Indian Journal of Science and Technology, Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/105827, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Significance of Hybrid Feature Selection Technique for Intrusion Detection Systems

Vaishali Chahar*, Rita Chhikara, Yogita Gigras and Latika Singh

Department of CSE/IT, The NorthCap University, Gurugram - 122017, Haryana, India;
tn.chahar@gmail.com, ritachhikara@ncuindia.edu, yogitagigras@ncuindia.edu, latikasingh@ncuindia.edu

Abstract

Objectives: Intrusion detection is the need of technical world where data is generating and changing at a very rapid rate. In last decade feature selection is the science that has given a new perspective to research in the area of Intrusion Detection System. Objective of this paper is to perform an analysis and comparison of various feature selection techniques with a new technique of hybrid Particle Swarm Optimization (PSO). **Statistical Analysis:** In this paper well known filter and wrapper feature selection techniques have been explored along with a hybrid PSO technique on the standard KDDCup99 dataset. A comparative analysis is performed over four filter techniques and two wrapper based techniques. Four different classifiers are compared to select the one providing good accuracy on the dataset. **Findings:** The hybrid PSO feature selection technique gives significant improvement in prediction capability as compared to traditional feature selection approaches. Analysis shows that SVM classifier provides better classification results. SVM is used as classifier because of its high accuracy. The analysis over 4 filter and two wrapper techniques shows that Hybrid PSO provides better results with 98.6% accuracy and 24 feature subset. **Application/Improvements:** Analysis provides importance of hybrid PSO, which may be applied to not only intrusion detection but also various other areas where feature reduction is required.

Keywords: Binary Particle Swarm Optimization (BPSO), Hybrid PSO, Intrusion Detection System (IDS), Support Vector Machine (SVM)

1. Introduction

With the increase in usage of internet users are facing challenges of known and unknown vulnerabilities. Traditional systems of firewall, antivirus etc. are not capable enough to prevent intruders and attackers due to high complexities of software applications. Hence, a technique is required which can prevent the system from various intrusions and unauthorized access. Intrusion can be performed by various ways such as stealing someone's password, unauthorized access of a system or network and acting like a legitimate user trying to access non permitted data. Intrusion Detection System (IDS) is one of the techniques which can detect attacks on a system and helps in handling the security problems. Intrusion Detection System (IDS) plays an important role to keep our network secure by detecting the malicious data coming to the network and generating alarm and taking specified action[1].

IDS can operate in different modes. The modes of operation are to have a basis for analysis of network packets. These metrics can be used to deduce whether a particular network or a system has been compromised or not. In most of the cases, the information collected indicates whether further action needs to be taken or not.

IDS can operate in two different modes: Anomaly detection based and Misuse detection based systems[2]. Anomaly detection method detects abnormal behavior in

*Author for correspondence

any network by checking a pool of normal behavior operations and it can detect new attacks in the system without any prior knowledge. It checks data with the already existing pool of normal data collected and if it senses any abnormality then it declares a new type of attack in system. It deals with variation of user behavior, also known as behavior based detection. The major advantage of anomaly detection is it has ability to detect novel or unknown attacks based on audit data[2].

Misuse Detection or Signature based is another technique where attacks are classified based on the attack signatures. This system works by comparing the activities with the already generated set of signatures. In this method, IDS inspects to detect abnormal behavior by analyzing the given traffic based on several rules and by comparing these rules the system can detect type of attacks. The advantage of misuse detection is accuracy in results, lesser false alarm rate. The disadvantage of misuse detection is that it cannot detect unknown attacks[2].

The attacks on the network can be broadly studied under four headings; Denial of Service (DoS), Probing attack, Remote to Local (R2L) and User to Local attacks (U2R)[3]. In Denial of Service the attacker disrupts the network services to a user by flooding the host server with unnecessary requests. The purpose of this is to make the host server too busy such that user is not able to access the network. The attackers of Denial of Service mainly target web related to banks or credit card payments. For example is Syn Flood, attacker sends a flood of TCP/SYN packets, mostly with a fake sender address. Another example of Denial of Service is Tear drop which works by sending IP fragments combined with a bug to the destination machine which can the operating system[3].

