An Efficient Approach to identify an optimal feature selection method using improved Principle Component Analysis in supervised learning process

D.Hemavathi [1], H.Srimathi [2]
Department of Information Technology, School of computing
SRM Institute of Science and Technology
India
hemavathi.d@ktr.srmuniv.ac.in [1], srimathi.h@srmuniv.ac.in [2]

Abstract

Feature selection is an important process in various data relevant to the domains like financial organizations, marketing, healthcare, education etc. Identifying the relevant feature is the major impact in obtaining the accurate results. Optimal feature selection plays a vital role in predictive analysis. Feature selection or dimensionality reduction process are useful in identifying the most important variables or parameters which is useful in predicting the outcome. Feature selection helps to train the machine algorithm faster as well as reduced the complexity of the model. Classification plays an important role to identify the pattern analysis. Classification algorithms helps to group the major categories of data sets in supervised learning. Even though there are many number of attributes are available, only some features helps to improve the accuracy of the model. In the real world scenario filter methods are used to extract the features without outliers. Nevertheless Multi collinearity problem is not overcome by the filtering techniques. So that there is lack of identifying the best subset of features for better analysis. Existing systems use various learning algorithms (Feature Selection Algorithms) like fast forward selection and backward Elimination algorithm, Recursive feature elimination approach, Focus and relief algorithms. Principle component analysis (PCA) is well known technique to form the finest subset of features with the Neural Network classifier in supervised learning techniques. To improve the accuracy further we have proposed the new approach Borrowed PCA as an optimal feature selection method to interpret the best subset of features by using artificial neural networks classifier. And the method has been validated with the KDD Cup 1999 dataset used for measuring intrusion detection problems.

Keywords: Feature selection methods, feature selection Algorithms, Principle component analysis, Borrowed Principle component analysis, classification methods, Artificial Neural Network.

1 Introduction

The feature selection process in terms of supervised learning follows the steps i. size that optimizes the measure of evaluation, ii. The set of smaller size that satisfies an explicit conditions on the measure of evaluation iii. The created subset with the best subset formation based on the size and its measure. Learning speed, generalization capability or simple representation are the general purpose pursued is the improvement of the inductive learner. Irrelevant features, noisy features, redundant information are to be eliminated by using proper filtering methods. Major purpose of feature selection techniques are to train the machine algorithms in faster manner. In such a way overall complexity of the model to be reduced and accuracy to be improved.

Filtering methods are generally used in pre-processing step. Basically features can be selected in an independent way. Feature score is useful for selecting the best feature. Various statistical test can be conducted to identify their correlation. The following statistical test are useful to define correlation coefficients. Linear discriminant analysis is used for Identify the linear combination of features. It is for categorizing two or more classes of categorical variables. Mean Values of various groups of data are same or different can be checked by using the stastical test in ANOVA method. Chi square test are used to find the correlation or association between the groups of categorical features. But highly correlated factors such as height and weight, car sale amount and total number of years from the purchase of car / Kilometres running etc. Can't be removed by filter methods it is referred as multicollinearity of features. Before training models of the data separate attention to be given for the multicollinearity of features.

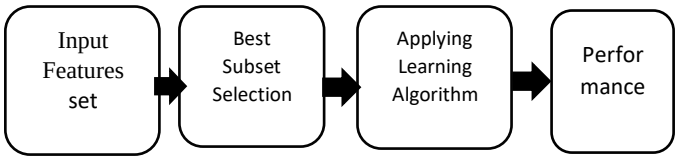


Figure 1 Filtering Methods

Wrapper methods are useful to create subset of features. By creating the subset of features the search

problem could be reduced. Forward selection, backward elimination and recursive feature elimination algorithms are used to form the subset of features. Depending on ranking the features the optimization process could be performed for obtaining the best subset of features.

Parameters	Filter methods	Wrapper Methods
Feature Relevancy	Relevancy calculated by their correlation and dependent variable	Calculates Usefulness of subset of features for training the model
Speed	Faster	Slower than filter methods due to train the model
Evaluation	Statistical methods for evaluation	Use cross Validation
Best Subset Identification	Failed to do	Can Identify best suset
Model building	Over fitting problem might not be overcome	Make the model more prune to over fitting
Multicollinearity	Do not remove multicollinearity features	Multicollinearity features dealt with training the model

Table 1 Comparison of Filter and Wrapper methods

There are different types of features selection methods like continuous and binary.Let S -> Original set of features, Continuous feature selection based on the assignment of weights w_i to every feature s_i from the original set S. i.e $S_i \in S$. In binary selection technique, binary weights to be assigned directly or filtering the outcome of the continuous problem. General characteristics of feature selection process includes Search organization, successor generating method, Measure for evaluation. Search algorithm is very important to drive feature selection techniques. Efficiency in searching process can be defined with the two methods like continuous and binary. In binary feature selection methods the particular feature can be represent to be selected or not for concluding relevancy but only in the continuous method it can represents that how much percentage that the feature can be used i.e. the degree of filling the specific attribute.

