A Comprehensive Review of Subspace Clustering in the Analysis of Big Data

T. Gayathri\(^1\), D. LalithaBhaskari\(^2\)

\(^1\)Assistant Professor, Department of CSE, Shri Vishnu engineering college for women, Bhimavaram, Andhra Pradesh, India
\(^2\)Professor, Department Of CS&SE, College of Engineering, AUCE(A), Andhra University, Visakhapatnam, Andhra Pradesh, India.

Abstract - In this age of information technology, statistics holds the key to development in different paradigms. The knowledge has to be extracted from huge treasure trove of database. With advent of superior computing platforms, increased storage spaces and shift towards e-platforms have generated huge volumes of data that has to be analyzed and interpreted. This advent of Big data has thrown in challenges that have to be surmounted through improved clustering and analysis approaches. In this work a detailed review about Big data, its attributes and significance of Subspace clustering algorithm are presented. The primary contribution of this paper is in sifting through literature and bringing about a detailed presentation about how different authors have classified Subspace clustering approaches. Important algorithms that can serve as a benchmark for any future development have also been explained briefly.

Keywords — Big Data, Clustering, Subspace clustering, Classification.

I.INTRODUCTION

Data is one of the most cherished entities in this age of communication and ubiquitous computing. It is the collection of variable and values that are related by a certain degree and in certain other cases differ by a certain degree. With the explosion in size of storage devices the size of the database has also increased dramatically. With the penetration of pervasive computing platforms in the form of smart phones, huge treasure trove data is collected and stored inadvertently. This data can provide valuable insight in terms of its content which has led to the growth of tools that automatically extract the knowledge from the data [1]. A database can be termed as an organized collection of data enabling easy manage, access and update. Data mining can is the process of discovering interested knowledge. This includes extraction of information about associated patterns, anomalies and significant structures that can be interpreted different forms of information repositories like databases and data warehouses. Data mining (DM) is also called as Knowledge Discovery in Databases (KDD) or Knowledge. It involves application of many computational techniques from statistics, information retrieval, machine learning and pattern recognition. The main aim of Data mining is to extract only the required patterns from the database in a short time span. Depending on the type of patterns that has to be mined, data mining tasks can be classified into summarization, classification, clustering, association, and trends analysis [2].

Enormous amount of data are generated every minute and in fact every second, so improved methods of analysis are required to extract information that is of best interest to the user. In this context, Big data refers to rapidly growing datasets with sizes that are beyond the capability of traditional data base tools for storing, managing and analyzing the data. Big data can be termed as a heterogeneous collection of both structured and unstructured data. Some of the reasons that can be attributed to the tremendous growth in big data include availability of data, increased storage capability and the exponential increase in the processing power of the computing platforms. Big data involves the use of large data sets to help in the collection or reporting of data that can aid in the decision making process. The data may be enterprise specific or general and private or public [3]. Big Data mining refers to the process of going through big data sets in an effort to look for and extract relevant information. Big data samples are available in different streams and domains like astronomy, atmospheric science, social networking sites, life sciences, medical science, government data, natural disaster and resource management, web logs, mobile phones, sensor networks, scientific research, telecommunications etc. [4]. Clustering is one of the most important tools in the quest to analyze and interpret Big data. Owing to its tremendous applications in summarization, learning, segmentation, and target marketing [5, 6, 7] the problem of data clustering has been investigated widely in the literature. In the absence of specific
labeled information, clustering can be considered as a concise model of the data which can be interpreted in the sense of either a summary or a generative model. The basic problem of clustering may be stated as ‘Given a set of data points, partition them into a set of groups which are as similar as possible’ [7].

With the advent of Big data and as the datasets become larger and more varied, it’s imperative to make adaptations to existing algorithms to maintain cluster quality and speed. In the case traditional clustering algorithms, they consider all of the dimensions of an input dataset. This is employed to enhance the possibility of learning as much information from each of the object described. In high dimensional data in circumstances many dimensions are often irrelevant. These irrelevant dimensions that may possibly hiding in a noisy data can result in erroneous clusters being generated by the clustering algorithms. In very high dimensions it is common for all of the objects in a dataset to be nearly equidistant from each other, completely masking the clusters. In order to improve the quality of the clusters feature selection methods have been employed successfully to a certain extent. An efficient way of clustering high dimensional data prevailing in Big data is to employ Subspace [6] clustering approach. Unlike feature selection methods which examine the dataset as a whole, subspace clustering algorithms localize their search and are able to uncover clusters that remain in multiple and overlapping subspaces.

