Building Efficient Intrusion Detection System Using Factor Analysis and Support Vector Machines

P Indira Priyadarsini
Dept of Computer Science & Engg.
Acharya Nagarjuna University
Guntur,A.P.,India

I Ramesh Babu
Dept of Computer Science,&Engg.
Acharya Nagarjuna University
Guntur,A.P.,India

***Abstract* - Intrusion detection is a critical issue in network security, for protecting network resources. Therefore an accurate system of detecting intrusions is to be built to give assurance for information in any organization either public or private. The main goal is to increase the detection rate and reduce the false alarm rate. Since existing Intrusion Detection Systems (IDSs) use all the features to detect known intrusions, they achieve depressed results. We have proposed an algorithm Factor Analysis based Support Vector Machine (FA-SVM) for developing efficient IDS by making use of popular statistical technique called Factor Analysis (FA) through which the features are analyzed as factors. To design more effective and efficient IDSs it is very essential to select the best classifiers. Therefore we used Support Vector Machines (SVMs) which are good enough with high generalization ability. This work is done on knowledge discovery and data mining cup dataset for conducting tests. The performance of this approach was analyzed and compared with existing approaches like Principal Component Analysis (PCA) using SVM and also classification with SVM itself without feature selection. The results proved that the proposed method enhances the intrusion detection and outperforms existing approaches thus modeling computationally efficient IDS with minimum false positive rates.**

Key words: Intrusion Detection System (IDS), Network Security, Factor Analysis (FA), Support Vector Machines (SVMs), Principal Component Analysis (PCA).

1. INTRODUCTION

Intrusion detection is a critical issue in network security, for protecting network resources. Therefore an accurate system of detecting intrusions is to be built to give assurance for information in any organization either public or private. The main goal is to increase the detection rate and reduce the false alarm rate. Intrusion Detection System (IDS) is a method which dynamically monitors the events occurring in a system, and decides whether these events are signs of an attack or constitutes an authorized use of the system [1] [2] [3].There are many types of IDSs in terms of monitoring the network traffic such as Network Intrusion Detection System (NIDS), Host Based Intrusion Detection System (HIDS) and Hybrid Intrusion Detection System.

IDS has to monitor large amount of audit data even for a small network, therefore analysis becomes more difficult, which leads to poor detection of suspicious activities. There are diverse affinities between features. So, IDS has to decrease the quantity of the data to be processed by removing the features that contain false correlations and redundant information. This results in gaining better accuracy and lower computation time. IDS task is commonly modeled as a classification procedure in a machine-learning context. Many methods were proposed to develop an efficient IDS, among those Support Vector Machines(SVMs) have gained a significant importance using intrusion detection system using various kernels [4].In modeling efficient IDS,it is necessary to reduce the features which showed a great change in the performance[5].

For constructing an Intrusion Detection System the research mainly falls in two ways: detection model generation and intrusion feature selection. In achieving best accurate results preprocessing techniques like feature selection, feature reduction have become crucial in Intrusion Detection Systems [6].The recent study illustrated an improved false positive rate using Artificial Neural Networks (ANN) in Intrusion Detection mechanism with Principal Component Analysis(PCA) as a feature selection strategy [7].There are numerous studies which show reasonably good results with feature reduction using Support Vector Machine(SVM) as a classifier tool[8][9][10][11].In another study using Classification and Regression Trees (CART) and Bayesian Networks(BN) Chebrolu et. al has given ensemble feature selection algorithms which results in lightweight IDS[12].More recently a study on Generalized Discriminant Analysis as a feature selection technique achieved good results[17].

Even though SVM is a good classification technique, when applied to massive datasets many problems will be occurring. Since solving SVM is similar to solving a quadratic optimization problem, when the dimensionality increases it needs a large computational time and memory. Meanwhile for a pattern classification problem e.g.: intrusion detection, it is difficult to decide which features are useful for classifying attack or normal activity. But with IDS there are large amount of dimensions *d* as well as examples *k* which leads to inaccurate results. Therefore there is a need to select most significant features and apply high performance classifiers like SVMs which results in low false alarm rates.

Here in this paper we have taken a popular statistical technique called Factor Analysis (FA) as a dimensionality

reduction technique through which the features are analyzed as factors. The rest of the paper is organized as follows. Section 2 describes An Overview of Support Vector Machines and Factor Analysis. Section 3 will describe the Proposed IDS Model with a novel algorithm and Section 4
give Experimental results followed by Conclusions with future work.