S-> {s1,s2,s3…..sn) (1)

Continuous : wi ∈ [0,1] (2)

Binary : wi ∈ {0,1} (3)

Generating successor includes the possible variants of the current hypothesis to be proposed. The set of attributes are relevant for prediction by creating the subset with adding a new attribute to the group(forward selection) or removing the attributes one by one from the existing subset (backward selection). And the goodness of fit can be measured using evaluation metrics.

2 Related Work

Based on linear dependency similarity can be measured between two random variables. They proposed the approach Maximum Information Compression Index (MICI) [1] used for feature selection. To measure the similarity in between 2 random variables there are two category of approaches are used one is non-parametrically test the relativeness of probability distribution of variables. It may not be useful for feature selection because it is delicate to location and dispersal of distributions. Second approach is amount of functional dependency among variables (linear dependency) can be measured.

Linear dependency measures are correlation coefficient and least square regression error.

Correlation coefficient ρ between two random variables a,b are: [1]

$$\rho(a,b) = \frac{Cov(a,b)}{\sqrt{Var(a)Var(b)}} \qquad (4)$$

If a & b completely correlated, exact linear dependency exists (-1 or 1)

$\rho(a,b)$ totally uncorrelated $\rho(a,b)$ =0 ;

[1- $\rho(a,b)$] can be used for measure similarity

Principal component analysis used to reduce the data dimensions [9] . Mainly it is used to decrease the noisy data. Initial PCA technique is not well suited for achieving robustness in case of high dimensional space of data. [9] So that this paper describes about SPCA Simple Principle component analysis which is less sensitive to the outliers part. Three approaches have been identified to increase the robustness of PCA to outliers. Elimination of outliers, modification of outliers with appropriate values and finding the robust version of covariance and correlation matrices. SPCA has been developed with the normalization process of L1 (Least absolute deviation) and L2 (Least absolute sum of square) norms. This paper concludes that L1 norm of projection is more robust to data lies further from central value than L2 projection norm.

Performance and efficacy of various feature selection algorithms has been analysed [2]. Various methods to identify the relevant features and the classification methods performance with reduced number of features has been discussed in this study. Different types of Feature selection algorithms [2] and the interactions among the algorithms were compared in this article. Proportion of selected features and the execution time (CPU time) has been analysed with various filter and wrapper feature selection algorithms against the various classification methods like CART, PNN, SVM, Linear Regression.

We observed that this paper[10] explains a Las Vegas Feature selection algorithm [10] that creates probabilitic choices to support the search more quickly to find the exact subset of features. LVF is a filter based scheme to find the relevant features and also removes the noisy and irrelevant features. And the inconsistency criterion rate calculation has been proposed to understand that to the amount of degree that the dimensions reduced in the data can be accepted. This article proposed the next version of LVI (Enhanced Las Vegas) [10] which helps to increase the running time of the LVF Algorithm. And both the algorithms were validated with various artificial data sets to show the average execution time.

An integrated data mining model [3] of feature selection algorithms and ensemble learning classification has been proposed in this article. It includes genetic algorithm, Principal component analysis (PCA), information gain ratio and relief attribute evaluation. Classification accuracy has been tested with support vector machine algorithm. Once the model has been selected for each feature it was applied to base and ensemble classification algorithms. Based on the number of factors, Accuracy and AUC (Area Under the Curve as a criteria for comparing learning algorithms) measures [3] were compared among various classifiers. Best accuracy was obtained by PCA algorithm. And ANN Adaboost classifiers obtained the best accuracy and AUC values. Optimal parameters have been identified for each model for finding the credit score of bank customers.

Data envelopment analysis [12] and entropy method are used in an integrated approach which is used for selecting the features. Determining the efficiency of algorithms and weighting the criteria for selecting the features have been analysed. This paper [12] proposed the integrated model is a sample of applying Multiple Criteria Decision making method in various scientific domain.

Best subset of features can be identified using feature ranking methods [4]. This Paper [4] proposed the optimal feature selection process by feature ranking algorithm. Six different ranking algorithms were compared with 2 different real world datasets. [4] And they have implemented the selected features with four different classification algorithms like IB1, decision tree, Naïve Bayes and RBF network. Higher ranking features considered for subset selection. Almost all the ranking algorithms provides more or less similar ranks for the features considered as high ranked feature. And the classification accuracy has been calculated with different classifiers. Authors observed that the highest accuracy can be obtained by testing the given classifier with various feature subsets obtained from different ranking indices.

Principle Component Analysis (PCA) does not guarantees a hundred percent dimensionality reduction with complete features. To improve the performance it is essential to eliminate the irrelevant/noisy features. This paper [6] proposed an approach called PCA-based optimal FSS for fuzzy extreme learning machine (PF-FELM) that can manage weighted classification problem. Based on highest occurrences of different filter-ranking algorithms , Feature Selection System chooses an order of finest features.

