International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Intrusion Detection System Using Weka Data Mining Tool

Asma Abbas Hassan[1], Alaa F. Sheta [2], Talaat M. Wahbi[3]

[1]Computer Science, Sudan University of Science & Technology Khartoum, Sudan

[2]Computer Science Department, Taif University, Taif , Saudiarabia

[3]Computer Science, Sudan University of Science & Technology Khartoum, Sudan

Abstract: ***Traditional intrusion prevention techniques, such as firewalls, access control or encryption, have failed to fully protect networks and systems from increasing attacks.Therefore an intrusion detection system (IDS) has become an important component of security infrastructure and a key part of system defense to detect these attacks before they make a disaster in the system. In this paper, we are going to design an intrusion detection system using Weka Data Mining Software, to check the existence of intrusion, and classify it when detected, to know the type of intrusion, according to attack types, this will implemented in Weka 3.6 Software, with KDD CUP 99 intrusion detection dataset. We construct a system with a very accurate, flexible and effective results when compare with other systems***

Keywords: intrusion detection, Naive bayes, Decision Trees, Random Forest, Random Tree, Confusion Matrix

1. Introduction

All of our needs is become on network, and the wide spread of computer network and its application lead to various types of intruders and hackers and they attacks for different purposes, and this intrusion may lead to terrible disaster for the user of the network, that is why the network connection has to be very secure and protected every ones privacy, doing this by computer manufactures in hardware is very difficult from both technical and economical views, in this paper we introduce a software solution, which is a new intrusion detection system used to find a way to detect the intrusion, and classify it according to attack types, the experiments are done by using Weka Data Mining Software.

1.1. Classification Techniques For Intrusion Detection

1.1.1. Classification

Classification algorithms in data mining are used in Intrusion Detection System to classify attacks or intrusions from ordinary things happen in systems. Classification algorithms are supervised learning approach, it doesn't require class names for the prediction reason. Classification algorithms are primarily fall into two classifications, one is binary classification and other one is multiclass classification.[1] Binary classification technique classifies the element of a given set into two groups on the basis of whether they have some characteristic or not, while multiclass classification technique classifies instances into more than two classes. Some classification algorithms naturally allow the utilization of more than two classes; others are by nature binary algorithms.Here are some differnet classification techniques:

1.1.1.1. Support Vector Machine

Support vector Machine (SVM), a promising pattern classification technique, [21]. SVMs are supervised learning models with associated learning algorithms that have been applied increasingly to misuse detection in the last decade.

1.1.1.2. Decision tree

Quinlan [22] proposed a decision tree classifier which is one of the most known machine learning techniques. A decision tree composed of three basic elements [23]:

- A decision node representing test or condition on data item.
- An edge or a branch which corresponds to the one of the possible attribute values whichmeans one of the test attribute outcomes.
- A leaf which determines the class to which the object belongs

1.1.1.3. Naïve Bayes

Naïve Bayes can be considered as an upgraded version of Bayes Theorem as it assumes strong independence among attributes. Bayesian classifier encodes probabilistic relationships among variables of interest. This means that the probability of one attribute does not affect the probability of the other.

Mrutyunjaya Panda and Manas Ranjan Patra [24] proposed a framework of network intrusion detection system based on naïve bayes algorithm.

1.1.1.4. Neural Networks

A neural network consists of a collection of processing elements that are highly interconnected and transform a set of inputs to a set of desired outputs. The result is determined by the characteristics of the elements and the weights associated with the interconnections between them. By modifying the connections between the nodes, the network can adapt to the desired outputs. Neural networks have been used in both anomaly detection and misuse detection. For anomaly detection, neural networks were modelled to learn the typical characteristics of system users and identify significant variations from the user's established behaviour as anomaly. In misuse detection, the neural network would receive data from the network stream and analyze the information for instances of misuse [25].