2. Machine Learning Perspective: AN Overview

2.1. Support Vector Machines

Mainly classification in IDS deals with false positive reduction and classifying between normal and attack patterns, therefore Support Vector Machines (SVMs) are best classifiers. SVMs are supervised learning techniques.SVM is based on statistical learning theory and is developed by Vapnik [13][14][15].These are built using support vectors, which are responsible for classification of data points with Maximal Marginal Hyper plane(MMH). The main aim is to classify the data points using MMH by solving quadratic optimization problem [16].SVMs have smaller running times and give high accurate classification results. The attractiveness of SVMs lies in its mathematical equations and pictorial illustrations.

SVM is a machine, constructed based on support vectors which are decisive points in both of the classes. Once support vectors are identified then it is easy to draw the hyper plane which separates both positive and negative classes. In this way classification process is done in SVM. It uses class label, so they are called as supervised learning techniques. By training the model we used to get the weight vector and bias vector values which are used to identify support vectors.SVM construction can be done both in data linearly separable case and linearly inseparable case. When the data is linearly separable, MMH is constructed based on training points and class boundary. When the data is linearly inseparable, the data is mapped to a high dimensional feature space and classification is done. The process of mapping to a high dimensional feature space is called kernel function. The Figure 1: given below illustrates the classification of SVM.

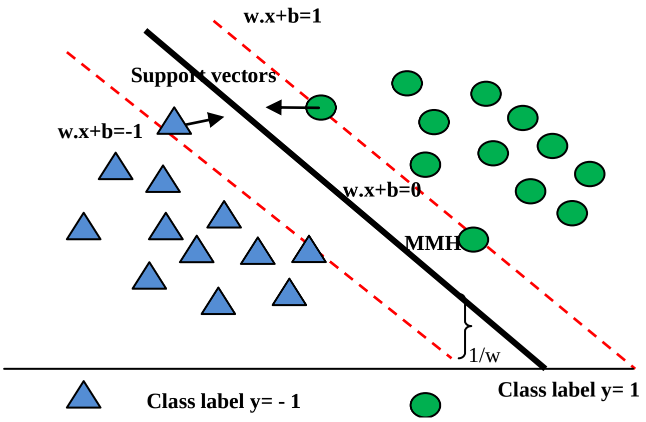

Figure1: General idea of SVM organization

But training with SVMs on huge datasets, is time consuming. In recent years, there has been a lot of work done to improve learning methods using SVMs. One approach is to optimize the SVM algorithm [20, 21] to solve the convex optimization problem. Other approaches include simplification phase in reducing the training set size [22, 23]. To perform training using SVM, model selection is crucial. Even though the SVM algorithms are lesser sensitive to curse of dimensionality, dimensionality reduction techniques can enhance the efficiency of SVMs. In SVM, generalization ability depends on the choice of SVMs parameters.

In Training the dataset using SVMs, the user should provide the type of kernel function to be applied [21]. There are several kernel functions namely linear, sigmoid, polynomial, radial basis and Gaussian and so on. The performance of SVM depends mainly on the kernel selected. More general studies showed that Radial Basis Function (RBF) is most popular choice of Kernel option because of their localized and finite responses across the entire range of the real x-axis [2]. The SVM work flow is given with the following algorithm [16].

SVM Algorithm

Input:D={$(x^l,y^l),(x^l,y^l),\ldots..,(x^l,y^{l)})$},x€$R^n$, y€{-1,+1}

Define: w_i,$b_{i,}\lambda_j$ where w is the weight vector,b is the bias vector, λ_j is the lagrangian multiplier and i=number of attributes and j=number of intstances.

Solve: $\mathbf{L_D} = \sum_i^N \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i\lambda_j x_i\, x_j\, y_i\, y_j$ where L_D is the dual form. It must be solved to obtain $\lambda_{j.}$

Calculate: w,b are obtained by substituting $\lambda_{j.}>0$ values in the equations $w=\sum_i^N \lambda_i y_i x_i$ and for getting b,in $\lambda_i(\, y_i(\, w.x_i +b) -1) =0$

Classifier: $f(x)=sgn(w^*.x+b^*)$ if sgn is' +' then class is positive,if sgn is' –' then class is negative.

