Webology (ISSN: 1735-188X)
Volume 17, Number 2, 2020

Classification Of Nsl Kdd Dataset Using Genetic Algorithm Based Feature Selection And Ensemble Approaches

Reshamlal Pradhan[1] , Dr. S.R.Tandan[2]

[1]Dr. C. V. Raman University Bilaspur, India**.**

[2]Dr. C. V. Raman University Bilaspur, India.

Article History: Received: 03.01.2020 Accepted: 29.01.2020 Published: 24.02.2020

ABSTRACT

Information or network security has become a crucial concern for every organization and individuals, aiming to secure data and information from hostile operations or invasions. Intrusion Detection Systems (IDS) are vital in network security. It identifies network security breaches. IDSs monitor computer networks for malicious activity. This work presents an ensemble-based intrusion detection system with evolutionary search-based feature selection. WEKA was used to explore stacking, bagging, and boosting on the NSL-KDD dataset. Ensembles learning mechanisms utilized decision tree methodologies. In the developed approach, the NSL-KDD dataset is preprocessed and subsequently feature-selected to reduce its size. Evolutionary search mechanism is utilized for feature selection. Selecting relevant machine learning algorithms based on attack type creates the model. The suggested system's accuracy is tested using the NSL-KDD dataset. The suggested method provides excellent accuracy for all assaults. Empirical results show that using ensemble techniques with selected reduced features enhances the performance of network intrusion detection system.

Keywords: IDS, NSL KDD, Feature selection, Genetic algorithm, Classification techniques, Ensemble technique.

1. INTRODUCTION

In the last two decades, information technology has advanced at a breakneck pace. Industry, business, and different aspects of human existence all use computer networks. As a result, establishing a trustworthy network is a critical responsibility. The development of information technology has posed several issues. Creating a trustworthy network is a demanding endeavor. The widespread use of computers and easy access to the internet has increased the variety of methods for attacking and deceiving a system. Intrusion is defined as the unauthorised access to, or seizure or possession of the systems of an individual or corporate organisation (Zhu, D. et al, 2001).

Despite the fact that network-based systems can be secured using a range of security methods such as information encryption, and intrusion prevention, many intrusions go unnoticed. Internal attacks, for example, are not prevented by firewalls. As a result, intrusion detection systems (IDSs) are crucial in network security.

The process of finding knowledge and valuable patterns is called data mining (Nejad et al., 2008). Today's research institutes and organisations deal with large amounts of structured and unstructured data. Organizations and researchers must better detect intrusions (or attacks). These types of data require machine learning approaches. Machine learning uses Classification, Clustering, and Regression to ensure security. Effective computer security requires intrusion detection and prevention systems.An intrusion detection system (IDS) is used to protect data integrity, confidentiality, and system availability from attacks. Anomaly detection and misuse detection are two types of intrusion detection algorithms. Anomaly detection, which searches for patterns that are different from normal behaviour, is used to detect attacks. Misuse detection identifies attacks by analysing patterns found in past invasions. Current IDSs have a significant problem in that they do not generalise to identify unknown-signature attacks. Anomaly detection systems are adaptive, which means they can handle new attacks but not the specific type of attack. As a result, many supervised and unsupervised intrusion detection methods have been created (Lee et al., 1999). Classification is a method of supervised learning. An IDS that uses classification techniques tries to categorise all traffic as either normal or malicious. It has the ability to help security specialists save time and analyse attack data. This study examines the application of ensemble classification to intrusion detection system. We used multiple classification methods and evolutionary feature selection on the NSL-KDD dataset to assess performance in terms of accuracy and computation time. According to the study, classification approaches outperform others in the intrusion detection dataset.

The paper is organized in following sections. The second portion discusses the background of the study, which includes work in the field, Classification techniques, feature selection techniques and NSL KDD dataset. Section three present methodology and framework. In section four, experimental findings are described; and finally, in section V, the conclusion is presented.

2. BACKGROUND

2.1Work in the field

Anderson (1980) expanded the idea of intrusion detection by presenting a threat categorization model that generates a security monitoring surveillance system based on detecting anomalies in usage patterns. Lippmann et al. (2000) compared data mining classification techniques for intrusion detection. Schultz et al. (2000) proposed a framework for finding new examples that employs data mining algorithms to train multiple classifiers.