2. Related Works

There are many researches works on intrusion detection with various techniques, they all told us about the importance of this problem and try to solve it in various ways, some of those researches are summarized as follows:

- L. Portnoy, E. Eskin, and S. Stolfo, used k-means and the fuzzy c-means clustering methods in intrusion detection. The clustering techniques has a disadvantage, that it is based on calculating numeric distance between the observations, so, the observations had to be numeric, and another disadvantage is that, the clustering methods unable to capture the relationship between different features of a single record, which degrades attack detection accuracy.[10]
- N.B. Amor, S. Benferhat, and Z. Elouedi, used Naive Bayes classifiers for intrusion detection. But in this the size of a Bayesian network increases rapidly as the number of features and the type of attacks modeled by a Bayesian network increases.[3][6]
- Yusufovna, S.F, wrote a paper "Integrating Intrusion Detection System and Data Mining" in this paper, he used data mining approaches for intrusion detection. The method were used can deal with symbolic data and the features can be defined in the form of packet and connection details. [5]
- Oludele Awodele, Sunday Idowu, Omotola Anjorin, and Vincent J. Joshua, wrote a paper presents an Intelligent Intrusion Detection and Prevention System (IIDPS), which monitors a single host system from three different layers; files analyzer, system resource and connection layers. The approach introduced, a multi – layered approach, in which each layer harnesses both aspects of existing approaches, signature and anomaly approaches, to achieve a better detection and prevention capabilities.[14]
- Mansour Sheikhan and Amir Khalili [5] proposed an algorithm to develop IDS and classify the patterns of intrusion. To evaluate the performance of their system with other machine learning algorithms, multi-layer perception (MLP). They get better performance on their result.
- Kyaw Thet Khaing [6] proposed a model using SVM consisting of Recursive Feature Elimination and a k-Nearest Neighbor (KNN) technique to carry out a feature ranking and selection job of the model. He gets efficient model in his result.
- Gang Wang et al.[7] Proposed an approach, called FC-ANN, based on ANN and fuzzy clustering, he gets better performance on his result.
- Muna Mhammad T. Jawhar and Monica Mehrotra [8] presented an intrusion detection model using neural network and hybrid fuzzy logic. The model proposed to identify the attack and classify attacks and the data was from KDD CUP intrusion detection data set.
- Pohsiang Tsai *et al.*[9] Suggested a Machine Learning (ML) framework that detect the type of intrusion detection and classify it using the KDD CUP data set. Their approach obtained higher performance in comparison with other state-of-the-art detection methods.

3. Methodology

3.1. Intrusion detection definition

Intrusion detection is used to catch intrusions or attacks into computer and network systems when they want to violate the system security and privacy [1], as illustrated in figure (1) below:

Figure 1: intrusion detection

3.2. Intrusion Detection Overview

Any intrusion detection system has four major categories that must be respected and put in mind when intrusion detection system mentioned.

These categories are:
1) Intrusion type
2) Detection behavior.
3) Detection approaches.
4) System types.

The categories are illustrated in the figure (4) below:

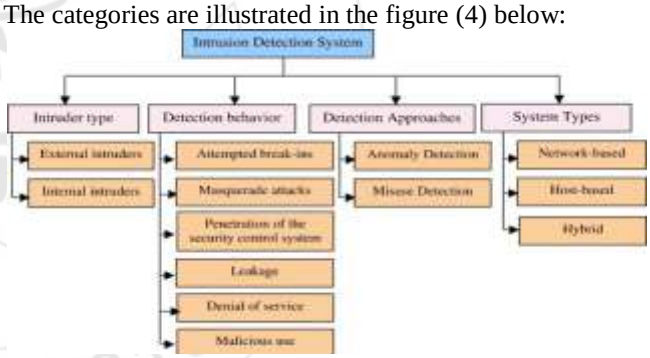

Figure 4: Intrusion detection system categories

3.2.1. Networking Attacks

There are four major categories of attacks that KDD data set covers:
1) Probing attacks (information gathering attacks).
2) Denial of- Service (DoS) attacks (deny legitimate requests to a system). E.g. guessing passwords
3) User-to-root (U2R) attacks (unauthorized access to local super-user or root). E.g. various buffer overflow" attacks;
4) Remote-to-local (R2L) attacks (unauthorized local access from a remote machine). E.g. port scanning [11]

