STUDY ON FEATURE SELECTION AND FEATURE EXTRACTION TECHNIQUES IN DATA MINING

Wasim Akram[1], Dr. Avinash Sharma[2]

[1]Wasim Akram, M. Tech. Scholar, CSE, MITS Bhopal
[2]Dr. Avinash Sharma, Assistant Professor(HOD)

ABSTRACT: Dimensionality reduction in data mining focuses on representing data with minimum number of dimensions such that its properties are not lost and hence reducing the underlying complexity in processing the data. Principal Component Analysis (PCA) is one of the prominent dimensionality reduction techniques widely used in network traffic analysis. In this paper, efficiency of PCA and SPCA has been emphasized for intrusion detection and its Reduction Ratio (RR) has been determined, ideal number of Principal Components needed for intrusion detection and the impact of noisy data on PCA. Feature selection and Feature Extraction are one of the methods used to reduce the dimensionality. Till now these methods were using separately so the resultant feature contains original or transformed data. An efficient algorithm for Feature Selection and Extraction using Feature Subset Technique (FSEFST) in High Dimensional Data has been proposed in order to select and extract the efficient features by using feature subset method where it will have both original and transformed data. The results prove that the suggested method is better as compared with the existing algorithm

KEYWORDS : Dimensionality reduction, Data Mining , Principal Component Analysis (PCA), Reduction Ratio (RR), Feature selection and Feature Extraction.

1. INTRODUCTION

The High dimensional data can be transformed into low dimensional by Feature Extraction. Feature Extraction methods are also characterized into supervised and unsupervised methods. The familiar feature extraction methods are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The PCA and LDA methods will find reduced set in original or converted but not both. An approach is developed here where both original and transformed features can be obtained by considering the suitable threshold.

1.1 Principal Component Analysis (PCA) and Supervised Principal Component Analysis (SPCA)

Principal Component Analysis is a feature extraction technique that generates new features which are linear combination of the initial features. PCA maps each instance of the given dataset present in a d dimensional space to ak dimensional subspace such that $k < d$. The set of k new dimensions generated are called the Principal Components (PC) and each principal component is directed towards maximum variance excluding the variance already accounted for in all its preceding components. Subsequently, the first component covers the maximum variance and each component that follows it covers lesser value of variance. The Principal Components can be represented as the following where *PCi*— Principal Component '*i*'; *Xj*— original feature'*j*'; *aj*— numerical coefficient for *Xj*. PCA is one of the most prominently used feature extraction methods for traffic analysis. Brauckh off et al. (2009) discussed about implementing PCA with KL expansion method for anomaly detection and issue of right number of PC for analysis. Ringberg et al. (2007) discusses the sensitivity of PCA for anomaly detection, an issue related to number of PC, impact of anomaly size and gives a comprehensive study of the related issues on Abilene and Geant networks. Issariyapat and Kensuke (2009) discuss about using PCA for MAWI network and using the information from packet header for detecting anomaly.

PCA finds the "principal components" in the data which are uncorrelated Eigen vectors each representing some proportion of variance in the data. PCA and many variations of it have been applied as a way of reducing the dimensionality of the data .Supervised principal component analysis (SPCA) was proposed, which selects the PCs based on the class variables. PCA has an important limitation: it cannot capture nonlinear relationships that often exist in data, especially in complex biological systems. SPCA works as follows:

(1) Compute the relation measure between each micro-array data with outcome using linear, logistic, or proportional hazards models.

(2) Selected micro-array data most associated with the outcome using cross-validation of the models in step1.

(3) Principal Component Scores estimated using only the selected micro-array data.

(4) Regression fitted with outcome using model in step 1.

A similar linear approach is Classical Multidimensional Scaling (CMDS) or Principal Coordinates Analysis which calculates the matrix of dissimilarities for any given matrix input. It was used for large micro-array datasets because it is efficient in combination with Vector Quantization or *K*-

Means which assigns a similar linear approach is Classical Multidimensional Scaling (CMDS) or Principal Coordinates Analysis that calculates the matrix of dissimilarities for any given matrix input to a class, out of a total of K classes.