2.2 Factor Analysis

Factor Analysis is a popular statistical technique, which is the extension of Principal Component Analysis (PCA).It is useful in overcoming the shortcomings of PCA.It is also called Multivariate Statistical Analysis. Factor Analysis specifies that the attributes can be grouped by their correlations [24].Its significance is to find the intercorrelations between *n* attributes by deducing them into a set of factors *f*, which are relatively lesser than *n*, the number of attributes [25] [27]. It can be viewed as an attempt to approximate the covariance matrix ∑. Therefore it reduces the dimensionality of the dataset. Factor analysis produce a table in which the rows are obtained as raw indicator variables and the columns are factors that exhibit as much of the variance in these variables as possible. The cells in the table contain factor loadings and the importance of the factors lies in observing which variables are heavily loaded on certain factors. Therefore factor loadings are nothing but the correlation between the variables and factors.

There are three main steps involved in Factor Analysis.

A. Calculate initial factor loading matrix: This can be done by using two approaches: Principal component method and principal axis factoring.

B. Factor rotation: The objective of the rotation is to try to make sure that all variables have high loadings only on one factor. There are two types of rotation methods namely orthogonal and oblique rotation. Generally orthogonal rotation is used when the common factors are independent.

C. Calculation of factor scores: When calculating factor scores, (m factors as f_1, f2…f_m) a decision has to made as how many factors to include. One vital thing is check the total variance of original variables is more than 75%.Then choose m to be equal to the number of eigenvalues over 1.

3. Proposed Intrusion detection System

SVMs are powerful classifiers; they yielded good results when applied to intrusion detection. They are applied to data with a large number of features, but their performance has been drastically increased by reducing the number of features [19].In building IDS, KDD Cup 99 dataset which is a bench mark in the area of intrusion detection and security evaluation frameworks is used. Generally IDS is a classification technique in a machine-learning framework. Here in the proposed model we have added another phase to reduce the number of features and then perform classification task. The key objective is to increase the detection rate and reduce the false alarm rate**.** It consists of five phases: collection of raw KDD cup 99 dataset, pre-processing, and feature reduction scheme, parameter selection using SVM and testing. The proposed model of IDS is described in the figure below.

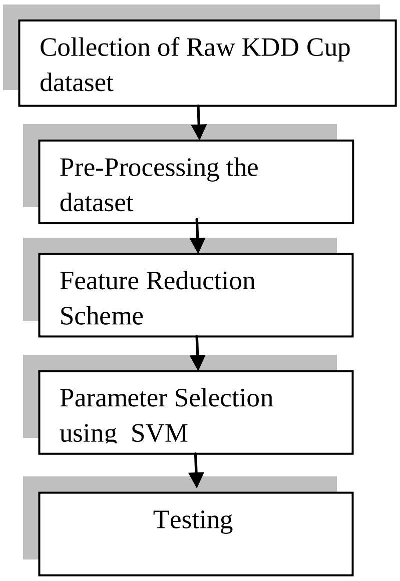
Collection of Raw KDD Cup dataset
Pre-Processing the dataset
Feature Reduction Scheme
Parameter Selection using SVM
Testing

Figure 2: Proposed Model of Intrusion Detection System

3.1. Collection of Kdd Cup 99 Dataset

The KDD cup 99 dataset which is 10% of TCP/IP dump data collected from USAir force LAN in the year 1998.It contain 4, 94,020 records of which 97277, 391458, 4107,1126,52 are Normal, DOS, Probe, R2L, U2R respectively. Because the dataset cannot be processed in its format various pre-processing techniques have to be applied to the dataset.

3.2. Pre-Processing Phase

Given dataset is not ready for the detection process. It has to be processed before undergoing intrusion detection procedure. Then pre-processing techniques are applied to the dataset in order to improve the time, cost and quality of results. After completing the Pre-processing phase, it is necessary to collect the necessary or important attributes, to do this we go for feature reduction phase.

3.3. Feature Reduction Phase

For the system to be perfect it is necessary to reduce the features in the raw data based on data analysis. To obtain most essential attributes and remove that are worse associated, feature reduction is done. In the Feature reduction phase a popular statistical technique called Factor Analysis (FA) is used as a dimensionality reduction technique. Factor Analysis is responsible for binding the number of features or attributes to the number of factors based on the correlation between the features. Therefore, a novel feature reduction scheme based on FA-SVM algorithm is proposed in which the factor analysis and SVM are applied. The algorithm is given as follows.