3.2.2. Classification of Intrusion Detection

We can classify the Intrusions Detection into two main categories. They are as follows:
(1) Host Based Intrusion Detection systems(HIDSs), which evaluate information found on a single or multiple host systems, including contents of operating systems, system and application files. as illustrated in fig (5) [12]

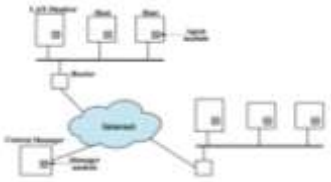

Figure 5: Host Based Intrusion Detection systems

(2) Network Based Intrusion Detection systems (NIDSs), which evaluate information captured from network communications, analyzing the stream of packets which travel across the Network. as illustrated in fig (6) [12]

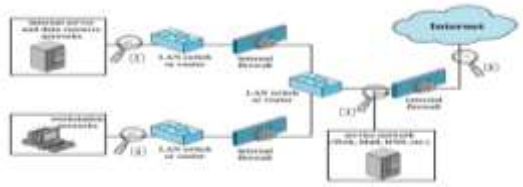

Figure 6: Network Based Intrusion Detection systems

(3) Hybird Intrusion Detection Systems

Some of the most current intrusion detection system only uses one of the two detection methods, misused detection or anomaly detection both of them have their own limitations, this is the technique which combines misuse detection system and anomaly detection system is known as hybrid intrusion detection system or we can say that the technique [16] which combines the network intrusion detection system and host intrusion detection system is known as hybrid intrusion detection system. as illustrated in fig (7)

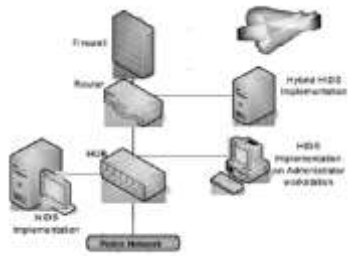

Figure 7: Hybird Intrusion Detection Systems

3.2.3. Components of Intrusion Detection System

Any intrusion detection system must consist of three main components [4]

(1) The data source(event generator), it can be categorized into four categories, namely Host-based monitors, Network-based monitors, Application-based monitors and Target-based monitors.

(2) The analysis engine. Which takes information from the data source and examines the data for attacks. There are two ways of analysis attacks:

- Misuse/Signature-Based Detection: This detects intrusion that follows well-known patterns of attacks (or signatures). [5,6]
- Anomaly/Statistical Detection: An anomaly based detection engine will search for something rare or unusual. [7]
- The response manager. Which works only when finding possible intrusion attacks on the system, as illustrated in fig (8)

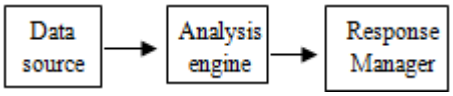

Figure 8: Components of Intrusion Detection System

3.3. Weka Data Mining Software

Weka (Waikato Environment for Knowledge Analysis), is a collection of machine learning algorithms for data mining tasks, it's afree software available under the GNU (General Public License), it's developed at the University of Waikato, New Zealand. [17,18]

The algorithms can either be applied directly to a dataset or called from your own Java code. [12,13] Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

3.4. Why use the Weka Data Mining Software

The WEKA tool incorporates the following steps [19,20]

- Analysis and pre-processing of the features in the database and assessing the correctness of the data.
- Definition of the class attributes which divide the set of instances into the appropriate classes.
- Extraction of the potential features to be used for classification.
- Selection of a subset of features to be used in the learning process.
- Investigation of a possible imbalance in the selected data set and how it may be counteracted.
- Selection of a subset of the instances, i.e. the records that learning is to be based on.
- Application of a classifier algorithm for the learning process.
- Decision on a testing method to estimate the performance of the selected algorithm.

3.4.1. The Main Advantages Of Weka Data Mining

Weka data mining can truly aid an enterprise attain its fullest prospective. It is an approach to evaluate how business is becoming impacted by particular qualities, and may assist company entrepreneurs improve their earnings and steer clear of generating company mistakes down the line. Fundamentally, through this process, a company is analyzing specific information from distinct perspectives to be able to obtain a total rounded watch of how their business is performing. Enterprise proprietors can get a broad point of view on points these as client trending, where they may be shedding cash and where they're creating cash. The knowledge may also reveal methods that may help a business lower unneeded fees and may aid them boost their overall income.

3.4.2 . Characteristics Of Weka Data Mining

Weka data mining, computer software, can let any organization to evaluate and analyze their information in more effective ways. It could deal with users in a friendly way, that is by allowing them to look into their information individually from a variety of distinct angles and factors to watch. In addition to that, weka data mining software allows the user to find out the correlations and patterns of his

respective personal information in contrast to individuals across numerous other localized directories.

4. Results and Discussions

4.1 Classification Model

Here we take nine algorithms in Weka for the classification,the test option which used in all techniques is cross-validation with 10 folds.

4.1.1 Trees

Decision Trees

The Decision tree is a classifier algorithm .

- It uses training data to build a decision trees.
- It chooses one attribute of the data at each node of the tree, that splits its set of samples into subsets enriched in one class or the other.
- The normalization is the criterion that decision trees based on in getting the results by choosing an attribute from the splitting data.
- The attribute with the highest normalized information gain is chosen to make the decision.
- For each attribute, the gain is calculated and the highest gain is used in the decision node. There are many algorithms in decision tree, we examine six of them in our experement (J48, J48 graft, Random forest, Simple chart, Rep tree and Random tree), and gain different results.

Random Tree

Random Tree is a supervised Classifier; it is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree. Random trees have been introduced by Leo Breiman and Adele Cutler. The algorithm can deal with both classification and regression problems. Random trees is a collection of tree predictors that is called forest . [26], [27]

Rep tree

RepTree uses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree. Basically Reduced Error Pruning Tree ("REPT") is fast decision tree learning. [26] [28] [29]

J48

J48 is a version of an earlier ID3 algorithm [30] developed by J. Ross Quinlan. J48 is an open source java implementation of the C4.5algorithm in the weka data mining tool. The J48 Decision treeclassifier follows the following simple algorithm. In order toclassify a new item, it first needs to create a decision tree taken six different data sets based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminatesthe various instances most clearly [31].

Simple chart

Simple Chart method is CART (Classification And Regression Tree) analysis.It was developed by Leo Breiman in the early 1980s. It is used for data exploration and prediction also. Classification and regression trees are classification methods which in order to construct decision trees uses historical data.[32]

Random Forest

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large.[33]

J48 graft

J48graft generates a grafted DT from a J48 tree. The grafting technique adds nodes to an existing decision tree with the purpose of reducing prediction errors.[34] These algorithm identifies regions of the instance space that are not occupied by training instances, or occupied only by misclassified training instances, and consider alternative classifications for those regions. In other words, a new test will be performed in the leaf, generating new branches that will lead to new classifications. Grafting is an algorithm for adding nodes to the tree as a post-process. Its purpose is to increase the probability of rightly classifying instances that fall outside the areas covered by the training data. Grafting is a post-process that can be applied to decision trees. Its aim is to decrease prediction error by reclassifying regions of the instance space where no training data exists or where there is only misclassified data. Its aim is to find the best matched cuts of existing leaf regions and branches out to create new leaves with other classifications than the original. Though tree becomes more complex, but here only branching that does not introduce any classification errors in data already rightly classified is considered. [35]

4.1.2 Bayes

(Naive bayes)

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes‟ theorem with the “naive” assumption of independence between every pair of features.

4.1.3 Function

(SMO, RBF Network)

(SMO)

The new SVM learning algorithm is called Sequential Minimal Optimization (or SMO). Instead of previous SVM learning algorithms that use numerical quadratic programming (QP) as an inner loop, SMO uses an analytic QP steps.[36]

(RBF Network)

RBF network consists of three layers, an input layer, which reads n inputs, a hidden layer that consists of m radial basis functions, and an output layer consisting of a linear additive function, which produces the response. The input neurons are linear and pass the input along to hidden neurons without any processing. The input feeds forward to each hidden

neuron. Using radial basis function the hidden neuron computes the signal and pass on these signals through weighted pathways to the linear output neuron which sums these up and generates an output signal. [37]. The difference between the standard feed forward neural network and the radial basis function network is that hidden neurons compute signals based on distances to training data points rather than inner-products with weight vectors, and the signals from each of the hidden neurons are linearly superposed at an output neuron using tunable weights.[38],figure(9)illustrstes.