Probing attack is carried out by a program or a device which is inserted at main juncture in a network for capturing information about various network activities. One of the common methods of Probe attacking is Port Scanning. In Port Scanning the intruder listen for the open port TCP/UDP and forwards the packet with different destination port to get the information related victim's machine[3].

Remote to Local Attack (R2L) is applied by an attacker when there is no access to a remote machine. In this attacker sends a malicious program through packets to a victim's machine over the network. This helps the attacker to identify the vulnerable points and enter into victim machine to gain access of important information like password etc[4]. In User to Root Attack (U2R) the attacker captures the normal account on some system and then it exploits the system vulnerabilities to gain root access. For example: Buffer overflow attack is an abnormality where the attacker targets the buffer of the machine and while writing the data into it crosses the buffer boundary and writes into the neighboring memory locations.

Intrusion detection nowadays deals with large amount of big data. Dealing with big data requires excess overhead and is difficult to work with, because of its size, variation and veracity[4]. In intrusion detection two phase processing is performed: In first step feature reduction is performed where from a large number of features set, a small number of features are selected which can help in speed up intrusion detection without compromising efficiency, provides low rate of false alarm of the system and then selected features with paired condition perform the classification of data as intrusion or normal. This paper focuses on the feature selection techniques. Several data mining and genetic based techniques have been used for performing feature selection.

Feature selection is a technique used to address Big Data challenges[4]; it helps in reducing classification processing time. Feature selection should be applied carefully by considering the factors like relevancy and efficiency of the reduced set. If the selected features are not relevant to the problem then instead of helping in intrusion detection they will increase the false alarm rate in the system causing Intrusion Detection System a failure. Also from time to time types of attacks are changing rapidly which may cause some of features to be relevant and some irrelevant so points needs to be considered during feature extraction.

Various techniques are being used for feature reduction one of them is *evolutionary techniques*, **genetic algorithm** base[5] some of them are discussed here:

Author[6] used an evolutionary technique of Particle Swarm Optimization (PSO) with rough set concept for feature reduction. Where[7] used a Simplified Swarm Optimization (SSO) technique which is advancement over PSO for intrusion detection. The authors have applied proposed technique over different data sets. A modified PSO is used in[8] for feature selection. Author[9] uses Particle Swarm Optimization technique for feature reduction. By reducing the feature set it reduces the processing time. But a very small feature set can also leads to problem of high false positive rate of the system which will degrade system performance. In another research[10] proposes a nature inspired approach based on cuttlefish algorithm. Fuzzy

based[11] techniques are also used to perform the feature selection on KDDCup99 data set. In[12] a fuzzy based feature selection technique using Support Vector Machine is used. Other techniques include use of snort, fuzzy based rule system, Support Vector Machine[13] and various other data mining based techniques. In another research[14,15] proposed a feature selection based on neural network. In another research[16] introduced a Genetic Algorithm based on fuzzy to detect the network traffic.

So feature selection is important in Intrusion Detection but it should be performed with care so that it does not reduce the efficiency of the system. Different measures are used to check the efficiency and accuracy of an IDS system which includes prediction performance, time performance and fault tolerance[17]. Prediction function involves the classification rate of true and false classified network traffic. Systems performance depends on correct prediction rate. If rate decreases then false positive rate will be high. Time performances is the rate at which IDS is generating and propagating the results to resolve the attacks. The other factor is fault tolerance which requires IDS to be robust and ability to recover from attacks. In this paper a detailed analysis on various feature reduction techniques is performed over a new hybrid technique of GLBPSO[18] which are explained below.

2. Proposed Work

In the proposed the IDS is tackled as a pattern recognition problem consisting of two classes 'normal' and 'anomaly'. The purpose of this work is to find relevant features that help distinguish between the two classes. To classify data following three steps are performed:

- Feature Extraction.
- Feature Selection.
- Classification.

2.1 Feature Extraction

The features employed for the experiments are the existing 42 features available with NSL KDDcup99 dataset[19]. The 42 features can be classified broadly into four categories of Basic, Content, Traffic and Host features depending upon their origin and the class label which consists of two class labels 'normal' and 'anomaly'. The anomaly involves following attacks; DOS, Probe, R2L, U2R. A list of features is given in Table 1.

All the features extracted may not be significant hence the need for feature selection techniques evolved. Different feature selection techniques employed in the study are as discussed below.