FA-SVM Algorithm

Input: *D_{ij} is a dataset where 'i' is the number of instances and 'j' is the number of features.*

Step1: *normalize the dataset using suitable pre-processing techniques.*

Step2: *Then calculate the factor loading matrix of D_{ij}.*

Step3: *find the cumulative variance and determine principal factors using Eigen values*

Step4: *Now rotate the factor loading matrix, then compute the factor score and make it the new feature.*

Step 5: *Then train the dataset with the transformed features using SVM classification with c=10 and γ=0.01 and using Radial Basis Kernel function.*

Step 6: *Use the trained FASVM algorithm to predict either normal or attack traffic.*

Output: *D_{ik} is the resultant dataset with 'k'number of features reduced and k<j with the classified results.*

3.4. Intrusion Detection: Parameter Selection Using Svm

The dataset is classified using SVM, it will dispose the classes either attack or normal data. Since SVMs are simply capable of binary classifications, we will need to use five SVMs, for the 5-class classification in intrusion detection. We separate the data into the two classes of "Normal" and "Others" (Probe, DOS, U2Su, R2L) patterns, where the Others is the group of four classes of attack instances in the data set. The objective is to divide normal and attack traffic. We repeat this process for all classes. Training is conducted using the RBF (radial basis function) kernel.

3.5. Testing

Here in this approach, we conduct 10 fold cross validation. The dataset is partitioned at random into 10 equal parts in which the classes are taken approximately as same scope as in the full dataset. Each part is held out in turn and the training is conducted on remaining 9 parts, then its testing (error rate) is conducted on holdout set. The training procedure is conducted in total of 10 times on different training sets and finally the 10 error rates are averaged to fetch overall error estimate.

4. Experiments Conducted

4.1. Dataset Description

The Knowledge Discovery and Data Mining (KDD) Cup 99 dataset [18] was used in conducting the experiments and examining the results. It was taken from the Third International Knowledge Discovery and Data Mining Tools Competition. Each connection record in the data set constitutes 41 attributes [2] which are of both continuous and discrete type variables. There are 22 categories of attacks from the following four classes: Denial of Service (DOS), Root to Local (R2L), User to Root (U2R), and Probe. The dataset holds 391458 DOS attack records, 97278 normal records, 4107 Probe attack records, 1126 R2L attack records and 52 U2R attack records [17].

4.2 Data Disposing

In our Experiments, we conduct 5-class classification. Here we have taken a subset of KDD cup 99 dataset containing 14207 records of which it is considered as training and testing dataset after eliminating most of the redundant records. The dataset size is taken as proportionate to the relative size as in KDD cup 99 dataset. Therefore in the resulting dataset there are 3000 normal records, 10000 DOS records, 574 Probe records, R2L 401 records and 52 U2R records. Where the four attacks are combination of 22 different types of records that belong to the four different classes described in section 4.1, and the last one is the normal data.

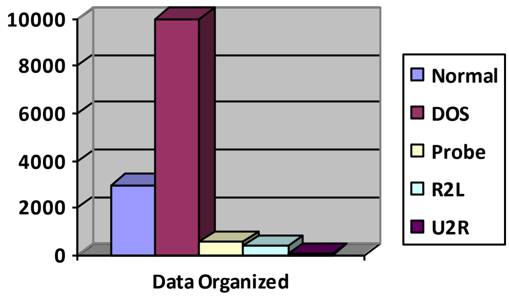

Figure 3: The Distribution of the dataset

All the symbolic attributes are converted to numeric. Therefore the attributes protocol_type, service and flag are converted to numeric. Now the redundant records are removed from all the classes. Then attributes duration,src_bytes,dst_bytes are discretized**.**Then sampling is applied to take subset of it, since using the entire set of data is too expensive and time consuming for processing.

Then feature reduction scheme is applied which involves finding the factors that dominating the attributes in the dataset. This can be done by applying our algorithm FA-SVM. Then the outcome is obtained with 12 factors (where 41 features are transformed as 12 different factors), which are fed to SVM classifier.

4.3 Results

We have performed three types of experiments.

1) The dataset taken containing 14027 records with no feature selection, i.e. taking 41 attributes, we applied SVM.

2) In the second experiment we have applied Principal Component Analysis as feature selection, through which 19 attributes are obtained and then SVM is applied.

