Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 25 (2016) 208 – 215

Procedia Technology

Global Colloquium in Recent Advancement and Effectual Researches in Engineering, Science and Technology (RAEREST 2016)

An effective multi-clustering anonymization approach using discrete component task for non binary high dimensional data spaces

Arun Shalin L.V*,Dr.K.Prasadh[b]

* Research Scholor,ManonmaniumSundarnar University,Tirunelveli,India

[b] *Principal,Mookambika Institute of Technology,Kerala,India*

Abstract

Clustering in common is a process of grouping elements together, so that the elements assigned to the same cluster are more comparable to each other than the remaining data points. Certain difficulties related to dealing with high dimensional data are ubiquitous and abundant. Research works conducted using anonymization method for high dimensional data spaces failed to address the problem related to dimensionality reduction for non binary databases. In this paper, Discrete Component Task Specific Multi-Clustering (DCTSM) approach is presented for dimensionality reduction on non binary database. To start with the analysis of attribute in the non binary database takes place and the process of projecting clusters identifies sparseness degree of dimensions. Then with the quantum distribution on multi cluster dimension, the solution for relevancy of attribute and redundancy on non-binary data spaces is provided. As a result, dimensionality reduction on non binary data leads to performance improvement on the basis of tag based feature. Multi clustering tag based feature reduction extracts individual features and are correspondingly replaced by the equivalent feature clusters (i.e.) tag clusters. During training, the DCTSM approach, multi clusters are used instead of the individual tag features and then during decoding the individual features are replaced by the corresponding multi clusters. To measure the effectiveness of the method, experiments are conducted on existing anonymization method for high dimensional data spaces and compared with the DCTSM approach using Statlog German Credit Data Set. DCTSM approach obtained results of 7.05 % improved accuracy and was observed that it took minimal time during tag feature extraction and resulted in lesser error rate.

Peer-review under responsibility of the organizing committee of RAEREST 2016

*Keywords:*High-Dimensional Data Space;Non-Binary Database;Discrete Component Task Specific;Quantum Distribution; Dimensionality Reduction

Peer-review under responsibility of the organizing committee of RAEREST 2016
doi:10.1016/j.protcy.2016.08.099

1. Introduction

In the recent years, different types of clustering algorithms are introduced which are approximately separated into four types ranging from projection, hierarchical, density-based to subspace algorithms. The different types of algorithm as mentioned investigate the clusters in lower-dimensional projection of the original data. It is normally favored when dealing with information that is high dimensional. Motivated by the fact of high dimension, with the preference of observation that has more dimensions regularly leads to the so called curse of dimensionality, where the performance of many normal machine learning algorithms becomes impaired. This is frequently due to two pervasive effects such as the empty space incident and deliberation of distances.

The term 'curse of dimensionality' refers to the fact that all high dimensional data sets tend to be sparse, because the number of points necessary to symbolize any distribution grows exponentially with the number of dimensions. This results in bad density estimates for high-dimensional data, causing complexity for density based approaches on non- binary database. The latter is a rather counterintuitive property of high dimensional data point representations, where all distances between data points tend to turn out to be harder to differentiate with the increase in dimensionality which is omnipresent and copious.

Novel anonymization methods for sparse high-dimensional data [1] were based on estimated Nearest Neighbor (NN) search in high-dimensional spaces, which was evaluated using Locality Sensitive Hashing (LSH). The data transformation involved in it extracts the establishment using the underlying reduction into a band matrix and gray encoding-based sorting. These band matrixes and gray encoding made the establishment of anonymization in groups resulting in lesser information loss with the help of an efficient linear-time heuristic but problem related to non binary databases was not solved. Anonymization methods for sparse high-dimensional data do not use dimensionality reduction techniques for more effectual anonymization.

The idea of selecting subset of good features with high variance and feature subset selection are proved to be certain efficient methods for dimensionality reduction. With the selection of good feature, the irrelevant data are removed that increases the accuracy related to learning and maximizing the comprehensibility for non-binary database. The feature subset selection methods for non binary database are divided into four types namely, extensive category explicitly embedded, wrapper, filter, and hybrid techniques. The embedded methods integrate feature selection as a part of the training process and are usually precise, and therefore proved to be more efficient than the other three types.

On the basis of the aforementioned techniques and methods applied, the proposed work uses Discrete Component Task Specific Multi-Clustering (DCTSM) approach for non binary high dimensional data spaces to improve accuracy. DCTSM first clusters different types of pertinent attributes to recognize the constituent records of it. The problem of projected clustering is addressed by identifying the clusters and its appropriate attributes extracted from statlog german credit dataset. Subsequently, discrete dimensional projection clusters using the quantum distribution model are evolved that also considers the problem related to attribute relativity and redundancy.

DCTSM approach offers a multi cluster formation based on objective function and evolve a discrete dimensional projection clusters. Finally, dimensionality of the data is reduced by ignoring the lower Eigen value components. On the other hand, DCTSM approach uses the class information to perform a projection of the features which best separate two or more classes.

Experiments using datasets Statlog German Credit Data Set extracted from UCI repository confirm that, the DCTSM approach facilitates higher level of accuracy and also the multi-clustering process is considerably efficient. Empirical studies show that the adoption of the DCTSM approach improves the level of accuracy and minimizing the error rate compared to the significantly more efficient state of art technique. The contribution of Discrete Component Task Specific Multi-Clustering (DCTSM) approach on non binary database for dimensionality reduction mining includes the following:

(1) To identify sparseness degree of dimensions using Discrete Component Task Specific Multi-Clustering (DCTSM) approach
(2) To provide solution provide solution for relevancy of attribute and redundancy on non-binary data spaces