Table 1. List of features in KDDcup99 dataset

Basic Features		Content Features		Traffic Features		Host Features	
Sr. No.	Attribute Name	Sr. No.	Attribute Name	Sr. No.	Attribute Name	Sr. No.	Attribute Name
1	Duration	1	Num_failed_logins	1	Count	1	dst_host_count
2	protocol_type	2	Logged_in	2	Serror_rate	2	dst_host_srv_count
3	Service	3	Num_compromised	3	Rerror_rate	3	dst_host_same_srv_rate
4	src_bytes	4	Root_shell	4	srv_diff_host_rate	4	dst_host_diff_srv_rate
5	dst_bytes	5	Su_attempted	5	srv_count	5	dst_host_same_src_port_rate
6	Land	6	Num_root	6	same_srv_rate	6	dst_host_srv_diff_host_rate
7	Flag	7	Num_file_creations	7	diff_srv_rate	7	dst_host_serror_rate
8	wrong_fragment	8	Num_shells	8	srv_rerror_rate	8	dst_host_srv_serror_rate
9	Urgent	9	Num_access_files	9	srv_serror_rate	9	dst_host_rerror_rate
		10	is_hot_login			10	dst_host_srv_rerror_rate
		11	is_guest_login				
		12	Num_outbound_cmnds				
		13	Hot				

2.2 Feature Selection Techniques

Feature extraction techniques are structured into two types: Filter and wrapper method. Filter method uses the inherent properties of the training set and does not require any learning algorithm. It is further of two type's multivariate and univariate methods. Univariate works on each feature separately where multivariate considers the relevance between multiple features[20].

In this paper two filter techniques are used which are Gain Ratio and ReliefF. ReliefF[21] is a multivariate method that works by weighting different features and it chooses the features that are most distinguishable or most irrelevant among different classes. ReliefF method is very robust to noise and can also deal with missing values in dataset. The Gain Ratio[22] approach is used with multivalued attributes to reduce the bias. It is the ratio of Information Gain to the integral information. Minimum redundancy maximum relevance (mRmR)[23] is a feature selection method which identifies variable which are distinct from each other but are highly relevant to the classes used for classification. It minimizes the redundancy in feature subset. Another technique used is Fisher Score[24] technique which is a supervised feature selection technique which finds a subset of features selected independently using a Fisher Score. Fisher Score technique on the other hand does not handle the redundancy issues in features like mRmR. The value of performance metrics[25] like TP rate, accuracy is measured.

Wrapper method works by learning however this method requires high computational time as compared to filter method and it also considers dependencies among features. Two wrapper based techniques applied are PSO[26] and GLBPSO[18].

2.2.1 Hybrid PSO - GLBPSO

PSO is a metaheuristic computational method proposed by Dr. Russell C. Eberhart and Dr. James Kennedy[27] in 1995. It is based on the concept of bird flocking. PSO works by randomly initializing of the population known as particles. Each particle is represented as a point in an X-dimensional space which moves around with certain velocity and updates their position based on their previous movements. To reach at an optimal solution each particle in X-Dimensional space moves in the direction of its previous best positions (pbest) and global best position (gbest). For each particle i and dimension j space x_i x_i is represented as $x_i = (x_i(i1), x_i2, \dots \dots x_ij)$ and velocity $v_i = (v_{i1}, v_{i2}, \dots \dots v_{ij})$ denotes the fly velocity of particle. The velocity and position vector are updated according to the equations:

$$v_{ij}^{t+1} = \mathbf{w} * v_{ij}^{t} + c_1 rand\left(pbest_{ij}^{t}\right) + c_2\ rand\left(gbest_{ij}^{t} - x_{ij}^{t}\right)$$

And, $x_{ij}^{t+1} = x_{ij}^{t} + v_{ij}^{t+1}$

In above formula is the velocity of i^{th} particle in j space whose value is limited to the range $[-v_{max}, v_{max}]$ and x_{ij} the position of particle and c_1 and c_2 are the constant social and personal learning factors acceleration coefficients[27].

