Integrated Modified PC-LD Analysis Based Dimensionality Reduction Techniques

Dr. P. SARASWATHI,
Assistant Professor, Dept of Computer Science
Sindhi College, Chennai – 600 077.Tamil Nadu, India.

Abstract- *In the field of machine learning and pattern recognition, dimensionality reduction is an important research area. In recent years, dealing massive dataset poses a severe challenge to many existing feature extraction and selection methods in terms of efficiency and effectiveness. Feature extraction and selection addresses the difficulty of identifying the most useful set of features. These methods aim to remove redundant and irrelevant features to achieve more accurate results. The existing work has the problem in high dimensionality of feature space. The paper fills the gaps by integrating the PCA and LDA methods in the dataset. In this paper, proposed two features extraction approaches that integrated as modified PC-LD analysis through some statistical criterion. Our experiment result shows that modified PC-LD analysis has achieved better accuracy compared to existing work.*

Keywords- Feature Extraction, Pattern Recognition, Feature Selection, Machine Learning, Dimensionality Reduction.

I. INTRODUCTION

As the dimensionality of the data rises in machine learning, the quantity of data essential to give a reliable analysis grows exponentially. The machine learning application depends on various factors, in that the important one is the data quality. The knowledge discovery will be difficult if the data contains irrelevant, redundant or noisy elements. In machine learning technique, there are often too many factors leads the performance of better result. A dataset contains a massive amount of input attributes in a variety of cases that make the task of predictive modeling more complicated. Because it is very complex to visualize or make predictions for the dataset with a high number of features, for such cases, dimensionality reduction techniques are required to use. The process of reducing the random attributes under dimensionality reduction obtains a set of principal attributes. It can be divided into feature extraction and feature selection.

Feature extraction is a part of the dimensionality reduction techniques in which, an initial set of the raw data is separated and summarized to more convenient groups. Feature selection chooses a subset of original features according to a well-defined evaluation criterion. It is frequently used dimensionality reduction techniques which removes irrelevant and redundant features. This method has more helpful for real applications because it accelerates the algorithm that improves the performance of the model. The other dimensionality reduction techniques like principal component analysis (PCA) or linear discriminant analysis (LDA) are based on projection and that applied the feature selection technique which will not alter the original illustration of attribute set.

In Principal Component Analysis (PCA), a pattern recognition technique is applied to analyze the high dimensional data. For data analysis, needed to reduce the high dimension of the data into low dimension and then interpret the results. In the field of research, it is very difficult to analyze large amount of data. PCA algorithm is used to compute relation between the huge correlated dataset. In LDA, a dimensionality reduction technique is normally used for supervised classification problems. It is used for modeling that separate two or more classes. It is used to project the features in higher dimension space into a lower dimension space. In the paper, proposed a modified PC-LD analysis for better accurate results. It is an integrated technique of PCA and LDA. Compare to existing work the ensemble technique of combined PCA and LDA leads a better performance.

In this paper an ensemble method based on Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) is introduced to maximize the effectiveness of single feature extraction algorithm and to develop an efficient intrusion detection system. The minimum information loss for PCA of preferred class for LDA leads the best challenge to join these two techniques and promote from their optimistic aspects returns with valuable improvement for classification performance.

II. REVIEW OF LITERATURE

Erwin Hidayat et. al. [1] presented a comparative study of feature extraction using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for face recognition. The experimental results showed that LDA is much better than PCA. Safae Lhazmir et al. [2] reduce their dimensionality to a size more compatible with the resolution methods. The aim of this paper is to study the potential of dimensionality reduction in text categorization of a publicly available dataset.

Binjie Xiao et. al. [3] presented a method to extract features by PCA (Principal component analysis) from a series of woods surfaces' images. The results show that can fuse the series and extract features from the same surface by using this proposed method. Fengxi Song et. al. [4] applied principal components analysis (PCA) to feature selection from a viewpoint of numerical analysis. The proposed method takes a number of eigenvectors into account and uses a reasonable scheme to perform feature selection. The proposed method reduces the dimensionality of the original samples in the recognition accuracy.

Wu Xiao-Jun et. al. [5] determined the optimal set of discriminant vectors for feature extraction in pattern recognition. The proposed a method method used small sample size (SSS) problems shows the effectiveness. Pooja Manghirmalani Mishra et. al. [6] applied large amount of unlabelled data for feature reduction that involves classification. In the result shows the better accuracy to classify the data.