Hwang et al. (2007) developed a three-tier intrusion detection system architecture that comprises a blacklist for detecting known assaults and a white-list for detecting routine traffic. The remaining traffics, which were identified as anomalies by the white-list, were classified using a multi-class SVM classifier.Srinivasulu et al. (2009) employed a confusion matrix to evaluate the performance of three data mining classification techniques: CART, Naive Bayesian, and Artificial Neural Network Model.The value of each attribute in the KDD '99 intrusion detection dataset for class categorization was examined by Tavallaee et al. (2009).Bolon-Canedo et al. (2011) proposed a strategy using discretizers, filters, and classifiers. It aims to improve classifier performance with less features. Both binary and multiple-class classification issues have been solved. It was evaluated using the KDD Cup 99 benchmark. The findings showed that the suggested technique enhanced performance while reducing features by 80%.Reddy et al. (2011) also provided an overview of a number of data mining methodologies that have been proposed to improve IDSs.

Mukherjee et al. (2012) suggested an FVBRM model for feature selection and compared it to three feature selectors: CFS, IG, and GR. Compared to IG and GR, experimental data suggest that CFS enhances Nave Bayes classification accuracy. Although GR is an extension of IG, both algorithms for feature selection are used in research and IG outperforms GR. Compared to CFS, FVBRM improves classification accuracy but takes longer.Principal Component Analysis and Naive Bayes are two machine learning algorithms discussed by Neethu (2012) for dimensionality reduction and attack categorization. Reduced dimensionality lowers memory needs and speeds up performance.Kalyani et al. (2012) compared classification techniques such as Naive Bayes, J48, OneR, PART, and RBF Network algorithm using the NSL-KDD dataset. The advantages of the NSL-KDD dataset over the KDDCUP'99 dataset are also discussed.

Masarat et al. (2014) presented a multi-step IDS methodology. KDD has several qualities, however not all are beneficial for classification tasks. Gain Ratio-based feature selection is given. J48 trees are trained with optimal features using a Roulette Wheel based on feature gain ratios, which favours high-gain features in random feature selection. Final step: fuzzy weighting ensemble classifiers. With the fuzzy weighted combiner, classifiers' cost and performance are weighted. Results show suggested strategy outperforms similar methods. It takes longer but improves accuracy.Subramanian et al. (2012) intend to use decision tree methods to categorise the NSL-KDD dataset in order to develop a model based on their metric data and test the efficacy of decision tree algorithms. Nadiammai et al. (2012) compared the accuracy, sensitivity, specificity, time, and error of all rule-based and certain function-based classifiers in order to forecast their effectiveness.

A fuzzy evolutionary method for intrusion detection is created using the KDD99 dataset. A sequential intrusion detection system's results are compared. The data show that the suggested system has a greater DoS, probing, R2L, and U2R detection rate than existing systems. The proposed system takes less time to train and test the dataset than current methods, and it consumes less memory (Danane et al., 2015).

The literature works illustrated here demonstrate how various authors use various standalone machine learning approaches as well as ensemble techniques. They used various datasets, including KDD Cup 1999 and NSL KDD, to train their model. The improvement of performance in terms of computation time, accuracy, memory and CPU time, or feature reduction and efficiency, is one of the common observations from all of the research mentioned.

2.2 Classification Techniques

The process of finding knowledge and valuable patterns is called data mining. When it comes to dealing with organised and unstructured data as well as security issues, machine learning techniques are crucial. Machine learning approaches come in a variety of forms, and they all perceive and provide knowledge for fraud detection.Many machine learning techniques, including Classification, Regression, and Clustering, are used to find and ensure security measures. Intrusion detection systems (IDSs) and intrusion prevention systems are the most important instruments for preventing these sophisticated attacks and security issues (IPSs). classification is a supervised learning mechanism. Using classification algorithms, an IDS attempts to classify all traffic as either benign or malicious. Security professionals can use it to analyse attack data and save time. The classification techniques employed in this paper are J48, Random Tree, Bagging, Boosting and Stacking. Here Bagging, Boosting and Stacking are ensemble approaches.