For calculating the reduced feature set a technique hybrid PSO-GLBPSO[18] is applied which works in two phases where in first phase features are ranked based on the univariate and multivariate filter approaches and redundant features are removed. The ranked features are then applied to second phased which is a wrapper approach based on improved PSO, where the concept of gbest and pbest both are considered to find the best reduced feature set. If the value of gbest does not change for any two successive iterations then pbest value is considered. Also a new concept of hope/rehope is introduced where if the particle's fitness value decreases continuously then it will be discard in future iterations results in reducing swarm size. Value of two acceleration coefficients *c1* and *c2* is set to 2 and inertia weight *w* value is set to 1 for GLBPSO.

3. Experimental Results and Analysis

For observing the performance of different techniques, NSL-KDDCup99 dataset is used which is a refined form of standard KDDCup99 dataset[28]. Experiments are performed by using MATLAB 7.10.0 (R2010a) on windows7 PC with 8GB RAM.

3.1 Data Set

KDD Cup dataset is a standard dataset which is prepared by the MIT Lincoln lab, KDD dataset is derived from the standard DARPA'98 dataset. In this experiment setup NSL-KDDCup99 dataset is used which is improved version of KDDCup99 where duplicate data are removed which helps in unbiased results[29]. It has 41 features with a class label which categorize the data as normal or attack type. There are total 37 attacks

specified in the dataset which are categorized into one of four categories of DOS, Probe, R2L and U2R. KDD contains attributes values both as continuous and symbolic. Before performing feature reduction preprocessing of dataset is performed by converting symbolic attribute into continuous. The dataset is divided into two sets training set and test set. The training set consists of 22000 samples and test set consists of 2000 samples.

3.2 Experimental Results

Various experiments have been performed to evaluate the KDDCup dataset and identify relevant and non-redundant features. Three classifiers SVM[30], C4.5[31], Naive Bayes[32] and Random Forest[33] are employed to test the performance of the dataset and differentiate between the data as anomaly or normal. To find relevant features two feature selection techniques ReliefF[21], Gain Ratio[22], mRmR[23], Fisher Score[24] are applied. The results of these techniques are then compared with the new hybrid technique GLBPSO.

3.2.1 Performance of Different Classifiers

The performance of dataset with four different classifiers[34]; SVM, C4.5, Naive Bayes and Random Forest is tested. As observed from in Table 2 SVM[35] gives the best accuracy. The reason behind this is the strong mathematical foundation of SVM[36].

Table 2. Performance with different classifiers

Classifier	Accuracy
Naïve byes	54.9%
C4.5	60.3%
SVM	64%
Random Forest	58.8%

3.2.2 Performance of Feature Selection Techniques in Terms of Classification Accuracy

Six different filter feature selection techniques are applied on the dataset. The classification is performed with SVM as it is giving best prediction results as discussed in Section 4.2.1. Four of these filter techniques (ReliefF, GainRatio, mRmR, Fisher) are filter approaches and two wrapper approaches (PSO, Hybrid PSO-GLBPSO). As observed from in Figure 1 Hybrid GLBPSO is giving the highest classification accuracy with just 24 features. The features selected by each technique are depicted in Figure 2. The reason behind this is that filter approaches are able to remove irrelevant features in Phase I and modified wrapper PSO approach removes redundant features in the second Phase.

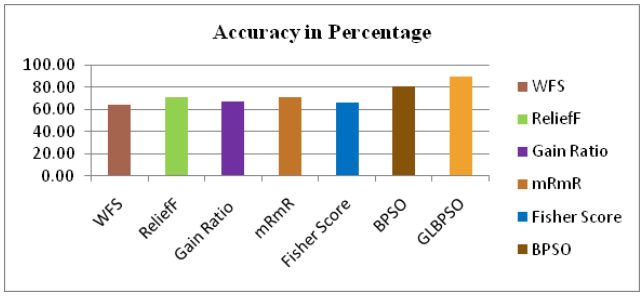

Figure 1. Classification accuracy of feature selection techniques.

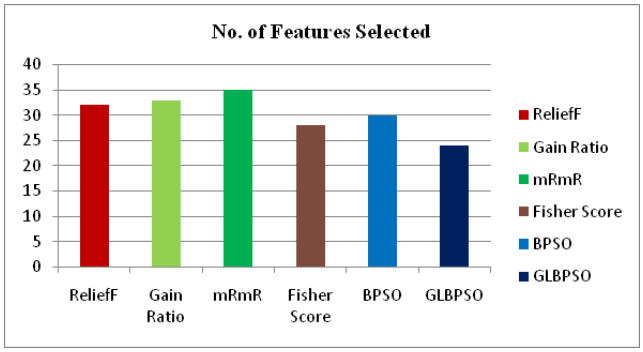

Figure 2. Reduced features of feature selection techniques.