2. RELATED WORK

Hoang Vu Nguyen et al., [11] introduced efficient feature extraction method which improves the detection accuracy when applied on two detection techniques. Yue WU et al., [12] applied maxHeap based approach instead of OFS (Online Feature Selection) to extract more efficient features compared to existing batch learning methods. Agnieszka Wosiak et al., [13] used SVMREF (Support Vector Machine-Recursive Feature Elimination) technique to classify both binary and multiclass dimensional set for a specific application. ZHAO Zhongwen et al., [14] employed PCA and SVM to reduce dimension, classify the data and project the classified data to two-dimensional features. Manikandan G et al., [15] proposed a method to discover the optimal threshold value and feature subsets are given to the classifier to obtain maximum accuracy. Ding et al., [16] discuss about the novel feature selection framework for generally minimizing the feature redundancy to increase the ranking score for the given feature, which can originate from any supervised or unsupervised methods. W. Sheng et al., [17] proposed a technique to refine feature selection and to obtain the cluster centers prearranged in the chromosomes called local search operations. D. Cai et al., [18] discussed about unsupervised feature selection to select those features which are having multi-cluster structure where the data can be preserved by the method called Multi-Cluster Feature Selection (MCFS). Z. Xu et al., [19] proposed a novel discriminative semi supervised feature selection method in order to maximize the classification margin to differentiate between labelled and unlabelled data. K. Benabdeslem et al., [20] describes about two important processes to provide an effective selection of semi-features in semi-supervised environment. P. Mitra et al., [21] introduced a method called maximum information compression index, which describes how feature selection and extraction can be improvised with an entropy measure.

Dash M et al., [22] recommended a filter technique that is independent of any clustering algorithm. The method has entropy calculation that is low if clusters are distinct and high clusters are not distinct. Song Q et al., [23] proposed a FAST algorithm which makes use of the Minimum Spanning Tree (MST) clustering means to find the efficient features. Different clusters will have relatively independent Features and FAST algorithm has a high possibility of generating a subset of useful and independent features. .

3. METHODOLOGY

The normal procedure in any feature extraction involves finding the Eigen value and Eigen vector by framing criterion function. Highest eigenvectors comprise most of the information. Least Eigen value means the little information along side its resultant Eigen vector.

Chart -1: Name of the chart

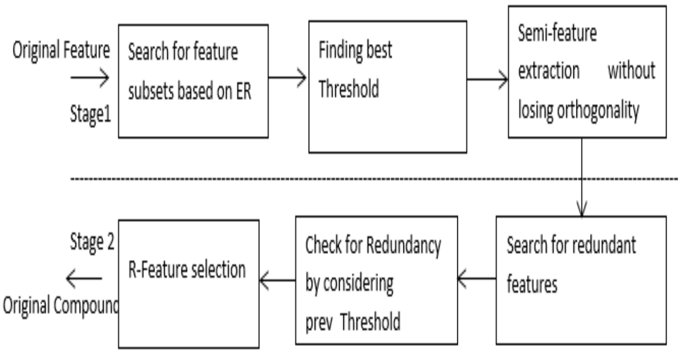
Original Feature
Stage1
Search for feature subsets based on ER
Finding best Threshold
Semi-feature extraction without losing orthogonality
Search for redundant features
Check for Redundancy by considering prev Threshold
R-Feature selection
Stage 2
Original Compound

Fig. [1] Stages of finding the compound feature generation

TABLE I. Notations used in the algorithm

Symbols	Definitions
N	Number of Original Features
T1, T2	Predefined Threshold values
D	Set of semi-features
ER i , j	Error rate
O	Original Feature set {f1, f2, …..fn}

The Notations used in algorithm are given in Table I. Algorithm 1 is to find the features and the best Threshold. Initially the set D will be empty and the selected semi-features are stored in D by considering the pair Vi and Vj and if the considered pair threshold is less than T1 then check for the Prev Threshold. If it is less than Prev Threshold then the pair which considered the best pair.

Algorithm 1: To search for features based on (Error Ratio) ER and to get best Threshold

```
Begin
Step 1: for each($v_j \neq v_i$) ∈O∪ D do
Step 2: if Red($v_i$>$v_j$)>$T_2$then
Step 3: Red ($v_i$>$v_j$)>prev Threshold then
Step 4: prev Threshold ←Read ($v_i$,$v_j$)
Step 5:            $v_i$ ←i
```

```
 Step 6:             $v_j$ ←j
 enf if
 end for
Step 7: u'← R-Feature of {i,$v_j$}
Step 8: u" ←weak feature of {i,$v_j$}
Step 9: if u" ∈O then
 O ←O – {"}
else
 D ←D–{"}
if (Dim ==| O ∪D|) then
 flag←FALSE
else
 flag←TRUE
end if
 endif
step10 : return C← O∪D
end
```