This work explores the varied and different definitions of Big data and the specific issues that can be attributed to Big data analysis. These includes the attributes that are used to define Big Data and their significance. The primary objective of this paper is to shed light on different types of Subspace clustering techniques available in the literature. It’s widely accepted in the research community that like multiple definitions for Big data, there are numerous classification of Subspace clustering algorithm. This paper dwells in to methods of classifications employed by different researchers over a period of time in classifying and categorizing Subspace clustering methods. This kind of understanding is very essential to develop any new techniques that can help in the analysis and interpretation of Big Data. A brief explanation about important Subspace algorithms that can serve as a benchmark for any further development of algorithms is also given. After introduction in section 1, different definitions of Big data is briefed in section 2 followed by the challenges experienced by the high dimensional data clustering in section 3. A review of literature about classification of Subspace clustering approaches is explained in section 4 and discussion about important existing approaches is presented in section 5. The conclusions and references used are presented in section 6 and section 7 respectively.

II.DEFINITION OF BIG DATA

There are multiple approaches and definitions available for defining Big data, obviously size is the first characteristic that frames the question “what is big data?” But in recent years other characteristics of big data have emerged. Laney [8] suggested that Volume, Variety, and Velocity (or the ThreeV’s) are the three dimensions of challenges that have to be encountered in data management. The Three V’s have emerged to define the framework of big data [9, 10]. May leading institution and corporations have played a key role in defining the attributes of Big data. Gartner, Inc. defined Big data as: “Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making [11]. Likewise Tech America Foundation defined big data as “Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.” [12].It can be inferred from these definitions that volume, variety and velocity are the key components of Big data. These three important attributes define and determines the course of strategies that can be employed for analysis of big data.

Volume refers to the magnitude of data. Big data sizes are reported in multiple terabytes and petabytes. One petabyte equals 1024 terabytes. Definitions of big data volumes are relative and vary by factors, such as time and the type of data. What can be construed as big data today may not meet the criteria for future. It is natural that in the future because storage capacities will increase, allowing even bigger data sets to be captured. It’s also important to understand that variety also plays a key role in defining how big a data is. The type of data brings about a new challenge in defining the size. Two different datasets of similar size may require completely different management approaches and thus the ensuing issue how to deem a data is big or small. As pointed earlier industry being a key contributor in the development of Big data the type of industry also plays a significant role in defining the thresholds for Big data volume.
Variety refers to the structural heterogeneity in a dataset. The advent of technology has allowed the use of various types of structured, semi-structured, and unstructured data. Structured data, which constitutes only 5% of all existing data [13] primarily, refers to the tabular data found in spreadsheets or relational databases. The bulk of unstructured data include text, images, audio, video etc. At times these data types lack the structural organization that is very much essential for automated machine analysis. The format of semi-structured data does not conform to strict standards. Extensible Markup Language (XML), a textual language for exchanging data on the aWeb, is a typical example of semi-structured data. XML documents contain user-defined data tags which make them machine-readable.

Velocity subjectively refers to the rate at which data are generated and the speed at which it has to be processed, analyzed and interpreted. The need to have real time analytics and situation based planning coupled with the proliferation of digital devices has led to an unprecedented rate of data creation. Even conventional retailers are generating high-frequency data. Wal-Mart, for instance, processes more than one million transactions per hour [13]. The data emanating from mobile devices and flowing through mobile apps produces torrents of information that are currently used to generate real-time, personalized offers for everyday customers.

In addition to the three V’s, other dimensions of big data have also been mentioned in the literature. These are specifically coined by IT giants who play a stellar role in defining the generation, need and use of Bigdata. Veracity, the fourth V coined by IBM represents the unreliability inherent in some sources of data. This is to represent uncertainties in natural decision making process and issues like human judgment. This type of uncertainty is bound to be experienced in most of the data generated by social media. In spite of these uncertainties they contain valuable information, so it is imperative to deal with imprecise and uncertain data. This is another attribute of big data, that has to be addressed using tools and analytics that are deployed for management and mining of uncertain data.

Variability (and complexity) was introduced SAS as two additional dimensions of big data. Variability refers to the variation in the data flow rates as often it can be observed that big data velocity is not consistent and has periodic peaks and troughs. Complexity refers to the fact that big data are generated through a myriad of sources. This imposes a critical challenge: the need to connect, match, cleanse and transform data received from different sources. Value introduced Oracle is used as a one of the defining attribute of big data. Based on Oracle’s definition, big data are often characterized by relatively ‘low value density’. That is, the data received in the original form usually has a low value relative to its volume. How-ever, a high value can be obtained by analyzing large volumes of such data.

