iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** III **Month of publication:** March 2019

DOI: http://doi.org/10.22214/ijraset.2019.3434

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue III, Mar 2019- Available at www.ijraset.com

A Review on Feature Subset Generation Techniques

Kajal S. Mahale[1], Dr. Madan U. Kharat[2]
[1]*Student, M.E Computer Engineering, MET's institute of engineering, Nashik, India*
[2]*H.O.D, Computer Engineering, MET's institute of engineering, Nashik, India*

Abstract: High dimensional dataset requires high processing time and memory. To efficiently execute high dimensional dataset dimensionality reduction technique is used. The dimensionality reduction technique removes unused and redundant attributes from a dataset. In dimensionality reduction, feature selection and feature extraction are two important techniques. Most of the existing work focuses on individual dimensionality reduction technique i.e. feature selection or feature extraction is studied and implemented independently. The combined study of these two techniques can create compound feature subset containing merged as well as original features in feature subset. There is need to bridge the gap between these two strategies. In this work, strategies for feature selection and feature extraction are studied with their advantages and limitations. A new technique is proposed by analysing the problems in existing system.

Keywords: Dimensionality reductions, Feature selection, Feature extraction, LDA, PCA, Correlation coefficient, Variance, Clustering, Classification.

I. INTRODUCTION

The growing use of computers and internet systems generates large amount of data every day. To store such bulk data is challenging task. The processing also requires high computational power. To overcome this challenge, the data is mined and important data is extracted from this data. The mining technique removes the noise and redundancy and preserves some important aspects and patterns.

The high dimensional data contains large number of attributes. Such multidimensional datasets are generally processed using classification or clustering techniques. The attributes may contain redundancy or noise. Such noisy attributes may hamper the classification or clustering accuracy whereas redundant attributes simply degrades the efficiency of algorithms.

To improve system efficiency dimensionality reduction is important technique in data mining domain. In this technique redundant and noisy attributes are removed from the dataset. This process generates subset of dataset based on selected important attributes from the dataset. The generated subset improves the classification algorithm accuracy and Normalized mutual information score in case of clustering process.

The dimensionality reduction is performed using following 2 ways:

A. Feature Selection

In feature selection process the important attributes in a dataset are filtered and new feature subset is generated. This selection process is executed in two way: It either selects one by one relevant important feature from the dataset and generates new feature subset or it removes one by one irrelevant unused feature from the dataset to reduce the dimensional space.

B. Feature Extraction

In feature extraction process two or more attributes are combined to generate new attribute. This technique generates such compound attributes based on some transformation techniques.

The feature selection and extraction algorithms are again classified in categories based on the nature of dataset such as:

1. Supervised approach**:** In supervised approach class attributes plays an important role. The attribute relevance is compared with the class attribute of dataset. The dataset containing class attribute are generally used for classification technique.
2. Unsupervised Approach: In this approach no class attribute is required. The relevance of attribute is compared with all other dataset attributes. The dataset without class attribute are generally used by clustering technique.

The Following section includes the detailed study of feature selection and feature extraction techniques with supervised and unsupervised approaches.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue III, Mar 2019- Available at www.ijraset.com

II. LITERATURE REVIEW

The feature selection technique is further categorized in three sections:

1) *Filter:* This is attribute selection method where attributes are selected based on some statistical techniques such as: correlation coefficient, mutual information, etc. The filter technique is independent of any machine learning algorithm. Prabitra Mitra proposes a unsupervised model technique based on similarity based feature selection. It uses maximum information compression index measure as statistical measure for evaluating attribute usefulness. This measure finds the similarity among attributes and removes the redundancy [2]. Q. Song, J. Ni, and G. Wang propose a new mechanism based on supervised model. It uses fast clustering-based feature selection algorithm. This algorithm generates attribute clusters using graph-theoretic clustering technique and then selects the most relevant attribute to the class label that represent the cluster of attribute [3].
2) *Wrapper:* This method follows the supervised learning model. It uses some machine learning algorithms. The classifier such as KNN, neive bays are used to evaluate inferences. It uses Forward Selection, Backward Elimination and Recursive Feature elimination techniques. The wrapper methods are time consuming and computationally expensive and hence not applicable in real world examples.
3) *Embedded:* Embedded method uses objective function or performance of learning algorithm along with intrinsic model building metric. This technique uses feature selection and wrapper strategy simultaneously. B. Efron, T. Hastie, I. Johnstone proposes a new technique named as Least Angle Regression (LARS). This technique proposes few modifications in LASSO regression technique. It is efficient one and follows less greedy selection technique than existing embedded systems [3].

A. Feature Extraction

This technique transforms the high dimensional dataset to low dimensional dataset by transforming the features linearly or non-linearly. For transformation this technique combines to or more features and generates new feature subset. The techniques are classified on the basis of nature of dataset. Linear Discriminant Analysis and principal Component analysis are two widely used feature extraction techniques[4].

Principal Component Analysis is a unsupervised feature extraction technique. It transforms the original features to uncorrelated and orthogonal principal components. This technique finds the eigenvectors of a covariance matrix for top k highest eigenvalues[5].

