2018-19

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.