Consumption pattern of a particular state (Kerala) of people in rural and urban sectors analysed in the paper [7] . MPCE (Monthly Per Capita Expenditure) value per person over the period of 5 years have been analysed. Consumption pattern of Kerala was compared with 29 states of India. They have observed [7] that Kerala consumption pattern is higher than India in rural and urban sectors. It was observed that the behaviour analysis of consumers can be predicted based on their consumption pattern [8].

3 Feature Selection Methodology

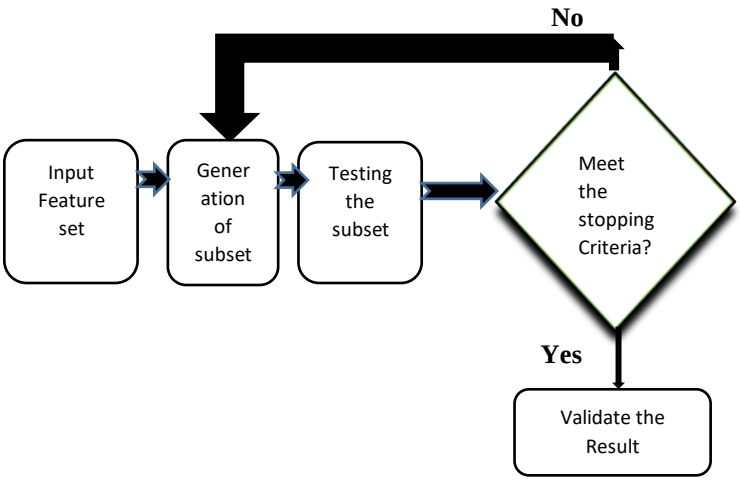


Figure 2 Structure of Feature Selection process

Wrapper Methods [11] are well suited in the training models by removing multi collinearity features. Until selecting the best subset it repeats the selection process to get the desired performance.

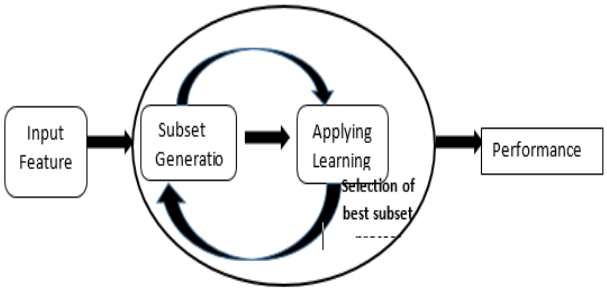


Figure 3 Wrapper Methods in Feature selection

Embedded methods are the advanced methods [11] which includes the advantages of filter and wrapper methods. It can be implemented by learning algorithms which contains their own built in selection methods. LASSO and Ridge regression methods performs L1 which minimize the sum of absolute difference between final objective value and approximate value and L2 which denotes Least Square sum of squared variation between target and approximate value. These two norms regularized to improve the robustness of identifying the principle component in the feature set.

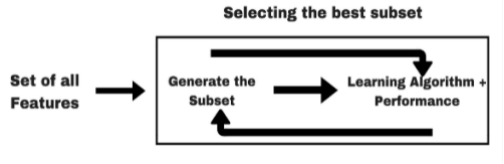


Figure 4 Embedded Methods in Feature selection [17]

3.1 Search Strategy for Feature subset Selection

There are three important feature selection strategies: Complete, Sequential and heuristic/ Random.

Complete Search. Based on the evaluation measure, it provides assurance to find the optimal solution While the complete search is complete there is no chance to miss the optimal subset.

Sequential Search. It provides completeness but may be the risk obtained by losing optimal subsets [13].

Random Search. There are 2 ways to perform the random search with randomly selected selected subset and proceeds in two different ways. One is to monitor sequential search, which vaccinates random property into the sequential approaches [13]. The next approach is to produce the next subset in a absolutely random way.

Search Strategy	Order of Search space	Key Benefit	Example
Complete	$O(2^N)$	Guarantees Optimal solution	Branch and Bound algorithms
Sequential	O(N)	Simple implementation and fast to express results	Sequential Forward/ Backward search algorithm
Random	O(N)	Optimality based on resources available.	Las Vegas Filter/ Incremental algorithms
Heuristic	Based on heuristic function	Domain specific knowledge while searching	hill climbing, best first search

Table 2 Comparison of search strategy for feature subset selection

Dependency, information, Distance, Consistency and classification accuracy are the important evaluation measure in feature selection.

4 Feature Selection algorithms

Various Feature subsets were evaluated using indexes and selects the finest one among them in various methods of feature selection. Depends on the selection process is supervised or unsupervised , Index measures the capacity of subsets[5] in classification or clustering. For high dimensional data, high computational complexity is involved during searching process. In branch and bound techniques, feature evaluation criteria is monotonic in nature. Forward and backward search process in greedy algorithm will provide the better results but increases computational cost [12]. An alternate way is to develop Genetic algorithms which provides robust methods to find out optimal subset in evaluating the indices.