Linear Discriminant Analysis is a feature extraction technique. This technique is supervised technique. This technique reduces the dataset dimensions with good class-separability. This technique avoids the class overfitting problem and preserves the discriminatory information [6].

Sreevani and and C. A. Murthy proposes a new technique based on feature selection and extraction strategy. In the existing works these two techniques are studied independently. The proposed technique bridges the gap between these two techniques. These two techniques are applied simultaneously on the dataset to generate the new feature subset. The feature subset includes the original features as well as the transformed features [1].

III. PROPOSED METHODOLOGY

Combined Minimum Projection error Minimum Redundancy C-MPeMR framework is proposed. It uses feature extraction and feature selection technique to generate feature subset. The generated feature subset contains original as well as the transformed features. This technique is classified in 2 types supervised – SC-MPeMR and unsupervised and UC-MPeMR. In supervised method, the supervised feature extraction technique Linear Discriminant Analysis LDA is used, Whereas in unsupervised method Principal component analysis PCA technique is used for feature extraction. For feature selection Correlation coefficient and variance is used. The C-MPeMR technique uses feature extraction and selection one after other to maintain orthogonality among generated and existing features. Based on the technique classification or clustering is used for quality analysis of generated feature subset. For supervised technique classification accuracy of KNN is evaluated whereas for unsupervised dataset kmeans clustering is used. Using kmeans clustering results the Jaccard coefficient (Jacc), Fowlkes-Mallows index (FM) index are evaluated for clustering performance analysis.