Algorithm 2: For Adding transformed Features to Semi-feature set

```
Begin
Step 1 : dk←First eigen component $v_i$ &$v_j$
Step 2 : D← D ∪ {K}
Step 3 : K← $K + 1$
Step 4 : if $v_i, v_j$ ∈  O then
 O ← O – {i,$v_i$}
 else if $v_i$ ∈ O,$v_j$ ∈ D then
 O ← O – { i }
 D←D–{j}
 else if $v_i$ ∈ D, $v_j$ ∈ O
then
 O←O–{ j}
 D ← D –{i}
 end if
 end if
end if
else
 D ← D –{i,$v_i$ }
 for each$v_i$ ∈ O ∪ D do
 prev Threshold then←  $T_2$
 end for
end
```

Algorithm3: For Checking Redundancy

```
Begin
Step 1: Consider Set of semi –feature D (Initially Empty)
Step 2: for a pair ( $v_i$,$v_j$) where $v_i$ ∈ O ∪D
 Prev Threshold=$T_1$
 for each($v_j \neq v_i$ ) ∈ O ∪ D do
 if $ER_{i,j}$<$T_1$ then
 if $ER_{i,j}$<Prev Threshold
 Prev Threshold← $ER_{i,j}$
                          $v_i$ ←i
                          $v_j$ ←j
 end if
 end if
 end for
 end for
 end
```

If the redundancy between (vi, vj) amongst the pair is better than a threshold T2 from the available features (vi, vj) , then R-feature is calculated using Ranking Criterion of both the features. Algorithm 3 gives one of the weak features that are rejected to get the reduced set.

4. Results and Discussion

TABLE II. Data Description

Dataset	Instances	Attributes	Classes
Abalone	4177	8	3
Ecoli	332	7	6
Pageblocks	5473	10	5

TABLE III. Ecoli Dataset for Classification Accuracy

EColi (d=5),T[(0.1, 0.7), (0.15, 0.85)]			
Sl.No.	Algorithm	Classification Accuracy	
		MPEMR	FSEFST
1	KNN1	0.8089	0.8126
2	KNN3	0.8165	0.854
3	KNN5	0.8292	0.8558
4	SVM	0.8548	0.8732

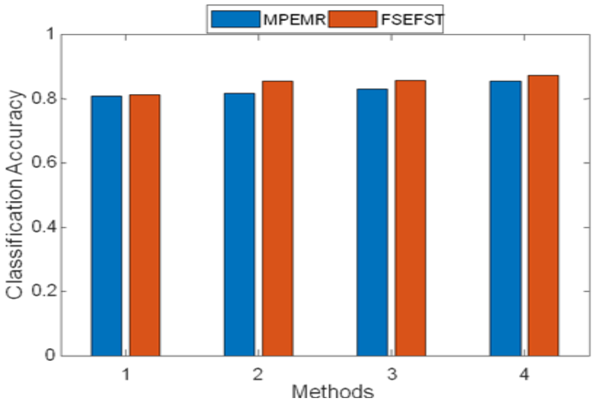

Fig.[2] EColi Dataset for Classification accuracy

TABLE IV. Ecoli Dataset for Redundancy checking

EColi (d=5),T[(0.1, 0.7), (0.15, 0.85)]			
Sl.No.	Algorithm	Redundancy checking	
		MPEMR	FSEFST
1	KNN1	0.76688	0.766
2	KNN3	0.78816	0.788
3	KNN5	0.81429	0.814
4	SVM	0.8244	0.824

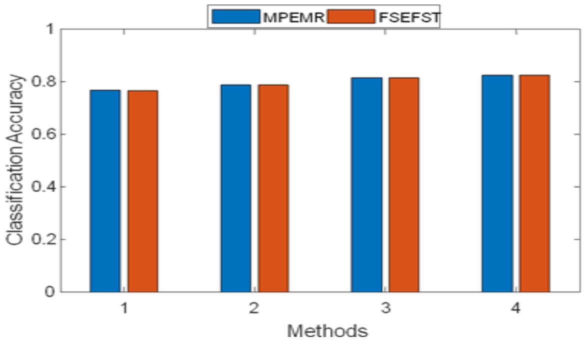

Fig.[3] Ecoli Dataset for Redundancy checking

TABLE V. Pageblocks Dataset for Classification Accuracy

Pageblocks (d=6), T[(0.08, 0.95), (0.1, 0.75)]			
Sl.No.	Algorithm	Classification Accuracy	
		MPEMR	FSEFST
1	KNN1	0.9632	0.969
2	KNN3	0.9674	0.9690
3	KNN5	0.966	0.9670
4	SVM	0.953	0.9513