2.3Feature Selection Techniques

Feature selection is now a crucial component of intrusion detection in order to achieve good performance. Present datasets available on various repository are made up of a variety of features. Relevance levels for features vary. The right features must be chosen in order to achieve higher performance (Hota, 2015). However, If the data is already present elsewhere, it can be redundant or unnecessary to incorporate any of these elements. IDS performance is slowed down as a result of the affected computation time. Finding the most pertinent characteristics for the categorization of training and test data is done through feature selection (Bolon-Canedo et al, 2011). General Feature selection techniques used now a days are Correlational Feature Selection, Information Gain, Gain Ratio etc. Apart from these various novel feature selection techniques are being developed by many researchers. In this study, Genetic Algorithm based feature selection technique is used to best classify the NSL KDD dataset.

3. METHODOLOGY

In this study the presented work is explained in three phases, First NSL KDD Dataset, Second Classification method and third Feature selection based on Genetic Algorithm. Dataset is described in order to understand various features available in the NSL KDD. For the classification purpose

decision tree and ensemble methods are used, while for the feature selection, Genetic algorithm method is used.

3.1 Classification Method

The research, which entails the development and validation of models, is designed within a framework. Decision trees and ensemble approaches are employed in this study as machine learning techniques during the model-building process.

A training dataset is used to develop the model, and a test dataset is used to evaluate the model. A training dataset consists of labelled data with precisely specified data definitions for average and attack-type data. As a result, these data are referred to as training data and are utilised to create classification models.

Classification models are validated using the test set of data once they have been built with the training dataset. Unlabelled data make up the testing set of data. This test batch of unlabelled information is classified by a trained model. with this categorization, model is validated. Validated models are taken into account for real-time implementations. Figure 1 represents the proposed classification methodology.

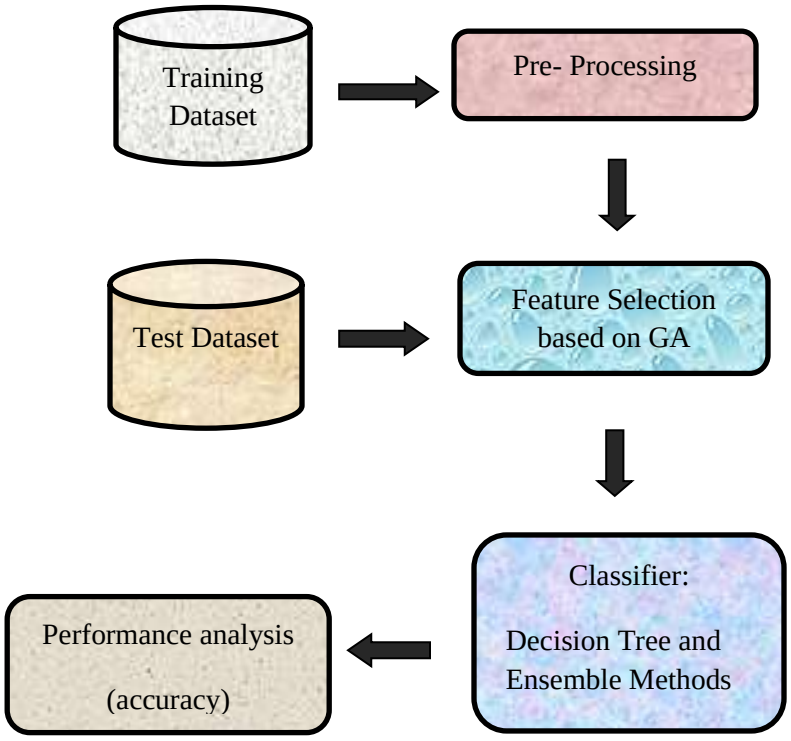

Figure 1: Classification Model

3.2NSL KDD Dataset

The NSL-KDD intrusion dataset, a publicly accessible benchmark dataset for intrusion detection, is used in this study for classification purposes. The dataset's data repositories contain a variety of NSL-KDD datasets; "KDDTrain" and "KDDTest" were selected from these datasets for training and testing, respectively. The dataset's connection patterns are described by 41 features and one class attribute, a total of 42 attributes. Class attributes further classified as either normal or attack type. Decision trees have the benefit of taking categorical attributes. The four categories of intrusions in the NSL KDD data set include DoS, Probe, U2R, and R2L attack types (Ingre, B.et al, 2011). Features of NSL KDD dataset are depicted in Table 1.