4.1 Las Vegas filter algorithm (LVF) & Las Vegas Incremental Algorithm (LVI) algorithm

LVF algorithm [2] is used for the relevant analysis to be accomplished on the data to recognize or eliminate redundant or unrelated attributes from the learning process. It helps to make our search rapidly to identify the right subset of features. The Las Vegas incremental algorithm (LVI) extension of LVF algorithm it is also referred as ELV(Enhanced Las Vegas Algorithm). LVI use a slice of the dataset called training data of slowly improved the size depends on the excellence of selected features. Let us consider Parameter p controlling the portion of training data. P size can be set as 10% or 20 %. As a part LVI uses LVF for feature selection process. In order to improve the pre-processing in case of huge number of instances in the dataset, LVI helps to reduce the search time than LVF because of the training data.

4.2 Branch and Bound algorithm (BB)

Branch and bound method is used for optimisation problems.[11] Even in case of greedy and dynamic approaches fails branch and bound can prove that helpful for optimized solutions. Also backtracking approach is possible in Branch and Bound method but greedy and dynamic approaches does not allow backtracking. BB is a slower method because it takes all possible subsets are implicitly inspected.

4.3 Sequential Forward Selection (SFS) / Sequential Backward Generation (SBG)

Forward selection approach [11] is an interesting approach to select the subset of features. Initially there is no idea about the feature group. At the time of starting the creation of subset, there is no feature in the model. At each iteration, add one feature to get improvement of the model for the accuracy point of view till an addition of new attribute does not affect the performance of the model.

Initial value X'=0 ;
J- Evaluation measure ;
n –number of features;
X'=X' U $\{x_i \in X \backslash X' \mid J(X' \cup \{x_i\})$ is bigger$\}$ (5)
Stop | X' | =n'

Forward selection approach

Step 1: Let S be the set of features {s1,s2,…sn}
Step 2: Assume the feature subset $S_A \leftarrow \{ 0 \}$
$S_A \leftarrow S_i$
for i:= 1 to n do
$S_A' \leftarrow S_i + S_{i+1}$
Step 3: Check the Accuracy if $S_A == S_{A'}$
end for
else
Go back step 2 until $S_A = S_{A'}$
end

Backward Elimination approach

Backward Elimination starts with all n features and remove the feature at each iteration which consider as a least significant one. And stop the iteration if there is no improvement is identified in the model by removing the feature.

Initial value X'= X;
J- Evaluation measure ;
n –number of features;
X'=X' \ $\{x_i \in X' \mid J(X' \backslash \{x_i\})$ is bigger$\}$ (6)
Stop | X' | =n'

Step 1: Let S be the set of features {s1,s2,…sn}
Step 2: Assume the feature subset $S_A \leftarrow \{ n \}$
$S_A \leftarrow S_n$
for i:= 1 to n do
$S_A' \leftarrow S_A - S_i$
Step 3: Check the Accuracy if $S_A == S_{A'}$
end for
else
Go back step 2 until $S_A = S_{A'}$
end

Recursive Feature Elimination approach

It is based on Greedy optimization algorithm which is used to find the best feature subset. At each iteration while creating the model, it repeatedly creates one best and worst performing feature. Till all the features have been exhausted, it creates the next model with the left out features.

Step 1: S←S1,S2,s3..Sn
Step2: i= 1 to n
For each iteration i
S' ← best F of S
S''← worst F of S
End for
Step 3: Rank the Feature F_r based on their elimination order.
F_r=1 if F is eliminated at the end of iteration

4.4 Relief Algorithm

Relief algorithm is tracks the filter structure. It uses the random search to find the features based on Euclidean distance to separate the classes [2]. This algorithm executes in a way that once the random choice of an object from a particular class identifies the closese neighbour within class is called 'near hit' and from another class is called 'near miss'. Objective of RELIEF algorithm is to calculate the excellence of features based on their values discriminate instances that are close to each other. Relevancy of the feature can be identified real value weight vector w. RELIEF is not well suited for large data sets, due to computational complexity.

4.5 Genetic Algorithm

One of the most progressive algorithms for feature selection is the genetic algorithm [16]. Genetic algorithms can be applied to enhance the performance of a predictive model, by choosing the best relevant features. Genetic algorithms are analysed and modifies the set of solutions at the same time.

Genetic algorithms are performs better than traditional feature selection techniques. Many features contained datasets can also be easily managed by Genetic algorithms. These algorithms does not require precise information about the problem further study. Finding the most discriminative subset of transmuted features is an optimization problem so that Genetic algorithm plays an important role in subset selection [15].