A. Jović et. al.[7] useful for finding accurate data models. Since attribute subset is infeasible, many search strategies have been proposed. The paper provides insight into FS for current hybrid approach. Thomas Rincy N et. al. [8] removes the redundant, irrelevant and noisy features from the original features of dataset by choosing the relevant features having the smaller subdivision of dataset.

S. Gayathri Devi et. al. [9] reviewed few of the available and well-known FS by pointing out the pros and cons of those techniques. This technical work studies the details of traditional techniques depending on evolutionary computation that is helpful in getting the subsets of features from huge datasets. A.K. Shafreen Banu wt. al [10] finds useful features and removes non-relevant features, and reduces the input dimensionality which simplifies the implementation of the classifier and speed up the processing rate. The merits of this method include multiple feature selection criteria and find small subset of features that perform well for the target algorithm.

III. METHODOLOGY

The drawback of the existing work has the problem in high dimensionality of feature space. The paper fills the gaps by integrating the PCA and LDA methods in the dataset. To overcome the existing problem, proposed work develops two features extraction approaches that integrated modified PC-LD analysis through some statistical criterion. In the Figure 1. shows the framework for modified PC-LD analysis.

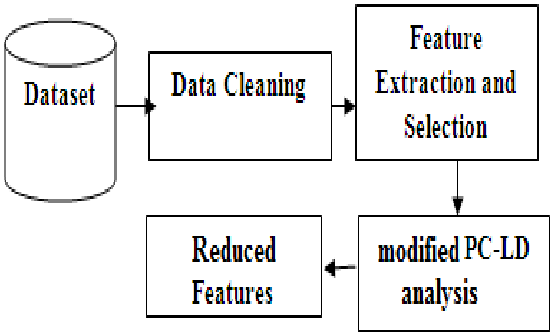


Figure 1. Framework for modified PC-LD analysis

A. Data Collection

The real world dataset are collected in the form of questionnaire contains demographic and geographic details. The collected dataset contains 653 instances with the related features The sample contains 73 features and it is applied to the proposed techniques and extracted 23 most features with original representation of feature set.

B. PCA Feature Extraction

Principal component analysis (PCA) is a method to decrease data dimensionality. It projects high dimensional data to a lower dimension in the least square sense. It captures big principal variability in the data and ignores small variability and reduces the dimensionality of a dataset by finding a new set of features, smaller than the original set of features. PCA is simple, non-parametric method for dimension reduction. For this reason it is useful for the compression and classification of data.

Algorithm: PCA-FE

```
Step 1: Getting the dataset
Step 2: Representing data into a structure
Step 3: Standardizing the data
Step 4: Calculating the Covariance of Z
Step 5: Calculating the Eigen Values and Eigen Vectors
Step 6: Sorting the Eigen Vectors
Step 7: Calculating the new features Or Principal Components
Step 8: Remove less or unimportant features from the new dataset.
```

C. LDA Feature Extraction

LDA focuses mainly on analytical the attributes in upper aspect space to lesser size. You can attain this in three steps:

- First, need to calculate the distance between the mean of different classes.
- Second, compute the distance among the mean and sample of each class.
- Finally, create the lower-dimensional space which maximizes the between-class variance and minimizes the within-class variance.

$$w^* = \underset{w}{\operatorname{argmax}}\, j(w), J(w) = \frac{w^T s_b w}{w^T w_w w}$$

while

$$S_b = (m_1 - m_2)(m_1 - m_2)^t$$

$$S_w = \sum_{i\in 1,2} \sum_{x\in z_i} (x - m_i)^2$$

D. Proposed modified PC-LD analysis for Feature extraction and selection

The ensemble PCA and LDA feature extraction is used, to measure the performance based on the accuracy of class pair binary classification using traditional method. In the LDA and PCA extract the dataset which is trained and divide the dataset that is test dataset into two equal subsets.

First control group for determination of best feature extraction algorithm, while second would be used for actual classification testing. Then, the two subsets performed on the first subset using either LDA or PCA according to performance on the second subset. The experimental results test on both first and second subsets of the testing dataset.