3.2.3 Feature Analysis

A total of 24 features are selected by the GLBPSO approach which is listed in Table 3 given.

The 41 features of dataset can be classified into four categories as mentioned in Section 3.1. From feature selected by GLBPSO mentioned in Table 3 first six features selected are of Basic category, next six are from Content, five are from Traffic and maximum seven features are selected from Host category of classification. Hence it is concluded that features from Host category are more significant in classifying the data under two categories of 'normal' or 'anomaly'.

Table 3. List of features selected by GLBPSO

Sr. No.	1	2	3	4	5	6	7	8	9	10	11	12
Feature Selected	duration	service	flag	dst_ bytes	land	wrong_ fragment	hot	root_ shell	num_ root	num_ outbound_ cmds	is_host_ login	is_ guest_ login
Sr. No.	13	14	15	16	17	18	19	20	21	22	23	24
Feature Selected	serror_ rate	srv_ serror_ rate	rerror_ rate	srv_ rerror_ rate	srv_ diff_ host_ rate	dst_host_ count	dst_ host_ srv_ count	dst_ host_ same_ src_ port_ rate	dst_ host_ srv_ diff_ host_ rate	dst_host_ serror_rate	dst_ host_ srv_ serror_ rate	dst_ host_ srv_ rerror_ rate

4. Conclusion and Future Work

A detailed analysis on various feature selection techniques are performed where accuracy and number of feature selected are considered as measures. SVM is the classifier giving best results on KDDCup99 dataset when compared to three other well-known classifiers. The experimental result on feature selection shows that Hybrid GLBPSO is superior to other five feature selection techniques. The features from Host category are found to be more relevant for classification in IDS. Future work could involve developing new hybrid feature selection techniques with other evolutionary algorithms such as Firefly etc. Multi-class classifiers can be explored to work on dataset with all the class labels.

5. References

1. Lazarevic A, Kumar V, Srivastava J. Intrusion detection: A survey. Managing Cyber Threats; Springer US. 2005. p. 19–78.
2. Amudhavel J, Brindha V, Anantharaj B, Karthikeyan P, Bhuvaneswari B, Vasanthi M, Nivetha D, Vinodha D. A Survey on Intrusion Detection System: State of the Art review. Indian Journal of Science and Technology. 2016 Mar; 9(11):1–9.
3. Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Computing Surveys (CSUR). 2009; 41(3):15.
4. Zuech R, Khoshgoftaar TM, Wald R. Intrusion detection and big heterogeneous data: A survey. Journal of Big Data. 2015 Feb; 2(1):1.
5. Effatnejad R, Rouhi F. Unit commitment in power system by combination of Dynamic Programming (DP), Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). Indian Journal of Science and Technology. 2015 Jan; 8(2):1–8.
6. Wang X, Yang J, Teng X, Xia W, Jensen R. Feature selection based on rough sets and Particle Swarm Optimization. Pattern Recognition Letters. 2007; 28(4):459–71.
7. Chung YY, Wahid N. A hybrid network Intrusion Detection System using Simplified Swarm Optimization (SSO). Applied Soft Computing. 2012; 12(9):3014–22.
8. Gong S, Gong X, Bi X. Feature selection method for network intrusion based on GQPSO attribute reduction. IEEE 2011 International Conference on Multimedia Technology; 2011. p. 6365–9.
9. Elngar A, Mohamed DA, Ghaleb F. A real-time anomaly network Intrusion Detection System with high accuracy. Information Sciences Letters International Journal. 2013 May; 2(2):49–56.
10. Kaur R, Sachdeva M, Kumar G. Nature inspired feature selection approach for effective intrusion detection. Indian Journal of Science and Technology. 2016 Nov; 9(42):1–9.
11. Tsang CH, Kwong S, Wang H. Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection. Pattern Recognition. 2007 Sep; 40(9):2373–91.
12. Suganthi J, Malathi V. Fuzzy based feature selection scheme through transductive SVM technique for cancer pattern classification and prediction. Indian Journal of Science and Technology. 2016 May; 9(16):1–12.
13. Renjit JA, Shunmuganathan KL. Network based anomaly Intrusion Detection System using SVM. Indian Journal of Science and Technology. 2011 Sep; 4(9):1105–8.
14. Wang H, Liu X, Lai J, Liang Y. Network security situation awareness based on heterogeneous multi-sensor data fusion and neural network. IEEE Second International