Algorithm	Search strategy	Subset Generation	Subset Evaluation
LVF	Random	Forward selection	Divergence
LVI	Random	Forward selection	Divergence
Branch and Bound	Complete	Weighted	Consistency
SFS	Sequential	Forward selection	Divergence
SBG	Sequential	Backward elimination	Divergence
Relief	Random	Distance	Distance
Genetic algorithm	Random	Weighted	Consistency

Table 3 Comparison of Feature selection algorithms

4.6 Linear Discriminant Analysis (LDA) Vs Principal Component Analysis (PCA)

Linear Discriminant Analysis plays an important role in linear transformation of data. LDA is a supervised technique it attempts to search a subspace of features which maximizes the class separability. Whereas

Principal component analysis (PCA) ignores class labels it considered the entire samples as a whole data set. LDA creates assumptions about normally distributed classes and equal class covariance's.

The methodology which is used to reduce the dimensionality of the feature space and push round in this reduced feature space. An elementary technique exactly suited for this problem is the *Principal Component Analysis* which attempts to find the directions of most variation in our data set.

PCA has the prospective to make feature selection and is able to select a number of vital individuals from all the feature components. Enormous data sets were increased in common and regularly include measurements on several variables. It is quite possible to reduce the number of variables when more information is available in the original dataset. Principal component analysis (PCA) is an efficient linear transformation technique for dimensionality reduction.

5 Proposed Methodology

5.1 Principle Component Analysis (PCA) Feature Selection

5.1.1 Definition of PCA

The main objective of PCA technique is to discover low dimensional space which is transformed from high dimensional space. PCA space direction explains the maximum variance. The direction of the PCA space represents the direction of the maximum difference of the given data as shown in Figure 5. As shown in the figure 5, the PCA space is contains of a number of Principle Components. Based on the amount of variance in its direction every principal component has a different strength.

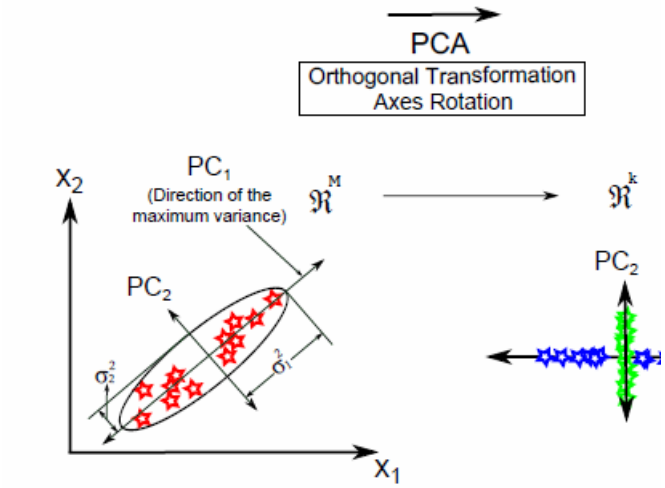


Figure 5 Example of the two-dimensional data (x_1, x_2).

5.1.2 Principal Components (PCs)

The PCA space contains k principal components. The characteristics of principal components are orthonormal, uncorrelated, and it represents the direction of the maximum variation. The first principal component $\left((PC_1 \; or \; v_1) \in \mathcal{R}^{M\times 1}\right)$ of the PCA space represents the direction of the maximum variance of the data, the second principal component has the second largest variance, and so on. Figure 5 shows how the original data are transformed from the original space $\mathcal{R}^M$ to the PCA space $\mathcal{R}^k$. Thus, the PCA technique is considered an orthogonal transformation due to its orthogonal principal components or axes rotation due to the rotation of the original axes. The following method is used to calculate the principal components.

5.1.3 Covariance Matrix Method

In this method, there are two main steps to calculate the PCs of the PCA space. First, the covariance matrix of the data matrix (X) is calculated. Second, the eigenvalues and eigenvectors of the covariance matrix are calculated. calculating the PCs using the covariance matrix method.

5.1.4 Calculating Covariance Matrix (**Σ**):

The variance of any variable calculates the difference of that variable from its mean value and it is defined as follows

$$\sigma^2(x) = Var(x) = E(x-\mu)^2) = E\{x^2\} - E(x)^2 \quad (7)$$

where μ is the mean of the variable x, and E(x) refers the expected value of x. When more number of variables are present then the covariance matrix is used and it is defined as follows

$$\Sigma_{ij} = E\{x_i x_j\} - E\{x_i\}E\{x_j\} = E[(x_i-\mu_i)(x_i-\mu_i) - (x_j-\mu_j] \quad (8)$$

The eigenvectors stand for the directions of the PCA space, and the corresponding eigenvalues represent the scaling factor, length, magnitude, or the strength of the eigenvectors. The highest Eigen value referred as the first principal component and it has the maximum variance.