3) In the third experiment, we used proposed algorithm FA-SVM, through which we achieved 12 new factors which are transformed from 41 attributes, then applied SVM classifier. We conduct all the experiments and obtained results using Java 1.6 and Weka 3.6.9 on the platform Windows 2007 with 3.40 GHz CPU and 2.0GB of RAM. WEKA is an open source Java code produced by researchers at the University of Waikato in New Zealand [26].

To evaluate our FA-SVM Algorithm, we computed three statistics variables in our experiments: the Accuracy (Acc), the detection rate (DR) and the false alarm rate (FAR). The Accuracy is defined as the percentage of instances that are classified correctly. The detection rate is defined as the percentage of records generated by the malicious programs, which are labeled correctly as anomalous by the classifier. The false positive rate is defined as the percentage of normal records, which are mislabeled as anomalous.

The time taken to conduct Experiment1 is 982.07 sec and it predicted 11244 instances correctly with 80% accuracy. The time taken to conduct Experiment2 is 803.5 sec and this one predicted 11863 instances correctly with 85% accuracy. While the time taken to build proposed model is 665 sec.It produced results with 92.5% accuracy, with classifying 12989 instances correctly out of 14027 instances.

The Detection rates and False Alarm Rates for SVM, PCA+SVM, FA+SVM are shown in the corresponding tables, Table I & II.

	Normal	DOS	Probe	R2L	U2R
SVM	93.5	79.4	77.7	9.4	9.6
PCA+SVM	95.2	84.5	84.4	16.4	17.3
FA+SVM	96.7	93.8	95.1	35	25

TABLE I: DETECTION RATE OBTAINED

	Normal	DOS	Probe	R2L	U2R
SVM	19.3	9.0	1.24	0.91	0.09
PCA+SVM	13.4	7.3	2.5	0.3	0.02
FA+SVM	6.03	5.5	0.9	0.15	0

TABLE II: FALSE ALARM RATE OBTAINED

The Detection Rates and False Alarm Rates of three experiments SVM, PCA+SVM, FA+SVM are depicted in the following charts in Figure 4 & Figure 5 for the evaluation of results in precise way.

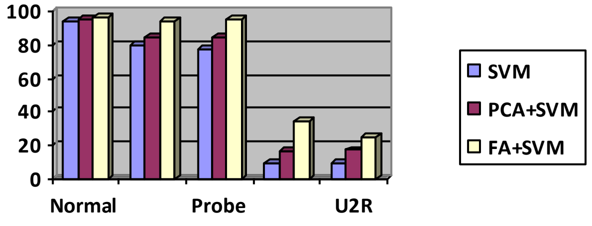

Figure 4: Comparision Of Performance Results: Detection Rate

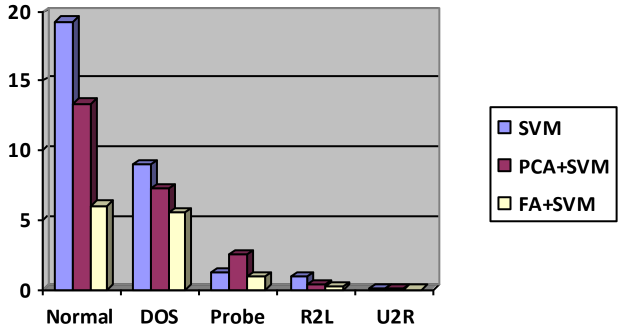

Figure 5: Performance Of Existing Techniques And Proposed Technique: Far Rate

5. Conclusion

Factor analysis its main goal is to reduce high-dimensional data, when the processing dataset is large with a more number of feature variables, it is advantageous. To design most efficient Intrusion Detection System it is necessary to go for dimension reduction, so the FA-SVM algorithm is best suited for detecting intrusive behavior. The results obtained in this study showed better accuracy and lower computation time. It is worth paying attention in using dimensionality reduction techniques for improving and building well proficient Intrusion Detection Systems (IDSs). Future research will employ alterations of the proposed method and upgrading to it to achieve enhanced performance and automation by developing classifiers that are more accurate for the detection of attacks.