III. HIGHDIMENSIONAL CLUSTERING CHALLENGES

There are 3 primary challenges to be surmounted in high dimensional data clustering, they are:

Curse of Dimensionality or Sparse Data: The Curse of Dimensionality, a term initially introduced by Richard Bellman [14], is a phenomenon that arises when applying machine learning algorithms to highly-dimensional data. The inability of clustering algorithms to cope with high dimensional data is referred to as curse of dimensionality. Figure (1) illustrates the curse of dimensionality and it can be inferred that there the number of regions grow exponentially with the increase in number of dimensions.

Fig. 1. Curse of Dimensionality

This can be attributed to the fact that as the dimension of dataset increases the relevance of distance measures between the data points decreases. The increased addition of dimensions results in spread of data to such an extent that in very high dimensions, they almost remain equidistant from each other. The problem is amplified when objects are related in different ways in different subsets of dimensions. Subspace clustering algorithm tends to address and uncover such type of relationship. The irrelevant features must be removed so that the clustering algorithm focuses on only the relevant dimensions. In the case of clusters found in lower dimensional space, they are easily interpretable and allow the user to have better study of the cluster for further application.
numbers of irrelevant attributes are present in the data these methods are rendered irrelevant as they preserve the distance between the objects. It will also be difficult to interpret the new features as they may be combination of originals. The feature selection methods target the most relevant dimensions from a dataset and reveal the group of objects that are similar. Even though this approach is effective with many types of database, it experience difficulty in identifying clusters that are found in different subspaces. Subspace algorithms are exactly capable of handling this type of data; these algorithms take the concepts of feature selection to the next level through selection of each cluster separately for relevant subspaces. Subspace clustering can be considered as an extension of feature selection that attempts to find clusters in different subspaces of the same dataset. Similar to feature selection, subspace clustering requires a search method and evaluation criteria. Subspace clustering should also limit the scope of the evaluation criteria so as to consider different subspaces for different clusters.

Existing subspace clustering approaches can be categorized under different schemes with the two major types of subspace clustering as top down and bottom up. Based on the search strategy were specified by P. Lance et al. [17]. The top down approach is further classified as per cluster weighting methods and per instance weighting methods. The authors have employed grid based classification, where the clusters are further sub classified based on the size of the grid as static and adaptive grid approaches. The bottom up approaches was not clearly classified as grid based or density based. Ilango et al [18] presented a classification where high dimensional clustering approaches were classified as partitioning approaches, hierarchical approaches, density based approaches, and grid based approaches and model based approach.

Karlton S. et al [19] classified subspace clustering into two categories like density based clustering and projected clustering. The authors stated that density based clustering approaches such as CLIQUE (Clustering In QUEst) [15], MAFIA (Merging Adaptive Finite Intervals And is more than a clique) [20], SUBCLU (density connected Subspace CLustering) [21] have their clustering rooted on density of data. Similarly projected clustering is observed in approaches such as PROCLUS (PROjectedCLUSTERing) [22], CLARANS [23], ORCLUS (arbitrarily Oriented projected CLUSTERing) [24], DOC (Density based Optimal projective Clustering) [25] etc.
H.P. Kriegel et al [26] high dimensional data clustering approaches based on the orientation of the data as subspace clustering based on axis parallel clustering or correlation clustering based on arbitrarily oriented clustering and pattern based clustering. The objective of the correlation based approaches is to find the clusters that may exist in arbitrarily oriented subspaces, one such example is ORCLUS [24]. A pattern based clustering approach p-Cluster [27] groups those objects exhibiting similar trends in a subset of attributes.

Problem oriented classification of Axis parallel subspace clustering approach results in classification like projected clustering, soft projected clustering and hybrid algorithms. PreDeCon [28] (subspace PREference weighted Density CONnected clustering) an example of projected clustering approach employs identification of unique assignment of each object to exactly one subspace cluster or noise. For example in the case of soft projected clustering algorithms, the number ‘k’ of clusters is known in advance and an objective function is formulated to optimize the generation of k-number of clusters. An example for soft projected clustering algorithm is COSA (Clustering Objects on Subsets of Attributes) [30]. Another Subspace clustering algorithm SUBCLU [31] intend to find all subspaces where clusters can be identified.