Figure 9: RBF Network

5. Evaluation Model

5.1 Cross Validation

Cross-validation, is a system for evaluating how the aftereffects of a factual dissection will sum up to an independent data set. It is mostly utilized within settings where the objective is predicted, and one needs to estimate how exactly a prescient model will perform in practice. 10-fold cross validation is ordinarily utilized. In stratified K-fold cross- validation, the folds are chosen so that mean response value is roughly equivalent in all the folds.
The test option which used in all techniques is cross-validation with 10 folds.

Technique		Result (Accuracy)	
		Correctly Classified Instances	Incorrectly Classified Instances
1	J48	95.9662 %	4.0338 %
2	Random tree	94.8641 %	5.1359 %
3	Rep tree	94.3398 %	5.6602 %
4	Random forest	97.5284 %	2.4716 %
5	Simple chart	95.0139 %	4.9861 %
6	J48 graft	96.0625 %	3.9375 %
7	NaïveBayes	85.8121 %	14.1879 %
8	SMO	91.6114 %	8.3886 %
9	RBFnetwork	86.8393 %	13.1607 %

That means the best technique to use in Spam detection is Random forest, J48 graft then J48 and Simple chart.

5.2. Confusion Matrix

A confusion matrix is a visualization tool used in supervised learning (in unsupervised learning it is called a matching matrix). A confusion matrix that summarizes the number of instances predicted correctly or incorrectly by a classification model.

6. Conclusion

The motivation behind this work is to watch how these calculation are utilized within the order the Intrusion Detection Attacks. In this manner the utilization of measurement lessening procedure is a vital assignment in this work to assess the execution. Here, nine characterization models, for example, j48 graft, J48 and Random forest are utilized and looked at their execution within three separate dimensionalities utilizing weka toolbox.From the result it is watched that Random forest performs better results in acuracy.

References

[1] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html Anazida Zainal; Mohd Aizaini Maarof ; Siti Mariyam Shamsuddin,(2009): Ensemble Classifiers for Network Intrusion Detection System, Journal of Information , Universtiy Teknologi Malaysia.
[2] Andrea Janssen , (2009): Hybrid Model International Journal of Computer Science and Network Security, October 2009.
[3] Hossein, M. Shirazi, (2009): Anomaly Intrusion Detection System Using Information Theory, K-NN and KMCAlgorithms "Australian Journal of Basic and Applied Sciences.
[4] James, P. Anderson(1980): Computer security threat monitoring and surveillance, Washington, Pennsylvania, Journal of Computer Science and Network Security, USA,
[5] Kristopher Kendall (1999): A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems "Massachusetts Institute of Technology, Journal of Computer Information technology..
[6] Li, Yuang, Guo, L.(2007): An Active Learning Based TCM-KNN Algorithm for Supervised Network Intrusion Detection. 26th Journal of Computers & Security.
[7] Sammany, M.; Medhat, T. (2007): Dimensionality Reduction Using Rough Set Approach for Two Neural Networks- Based Applications, Journal (RSE ISP) University, Tanta, Egypt.
[8] N.B. Amor, S. Benferhat, and Z. Elouedi, 2004 Naive Bayes vs. Decision Trees in Intrusion Detection Systems
[9] Yusufovna, S.F,Oct 2008 Integrating Intrusion Detection System and Data Mining.
[10] L. Portnoy, E. Eskin, and S. Stolfo, 2001 Intrusion Detection with Unlabeled Data Using Clustering.
[11] Christopher Kruegel ,Darren Mutz William ,Robertson Fredrik Valeu , Reliable Software Group University of California , Bayesian Event Classification for Intrusion Detection.
[12] J. P. Planquart, "Application of Neural Networks to Intrusion Detection", SANS Institute Reading Room.
[13] R. G. Bace, "Intrusion Detection", Macmillan Technical Publishing. 2000.
[14] Awodele, Oludele; Idowu, Sunday; Anjorin, Omotola; Joshua, Vincent J., "A Multi-Layered Approach to the Design of Intelligent Intrusion Detection and Prevention System (IIDPS)", Academic journal article from Issues