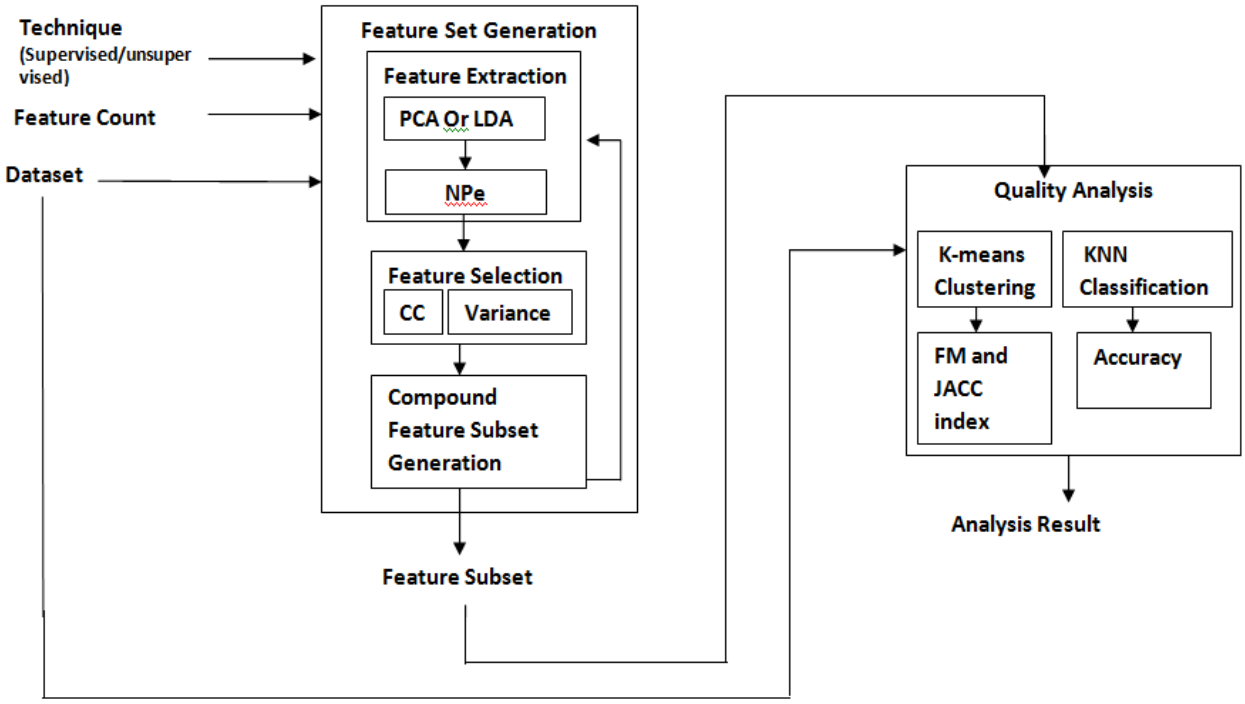


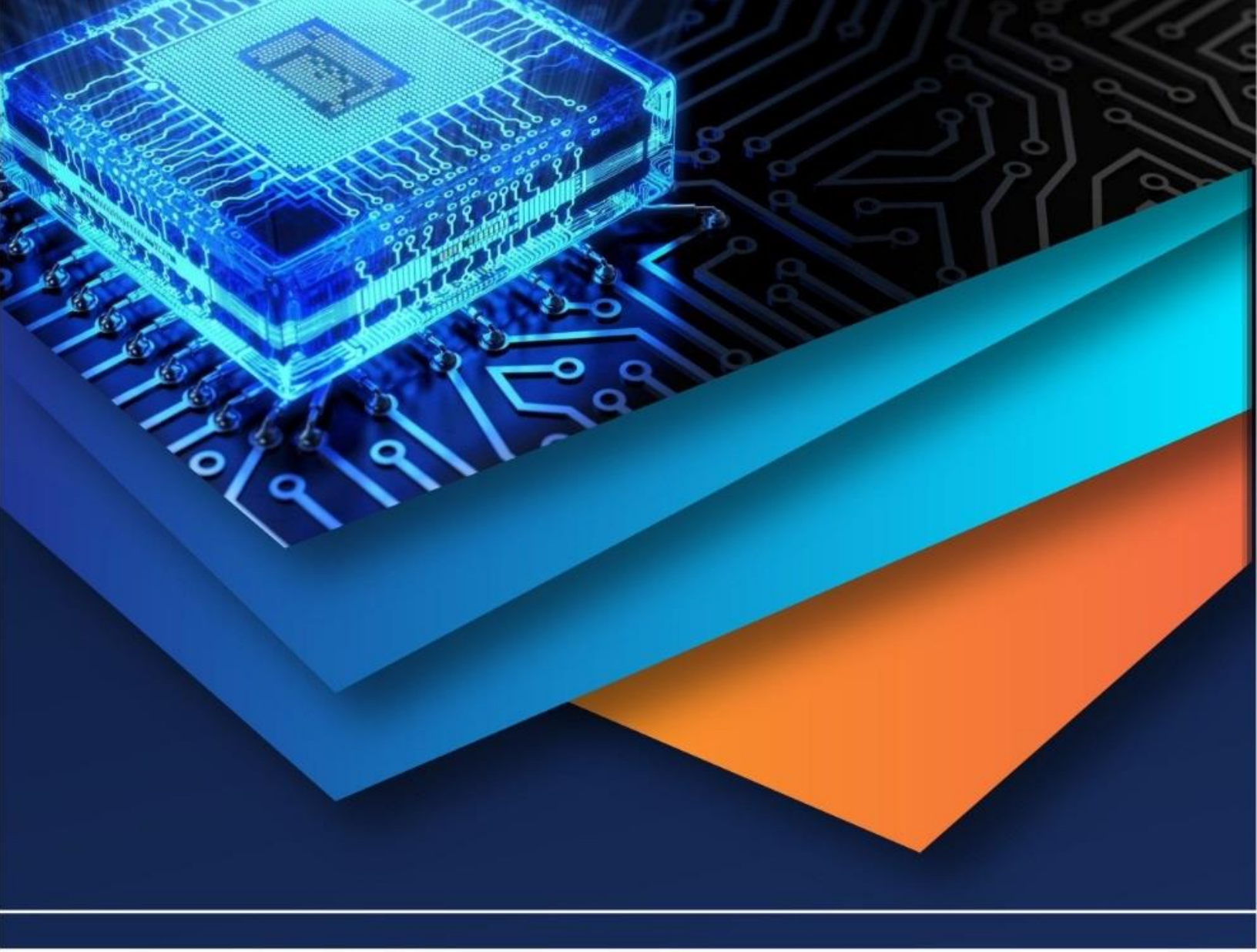
Fig.1: Proposed System Architecture

IV.CONCLUSIONS

In data analysis, dimensionality reduction is important technique for high dimensional dataset. For dimensionality reduction, Feature selection and feature extraction are two important techniques. In existing work these techniques are applied to the dataset independently. The generated feature subset contains either non-redundant original features or transformed features. By analysing problems in existing work, combined Minimum projection error with minimum redundancy C-MPeMR technique is proposed. In the proposed system, supervised and unsupervised approaches are handled. The technique uses feature selection and extraction simultaneously. The system executes the entire process without any feature selection or extraction pre-defined constant values. The generated feature set quality is measured using Jaccard coefficient (Jacc), Fowlkes-Mallows index (FM) index using k-means clustering for unsupervised mode whereas knn classification accuracy is evaluated for supervised mode.

REFERENCES

[1] Sreevani and C.A. Murthy, "Bridging Feature Selection and Extraction: Compound Feature Generation," IEEE Transactions on Knowledge and Data Engineering, Vol. 29, No. 4, pp. 757 - 770 , April 2017.
[2] P. Mitra, C. Murthy, and S. K. Pal, “Unsupervised feature selection using feature similarity,” IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 3, pp. 301–312, 2002
[3] Q. Song, J. Ni, and G. Wang, “A fast clustering-based feature subset selection algorithm for high-dimensional data,” Knowledge and Data Engineering, IEEE Transactions on, vol. 25, no. 1, pp. 1–14, 2013. B. Efron, T. Hastie, I.
[4] Johnstone, R. Tibshirani et al., “Least angle regression,” The Annals of statistics, vol. 32, no. 2, pp. 407–499, 2004.
[5] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” Journal of computational and graphical statistics, vol. 15, no. 2, pp. 265–286, 2006.
[6] P. A. Devijver and J. Kittler, Pattern Recognition: A Statistical Approach. London: Prentice-Hall International, Inc., 1982.
[7] K. Fukunaga, Introduction to statistical pattern recognition. Academic press, 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY