International Journal of Management, Technology And Engineering

ISSN NO : 2249-7455

Feature Grouping based on Supervised and Unsupervised Approach for Selection and Extraction Process

Kajal S. Mahale[1], Dr. Madan U. Kharat[2]

[1,2] Computer Engineering, MET's institute of engineering, Nashik, India

Abstract

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. In this proposed work, feature selection and feature extraction techniques are used in a single method named as: Modified-Minimum Projection error Minimum Redundancy technique- M-MPeMR. Based on the dataset information, M-MPeMR is executed in supervised and unsupervised mode. MPeMR technique combines 2 features at a time using feature extraction method. Feature extraction is applied using Linear Discriminant Analysis (LDA) technique for supervised mode whereas for unsupervised mode Principal component analysis (PCA) technique is used. For feature pair selection Normalized Projection error mechanism is used. Correlation coefficient and variance are used for feature selection process. The system executes the entire process of feature selection or extraction without any 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 and SVM classification accuracy is evaluated for supervised mode.

Keywords: *Classification, Clustering, Correlation coefficient, Dimensionality reductions, Feature selection, Feature extraction, Fowlkes-Mallows index, Jaccard coefficient, Variance.*

1. 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 data is generated mainly in neuroscience applications, Quantitative finance, Image processing, etc. To preserve such datasets, lower dimensional representation is generated by preserving maximally informative dimensions. Dimensionality reduction technique is used to reduce dataset dimension size. 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:

1.1 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 ways: it either selects one by one relevant important feature from the dataset or generates new feature subset or it removes one by one irrelevant unused feature from the dataset to reduce the dimensional space.

1.2 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.