since it can learn the hidden patterns from the network traffic data itself with no need of being supplied with any user defined rules. Out of the various unsupervised algorithms, K-Means clustering is the most popular technique due to its simplicity and efficiency, and this can be used to group similar network connections into a single cluster. However, there are several factors affect quality of clustering results by K-Means such as dimension and feature selection of input space. The presence of irrelevant and redundant or highly correlated attributes raised the

computational cost and also may affect the quality of the clustering. The KDD Cup 1999 dataset is still one of the most frequently cited benchmark datasets for intrusion detection algorithms. It has forty-one features that represent different properties of network connections such as the protocol information, traffic statistics, and host-based features. While the dataset has been instrumental to the advancement of IDS research, not all of its features are as crucial as others when it comes to clustering. Processing unnecessary features causes increase of execution time and computational complexity especially for large-scale

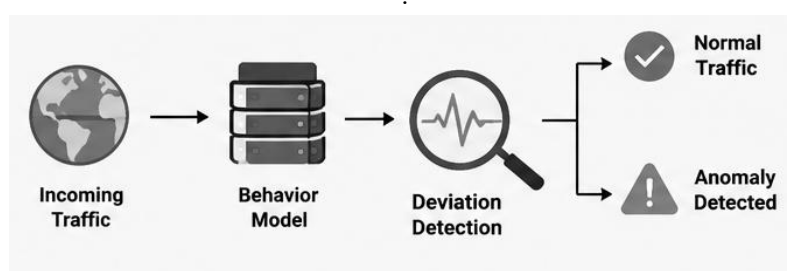


Fig 1.2 Categories of intrusion detection systems

Feature selection has been proven to be a successful technique for preprocessing which

selects the most significant features and removes the redundant or irrelevant ones. Feature selection

can enhance the efficiency of learning, reduce the computational cost, and improve the quality of cluster results by decreasing the dimension of the data set. In contrast to feature extraction techniques, feature selection maintains the original meaning of features and is hence more interpretable, which is of vital importance for security applications as they need to know the importance of features. This paper presents a feature selection-based intrusion detection system which combines feature optimization and K-Means clustering on KDD Cup 1999 dataset. The method is designed to retain the ability to distinguish between normal network traffic and suspicious activities and to significantly reduce the computational overhead by performing the selection of informative attributes prior to clustering. The focus of this work is to improve the efficiency of clustering rather than to propose a novel clustering algorithm.

The major objectives of this research are:

- The KDD Cup 1999 dataset was preprocessed and normalized to provide a good input for the clustering
- To select informative features of network traffic with a feature selection method.
- The goal is to reduce the dimensionality of the data set by discarding redundant features.
- To run the K-Means clustering algorithm on the reduced feature space.
- To assess the contribution of feature selection to the improvements in the cluster quality and in the computational performance.

2. Literature Survey

The phrase "Intrusion Detection System" (IDS) has established itself as a buzzword in today's cybersecurity world, but rather than just a buzzword, IDSs play a very crucial role in the modern cybersecurity ecosystem. With the ever-changing landscape of cyber threats, plenty of machine learning and data mining methods have been presented by researchers to gain better detection for malicious network behaviours while decreasing computational cost and false alarm rates. Dorothy E. Denning also proposed one of the first intrusion detection models, a behavior-based model which recognizes intrusions as

anomalous behaviors on the basis of monitoring audit data and user activity in real time. [1] This novel approach was the basis on which majority of today's anomaly based intrusion detection systems are built. [2] Lee and Stolfo presented a data mining-based approach to intrusion detection by using audit data to develop an automated detection model that can identify both known and novel attacks. Their results proved that data mining techniques were capable of automating intrusion detection and significantly enhancing the performance of network security. [3] Lee, Stolfo, and Mok also generalized this work by proposing feature construction methods which add features that capture properties of the data observed over time and thus improve the accuracy and efficiencies in the play of the machine learning based intrusion detection systems. [4] Chandola et al. presented a comprehensive review on methods for anomaly detection and classified the existing techniques into statistical, machine learning, information-theoretic, and clustering based techniques. Their work stress the need for unsupervised learning methods that can detect novel attack patterns in high-dimensional data. [5] Tavallae et al. investigated the KDD Cup 1999 dataset in depth and pointed out many flaws such as redundant records that bias the training and testing, and an extremely different distribution of the classes. Although admitting its weaknesses, the dataset is still considered one of the most credible benchmark datasets for testing intrusion detection methodologies due to its well-defined classes and its wide acceptance. [6] Hall proposed the Correlation-Based Feature Selection (CFS) algorithm that considers the predictive capability of a feature subset along with the degree of redundancy among the features. This approach substantially reduces the dimensionality of the dataset and the computational cost without sacrificing the learning performance. [7] Subsequently, Hall et al. integrated this algorithm into the data mining tool WEKA enabling researchers in data mining to apply this efficient method for pre-processing, feature selection, clustering, classification and visualization of datasets. [8] Witten et al. have described practical machine learning techniques in the WEKA system and have reviewed operations for preprocessing,

attribute selection, clustering and evaluation that can be applied to data mining tasks. [9] In a similar vein, Han, Kamber, and Pei also introduced some data mining techniques such as classification, clustering, association rule mining, and feature selection, which are also the basic methods applied to the new generation of intrusion detection system. [10] The K-Means clustering algorithm was introduced by MacQueen and it groups data into a predefined number of clusters (k) in which the distance between data points and their corresponding centroids is minimized. Owing to the simplicity of its procedures and efficiency of its computing, K-Means clustering can be considered as one of the most popular unsupervised learning algorithms in the application of network intrusion detection. [11] Kaufman and Rousseeuw made other contributions to clustering and these included different cluster validation measures and other means of assessing clustering goodness of fit. [12] Kohonen introduced the Self-Organizing Map (SOM), an unsupervised neural network that maps high-dimensional input data with preservation of its topological relationships to a low-dimensional grid, is used for data visualization and clustering. [13] Fiore et al. presented a network anomaly detection system based on Restricted Boltzmann Machines (RBM) to enhance the detection of malicious network activities. They showed experimentally that deep learning methods are well-suited for capturing the difference between normal and abnormal network traffic [13]. [14] Bhuyan et al. reviewed anomaly based intrusion detection techniques and illustrated the benefits of machine learning techniques in detecting complex cyber attacks. [15] Moustafa and Slay proposed the UNSW-NB15 as a new dataset to address the limitations of the KDD Cup 1999 dataset with contemporary attack behaviors and real modern network traffic. Their work provided a more realistic benchmark for exploring next-generation intrusion detection systems. [16] Axelsson also gave a full taxonomy for IDS by dividing the IDS methods into signature- and anomaly-based approaches. The analysis compares the positive and negative aspects of the different detection methods and addresses the challenges of detecting previously unknown attacks with a low false alarm

rate. [17] Yu and Liu introduced a fast correlation based filter feature selection method with the consideration of high dimensional datasets. Their method has established that a prior selection of a set of highly relevant features without training the model can save a lot of computational effort and improve the prediction accuracy. [18] In Wang et al. the use of Principal Component Analysis (PCA) for intrusion detection has also been pursued and it is shown that reduction in dimensions can effectively reduce computational cost with good detection rate. [19], [20] In a separate investigation, Wang et al. selected traffic to the essential attributes of a lightweight intrusion detector, and showed that the precision-selected features enable the speed-up of processing without losing detection capability. [21] A hybrid features selection approach for the KDD99 dataset using a combination of several selection techniques to choose the most informative features has been introduced by Araújo et al.. Their results showed that feature selection has a dramatic effect on improving the classification accuracy as well as on prediction time. [22] Mukkamala et al. presented a soft computing based ensemble intrusion detection system and showed that the concerted effect of multiple intelligent classifiers leads to superior detection performance as compared to the individual classification systems. [23] Wu and Banzhaf surveyed several computational intelligence techniques applied to intrusion detection systems and stated that hybrid machine learning based methods are more flexible in combating evolutions of cyber threats. [24] Nguyen and Choi studied the selection of classifiers in the context of network intrusion detection using data mining techniques and stated that the choice of an appropriate classifier is a key factor to enhancing the overall performance of an IDS. [25] Patcha and Park provided a survey on anomaly detection techniques and observed trends in the use of machine learning algorithms to detect previously unknown attacks in high-dimensional network data. [26] Sommer and Paxson made an in-depth assessment on the challenges of using machine learning in operational intrusion detection systems and indicated strong feature engineering and realistic datasets are crucial. [27] Sabhnani and

Serpen evaluated a number of learning algorithms on the KDD dataset, and showed the efficacy of different feature representation for different types of attacks having a significant impact on the classifier performance. [28] Lazarevic et al. performed a comparative analysis of anomaly detection techniques and found that ensemble learning methods give superior detection accuracy as compared to any individual algorithm. [29] Naïve Bayes, an algorithm based on Bayes theorem, was used by Panda and Patra for intrusion detection and they observed competitive classification results with less computational complexity. [30] Ambusaidi et al. In this paper we introduce a two stage filter feature selection method for high dimensional IDS. [31] Kamalov et al. have given an overview on recent FS approaches to IDS and pointed out that dimensionality reduction is gaining more and more attention in dealing with large-scale data on cybersecurity. [32] Di Mauro et al. have provided a critical review of supervised FS methods and how they affect the accuracy and computational cost of an intrusion detection system. [33] Madbouly et al. proposed a data mining based feature selection methodology to enhance intrusion detection performance by maximized feature selection. [34] Bridges and Vaughn developed a set of fuzzy data mining methods for intrusion detection, and showed the robustness of these methods to uncertain and imprecise network data. [35] Ester et al. proposed a DBSCAN to find arbitrary shape clusters in noisy data providing an alternative to partition based methods like K-Means. [36] Arthur and Vassilvitskii proposed the K-Means++ initialization scheme that improves on standard K-Means by speeding up the rate of convergence and providing a better solution on average. [37] Bezdek presented the fuzzification of clustering processes for pattern recognition and related areas the non exclusive option clustering model for the solution of pattern recognition problems with overlapping classes. [38] Bhavani and Thangavel also used the K-Means algorithm for detecting network intrusion and observed that clustering based methods can be used to separate normal and malicious traffic. [39] Jianliang et al. Further application of K-Means clustering for intrusion detection was studied by Jianliang et al. and its

activity was demonstrated to be effective for studying network traffic profiles. [40] Aghdam and Kabiri introduced a feature selection method based on Ant Colony Optimization to optimize attribute selection and enhance the performance of intrusion detection. [41] Sung and Mukkamala analyzed the correlation between the performance intrusion detection and feature selection and showed that removing redundant features can dramatically reduce computation time without impairing detection performance. [42] Xin et al. summarizes the application of machine learning and deep learning methods in the field of Cybersecurity and reports the related work in intelligent intrusion detection systems. [43] Khraisat et al. provide an in-depth survey for sn intrusion detection methods, benchmark data sets, and challenges in existing works and hence is useful for the future enhancement of IDS research. [44] In spite of the fact that more up-to-date datasets are available, The KDD Cup 1999 dataset remains one of the most widely used benchmark datasets to test the performance of intrusion detection algorithms. [45] Song highlighted that feature selection is very crucial for IDSs, and showed that the selection of an optimal subset of features can not only significantly enhance the detection accuracy but also reduce the computational complexity. Based on the existing literature, it is evident that although numerous studies have explored machine learning, clustering, and feature selection techniques for intrusion detection, comparatively fewer studies have focused on integrating **Correlation-Based Feature Selection (CFS)** with **K-Means clustering** to improve computational efficiency on the **KDD Cup 1999 dataset**. Therefore, the proposed work aims to reduce feature dimensionality using CFS before applying K-Means clustering, thereby decreasing execution time while maintaining effective intrusion detection performance.

3. Problem Recognition

The explosive expansion of network traffic has greatly raised the amount and complexity of data to be processed by the Intrusion Detection Systems. Current IDS solutions need to process in real-time thousands of network connections and achieve high detection accuracy with minimal

computational overhead. However, the performance of machine learning algorithms may be influenced by redundant and irrelevant features in the datasets of network traffic. The KDD Cup 1999 dataset, despite being one of the most widely used benchmark datasets for intrusion detection research, contains forty-one input attributes describing different characteristics of network connections. Although these attributes collectively represent network behavior, not every feature contributes equally to the clustering process. Certain attributes may provide overlapping information, while others have limited influence on distinguishing between normal and malicious network activities. Processing all available features increases execution time, memory consumption, and computational complexity, particularly when clustering algorithms are applied to large datasets. The K-Means clustering algorithm is highly dependent on distance calculations between data instances and cluster centroids. As the number of input features increases, the computational cost of these distance calculations also increases. Also, irrelevant features may distort the cluster boundary and degrade the quality of clustering, thus it is challenging to separate normal traffic and attack traffic precisely. Another issue of high-dimensional intrusion detection data is the curse of dimension. Due to the increased number of dimensions, data points become more alike, which results in the deterioration of the clustering methods' performance. This leads to a tighter cluster, longer system running time, and more inefficient IDS. Although modifications to algorithms (or manual removal of selected features) have been proposed in earlier work to increase the performance of intrusion detection, these methodologies are usually dataset-specific and do not offer a generalized approach to feature optimization. But an effective preprocessing scheme for selecting useful attributes systematically prior to clustering with no or negligible computational overhead is highly demanded. Hence, the main challenge in this work is to propose a computationally effective feature selection approach to reduce the dimensionality of the KDD Cup 1999 dataset before using the K-Means clustering technique. The objective is to decrease computational complexity while

preserving the essential information required for effective intrusion detection. The motive of this research is to enhance the computational efficiency of the intrusion detection on the KDD Cup 1999 dataset by feature selection without degrading the clustering performance. [The] specific objectives [of this paper] are:

1. To perform preprocessing on KDD Cup 1999 dataset which includes processing categorical attributes, normalization and data formatting.
2. A suitable feature selection method will be used to study the significance of network traffic features.
3. To reduce dimensionality by removing redundant and insignificant features from data.
4. To run K-Means algorithm on the recommended features.
5. The comparison of clustering results before and after feature selection is discussed.
6. To test the proposed approach in terms of Execution Time, Silhouette Score (SS), Davies–Bouldin Index (DBI), Calinski–Harabasz Index (CHI), Accuracy, Precision, Recall and F1-Score (here, ground-truth labels are employed in the assessment).
7. To show that suitable feature selection can increase computational efficiency with only minimal affect on intrusion detection effectiveness.

4. Proposed Methodology

In this paper, the K-Means clustering algorithm is applied on reduces the dimensions of the KDD Cup 1999 dataset, and this dimensionality reduction before applying the K-Means clustering algorithm is based on the proposed intrusion detection system that aims at enhancing the computational efficiency of network traffic analysis. Rather than using all the features in the dataset, the framework selects a subset of the most informative features by means of a two-stage feature selection process. This ensures elimination of redundant information, the computational complexity is reduced and the clustering is more efficient while the essential information, necessary for differentiating between normal and malicious network is retained.

The proposed framework consists of six major phases:

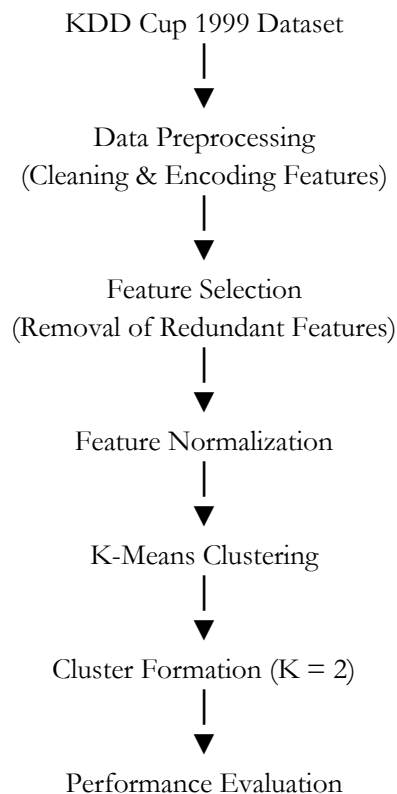
1. Dataset Acquisition
2. Data Preprocessing
3. Feature Selection
4. Data Normalization
5. K-Means Clustering
6. Performance Evaluation

Each phase contributes to improving the overall efficiency of the intrusion detection process. Apart any stage involved improves detects processing efficiency of intrusion. The KDD Cup 1999 benchmark dataset is utilized to carry out the experiments, which is the most widely used dataset for testing the performance of an IDS. The data set was originally created based on the DARPA Intrusion Detection Evaluation Program performed by MIT Lincoln Laboratory, and includes normal network activity along with multiple classes of cyber-attacks. Each connection

is represented by a 41- dimensional vector and a class label. The features are various properties of network traffic, such as protocol, service, and contain information about packets, hosts and connections. The attacks present in the dataset are broadly categorized into four major groups:

- Denial of Service (DoS)
- Probe Attacks
- Remote to Local (R2L)
- User to Root (U2R)

These attack categories enable researchers to evaluate intrusion detection algorithms across different network attack scenarios. The dataset contains both numerical and categorical attributes, requiring preprocessing before machine learning algorithms can be applied. The overall workflow of the proposed intrusion detection framework is illustrated below.



Raw network traffic data often contains categorical values and features with different numerical scales. Proper preprocessing is essential to ensure meaningful clustering results. The preprocessing stage consists of the following operations:

Processing Categorical Features

There are three categorical attributes in the KDD Cup 1999 dataset:

- Protocol Type

- Service
- Flag

And, since K-Means clustering is based purely on numerical values, these categorical features are converted to numeric values through encoding methods like Label Encoding or One-Hot Encoding. Encoding is the process of turning the textual network information into numerical representation that can be understood and interpreted by machine without losing the

semantic meanings. The magnitude of numerical attributes in the dataset is very diverse.

As an illustration

- src_bytes
- wrong_fragment

has values between 0 and 3 only. These changes could lead to poor computation of Euclidean distances. Then, Min-Max Normalization is used. The normalization equation is

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where

- X is the original feature value
- $X_{\min}$ is the minimum value
- $X_{\max}$ is the maximum value

After normalization, every feature lies within the interval $[0, 1]$ which prevents features with larger numerical ranges from dominating the clustering process. After feature selection, the optimized dataset is clustered using the K-Means algorithm. K-Means is an unsupervised machine learning algorithm that partitions data into K distinct clusters based on similarity. The algorithm minimizes the distance between each observation and its nearest cluster centroid. The objective function is

$$J = \sum_{i=1}^K \sum_{x \in C_i} \|x - \mu_i\|^2$$

Where

- K represents the number of clusters,
- C_i denotes the i th cluster,
- x represents a data point,
- μ_i denotes the centroid of cluster i .

The algorithm proceeds as follows:

Step 1

Initialize K random centroids.

Step 2

Calculate the Euclidean distance between every observation and each centroid.

Step 3

Assign every observation to its nearest centroid.

Step 4

Recalculate the centroid of each cluster.

Step 5

Repeat Steps 2–4 until centroid positions no longer change or the maximum number of iterations is reached.

For the proposed intrusion detection framework,

$$K = 2$$

representing

- Normal Traffic
- Intrusion Traffic

Benefits of our framework

Our method has numerous application benefits when compared to standard intrusion detection methods.

- Before clustering, it lowers the dimensionality of the network traffic data.
- Reduces the computational cost by discarding the redundant features.
- Enhances clustering performance without changing the K-Means.
- The proposed approach retains interpretability since original network attributes are used.
- Minimizes both time and space overheads.
- A scalable approach for very large intrusion detection datasets.
- May be generalized to other benchmark datasets like NSL-KDD, UNSW-NB15 and CICIDS2017.

Experiments and Results:

The experiments are carried out with on a Windows OS platform the WEKA 3.8 machine learning tool. The KDD Cup 1999 dataset in the ARFF format were taken as the test data for the proposed feature selection scheme. It contains a total of 25,192 network connections characterized by 41 input attributes and one output class attribute. To differentiate between normal and anomalous behaviour, Simple K-Means clustering algorithm was run on the network traffic to create two clusters with the first one as normal and the second as anomalous. Feature selection was applied on Correlation based Feature Selection (CFS) evaluator and Best First search strategy. The running time of the clustering algorithm is registered before and after the feature optimization to assess the computational efficiency of the proposed method.

Step 1:

Load the KDD Cup 1999 dataset in to weka.

Step 2:

Now run Simple K-Means on the data using all features available.

Step 3:

Running time, cluster size.

Step 4:

Repeat the clustering on the dataset after removal of the 'count' attribute.

Step 5:

Do a Correlation-based Feature Selection (CFS) with Best First search.

Step 6:

Now run Simple K-Means on the selected feature subset.

Step 7:

Evaluate time and complexity of computation of all the experiments.

The first experiment was done on the unprocessed KDD Cup 1999 dataset with all forty-one input attributes. In order to set a baseline for comparison, the Simple K-Means clustering algorithm was run with out any feature optimization. Two clusters are generated by the algorithm representing normal and anomalous network traffic. The proposed model takes 1.20 s to finish all the clustering processes.

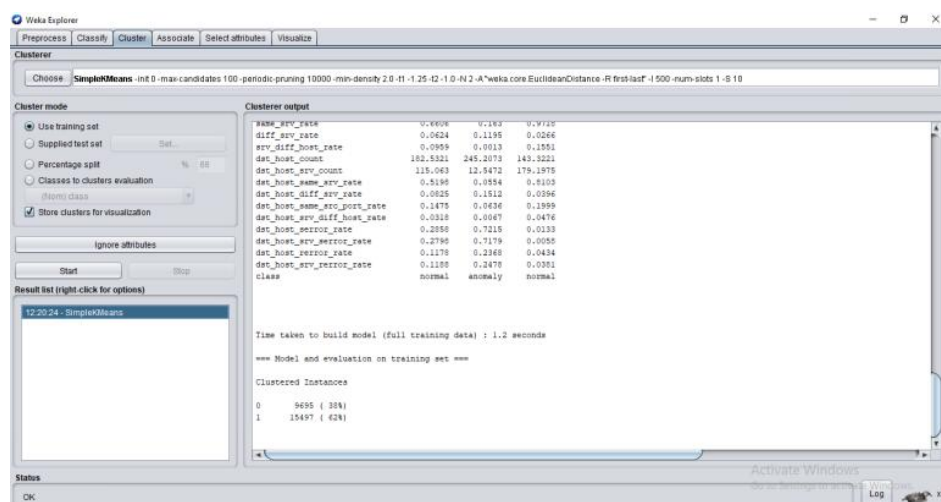


Figure 4.1 Simple K-Means clustering on the original KDD Cup 1999 dataset using all features.

In the second experiment, the count attribute was removed from the dataset before clustering. This experiment reproduces the approach used in previous studies to evaluate the influence of the

count attribute on clustering efficiency. The execution time decreased from **1.20 seconds** to **0.79 seconds**, indicating that removing redundant information can reduce computational overhead.

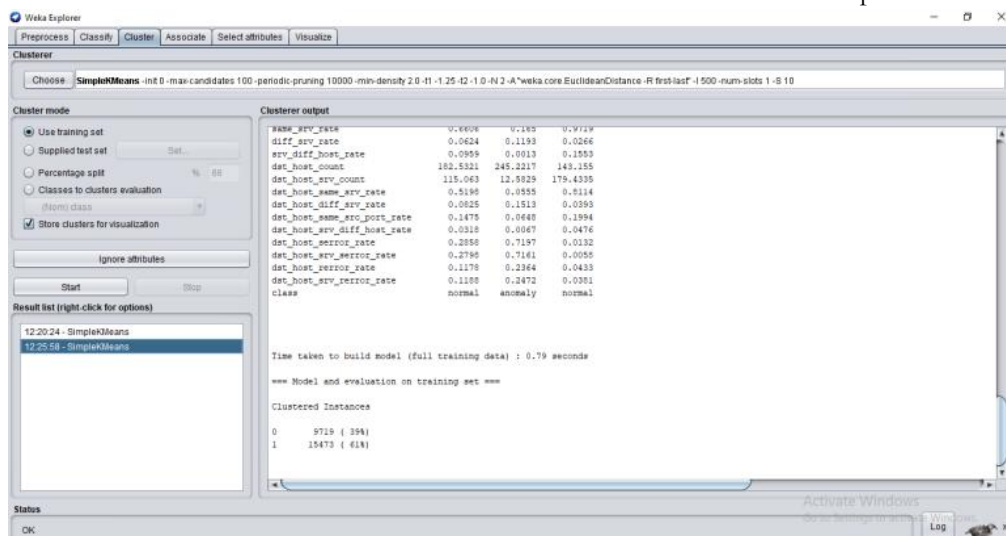


Figure 4.2 Simple K-Means clustering after removing the count attribute.

To develop a more systematic approach than manually removing a single attribute, Correlation-based Feature Selection (CFS) together with the Best First search strategy was applied. The evaluator examined 462 candidate feature subsets and selected the most informative attributes based

on their correlation with the class label and redundancy among themselves. The final subset contained eight features, reducing the dimensionality of the original dataset by approximately **80.49%**.

```

=== Attribute Selection on all input data ===

Search Method:
  Best first.
  Start set: no attributes
  Search direction: forward
  Stale search after 5 node expansions
  Total number of subsets evaluated: 462
  Merit of best subset found:    0.589

Attribute Subset Evaluator (supervised, Class (nominal): 42 class):
  CFS Subset Evaluator
  Including locally predictive attributes

Selected attributes: 4,5,6,12,26,29,30,37 : 8
    flag
    src_bytes
    dst_bytes
    logged_in
    srv_serror_rate
    same_srv_rate
    diff_srv_rate
    dst_host_srv_diff_host_rate

```

Figure 4.3 Feature selection using CFS evaluator and Best First search strategy.

Table 4.1 Selected Features

Attribute No.	Selected Feature
4	flag
5	src_bytes
6	dst_bytes
12	logged_in
26	srv_serror_rate
29	same_srv_rate
30	diff_srv_rate
37	dst_host_srv_diff_host_rate

The reduced dataset containing only the selected attributes was used as input to the Simple K-Means clustering algorithm. The clustering took only 0.17 second as opposed to clustering the original dataset, a huge saving in execution time.