Table 1: NSL KDD Dataset Attributres (Features)

Dataset	**Attribute Name**	**No. of Features**
NSL KDD DATA SET	1. Protocol_type, 2. Service, 3. Duration, 4. flag, 5.src_bytes, 6. dst_bytes, 7. wrong_fragment, 8. urgent, 9. hot, 10. num_compromised, 11. land,, 12. num_failed_logins, 13. logged_in, 14. root_shell, 15. num_shells, 16. su_attempted, 17. num_root, 18. is_host_login, 19. num_file_creations, 20. num_access_files, 21. srv_count, 22. num_outbound_cmds, 23. is_guest_login, 24. count, 25. serror_rate, 26. rerror_rate, 27. srv_serror_rate, 28. same_srv_rate, 29. srv_rerror_rate, 30. diff_srv_rate, 31. srv_diff_host_rate, 32. dst_host_count, 33. dst_srv_host_count, 34. dst_host_same_srv_rate, 35. dst_host_same_src_port_rate, 36. dst_host_diff_srv_rate, 37. dst_host_srv_serror_rate, 38. dst_host_serror_rate, 39. dst_host_srv_diff_host_rate, 40. dst_host_rerror_rate, 41. dst_host_srv_rerror_rate, 42. class.	42

3.3 Feature Selection based on GA

By identifying important attributes and excluding unimportant ones, feature selection aims to produce a subset of features that accurately describes the dataset with the least performance impact. The CFS, IG, and GR methods are several feature selection techniques. In this study, the feature selection method for classification is based on a genetic algorithm.

An AI search technique that makes use of the idea of natural selection and evolution is the genetic algorithm (GA). It is an effective evolutionary method for dealing with optimisation problems. The successful integration of GA parameters is essential for search.In addition to population, important parameters to take into account are mutation and crossover rates.Genetic search methods known as adaptive meta heuristics. Using biological genetic and evolutionary notions, Holland developed the first GA in 1975 to address a variety of optimisation problems (Holland, 1975).The effectiveness of GA searches has been shown to be significantly influenced by crossover and mutation rates. High crossover rates and low mutation rates are advised, according to several studies (Obitko, 1998) (Bernard et al., 2015).Two widely used techniques for

parameter tuning are the 50/50 crossover/mutation ratios and the common ratios with (0.03) mutation rates and (0.9) crossover rates (Chaiyaratana et al., 1999; Capraro et al., 2008). Fifty-fifty crossover and mutation ratio is considered here in the genetic algorithm for the feature selection on NSL KDD dataset.

4. RESULT ANALYSIS AND DISCUSSION

The experiment is performed on Weka machine learning tool. K-fold cross validation method is used for the model execution. Decision tree techniques J48, Random tree, and REP tree, Ensemble techniques Bagging, Boosting and stacking are used as classification techniques for the performance analysis.

With all 42 features of NSL KDD dataset, Decision Tree and Ensemble Methods are tested for the classification outcome. Results mentioned in table 2 depicts that decision tree and ensemble method maintained a good accuracy on NSL KDD dataset.

Table 2 : Performance of Classifiers with all Features of NSL KDD Dataset

Classification Techniques	**Accuracy**	**Computation Time (Sec)**
J48	98.08	6.17
Random Tree	97.68	0.94
REP Tree	97.97	2.55
Bagging	98.26	34.72
Boosting	98.44	51.66
Stacking	97.99	46.58

A graph plotting of aforesaid results, mentioned in table 2 are presented in figure 2. It is observed that Ensemble methods are maintaining good results over standalone decision tree techniques. However computational time of standalone decision tree techniques are significant than ensemble methods.