5.1.5 Borrowed PCA

In this borrowed feature selection double measures for selecting value in feature selection investigation, and classification. These two measures are based on the idea of "striking" unnecessary features. To cancel the effect of a features, they replaced all its values with the marginal mean (in the first procedure) or with the conditional mean (in the second). The bordering mean method is mainly intended to identify "noisy feature" unhelpful variables, but the provisional mean method could also deal with feature dependence. The idea behind the second procedure to PCA.

The PCA define variable feature $X \in R^p$ as a random feature vector with distribution P. The

coordinates of the vector X are defined as $X[i], i = 1.2 \ldots . p$. The covariance matrix of X is denoted Σ.

Borrowed PCA Algorithm

Step 1: Calculate dissimilar matrix

For a given random feature vector $\tilde{X}$, we say that it satisfies the assumption H1 if:

i) $E\left(\left\|\tilde{X}\right\|^2\right) < \infty$

ii) After the covariance matrix of $\tilde{X}$ is positive define.

iii) Eigenvalues of all the covariance matrix are dissimilar.

As is fine known, the first principal component related with the feature vector X is defined as

$$\alpha^1(P) = \alpha^1 \max Var(\alpha' X)$$
$$= \arg\max \alpha' \Sigma \alpha_{\|\alpha\|=1} \quad (9)$$

and the next principal components are defined as

$$\alpha^k(P) \coloneqq \alpha^k = \arg max_{\|\alpha\|=1, \alpha \perp [\alpha^1, \ldots . \alpha^{k-1}]} \; Var(\alpha' X)$$
$$= \arg max_{\|\alpha\|=1, \alpha \perp [\alpha^1, \ldots . \alpha^{k-1}]} \alpha' \Sigma \alpha \quad \forall \, 2 \le k \le p \; (10)$$

Where $\alpha^1, \ldots . \alpha^{k-1}$ is the subspace generated by the vectors $\alpha^1, \ldots . \alpha^{k-1}$.

From the borrowed theory, it tracks that, if $\lambda^1 < \lambda^2 < \cdots < \lambda^p$ are the Σ eigenvalues, the solutions to the PCA are the consistent eigenvectors, $\alpha^k, k = 1, \ldots p$.

Step 2: local based similarity calculation // **Local Borrowed Function**

The local borrowed objective function $H^l(I)$ as

$$H^l(I, P, P_{YI}) \coloneqq h^l(I) = \|\alpha^1(p) - \alpha^1(P_{YI}\|^2 \quad (11)$$

which procedures the formed distance between the oldest original principal component and the current principal component that is a function of the variables in the subset I.

Given a fixed variable of PCA $d, 1 \le d \ll p$, I_d, Id is the family of all subsets of $\{1, \ldots p\}$ with cardinality d and $I_{1,0} \subset I_d$is the family of subsets in which the minimum $H^l(I)$ is attained for $I \in I_d$

$$I_{1,0} = argmin_{I \in I_d} \, h^l(I) \quad (12)$$

or, consistently,

$$h^1(I_1) = min_{I \in I_d} \, h^1(I) for \; all \; I_1 \in I_{1,0} \quad (13)$$

Step 3: borrowed feature minimum distance calculation borrowed define

$$h^l(I, P, P_{YI}) \coloneqq h^l(I) = \|\alpha^l(p) - \alpha^l(P_{YI})\|^2 \quad (14)$$

and

$$I_{l,0} = argmin_{I \in I_d} h^l(I) \quad (15)$$

find a small cardinal subset I, for which the objective function is small sufficient, then the k-th principal component will be well clarified by a subset of the original variables.

6.2.2

Step 4: Global borrowed feature minimum distance calculation // **Global Borrowed Function**

Global borrowed feature objective function is

$$h(l) \coloneqq \sum_{l=1}^{q} p_l h^l(I) \quad (16)$$

with $p_l \ge 0, \sum_{l=1}^{q} p_l = 1, \; 2 \le q \le p \quad (17)$

$$\tilde{I}_{q,0} = argmin_{I \in I_d} h(I)$$

This time the objective function contracts with discovery a single subset features I to clarify the first q principal components on once. The q components are similarly significant then choose $p_l = \frac{1}{p}$. Otherwise, if some components are more important than others then we can put different weights. As a default choosing weights relative to the alteration that each component is illumination $p_l = \lambda^l / \sum_{l=1}^{q} \lambda^l$ (18).

6 Artificial Neural Networks

ANN is for nonlinear problems with flexibility. It follows three layered architecture: input layer is used for all input variables and output layer is for output neuron. Based on weights and bias the relationship between neurons can be identified. To increase the classification accuracy by using ANN, during training process weights and biases are adjusted. Artificial neural networks (ANNs) have been effectively used for classification, prediction, and association in different problem domains.