References

1. Ghosh A. K. (1999). Learning Program Behavior Profiles for Intrusion Detection. USENIX.
2. Mukkamala S., Janoski G., Sung A. H, “Intrusion Detection Using Neural Networks and Support Vector Machines,” Proceedings of IEEE International Joint Conference on Neural Networks, 2002, pp.1702-1707.
3. H. Debar, M. Dacier and A. Wespi, ”Towards a taxonomy of intrusion-detection systems” Computer Networks, vol. 31,pp. 805-822, 1999.
4. Wun-Hwa Chen, Sheng-Hsun Hsu,”Application of SVM and ANN for intrusion detection”, Computers & Operations Research, 2005 – Elsevier .
5. Rupali Datti, Bhupendra verma,”Feature Reduction for Intrusion Detection Using Linear Discriminant Analysis”, (IJCSE) International Journal on Computer Science and Engineering Vol 02, No. 04, 2010, 1072-1078
6. Andrew Sung,S Mukkamala.,”Feature Selection for Intrusion Detection using Neural Networks and Support Vector Machines”Transportation Research Record:Journal of the Transportation Research Board 1822.1,2003,pp.33-39.
7. Ravi Kiran Varma,V.Valli Kumari ,”Feature Optimization and Performance Improvement of a Multiclass Intrusion Detection System using PCA and ANN” , International Journal of Computer Applications (0975 – 8887) Vol 44 No13, April 2012.
8. Safaa Zaman and Fakhri Karray.,”Features Selection for Intrusion Detection Systems Based on Support Vector Machines”, Consumer Communications and Networking Conference, 2009. CCNC 2009. 6th IEEE
9. Gopi K. Kuchimanchi, Vir V. Phoha, Kiran S. Balagani, Shekhar R. Gaddam,”Dimension Reduction Using Feature Extraction Methods for Real-time Misuse Detection Systems”,Proceedings of the 2004 IEEE Workshop on Information Assurance and Security T1B2 1555 United States Military Academy, West Point, NY, 10,June 2004.
10. Heba F. Eid, Ashraf Darwish, Aboul Ella Hassanien, and Ajith Abraham,”Principle Components Analysis and Support Vector Machine based Intrusion Detection System”,ISDA 2010,363-367.
11. ZhangXue-qin, GU Chun-hua and LINJia-jun.,”Intrusion Detection System Based On Feature Selection And Support Vector Machine”,IEEE,2006
12. Srilatha Chebrolu, Ajith Abraham, and Johnson P. Thomas”Hybrid Feature Selection for Modeling Intrusion Detection Systems “Springer ,2004,pp 1020-1025.
13. Vapnik V.,” The Nature of Statistical Learning Theory”, Springer-Verlag, New York, 1995.
14. Cortes C.,Vapnik V.,“Support vector networks, in Proceedings of Machine Learning 20: pp.273–297, 1995.
15. Boser, Guyon, and Vapnik, “A training algorithm for optimal margin classifiers”,Proceedings of the fifth annual workshop on Computational learning theory.pp.144-152, 1992.
16. P Indira priyadarsini,Nagaraju Devarakonda,I Ramesh Babu,”A Chock-Full Survey on Support Vector Machines”, International Journal of Computer Science and Software Engineering,Vol 3,issue10,2013.
17. P Indira priyadarsini,I Ramesh Babu,”Modeling Intrusion Detection System based on Generalized Discriminant Analysis and Support Vector Machines”,International Conference on Recent Trends in Engineering and Technology Sciences-2014,pp 8-12.
18. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani “A Detailed Analysis of the KDD CUP 99 Data Set”, Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009).
19. Iftikhar Ahmad,Muhammad Hussain ,Abdullah Alghamdi,Abdulhameed Alelaiwi .,”Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components”Springer 2012.
20. Platt, J.: Fast training of SVMs using sequential minimal optimization, advances in kernel methods-support vector learning. MIT Press ,1999 ,pp.185–208
21. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. Sofware
Available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm ,2001
22. Yu, H., Yang, J., Han, J.: Classifying large data sets using SVM with hierarchical clusters. In: SIGKDD.,2003,pp.306–315
23. Lebrun, G., Charrier, C., Cardot, H.: SVM training time reduction using vector quantization. In: ICPR. Volume 1.,2004,pp. 160–163.
24. Nitin Khosla” Dimensionality reduction using factor analysis”MastersThesis,http://researchhub.griffith.edu.au/display/n26993f96c6bc6146d5444ea116009424,2006.
25. R. J. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey, 1998
26. M. Hall, et al., "The WEKA data mining software: an update," ACM SIGKDD Explorations Newsletter, vol. 11, pp. 10-18, 2009.
27. http://www.cs.cmu.edu/~pmuthuku/mlsp_page/lectures/slides/JFA_presentation_final.pdf.