The classification of hybrid algorithms refers to those algorithms that are capable of finding overlapping clusters and one example for such an algorithm is FIRES (FIltter REfinement Subspace clustering) [26]. Another classification subspace clustering algorithm based on parameterization of results is presented in [32]. The approaches were classified in to cell based, density based and clustering oriented approaches. CLIQUE [15] an example for cell based approaches look to identify r sets of fixed or variable grid cell that contain more than a certain threshold objects. Similarly Density based approach defines clusters as dense regions separated by sparse regions like in SUBCLU [31]. Similarly PROCLUS [22] a clustering oriented approach exhibit the complete properties of the entire set of clusters, these include the properties like the number of clusters, the average dimensionality and the properties that are statistically oriented etc..

V. IMPORTANT SUBSPACE CLUSTERING ALGORITHMS

This section dwells briefly about some important subspace clustering algorithms available in the literature. These algorithms were instrumental in the evolution of new approaches. It can be clearly observed that most of the algorithms are adaptations and modification of one of the existing approaches.

CLIQUE[15] one of the first algorithms algorithm combines density and grid based clustering and uses an APRIORI style to find clusters within subspaces of the dataset. This approach employs coverage, which is defined as the fraction of the dataset covered by the dense units in the subspace to identify the clusters. Once the dense subspaces are found they are sorted by coverage. The subspaces with the highest coverage are retained while the rest are removed. The algorithm then moves on to identify adjacent dense grid units in each of the selected subspaces using a depth first search approach. Clusters are then formed by combining these units using a greedy growth scheme. Initially the algorithm starts by having an arbitrary dense unit and then it greedily grows a maximal region. This growth is done in each dimension until the union of all the regions covers the entire cluster. The smallest redundant regions are removed until no further maximal regions can be removed through a repeated procedure. The hyper-rectangular clusters are then defined using a Disjunctive Normal Form (DNF) expression. CLIQUE is capable of finding many types and shapes of clusters and can present them in easily interpretable ways. The region growing density based approach of generating clusters helps CLIQUE to find clusters of arbitrary shape. CLIQUE is also equipped to find any number of clusters in any number of dimensions and the number is not predefined and limited by a parameter. Clusters may be found in the same, overlapping, or disjoint subspaces. The DNF expressions used to represent clusters are often very interpretable. This is advantageous in subspace clustering since the clusters often exist in different subspaces and thus represent different relationships. Like other bottom-up algorithms, CLIQUE scales well with the number of instances and dimensions in the dataset. CLIQUE (and other similar algorithms), however, do not scale well with number of dimensions in the output clusters. This is not usually a major issue, since subspace clustering tends to be used to find low dimensional clusters in high dimensional data.

ENCLUS [33] is a subspace clustering method based heavily on the CLIQUE algorithm. One primary difference between ENCLUS and CLIQUE is that it does not measure density or coverage directly, but instead it measures entropy. A subspace with clusters usually exhibit lower entropy when compared to a subspace without clusters. This forms the underlying
principle behind the operation of ENCLUS. Clusterability of a subspace is defined using three criteria: coverage, density, and correlation. Entropy can be used to measure all three of these criteria. It is also important to note that the entropy decreases as the density of cells increases. But under certain conditions, entropy can also decrease with the increase in coverage. Similarly Interest is a measure of correlation and is defined as the difference between the sum of entropy measurements for a set of dimensions and the entropy of the multi-dimension distribution. Larger values indicate higher correlation between dimensions and an interest value of zero indicates independent dimensions. ENCLUS employs the same APRIORI style, bottom-up approach as CLIQUE to mine significant subspaces. Pruning is accomplished using the downward closure property of entropy and the upward closure property of interest (i.e. correlation) to find minimally correlated subspaces.

PROCLUS

PROCLUS [22] was the first top-down subspace clustering algorithm. Similar to CLARANS [34], PROCLUS operates by sampling the data to select a set of k medoids and iteratively improving the results of the clustering. The algorithm uses a three phase approach consisting of initialization, iteration, and cluster refinement. A greedy algorithm is used for initialization to select a set of potential medoids that are far apart from each other. This ensures that each cluster is identified by at least one instance in the selected set. In the iteration phase a random set of k medoids are selected from this reduced dataset. The cluster refinement phase replaces bad medoids with randomly chosen new medoids, and determines if clustering has improved. Cluster quality is based on the average distance between instances and the nearest medoid. For each medoid, a set of dimensions is chosen whose average distances are small compared to statistical expectation. The total number of dimensions associated to medoids is calculated as 

\[ k^*l \]

where \( l \) is an input which defines the average dimensionality of cluster subspaces. Once the subspaces have been selected for each medoid, average Manhattan segmental distance is used to assign points to medoids, forming clusters. The medoid of the cluster with the least number of points is thrown out along with any medoids associated with fewer than \( (N/k) \times \text{min Deviation points} \), where \( \text{min Deviation} \) is an input parameter. The refinement phase computes new dimensions for each medoid based on the clusters formed and reassigns points to medoids, removing outliers. Like many top-down methods, PROCLUS is biased toward clusters that are hyper-spherical in shape. Also, while clusters may be found in different subspaces, the subspaces should be of same size as the user has to input the average number of dimensions for the clusters. Clusters are represented as sets of instances with associated medoids and subspaces and form non-overlapping partitions of the dataset with possible outliers.