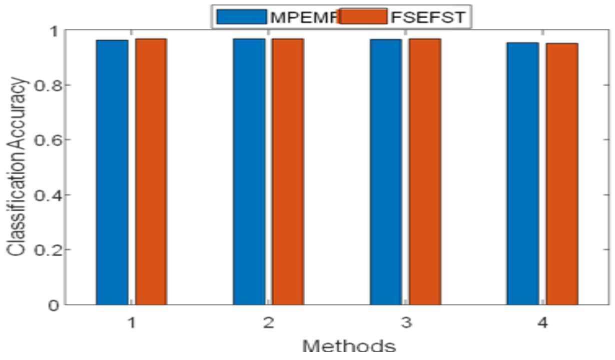

Fig.[4] Pageblocks Dataset for Classification Accuracy

TABLE VI. Pageblocks Dataset for Redundacy checking

Pageblocks (d=6), T[(0.08, 0.95), (0.1, 0.75)]			
Sl.No.	Algorithm	Redundancy checking	
		MPEMR	FSEFST
1	KNN1	0.9179	0.9640
2	KNN3	0.9198	0.9680
3	KNN5	0.9102	0.9666
4	SVM	0.8991	0.9519

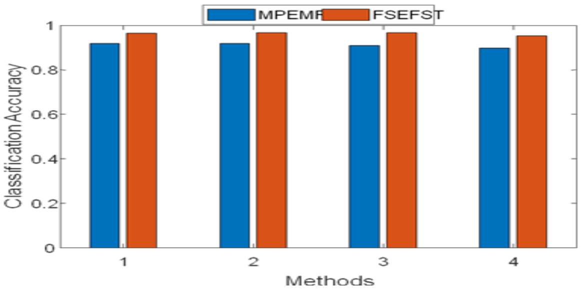

Fig.[5] Pageblocks Dataset for Classification Accuracy

TABLE VII. Abalone Dataset for Classification Accuracy

Abalone (d=2),T[(0.01, 0.95), (0.001, 0.5)]			
Sl.No.	Algorithm	Classification Accuracy	
		MPEMR	FSEFST
1	KNN1	0.4467	0.4819
2	KNN3	0.4764	0.4967
3	KNN5	0.4857	0.5190
4	SVM	0.5257	0.5353

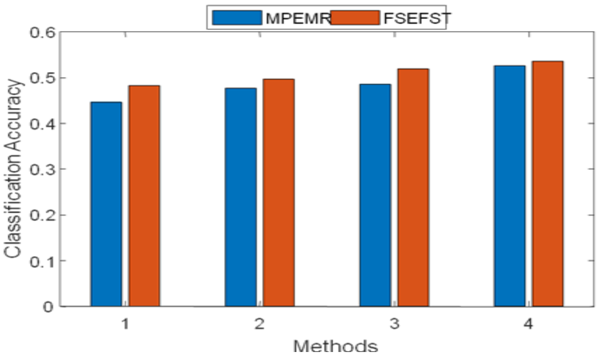

Fig.[6] Abalone Dataset for Classification Accuracy

TABLE VIII. Abalone Dataset for Redundancy checking.

Abalone (d=2),T[(0.01, 0.95), (0.001, 0.5)]			
Sl.No.	Algorithm	Redundancy Accuracy	
		MPEMR	FSEFST
1	KNN1	0.4715	0.4632
2	KNN3	0.4932	0.4856
3	KNN5	0.5078	0.4938
4	SVM	0.5460	0.5353

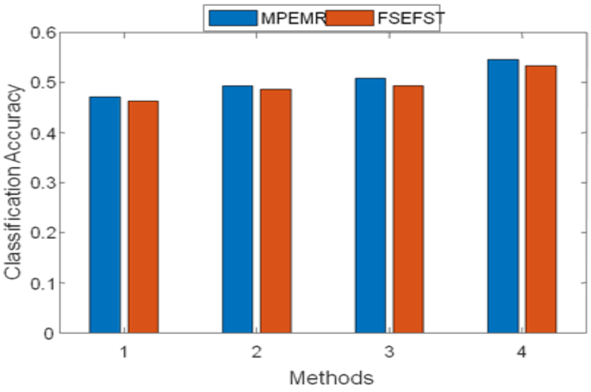

Fig.[7] Abalone Dataset for Redundancy checking.