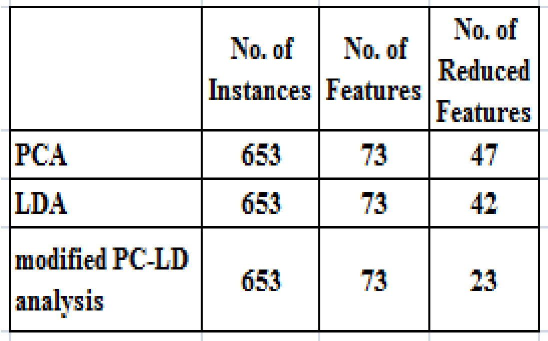
Algorithm: modified PC-LD analysis

1. Getting the dataset
2. Subtract the sample mean from the data.
3. Compute the scatter matrix.
4. Compute Eigen vectors corresponding to the largest k Eigen values.
5. Let the columns of Eigen vector matrix be A= [v1, v2 …vk].
6. The new projected data are listed.
7. Compute class each of mean sample and data.
8. Compute the class scatter matrix. Solve the generalized Eigen value problem.

IV. EXPERIMENTAL RESULT

The Experimental results for the proposed work are performed on WEKA tool, using the collected dataset with its full 73 features and 653 instances. From the above Table I, shows the reduced features for the proposed modified PC-LD analysis for feature extraction and selection**.**

Table I. Proposed modified PC-LD analysis for feature extraction and selection.



	No. of Instances	No. of Features	No. of Reduced Features
PCA	653	73	47
LDA	653	73	42
modified PC-LD analysis	653	73	23

In the Figure. 2, shows the experimental result for the proposed modified PC-LD analysis to extract and select the features.

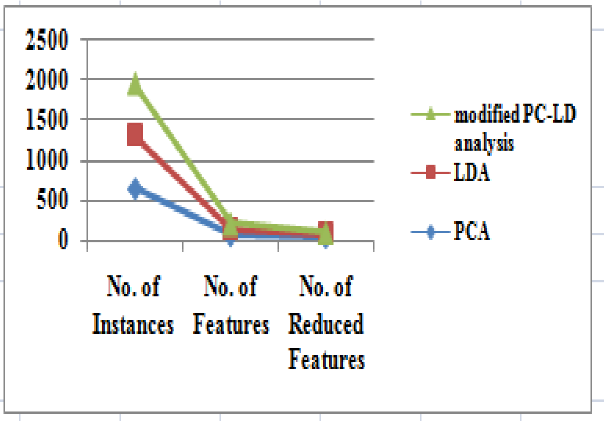


Figure. 2 Experimental result for the proposed modified PC-LD analysis for feature extraction and selection.

V. CONCLUSION

Based on the experimental result the new modified PC-LD analysis shows the feature extraction and selection in increased accuracy to build the model. The experimental result shows that ensemble modified PC-LD analysis has achieved better accuracy compared to existing work. The proposed work showed that feature reduction can progress better detection rate and instinctively the ensemble feature extraction methods showed a very good performance in feature extraction and selection.

REFERENCES

[1] Erwin Hidayat; Nur A. Fajrian; Azah Kamilah Muda; Choo Yun Huoy; Sabrina Ahmad, "A comparative study of feature extraction using PCA and LDA for face recognition", 2011 7th International Conference on Information Assurance and Security (IAS), IEEE,January 2012.

[2] Safae Lhazmir; Ismail El Moudden; Abdellatif Kobbane, "Feature extraction based on principal component analysis for text categorization", 2017 International Conference on Performance Evaluation and Modeling in Wired and Wireless Networks (PEMWN), IEEE,March 2018.

[3] Binjie Xiao, "Principal component analysis for feature extraction of image sequence", 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering, IEEE,June 2010.

[4] Fengxi Song; Zhongwei Guo; Dayong Mei, "Feature Selection Using Principal Component Analysis", 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization, IEEE,November 2010.

[5] Wu Xiao-Jun; J. Kittler; Yang Jing-Yu; K. Messer; Wang Shitong, "A new direct LDA (D-LDA) algorithm for feature extraction in face recognition", Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., IEEE,September 2004.

[6] Pooja Manghirmalani Mishra, Dr. Sushil Kulkarni, " Classification of data using semi-supervised learning", International Journal of Computer Engineering and Technology, Vol 4, Issue 4, IEEE,July-August 2013.

[7] A. Jović; K. Brkić; N. Bogunović, "A review of feature selection methods with applications", 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE,July 2015.

[8] Thomas Rincy N; Roopam Gupta, "Feature Selection Techniques and its Importance in Machine Learning: A Survey", 2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS), IEEE,February 2020.

[9] S. Gayathri Devi; M. Sabrigiriraj, "Feature Selection, Online Feature Selection Techniques for Big Data Classification: - A Review", 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), IEEE,November 2018.

[10] A.K. Shafreen Banu; S. Hari Ganesh, "A study of feature selection approaches for classification", 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), IEEE,August 2015.