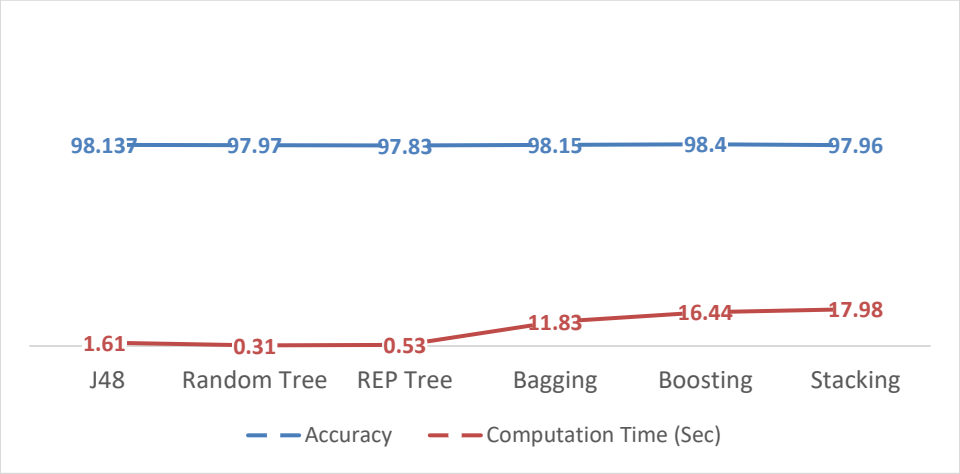

Figure 2: Performance of Classifiers with all Features of NSL KDD Dataset

For the feature selection on NSL KDD dataset, parameter tuning method of Genetic Algorithm is used. Fifty- fifty crossover and mutation ratio is considered in GA for the feature selection. By using this method 13 features are selected from NSL KDD dataset. List of selected features are given in the table 3.

Table 3: Selected Features fromNSL KDD Dataset

Dataset	**Selected Attribute(feature) numbers**	**No. of Features**
NSL KDD DATA SET	1,3,5,6,11,23,26,27,29,30,31,37,41	13

It is observed that very less quantity of relevant features, i.e. 13 out of 42 features are selected through the proposed genetic algorithm method. With the aforesaid selected feature set, Decision tree techniques and ensemble methods are tested on the NSL KDD dataset. The results are mentioned in the table 4.

Table 4: Performance of Classifiers with 13 selected Features

Classification Techniques	**Accuracy**	**Computation Time (Sec)**
J48	98.137	1.61
Random Tree	97.97	0.31
REP Tree	97.83	0.53

Bagging	98.15	11.83
Boosting	98.40	16.44
Stacking	97.96	17.98

A graph plotting of results mentioned in table 4 are presented in figure 3. It is observed that Ensemble methods are maintaining good results over standalone decision tree techniques. However computational time of standalone decision tree techniques are significant than ensemble methods. Compared to the classification with all 42 feature set, it is observed that performance of classifiers with 13 selected features is significant in computational time, while maintaining a good accuracy. Some classification techniques have achieved significant increase in accuracy.

Figure 3: Performance of Classifiers with 13 selected Features

5. CONCLUSION

In this study NSL KDD dataset is explored with genetic algorithm based feature selection technique. NSL KDD dataset consists of 42 features. It has 21 labels (attack type) in train set, while 37 labels (attack type) in test set of data. These labels are broadly falls under four category of attack types. With the presented feature selection method 29 feature are reduced, while 13 relevant features are selected for the classification process. It is found that with selected features decision tree and ensemble techniques have achieved significant decrease in computational time, while there are no major changes in accuracy compared to the classification with all feature set. However, J48 and Random tree techniques have achieved significant increase in accuracy. Which clearly denotes the efficacy of decision tree and ensemble technique based classification model with genetic algorithm based feature selection. The research can be explored in terms other measurement parameters of confusion metrics, i.e. precision, recall etc. Different types of attacks on NSL KDD data set is also needed to be explored.

Reference:

- Zhu, D., Premkumar, G., Zhang, X., & Chu, C. H. (2001). Data mining for network intrusion detection: a comparison of alternative methods. Decision Sciences, 32(4), 635-660.
- Lee, W., Stolfo, S. J., & Mok, K. W. (1999, May). A data mining framework for building intrusion detection models. In Proceedings of the 1999 IEEE Symposium on Security and Privacy (Cat. No. 99CB36344) (pp. 120-132). IEEE.
- Anderson, J. P. (1980). Computer security threat monitoring and surveillance. Technical Report, James P. Anderson Company.
- Lee, W., & Stolfo, S. (1998). Data mining approaches for intrusion detection, 7th USENIX Security Symposium, San Antonio, TX, 1998.
- Schultz, M. G., Eskin, E., Zadok, F., & Stolfo, S. J. (2000, May). Data mining methods for detection of new malicious executables. In Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001 (pp. 38-49). IEEE.
- Nadiammai, G. V., & Hemalatha, M. (2012, July). Perspective analysis of machine learning algorithms for detecting network intrusions. In 2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12) (pp. 1-7). IEEE.
- Hwang, T. S., Lee, T. J., & Lee, Y. J. (2007, June). A three-tier IDS via data mining approach. In Proceedings of the 3rd annual ACM workshop on Mining network data (pp. 1-6).
- Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In 2009 IEEE symposium on computational intelligence for security and defense applications (pp. 1-6). Ieee.
- Subramanian, S., Srinivasan, V. B., & Ramasa, C. (2012). Study on classification algorithms for network intrusion systems. Journal of Communication and Computer, 9(11), 1242-1246.
- NSL-KDD dataset, [Available Online] http://iscx.ca/NSLKDD.
- Lippmann, R., Haines, J. W., Fried, D. J., Korba, J., & Das, K. (2000). The 1999 DARPA off-line intrusion detection evaluation. Computer networks, 34(4), 579-595.
- Srinivasulu, P., Nagaraju, D., Kumar, P. R., & Rao, K. N. (2009). Classifying the network intrusion attacks using data mining classification methods and their performance comparison. International Journal of Computer Science and Network Security, 9(6), 11-18.
- Kalyani, G., & Lakshmi, A. J. (2012). Performance assessment of different classification techniques for intrusion detection. Learning, 2(1), J48.
- Reddy, E. K., Reddy, V. N., & Rajulu, P. G. (2021). A Detailed Study of Intrusion Detection in Data Mining. Research Issues on Datamining, 3-13.

- Neethu, B. (2012). Classification of intrusion detection dataset using machine learning approaches. International Journal of Electronics and Computer Science Engineering, 1(3), 1044-1051.
- Mukherjee, S., & Sharma, N. (2012). Intrusion Detection using Naive Bayes Classifier with Feature Reduction. Procedia Technology, 2012.
- Nejad, A.F., Kharazmi, S., & Bayati, S. (2008). Improving Admission Control Policies in Database Management Systems, Using Data Mining Techniques. International Conference on Computer Science and Software Engineering (ICCSSE) 2008.
- Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2011). Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. Expert Systems with Applications, 38(5), 5947-5957.
- Danane, Y., & Parvat, T. (2015, January). Intrusion detection system using fuzzy genetic algorithm. In 2015 International Conference on Pervasive Computing (ICPC) (pp. 1-5). IEEE.
- Masarat, S., Taheri, H., & Sharifian, S. (2014, October). A novel framework, based on fuzzy ensemble of classifiers for intrusion detection systems. In 2014 4th international conference on computer and knowledge engineering (ICCKE) (pp. 165-170). IEEE.
- Ingre, B., Yadav, A., & Soni, A. K. (2018). Decision tree based intrusion detection system for NSL-KDD dataset. In Information and Communication Technology for Intelligent Systems (ICTIS 2017)-Volume 2 2 (pp. 207-218). Springer International Publishing.
- Holland, J.H. (1975). Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence; MIT Press:Cambridge, MA, USA.
- Bernard, T., Gnauck, A., Jacobi, M., Karimanzira, D., Krol, O., Pfützenreuter, T., ... & Westerhoff, T. (2015). Modeling, Control and Optimization of Water Systems: Systems Engineering Methods for Control and Decision Making Tasks. Springer.
- Obitko, M. (1998). Introduction to genetic algorithms. Czech Technical University: Prague, Czech Republic.
- Capraro, C. T., Bradaric, I., Capraro, G. T., & Lue, T. K. (2008, May). Using genetic algorithms for radar waveform selection. In 2008 IEEE Radar Conference (pp. 1-6). IEEE.
- Chaiyaratana, N., & Zalzala, A. M. (1999, July). Hybridisation of neural networks and genetic algorithms for time-optimal control. In Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406) (Vol. 1, pp. 389-396). IEEE.
- Pujari, A. K. (2001). Data mining techniques. 4th edition, Universities Press (India) Private Limited.