The dimensionality reduction leads to a decrease in the amount of distance calculations in the centroid assignment phase, which increases the computational efficiency, while still preserving useful clusters.

Clusterer output			
num_shells	0.0004	0	0.0006
num_access_files	0.0043	0	0.007
num_outbound_cmds	0	0	0
is_host_login	0	0	0
is_quest_login	0	0	0
count	84.5912	166.3895	33.4178
srv_count	27.6988	9.9234	38.8191
error_rate	0.2863	0.7253	0.0117
srv_error_rate	0.2838	0.7212	0.0101
rerror_rate	0.1186	0.2467	0.0385
srv_rerror_rate	0.1203	0.2486	0.04
same_srv_rate	0.6606	0.163	0.9718
diff_srv_rate	0.0624	0.1195	0.0266
srv_diff_host_rate	0.0959	0.0013	0.1551
dst_host_count	182.5321	245.2073	143.3221
dst_host_srv_count	115.063	12.5472	179.1975
dst_host_same_srv_rate	0.5198	0.0554	0.8103
dst_host_diff_srv_rate	0.0825	0.1512	0.0396
dst_host_same_src_port_rate	0.1475	0.0636	0.1999
dst_host_srv_diff_host_rate	0.0318	0.0067	0.0476
dst_host_error_rate	0.2858	0.7215	0.0133
dst_host_srv_error_rate	0.2798	0.7179	0.0058
dst_host_rerror_rate	0.1178	0.2368	0.0434
dst_host_srv_rerror_rate	0.1188	0.2478	0.0381
class	normal	anomaly	normal
Time taken to build model (full training data) : 0.17 seconds			
=== Model and evaluation on training set ===			
Clustered Instances			
0	9695 (38%)		
1	15497 (62%)		

Figure 4.4 K-Means clustering using the optimized feature subset.

Table 4.2 Comparison of Experimental Results

Experiment	Features	Execution Time (s)	Cluster 0	Cluster 1
Original Dataset	41	1.20	9695	15497
Count Removed	40	0.79	9719	15473
Feature Selection (CFS)	8	0.17	9695	15497

Table 4.3 Feature reduction

Metric	Value
Original Features	41
Selected Features	8
Removed Features	33
Reduction (%)	80.49
Subsets Evaluated	462
Merit Score	0.589

The experimental results reveal that the computation overhead of the intrusion detection framework is reduced greatly with the help of feature selection. Even if only the count feature was removed, the execution time decreased by roughly a third, but the feature selection method proposed here obtained a far larger reduction by eliminating from 41 to 8 features. As a result, the running time of Simple

K-Means was reduced from 1.20s to 0.17s, with a total speedup of about 85.8%. The selected attributes mainly represent connection status, traffic volume, service behavior and network error characteristics, which also demonstrates that these features could contain enough discriminative information for clustering network traffic.

References:

- D. E. Denning, "An Intrusion-Detection Model," *IEEE Transactions on Software Engineering*, vol. SE-13, no. 2, pp. 222–232, Feb. 1987.
- W. Lee and S. J. Stolfo, "A Framework for Constructing Features and Models for Intrusion Detection Systems," *ACM Transactions on Information and System Security*, vol. 3, no. 4, pp. 227–261, Nov. 2000.
- W. Lee, S. J. Stolfo, and K. W. Mok, "A Data Mining Framework for Building Intrusion Detection Models," in *Proc. IEEE Symposium on Security and Privacy*, 1999, pp. 120–132.
- V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, Jul. 2009.
- M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in *Proc. IEEE CISDA*, 2009, pp. 1–6.
- M. Hall, "Correlation-Based Feature Selection for Machine Learning," Ph.D. dissertation, University of Waikato, New Zealand, 1999.
- M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
- I. H. Witten, E. Frank, M. Hall, and C. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Morgan Kaufmann, 2016.
- J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2012.
- J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.
- L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.
- T. Kohonen, *Self-Organizing Maps*, 3rd ed. Springer, 2001.
- U. Fiore, F. Palmieri, A. Castiglione, and A. De Santis, "Network Anomaly Detection with the Restricted Boltzmann Machine," *Neurocomputing*, vol. 122, pp. 13–23, 2013.
- M. Bhuyan, D. Bhattacharyya, and J. Kalita, "Network Anomaly Detection: Methods, Systems and Tools," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 303–336, 2014.
- N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," in *Proc. MilCIS*, 2015.
- S. Axelsson, "Intrusion Detection Systems: A Survey and Taxonomy," Technical Report 99-15, Chalmers University of Technology, 2000.
- L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in *Proc. ICML*, 2003, pp. 856–863.
- W. Wang, X. Guan, and X. Zhang, "A Novel Intrusion Detection Method Based on Principal Component Analysis," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8020–8024, 2009.
- W. Wang, S. Gombault, and T. Guyet, "Towards Fast Detecting Intrusions: Using Key Attributes of Network Traffic," in *Proc. ICIMP*, 2008.
- W. Wang, X. Zhang, S. Gombault, and S. J. Knapskog, "Constructing Important Features from Massive Network Traffic for Lightweight Intrusion Detection," *IET Information Security*, vol. 9, no. 6, pp. 374–379, 2015.
- N. Araújo, R. Oliveira, E. Ferreira, A. Shinoda, and B. Bhargava, "Identifying Important Characteristics in the KDD99 Intrusion Detection Dataset by Feature Selection Using a Hybrid Approach," in *Proc. ICT*, 2010.
- S. Mukkamala, A. Sung, and A. Abraham, "Intrusion Detection Using Ensemble of Soft Computing Paradigms," *Journal of Network and Computer Applications*, vol. 28, no. 2, pp. 167–182, 2005.
- S. Wu and W. Banzhaf, "The Use of Computational Intelligence in Intrusion Detection Systems: A Review," *Applied Soft Computing*, vol. 10, no. 1, pp. 1–35, 2010.
- H. Nguyen and D. Choi, "Application of Data Mining to Network Intrusion Detection: Classifier Selection Model," 2010.

- A. Patcha and J. Park, "An Overview of Anomaly Detection Techniques," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007.
- R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in *Proc. IEEE Symposium on Security and Privacy*, 2010.
- M. Sabhnani and G. Serpen, "Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset," in *Proc. IEEE SMC*, 2003.
- A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection," in *Proc. SIAM Data Mining*, 2003.
- M. Panda and M. Patra, "Network Intrusion Detection Using Naïve Bayes," *International Journal of Computer Science and Network Security*, vol. 7, no. 12, pp. 258–263, 2007.
- M. A. Ambusaidi, X. He, P. Nanda, and Z. Tan, "Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm," *IEEE Transactions on Computers*, vol. 65, no. 10, pp. 2986–2998, 2016.
- F. Kamalov, S. Moussa, R. Zgheib, and O. Mashaal, "Feature Selection for Intrusion Detection Systems," 2021.
- M. Di Mauro, G. Galatro, G. Fortino, and A. Liotta, "Supervised Feature Selection Techniques in Network Intrusion Detection: A Critical Review," 2021.
- A. Madbouly, A. Gody, and T. Barakat, "Relevant Feature Selection Model Using Data Mining for Intrusion Detection System," 2014.
- S. M. Bridges and R. B. Vaughn, "Intrusion Detection via Fuzzy Data Mining," in *Proc. ACSAC*, 2000.
- M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proc. KDD*, 1996.
- D. Arthur and S. Vassilvitskii, "k-Means++: The Advantages of Careful Seeding," in *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2007.
- J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer, 1981.
- T. Bhavani and K. S. Thangavel, "Cluster Analysis for Network Intrusion Detection Using K-Means Algorithm," in *Proc. IEEE*, 2008.
- M. Jianliang, S. Haikun, and B. Ling, "The Application of Intrusion Detection Based on K-Means Cluster Algorithm," in *International Forum on Information Technology and Applications*, 2009.
- M. H. Aghdam and P. Kabiri, "Feature Selection for Intrusion Detection System Using Ant Colony Optimization," *International Journal of Network Security*, vol. 18, no. 3, pp. 420–432, 2016.
- A. H. Sung and S. Mukkamala, "The Feature Selection and Intrusion Detection Problems," in *Proc. IEEE SMC*, 2003.
- Y. Xin, L. Kong, Z. Liu, et al., "Machine Learning and Deep Learning Methods for Cybersecurity," *IEEE Access*, vol. 6, pp. 35365–35381, 2018.
- A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges," *Cybersecurity*, vol. 2, no. 1, 2019.
- KDD Cup 1999 Dataset, University of California, Irvine. Available: KDD Cup 1999 Dataset Repository
- J. Song, *Feature Selection for Intrusion Detection System*, Ph.D. dissertation, Aberystwyth University, United Kingdom, 2016.