5. CONCLUSIONS AND SCOPE OF FUTURE WORK.

The Dimensionality reduction achieved by getting the reduced set from the combination of both original and the reduced set of the features. In order to get the reduced set a framework is proposed which generates the compound features by having minimum projection error and minimum redundancy. As the proposed approach gives the reduced set with the combination of both original and features in the condensed set, the results are compared with existing feature selection and extraction methods. The proposed method is showing the improvement in finding the projection error and redundancy check.

In future the FSEFST method can be applied on hyper spectral images classification to get reduced dimensions and reduction ratios of different hyper spectral images.

REFERENCES

[1] Abeer Alzubaidi and Georgina Cosma, "Efficient Feature Selection Algorithm for High Dimensional Data", International Journal of Electrical and Computer Engineering (IJECE), Vol. 6, No. 4, August 2016, pp. 1880-1888.

[2] Agnieszka Wosiak, Agata Dziomdziora, "Feature Selection and Classification Pairwise Combinations for High-Dimensional Tumour Biomedical Datasets", Schedae Informaticae, vol. 24, pp. 53-62, 2015.

[3] Bharat Singh, Nidhi Kushwaha and Om Prakash Vyas, " A Feature Subset Selection Technique for High Dimensional Data using Symmetric Uncertainty", Journal of Data Analysis and Information Processing, pp. 95-105, 2014.

[4] Brauckhoff, D., et al., 2009. Applying PCA for traffic anomaly detec-tion: problems and solutions. In: INFOCOM 2009, IEEE.

[5] D. Asir Antony Gnana Singh, S. Appavu and E. Jebamalar Leavline, "Literature Review on Feature Selection

Methods for High-Dimensional Data", International Journal of Computer Applications, vol. 136, No.1, February 2016.

[6] D. Cai, C. Zhang and X. He, "Unsupervised Feature Selection for Multi-Cluster Data", In Proceedings of 16th ACM SIGKDD International Conference Knowledge Discovery Data Mining, pp. 333-342, 2005.

[7] D. Sheela Jeyarani and A. Pethalakshmi, "Optimized Feature Selection Algorithm for High Dimensional Data", Indian Journal of Science and Technology, vol. 9, August 2016.

[8] Dash M, K Choi, P Scheuermann and H. Liu, "Feature Selection for Clustering a Filter Solution", In Proceedings of 2nd International Conference Data Mining, pp. 115-122, 2002.

[9] Ding, Hanuchuan Peng and Fuhui Long, "Feature Selection based on Mutual Information Criteria Max-Dependency, Max-Relevance and Min-Redundency", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, 2005.

[10] F Nie, S. Xiang, Y. Jia, C. Zhang and S. Yan, "Trace Ratio Criterion for Feature Selection", In Proceedings of 23rd National Conference in Artificial Intelligence, 2008, vol. 2, pp. 671-678.

[11] H. Liu and H. Motoda, "Computational Methods of Feature Selection", Boca Raton, FL, USA, CRC Press, 2007.

[12] Hoang Vu Nguyen and Vivekanand Gopalkrishnan, "Feature Extraction for Outlier Detection in High Dimensional Spaces, In Proceedings of Fourth Workshop on Feature Selection in Data Mining JMLR : Workshop and Conference vol. 10, pp. 66-75.

[13] Issariyapat, C., Kensuke, F., 2009. Anomaly detection in IP networkswith Principal Component Analysis. In: Communications and Information Technology, IEEE.http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html(18.02.16).

[14] J Liu, S Ji and I Ye, "Multi-task Features Learning via Efficient L2, L-noorm Minimization", In Proceedings of 20th International

[15] J Liu, S Ji and I Ye, "Multi-task Features Learning via Efficient L2, L-noorm Minimization", In Proceedings of 20th International Conference Machine Learning,vol. 3, pp. 856-863, 2003.

[16] K. Benabdeslem and M. Hindawi, "Efficient Semi-Supervised Feature Selection Constraint, Relevance and Redundancy", IEEE Transaction Knowledge Data Engineering, vol. 26, No. 5, pp. 1131-1143, May 2014.

[17] M. Robonik, Sikonja and I. Kononenka, "Theroretical and Empirical analysis of relief and rrelieff", Machine Learning, vol. 53, no. ½, pp. 23-69, 2003.

[18] Manikandan G, Susi E and Abirami S, "Feature Selection on High Dimensional Data using Wrapper Based Subset Selection", In Proceedings of Second International Conference on Recent Trends and Challenges in Computational Models, pp. 320-325, 2017.