6.1 Adam Optimization Algorithm

Based on adaptive valuations of lower-order moments,an algorithm is called Adam, [14] used for first-order gradient-based optimization of stochastic objective functions,. This method implementation part is direct and is computationally talented and has tiny memory requirements. Its property is never changed when the transformations applied to diagonal rescaling of the gradients. When the data or parameter are huge in size thene it is appropriate for such type of problems. From evaluations of first and second moments of the gradients the method calculates individual adaptive learning rates for different features.

6.2 Classification Algorithms

6.2.1 Naïve bayes

It is a probabilistic learning method. By Applying the Bayes theorem, probability of class label (Ci) can be calculated in specified all attributes Aj and predict the greatest posterior probability class.

For the observation count value n , with the available instance X, the Probability of class value calculated as:

$$p(C_i|X) = \prod_{j=1}^{n} p(A_j|C_i) \, . \; p(C_i) \quad (19)$$

6.2.2 Decision tree

Decision tree is a tree structure model which follows top down approach. which contains numerous branches, nodes and subnodes (leaf) . Each node

represents an attribute or variable. Entire dataset is divided into small data sets called branches. Leaves insists the class value that assigns each observation into one of the leaf value. There are different tree classifiers are available. In that CART classifies the entire transactions into subset of transactions which are treated as same values for the target variable. (BrownandMues,2012).

6.2.3 Support Vector Machine(SVM)

A Support Vector Machine (SVM) is an efficient classifier legitimately defined by separating hyper plane. SVM algorithm yields an optimum hyperplane which groups the new samples on, available. labelled training data which is referred as *supervised learning*. This hyper plane is a line separating a plane in to 2 parts in two dimensional space in which each class lay in each side.

7 Dataset Description

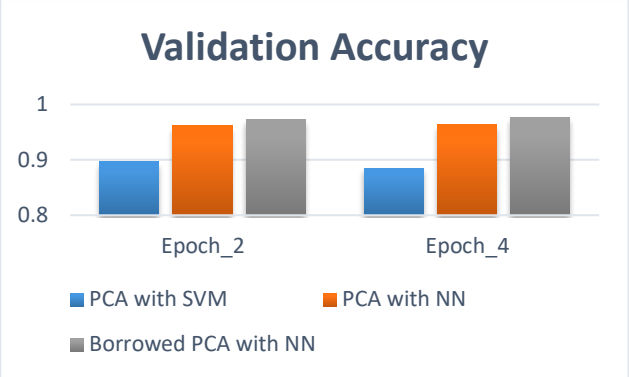
In our Experiment the KDD Cup 1999 dataset is used for measuring intrusion detection problems .Objective of the competition was to form a network intrusion detector, a analytical model accomplished the difference among bad connections, referred intrusions or attacks, and good connections. On a local area Network, over the periods of 9 weeks, the dataset was a collected. It is a pretend raw TCP dump data. Training data : Five million connections records from seven weeks of network traffic and Test data : two weeks of data generated near two million connection records. The training data is prepared with 22 different attacks out of the 39 available in the test data. The training dataset contains the known attack types but the novel attacks are included attacks in the test datasets not existing in the training data sets.

In total 494,021 records, the training dataset consists of 97,277 were ordinary which is 19.69 percentage, 391,458 Denial Of Service that is 79.24 percent, 4,107 which is 0.83 percent Probe, 1,126 R2L 0.23 percent and 52 i.e 0.01 percent U2R connections. In every connection there are 41 attributes relating different features of the connection also the attack type or as normal connection can be assigned as label.

8 Experimental Results and Findings:

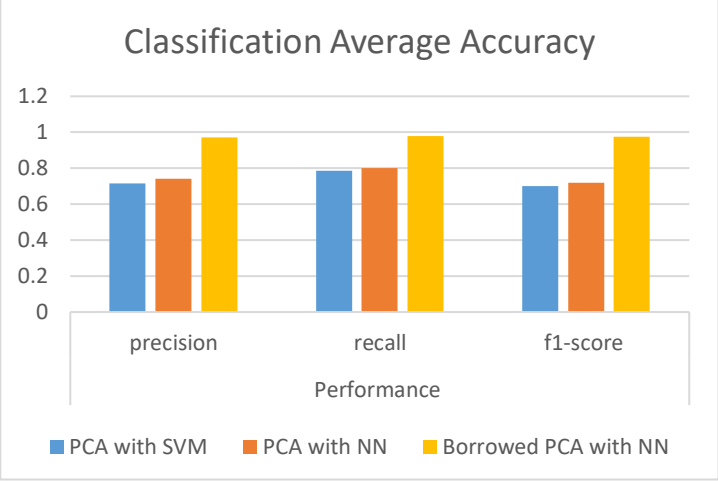
Algorithms	Epoch	
	2	4
SVM	0.8965	0.8853
PCA with NN	0.9615	0.9631
Borrowed PCA with NN	0.9725	0.9771

Table 4 Validation Accuracy



Algorithms	Performance		
	Precision	recall	f1-score
SVM	0.7164	0.7843	0.7011
PCA with NN	0.74	0.80	0.72
Borrowed PCA with NN	0.971	0.978	0.975