Multi-Symposiums on Computer and Computational Sciences; 2013 Aug. p. 352–9.
15. Jayasri T, Hemalatha M. Categorization of respiratory signal using ANN and SVM based on feature extraction algorithm. Indian Journal of Science and Technology. 2013; 6(9):5195–200.
16. Danane Y, Parvat T. Intrusion Detection System using fuzzy Genetic Algorithm. IEEE International Conference on Pervasive Computing (ICPC); 2015 Jan. p. 1–5.
17. Wahid N. A novel approach to data mining using Simplified Swarm Optimization. [Doctoral dissertation]. University of Sydney: 2011. p. 1–54.
18. Chhikara R, Sharma P, Singh L. A hybrid feature selection approach based on improved PSO and filter approaches for image steganalysis. International Journal of Machine Learning and Cybernetics. Springer. 2016; 7(6):1195–1206.
19. Nsl-kdd data set for network-based Intrusion Detection Systems. 2016. Available from: http://nsl.cs.unb.ca/NSL-KDD/
20. Mamoun A,S hamsul H, Jema LA, Rafiqul I, John Y, Sitalakshmi V, Roderick B. Hybrids of Support Vector Machine wrapper and filter based framework for malware detection. J Netw. 2014; 9(11):2878–91.
21. Kononenko I. Estimating attributes: Analysis and extensions of Relief. De Raedt L. and Bergadano F. editors. Machine Learning: ECML-94, Springer Verlag; 1994. p. 171–82.
22. Hall M. Correlation-based feature selection for discrete and numeric class machine learning. Proceedings of the 17th International Conference on Machine Learning; 2000. p. 359–66.
23. Peng HC, Long F, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005; 27(8):1226–38.
24. Lu JC, Liu FL, Luo XY. Selection of image features for steganalysis based on the Fisher criterion. Digit Invest. 2014; 11(1):57–66.
25. Bahl S, Dahiya D. Enhanced Intrusion Detection System for detecting rare class attacks using correlation based dimensionality reduction technique. Indian Journal of Science and Technology. 2016 Mar; 9(11):1–10.
26. Srinivas RS, Rao PR. PSO tuned fractional order PI controller for improvement of transient stability improvement in multi machine power system. Indian Journal of Science and Technology. 2016 Oct; 9(38):1–5.
27. Eberhart RC, Kennedy J. A new optimizer using particle swarm theory. Proceedings of the Sixth International Symposium on Micro Machine and Human Science; 1995 Oct. p. 39–43.
28. KDD Cup 1999. 2016. Available from: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
29. Aggarwal P, Dahiya D. Contribution of four class labeled attributes of KDD dataset on detection and false alarm rate for Intrusion Detection System. Indian Journal of Science and Technology. 2016 Feb; 9(5):1–8.
30. Cortes C, Vapnik V. Support vector networks. Machine Learning. 1995 Sep; 20(3):273–97.
31. Kohavi R, Quinlan JR. Data mining tasks and methods: Classification: Decision-tree discovery. Handbook of Data Mining and Knowledge Discovery; 2002 Jan. p. 267–76.
32. Jiang L, Cai Z, Zhang H, Wang D. Not so greedy: Randomly selected naive Bayes. Expert Systems with Applications. 2012 Sep; 39(12):11022–8.
33. Ho TK. Random decision forests - Document analysis and recognition. Proceedings of the IEEE Third International Conference; 1995 Aug. p. 278–82.
34. Meenakshi, Geetika. Survey on classification methods using WEKA. International Journal of Computer Applications. 2014; 86(18):16–9.
35. Ghate VN, Dudul SV. SVM based fault classification of three phase induction motor. Indian Journal of Science and Technology. 2009 Apr; 2(4):32–5.
36. Palanisamy R, Vijayakumar K, Nikhil K, Iyer M, Rao R. A proposed SVM for 3-level transformer-less dual inverter scheme for grid connected PV system. Indian Journal of Science and Technology. 2016 Nov; 9(42):1–7.