in Informing Science & Information Technology, Vol. 6.2009
[15] Intrusion Detection with Hidden Markov Model and WEKA Tool International Journal of Computer Applications (0975 – 8887) Volume 85 – No 13, January 2014
[16] S. Kumar, "Classification and Detection of Computer Intrusions", Purdue University, 1995.
[17] D. Patterson, F. Liu, D. Turner, A. Concepcion, and R. Lynch. Performance Comparison of the Data Reduction System. Proceedings of the SPIE Symposium on Defense and Security, Orlando, FL, March 2008.
[18] 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, Volume 11 , Issue 1, pp. 10-18, 2009.
[19] B.X.Wang, D.H.Zhang, J.Wang, et al, "Application of Neural Network to Prediction of Plate Finish Cooling Temperature", Journal ofCentral South University of Technology, 2008,15(1):136140.
[20] Ian H.Witten and Elbe Frank, "Datamining Practical Machine Learning Tools and Techniques", Second Edition, Morgan Kaufmann, San Fransisco, 2005.
[21] Cortes, Vapnik, Support-vector networks, Machine Learning, vol.20, 1995, pp.273–297
[22] Quinlan, C4.5: Programs for Machine Learning, 1993, Morgan Kaufmann Publishers, San Mateo,CA.
[23] Ben Amor, Benferhat, Elouedi, "Naive Bayes vs. Decision Trees in Intrusion Detection Systems,"Proc. of the 2004 ACM symposium on applied computing, 2004, pp. 420–424.
[24] Mrutyunjaya Panda, Manas Ranjan Patra, "Network Intrusion Detection Using Naïve Bayes,"International Journal of Computer Science and Network Security,vol.7 no.12, 2007, pp.258-262.
[25] J. Cannady, "Artificial Neural Networks for Misuse Detection," National Information Systems Security Conference, 1998.
[26] Ian H. Witten, Eibe Frank & Mark A. Hall., "Data Mining Practical Machine Learning Tools and Techniques, Third Edition." Morgan Kaufmann Publishers is an imprint of Elsevier.
[27] Bernhard Pfahringer, "Random model trees: ane_ective and scalable regression method" University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/~bernhard
[28] Dr. B. Srinivasan, P.Mekala, "Mining Social Networking Data for Classification Using REPTree", International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 10, October 2014 pp- 155-160
[29] Payal P.Dhakate, Suvarna Patil, K. Rajeswari, Deepa Abin, "Preprocessing and Classification in WEKA Using Different Classifier",Int. Journal of Engineering Research and Applications, Vol. 4, Issue 8(Version 5), August 2014, pp- 91-93
[30] J.R.Quinlan, Induction of decision trees, Machine Learning, vol. 1, no. 1, pp. 81–106,1986.
[31] Anil Rajput , " J48 and JRIP Rules for E-Goverence Data" International Journal of Compter Science & Security(IJCSS), Vol 5, Issue 2: 2011.
[32] Sunita B. Aher, Lobo L.M.R.J., "COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS", International Journal of Information Technology and Knowledge Management, July-December 2012, Volume 5, No. 2, pp. 239-243
[33] Brijesh Kumar Baradwaj, "Mining Educational Data to Analyse Students Performance" International Journal of Advanced Computer Science and Applications, Vol 2.
[34] J.Quinlan, Simplifying decision trees, Int. J. Human Computer Studies.
[35] Dipti D. Patil, V.M. Wadhai, J.A. Gokhale.Evaluation of Decision Tree Pruning
[36] Cortes, C., Vapnik, V., "Support Vector Networks," Machine Learning, 20:273-297, (1995).
[37] P. J. Joseph, Kapil Vaswani, Matthew J. Thazhuthaveetil,APredictive Performance Model for Superscalar Processors Microarchitecture,MICRO-39. 39th Annual IEEE/ACM International Symposium, on Volume, Issue, Page(s):161 -170](Dec. 2006).
[38] Satish Kumar, Neural Networks A Classroom Approach,Tata McGraw Hill, (2006).