Salud, Ciencia y Tecnología. 2023; 3:413
doi: 10.56294/saludcyt2023413

REVIEW

Big Data and Different Subspace Clustering Approaches: From social media promotion to genome mapping

Big Data y diferentes enfoques de clustering subespacial: De la promoción en redes sociales al mapeo genómico

Vijaya Kishore Veparala[1], Vattikunta Kalpana[1]

[1]Department of ECE, Mohan Babu University, Tirupati, A.P, India.

Cite as: Kishore Veparala V, Kalpana V. Big Data and Different Subspace Clustering Approaches: From social media promotion to genome mapping. Salud, Ciencia y Tecnología. 2023;3:413. https://doi.org/10.56294/saludcyt2023413

Received: 27-04-2023 **Reviewed:** 13-05-2023 **Accepted:** 19-06-2023 **Published:** 20-06-2023

Editor: Fasi Ahamad Shaik

ABSTRACT

In the present age of information technology, information is the most important factor in determining how different paradigms will progress. This information needs to be mined out of a massive computer treasure trove. The rise in the amount of data been analyzed and interpreted is a direct result of the proliferation of more powerful processing platforms, the increase in the amount of storage space available, and the transition toward the use of electronic platforms. A thorough study of Big Data, its characteristics, and the role that Subspace clustering algorithm plays is described in this work. The most important contribution that this paper makes is that it reads a lot of previous research and then makes a thorough presentation about the different ways that other authors have classified subspace clustering methods. In addition, significant algorithms that are capable of acting as a benchmark for any future development have been provided with a short explanation.

Keywords: Big Data; Clustering; Subspace; Classification; Integrative Review.

RESUMEN

En la era actual de las tecnologías de la información, la información es el factor más importante para determinar cómo progresarán los distintos paradigmas. Esta información debe extraerse de un enorme tesoro informático. El aumento de la cantidad de datos analizados e interpretados es consecuencia directa de la proliferación de plataformas de procesamiento más potentes, el incremento del espacio de almacenamiento disponible y la transición hacia el uso de plataformas electrónicas. En este trabajo se describe un estudio exhaustivo de Big Data, sus características y el papel que desempeña el algoritmo de clustering Subspace. La contribución más importante que hace este trabajo es que lee muchas investigaciones anteriores y luego hace una presentación exhaustiva sobre las diferentes formas en que otros autores han clasificado los métodos de clustering subespacial. Además, se han proporcionado, con una breve explicación, algoritmos significativos que pueden servir de referencia para cualquier desarrollo futuro.

Palabras clave: Big Data; Clustering; Subespacio; Clasificación; Revisión Integradora.

INTRODUCTION

In this era of constant communication and computing, data is widely recognized as one of the most valuable forms of property. One way to refer to it is as a collection of variables and values that, in some instances, are similar to one another to a certain degree and, in other cases, are dissimilar to one another to a certain degree. The database's size has significantly grown in tandem with the exponential growth in the capacity of recording devices. The spread of computing platforms in the form of smart phones has resulted in the accidental collection and storage of a vast treasury of data. Because the information contained in this data can be useful, there are now more instruments available than ever before that are able to rapidly extract useful information from large amounts of data. This is a direct consequence of the previous sentence.[1] One definition of a database describes it as an organized collection of data that is simple to manage, acquire, and keep updated on, while another defines data mining as the process of finding information that is relevant to the user and of relevance to them. This knowledge includes the ability to extract from large quantities of data details about related patterns, anomalies, and important structures that are stored in computerized formats such as data warehouses, databases, and other types of information archives. Therefore, data mining (DM), which is referred to as Knowledge Discovery in Databases (KDD) or Knowledge Discovery and Data Mining, is a method of looking through enormous amounts of data in an automatic fashion for structures that involve association rules.[2] It involves employing a large number of different computational methods, such as pattern identification, information extraction, machine learning, and statistics, amongst others. The primary objective of data mining is to quickly select only the essential patterns from a database while minimizing the amount of time spent doing so. Data mining activities can be broken down into the categories of summarization, classification, clustering, association, and trend analysis,[2] varying according to the kinds of patterns that are desired to be extracted.