DBSCAN [35], density-based clustering algorithms try to find clusters based on density of data points in a region. The primary idea behind density-based clustering is that foreach instance of a cluster the neighborhood of a given radius (Eps) should have a minnummum number of instances (MinPts). One of the most wellknown density-based clustering algorithms is the DBSCAN [35].

DBSCAN separate data points into three classes

Core points: These are points that are in the interior of a cluster. A point is considered to be interior if enough points are present in its neighborhood.

Border points: Border point is a point that is not a core point. In this case, a point does not have enough points in its neighborhood. It falls within the neighborhood of a core point.

Noise points: A noise point is one which does not belong to either core point or the border point.

In order to find a cluster, DBSCAN starts by choosing an arbitrary instance (p) in data set (D) and identifies all instances of D with respect to Eps and MinPts. The algorithm employs spatial data structure - R*tree [36] to locate points within Eps distance from the core point of the clusters. A modified version of DBSCAN, an incremental DBSCAN is presented in [10] and was proved to yield the same result as DBSCAN. In addition, another clustering algorithm (GDBSCAN) generalizing the density-based algorithm DBSCAN is presented in [37].

SUBCLU (density-connected SUB space CL Utering) [21] is the first density based subspace clustering approach which extended the concept of DBSCAN for high dimensional data. It employs greedy algorithm to detect the density connected clusters in all subspaces of high dimensional data and uses monotonicity property to move higher dimensional projections thereby ensuring the reduction of search space to a large extent. It avoids the limitations of grid-based approaches like the dependence on the positioning of the grids or fixed shape of clusters. The algorithm initially generates all the 1-dimensional clusters using input parameters, \( \mu \) - density threshold and \( \_ \)-distance (radius), by using DBSCAN [35] to each 1-dimensional subspace. It then checks for every k-dim cluster to included or pruned. The clusters are then
generated by the application of DBSCAN on each (k+1)dimensional candidate subspace. The steps are repeatedly executed as long as the set of kdimensional subspaces containing clusters is not empty.

INSCY (INdексing Subspace Clusters with in-process-removal of redundancy) [38] is another efficient subspace clustering algorithm which is based on the subspace clustering notion of [39]. It employs a depth first approach to mine recursively in a region of all clusters in all subspace projections and then further continue with the next region. It evaluates the maximal high dimensional projection first by rapidly pruning all its redundant low dimensional projections. This approach enhances the efficiency as it overcomes the drawbacks of breadth first subspace clustering and it reduces the runtimes substantially. It also allows indexing of promising subspace cluster regions. INSCY proposes a novel index structure SCYtree that has a compact representation which allows arbitrary access to subspaces. SCY-tree uses in-process redundancy pruning to have an effective subspace clustering.

 Scalable Density Based Subspace Clustering [40] reduces subspace processing by identifying and clustering promising subspaces and mines few selected subspace clusters only. It operates on the principle that any high dimensional subspace cluster will appear in many low dimensional projections. By mining only some of them, the algorithm avoids the in between subspaces by gathering enough information into jump directly to the more interesting high dimensional subspace. Database scans are completely avoided with this kind of approach for many intermediate, redundant subspace projections, steering the process of subspace clustering. It uses priority queue to initialize the information of density estimates. It gives a basis for selecting the best candidate from the priority queue. The priority queue is split into three levels for multiple density granularities. It skips intermediate subspaces in a best first manner and jumps directly to high dimensional subspaces.

VI. CONCLUSION

Big data is becoming the mainstay of many businesses and domains from social media marketing to genome mapping big data analysis has tremendous scope and challenges. Subspace clustering is one of the crucial methods that can aid in the analysis of the Big data. This paper presents an overview about Big data and issues that are specific to it. In order to understand clearly the different types of Subspace algorithms and its classification, a review of literature is presented about different types and classification of Subspace clustering approaches. A brief presentation about important Subspace clustering approaches has also been discussed. This paper can help the researchers in understanding the Big data, its challenges and existing approaches and aid in the design of novel clustering approaches for Big data.

REFERENCES


