Chapter 3
Subspace Clustering of High Dimensional Data Using Differential Evolution

Parul Agarwal
Jaypee Institute of Information Technology, India

Shikha Mehta
Jaypee Institute of Information Technology, India

ABSTRACT

Subspace clustering approaches cluster high dimensional data in different subspaces. It means grouping the data with different relevant subsets of dimensions. This technique has become very effective as a distance measure becomes ineffective in a high dimensional space. This chapter presents a novel evolutionary approach to a bottom up subspace clustering SUBSPACE_DE which is scalable to high dimensional data. SUBSPACE_DE uses a self-adaptive DBSCAN algorithm to perform clustering in data instances of each attribute and maximal subspaces. Self-adaptive DBSCAN clustering algorithms accept input from differential evolution algorithms. The proposed SUBSPACE_DE algorithm is tested on 14 datasets, both real and synthetic. It is compared with 11 existing subspace clustering algorithms. Evaluation metrics such as F1_Measure and accuracy are used. Performance analysis of the proposed algorithms is considerably better on a success rate ratio ranking in both accuracy and F1_Measure. SUBSPACE_DE also has potential scalability on high dimensional datasets.

1. INTRODUCTION

Clustering is one of the vital approaches in the field of data mining. It forms the groups of similar data on basis of certain properties. The most common property is distance measure. The criteria for assembling the similar datasets into one group and dissimilar in other groups vary from algorithm to algorithm. In today's world clustering is being used in number of fields like engineering, medicines, marketing, economics etc. In engineering field (Hans-Peter Kriegel, Kröger, & Zimek, 2009), clustering plays an important role in artificial intelligence, spatial database analysis, web mining, computer vision, pattern

DOI: 10.4018/978-1-5225-5852-1.ch003

recognition, face recognition, machine learning (Ira Assent, 2012) and many more. It also has its application in mechanical engineering, electrical engineering, medical sciences like microbiology, genetics, pathology etc. There are variety of clustering algorithms (Fahad et al., 2014) like partition based (K-Means, K-Mediods, K-Modes, CLARA, PAM, fuzzy c means), density based (DBSCAN, OPTICS, DENCLUE), hierarchical based (BIRCH, CURE, ROCK), grid based (STING, OPTIGRID, CLIQUE) and model based (EM, COBWEB).

The traditional clustering algorithms clusters data in full dimensional space i.e. considering all attributes while clustering. However, when number of attributes increases i.e. dimensions of data are amplified then traditional algorithms fails to give meaningful clusters. The reason behind this failure is that data becomes sparse in high dimensional space and distance measure becomes meaningless. This problem is coined as curse of dimensionality (Ira Assent, 2012). Traditional clustering algorithm breakdown when implemented on high dimensions. In high dimensional data, it is possible that there exists number of clusters for which only few dimensions are relevant instead of all dimensions. The subsets of dimensions are called subspaces. Clustering in high dimension is possible through subspaces and is called subspace clustering. However, there are number of challenges in subspace clustering (Parsons et al., 2004) like huge combinations of dimensions, overlapping subspaces etc. Due to these challenges, there has been scope of improvement in these algorithms. Subspace clustering not only determines the clusters in dataset but also the subspaces in which these clusters are present. There are two main search methods of subspace clustering: top down and bottom up search methods (Parsons et al., 2004).

Top down approach of subspace clustering method searches the subspaces and clusters in descending fashion. It follows three phases of clustering i.e. initialization phase, iteration phase and refinement of clusters formed in iterations. The input parameter for this approach is size of subspace and number of clusters. Algorithms of top down approach start from initializing all dimensions with same weight. Clustering in initialization phase is performed in full dimensional space. When clusters are formed, each dimension is assigned new weights for each cluster. The next iteration starts by considering the new weights of dimensions for clustering. This is an exhaustive process as it requires multiple iterations for best results. The one way to implement this approach is use of sampling. Sampling can improve the performance of top down subspace clustering algorithms. However, there are number of drawbacks in this approach. Top down approach is based on only partitioning of dataset that means each instance will belong to only one cluster. Overlapping subspaces could not be determined. The input parameters requires in top down approach are hard to determine prior of clustering. For unlabelled dataset or for unknown number of clusters, this approach is not suitable. Additionally, determining size of subspace priory is an intricate problem. If sampling strategy is used, then size of sample is also a critical parameter that should be known before clustering. Some algorithms of top down approach are PROCLUS, ORCLUS, FINDIT and COSA (Parsons et al., 2004).

Bottom up approach of subspace search method finds the subspaces and clusters in ascending fashion of dimensions. It uses APRORI principle to reduce the search space for subspace clustering. The algorithm starts clustering at lower dimensions. It determines the dense points at small dimensions and then move towards larger subspaces. The downward closure property is followed. If the point is dense at D dimensions that it is dense at (D-1) projections of dimensions. The algorithm based on bottom up approach terminates unless no dense unit is discovered. The advantage of bottom up approach is that it is capable of detecting overlapping subspaces. This implies that a data point or instance can be a part of