Owing to the fact that the quantity of data is growing at an exponential pace, improved evaluation is needed to extract data that most closely corresponds to the interests of individual users. Enormous amounts of data are created each minute and in fact each second. In the context of this discussion, the term “big data” refers to datasets which expand at an alarming rate and are of a size that exceeds the capacity of traditional database tools that can be used to store, organize, and conduct analysis on the data. The phrase “big data” refers to a compilation of data that contains information that is both structured and unstructured. Some of the factors that can be attributed to the tremendous development of big data include accessibility of information, a rise in the capability of storage, and a rapid rise in the processing capacity of computing platforms. These are just some of the factors that have contributed to the rise in popularity of big data. Big data refers to the practice of utilizing large data sets in order to efficiently handle gathering or publishing of information that assists businesses or other individuals in making decisions. The information could be specific to the business, or it could be general. It could also be private or open to the public.[3]

The process of finding by means of enormous amounts of data for information that is relevant to a particular question or problem and then collecting that information is known as "big data mining. “Big data samples are used in many areas, including astronomy, atmospheric science, social networking sites, life sciences, medical science, government data, natural disasters and resource management, web logs, mobile phones, sensor networks, scientific research, telecommunications, and more.[4] In the endeavor to analyze and understand big data clustering makes up one of the most important things you can have. Research has paid a lot of attention to the topic of data arrangement that has been done on it as a result of the enormous number of applications it has in summarizing, learning, dividing into groups, and selling to specific groups.[5,6,7] Clustering can be thought of as a simplified model of the data when no specific information is given. Depending on the situation, this model can be seen as either a recap or a way to come up with new ideas. Given a set of data points, the main task of clustering is to "split them up into groups those are as identical to each other as is humanly possible".[7]

It is imperative to make modifications to preexisting algorithms to maintain the quality and speed of clusters in light of the emergence of big data and the fact that datasets are growing both in size and in variety. When it comes to traditional clustering algorithms, each and every dimension of a data collection is examined and considered. This is done to increase the likelihood of gleaning as much information as possible from each of the objects being mentioned. However, in high-dimensional data, many of the dimensions are frequently extraneous to the problem. These unimportant dimensions when you try to conceal clusters by using data with a lot of noise, it can cause the clustering algorithms to produce incorrect clusters. When there are a lot of variables in a dataset, it is not uncommon for all the objects to be nearly located at the same distance from one another, which completely masks the clusters. Methods of feature selection have been successfully implemented, to a certain degree, to bring about an improvement in the overall quality of the clusters. The Subspace[6] clustering approach is a very good way to group the complex information that is often found in big data. In contrast to feature selection methods, which focus on analyzing the information in its entirety, subspace clustering algorithms narrow their focus to a specific region. As a direct consequence of this, these algorithms can identify clusters that can be found in a great variety of subspaces, some of which may intersect with one another.

The term "big data" can have a variety of meanings, and the author of this work investigates those meanings as well as the challenges that are associated with big data analysis. This contains the characteristics that are employed in defining "Big Data," as well as its significance and those characteristics. The primary purpose of this study is to cast light on various kinds of Subspace clustering techniques that are already present in the existing body of research. In the same way that there are multiple descriptions for "Big data," there are numerous classifications of "Subspace clustering algorithm," as is generally agreed upon in the academic research community. This article delves into the various classification strategies that have been utilized by a variety of researchers over the course of some time to classify and categorize different subspace clustering strategies. This level of comprehension is necessary for the development of any new strategies that might be of assistance in the examination and interpretation of big data. A concise description of significant Subspace algorithms, which can act as a benchmark for any further development of algorithms, is also provided in this article. After an introduction in the first section, a discussion of the various classifications of big data is presented in the second section, which is followed by a discussion of the difficulties presented by high-dimensional data clustering in the third section. Section 4 provides an explanation of a literature review pertaining to the classification of subspace clustering approaches, and section 5 provides a conversation about significant existing approaches. The references that were used are presented in section 6, while the conclusions themselves are presented in section 7.