Table 5 Classification Average Accuracy



9 Conclusion and Future Work

we proposed the efficient feature subset selection by using borrowed PCA technique with Artificial Neural Network. PCA is the best method in feature subset selection for both supervised and unsupervised technique. Irrespective of dependency of data, our proposed approach is well suited for supervised

techniques in various domains. The method has been tested with intrusion detection dataset and we used the best classifier ANN to improve the accuracy in case of noisy data. Proposed borrowed PCA does not eliminate the prominent features which are considered as an important feature to make the prediction in an accurate way. In our future work we aim to explore the possibility of applying borrowed PCA for unsupervised techniques to improve clustering accuracy. Also Linear Discriminant Analysis can be used for feature subset selection in supervised techniques.

References

(1) Journals:

[1] Pabithra mitra, CA murthy and sankar ,Unsupervised Feature selection using feature similarity, IEEE Transactions on Pattern Analysis and Machine Intelligence , Vol.24, No.3, March 2002.

[2] A. Salappa, M. Doumpos and C. Zopounidis*,Feature selection algorithms in classification problems:an experimental evaluation, Journal of Optimization Methods and Software, Vol.22,No.1,pgno:199-214, 2007.

[3] Fatemeh Nemati Koutanaei, Hedieh Sajedi, Mohammad Khanbabae , A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring, Journal of Retailing and consumer services pgno:11-23, 2015,Elsevier.

[4] Novaković, Jasmina. "Toward optimal feature selection using ranking methods and classification algorithms." *Yugoslav Journal of Operations Research* 21.1 (2016).

[5] Xu, Yan, Peng Qiu, and Badrinath Roysam. "Unsupervised discovery of subspace trends." *IEEE transactions on pattern analysis and machine intelligence* 37.10 (2015): 2131-2145.

[6] Kale, Archana Pritam, and Shefali Sonavane. "PF-FELM: A Robust PCA Feature Selection for Fuzzy Extreme Learning Machine." *IEEE Journal of Selected Topics in Signal Processing* 12.6 (2018): 1303-1312.

[7] Foxall, Gordon R., et al. "Consumer behavior analysis and social marketing: The case of environmental conservation." *Behavior and social issues* 15.1 (2006): 101-124

[8] Padma, P., et al. "Changing Scenario of Household Consumption Pattern in Kerala: An Emerging Consumer State of India." *Social Indicators Research* 135.2 (2018): 797-812.

(2) Conference Proceedings:

[9] Partridge, Matthew, and Marwan Jabri. "Robust principal component analysis." *Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop*. Vol. 1. IEEE, 2000

[10] Nandi, G. (2011). An enhanced approach to Las Vegas Filter (LVF) feature selection algorithm. 2011 2nd National Conference on Emerging Trends and Applications in Computer Science. doi:10.1109/ncetacs.2011.

[11] Molina, Luis Carlos, Lluís Belanche, and Àngela Nebot. "Feature selection algorithms: A survey and experimental evaluation." *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, 2002.

[12] Bamakan, Seyed Mojtaba Hosseini, and Peyman Gholami. "A novel feature selection method based on an integrated data envelopment analysis and entropy model." *Procedia Computer Science* 31 (2014): 632-638.

[13] Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. "A survey of feature selection and feature extraction techniques in machine learning." Science and Information Conference (SAI), 2014. IEEE, 2014.

[14] Diederik P. Kingma*, Jimmy Lei Ba, “ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION” International conference on Learing representations, 2015.

[15] Ahmad, Iftikhar, and Fazal e Amin. "Towards feature subset selection in intrusion detection." *Information Technology and Artificial Intelligence Conference (ITAIC), 2014 IEEE 7th Joint International*. IEEE, 2014.

(3) Books:

[16] W. Siedlecki And J. Sklansky , A Note On Genetic Algorithms For Large-Scale Feature Selection University of California, Irvine, CA 927J7, USA chapter 1.3.2, handbook of pattern recognition and computer vision.

(4) Blogs:

[17]https://www.analyticsvidhya.com/blog/tag/feature-selection

D.Hemavathi working with SRM Institute of Science and Technology, Chennai from 2004. Working as a Assistant Professor, Senior Grade in the department of Information Technology. She is very much interested in the area of Distributed systems, Big data analytics. She is one of the core member in Big data Analytics Research group. Research emphasis is on Data Analytics

H. Srimathi is having 18 years fteaching experience and 6 years of industrial experience. Her area of interest include Semantic Web, eLearning, Object Oriented software engineering, Cloud and Big Data. She has published a book on "Object Oriented Analysis and Design using UML". She has served as Senior Manager, Curriculum Development for the project of Curriculum Development of Computer science subjects for Form 4 and Form 5 students, Ministry of Education, Malaysia. She has published 16 papers in National and International Journals.