DEVELOPMENT

Definition of big data

Big data can be defined in a variety of ways, and the first quality that comes to mind when asking "what is big data?" is, without a doubt, its magnitude. But there are other aspects of big data that can be considered as well. In contrast, other aspects of big data have begun to appear over the past few years. According to Laney(8), the three dimensions that need to be dealt with in data administration are volume, variety, and velocity (also known as the Three V's). Big data can now be defined by its structure, which is referred to as the "Three V's".(9,10) Many of the world's most prestigious institutions and businesses have been instrumental in determining the characteristics of big data. Big data was defined by Gartner, Inc. The definition of "big data" is as follows: "High-volume, high-velocity, and high-variety data assets that call for cost-effective, innovative forms of processing of data for enhanced insight and decision making".(11) Tech America Foundation gave a similar definition: "Big data is a term for large amounts of fast-moving, complex, and different data that require modern techniques and methods to capture, store, distribute, manage, and analyze".(12) Based on these definitions, it is reasonable to conclude that the three most important aspects of big data are volume, diversity, and velocity. These three essential characteristics describe and establish the direction of the various approaches that can be taken to analyze large amounts of data.

The quantity of material is referred to as its volume. The sizes of large amounts of data are typically published in multiple terabytes and petabytes. One petabyte equals 1024 gigabytes. There is no definition of big data volumes that is universally recognized and acknowledged; rather, these definitions are relative and depend on a variety of variables including the time as well as the kind of info. What does "big data" mean in this day and age might not fulfill the requirements in the years to come? It is inevitable that in the future, because the amount of storage will increase, which will make it possible to record even more expansive data collections. In addition to this, it is absolutely necessary to have a solid comprehension of the role that variation plays when deciding the overall size of a data set. The nature of the data introduces a fresh obstacle when attempting to specify its quantity. The question of whether or not a data collection constitutes a "big" or "small" one arises whenever two datasets of comparable size call for management strategies that couldn't possibly be more dissimilar from one another. As was mentioned earlier a substantial amount of progress toward the development of big data was made by businesses. In addition to this, the kind of business also plays a substantial role in determining the benchmarks for the volume of big data.

The different kinds of structures that can be found in a dataset are referred to collectively as the "variety" of that dataset. The advancement of technology has made it possible to use a broad variety of data formats, including structured, semi-structured, and unstructured data. This has made it possible to store and retrieve data in a more efficient manner. A great example of structured data is the tabulated information that can be found in relational databases or spreadsheets. Structured data only accounts for 5 % of all the data that is currently available.(13) Textual content, pictures, music, video are all instances of the types of data that make up the majority of unstructured data. There are times when these kinds of data do not have the structural organization that is absolutely necessary for the analysis of automated machines. The format of the material that is semi-structured does not comply with any specific standards. A common illustration of data that is semi-structured is the Extensible Markup Language (XML), which is a written language used for the exchange of data on the web. Documents written in XML have data tags that are user-defined, which enables machines to understand them.

The term "velocity" is used in an idiomatic sense to refer to rate at which data are formed as well as the swiftness with which they must be processed, evaluated, and interpreted. In conjunction with the proliferation of digital devices, the requirement for real-time analytics and situation-based planning has resulted in a rate of data generation that has never been seen before. Even the most traditional shopping establishments are producing high-frequency data. For example, Wal-Mart is capable of handling nearly one million sales every

https://doi.org/10.56294/saludcyt2023413

single hour.[13] The torrents of information that are produced by the information that comes from smartphones and mobile apps is currently being used to make real-time, personalized offers for users as they go about their daily lives.

In the study that has been done so far, besides the three Vs, other aspects of big data were also talked around. Those terms were developed in a very specific way by titans of the information technology industry, who play a pivotal role in describing the generation, need, and use of big data. IBM came up with the idea for the fourth "V," which stands for "veracity," which describes the intrinsic unreliability of certain data sources. This is meant to symbolize the unpredictability that comes with the natural decision-making process, as well as other factors such as human judgment. There is a high probability that this kind of unpredictability will be present in the vast majority of the data that is produced by social media. Dealing about unclear and questionable data is important because, even though they hold important information, they can be wrong in a number of ways. This is another part of big data that requires to be dealt with, and it can be done with the help of tools and analytics that have been made just for managing and mining data that isn't clear.

The company SAS was the first to introduce the concept of changeable data and challenging information as two new categories of "big data." The word "variability" refers to how the rates at which data is collected change, which is flowing, as it is frequently observed that the velocity of big data is not constant but rather exhibits periodic peaks and valleys. The word "complexity" describes the fact that a lot of different types of sources make a lot of data. This is a big problem because it means that data from different sources needs to be linked, matched, cleaned, and changed in some way. Value incorporated Oracle is frequently cited as one of the distinguishing characteristics. The term "low value density" is used to describe the standard characteristics of big data sets, according to Oracle's definition. To put it another way, the worth of the data when it is still in its original form is typically quite low in comparison to the amount of space it takes up. On the other hand, one can acquire a high value by analyzing a large amount of data of this kind.

High dimensional clustering challenges

Clustering high-dimensional data is difficult for a number of reasons, the most significant of which are as follows:

Curse of Dimensionality or Sparse Data: when machine learning algorithms are used on data that has a high dimensionality, an occurrence that is referred to as the Curse of Dimensionality occur.[6] Was the first person to describe this phenomenon with the name? The inability of clustering algorithms to successfully deal with data sets that contain a large number of dimensions is what's meant when people talk about the "curse of dimensionality." Figure 1 illustrates the "curse of dimensionality," which states that the number of regions will increase at an exponential rate if the number of dimensions is increased. This can be seen by looking at the figure 1.

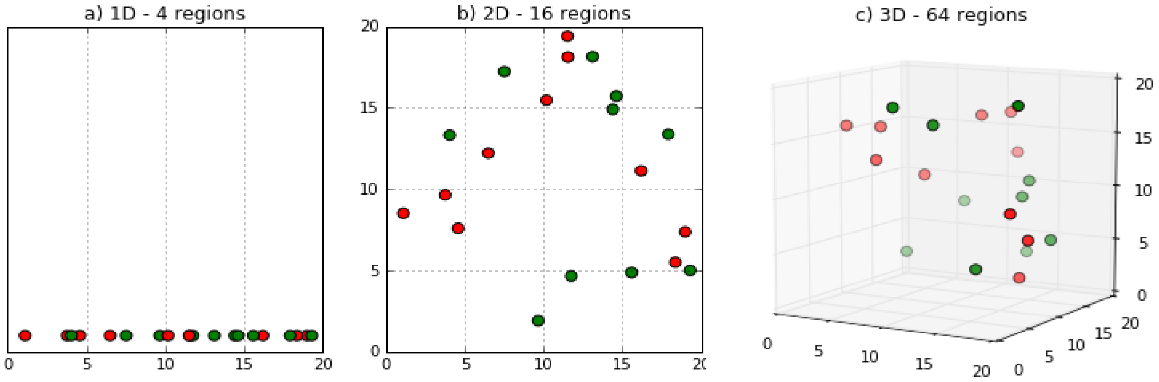

Figure 1. Curse of Dimensionality[14]

This can be understood by considering the fact that the significance of distance measures in data points decreases as the dimension of the dataset increases. This is something that can be explained. This is something that can be attributed to the fact that the spread of the data that occurs as a result of the increased addition of dimensions eventually reaches a place where, in very high dimensions, the data are nearly equally distant from one another. The issue becomes even more complicated when the objects in question are connected to one another in a variety of different subdivisions of dimensions. An algorithm called subspace clustering has a tendency to address and discover this kind of relationship. It is essential to get rid of any superfluous characteristics in order to guarantee that the method for clustering can concentrate solely on the important dimensions. This can be accomplished by removing any irrelevant characteristics. Clusters that are located in small dimensional spaces have a tendency to be easier to understand, which enables the individual using it to more effectively direct the path that the subsequent investigation takes.

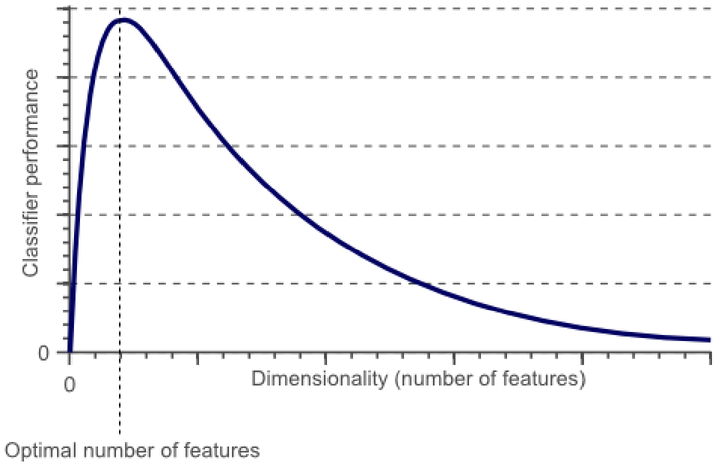

Figure 2. Classifier performance vs. Dimensionality[14]

In a similar fashion, the effectiveness of the classification rises along with the dimensionality of the data that is available to it. Up until the point where the optimal number of features is attained. Figure 2 depicts this truth for your viewing convenience. Any subsequent rise in dimension that fails to outcome in an increase in the overall number of training samples will result in a decline in the performance of the classifier. In the examination of big data, another essential facet that must be taken into consideration is the aforementioned matter.

Irrelevant Dimensions: while working with high-dimensional data it is quite normal for a significant number of the dimensions to be unnecessary for the clustering or analysis of the data.[15] In situations like this one, one of the most common solutions is to reduce the dimensionality of the data while maintaining the integrity of the essential information. As a result, the process of clustering generally begins with a phase that is known as "feature selection," that endeavors to strip the data of characteristics that are not important to its interpretation. However, because the clusters are embedded in a variety of subspaces, the global filtering approach to feature selection is not practical when working with high-dimensional data. When clustering the data, one dimension may be helpful in certain subspace formations; however, in other subspace formations, it may be completely meaningless.

Correlations among Dimensions: the presence of correlation between attributes is typically observed across a significant number of categories. Therefore, it's possible that the groups aren't lined up with their axes parallel. Instead, they could be arranged in any way that someone wants. Because of these issues, the data universe has an average density that is relatively low. Not only is the data space not very full, but the noise values are also spread out evenly across the high-dimensional space.[16] When applied to high-dimensional data, the conventional clustering methods will prove to be inefficient when used to search for clusters.

Classification of subspace clustering approaches

Both techniques for transforming features and techniques for selecting features are included in the methodologies that can be used to cluster high-dimensional data. In the case of Feature transformation approaches, a dataset is broken down by combining the original attributes in various ways so as to reduce the total entirety dimensions that are used. Methods are helpful in illuminating the hidden structure that exists within databases. Because this method keeps the distances between items the same, it works less well when there are a lot of irrelevant attributes in the data. This is the primary drawback of such an approach. Additionally, the new features are composites of the old ones, causing it to be potentially very difficult to comprehend new characteristics when placed within the framework of the domain. On the other hand, feature selection approaches center their attention on the dimensions of a dataset that are the most pertinent to the function of bringing to light groups of things that are comparable in only a selected subset of their properties or characteristics. This method works well with many different kinds of databases, however, it may be difficult to identify them using this technique if the clusters are distributed across multiple subspaces. The algorithms that deal with subspace are perfectly capable of processing this kind of input. The concepts of selecting features are expanded upon by these algorithms, taking them one step further, which selects pertinent subspaces for each cluster individually.

The process of subspace clustering is an addition to the feature decision procedure which searches for groups within the numerous subspaces that make up a single dataset. The process of subspace clustering requires a search technique in addition to evaluation criteria, just like feature selection does. In addition, the scope of the assessment criteria needs to be restricted in some way for subspace clustering in order to take into account different subspaces for each of the distinct clusters. Existing subspace clustering methodologies are amenable to categorization via a wide variety of categorization methods. Lance et al.(17) said that there are two main ways to cluster subspaces: top down and bottom up, which are distinguished by the direction in which the search is conducted. In addition, the top-down technique can be subdivided further according to both cluster weighting techniques and per instance weighting methods are available. The authors have utilized a classification strategy that is based on a grid. In this technique, the groups are further sub classified depending on the dimension of the array using either a static or an adaptive grid strategy. It was not possible to categorize the bottom-up approaches with absolute certainty using either the grid-based or density-based frame works Ilango et al.(18) provided a classification in which multidimensional clustering techniques were divided into the following categories: partitioning techniques, density-based techniques, hierarchical techniques, grid-based techniques, and model-based techniques.

Subspace clustering was divided into two categories by Karlton et al.(19) research: Clustering based on density as well as clustering that was anticipated. The researchers referenced density-driven clustering techniques such as CLIQUE in their work. (Clustering In QUEst),(15) MAFIA (Merging Adaptive Finite Intervals and is more than a clique),(20) SUBCLU (density connected SUBspace CLUstering)(21) have the abundance of the data as the basis for their clustering. Similar clustering predicted by models is seen in methods such as PROCLUS. (PROjected CLUStering),(22) CLARANS,(23) ORCLUS (arbitrarily Oriented projected CLUStering),(24) DOC (Density based Optimal projective Clustering),(25) etc.

Clustering techniques for data with high dimensions that are predicated depending on the perspective of the data, some clustering methods include subdomain clustering, which is predicated upon axes parallel clustering; correlation clustering, which is predicated on arbitrarily directed clustering; and pattern-based clustering, which is predicated on the data being organized in a certain way, which is predicated on axis parallel clustering, were discussed in Kriegel et al.(26) Finding clusters that can occur in arbitrarily oriented subspaces is the goal of the correlation-based methods, of which ORCLUS(24) is one example. These methods are used to accomplish this. p-Cluster(27) is a method for pattern-based clustering that groups together the objects that demonstrate a subset of characteristics that share comparable tendencies.

The problem-based categorization that utilizes the axis parallel subspace clustering method provides generates as a result leads to classifications including composite algorithms, projected clustering, and soft projected clustering. The method known as PreDeCon is an illustration of a projected clustering strategy,(28) which stands for subspace PREference weighted Density CONnected clustering. With the help of this method, a one-of-a-kind distribution of each object to precisely one subspace cluster or cacophony can be found.(29) The context of the algorithms for gentle extended clustering the overall amount of clusters, denoted by k, is chosen in preparation, and a target expression is constructed in order to maximize the formation of k-numbers of clusters. This is done in order to achieve the best possible results. This is done so as to ensure that the clustering process operates as successfully as possible.(30) The algorithm COSA (Clustering Objects on Subsets of Attributes) is a good illustration of a fuzzy projected clustering algorithm.(31) A different subspace clustering algorithm known as SUBCLU(31) works toward the goal of locating all subspaces that contain groupings that can be identified.

FIRES are an example of a hybrid algorithm, which is a classification that relates to the algorithms that are able to identify overlapping clusters. This category of algorithms is called hybrid algorithms. (FIlter REfinement Subspace clustering).(26) In Müller et al.(32), an additional classification subspace clustering technique that depends on the parameterization of the of results is described.

The methodologies were broken down into three categories: cell-based, density-based, and clustering-oriented methodologies. CLIQUE(15) is an example of a cell-based strategy that searches collection of fixed or variable cells in a grid that exceeds a certain threshold in terms of the number of objects they comprise. Similar to SUBCLU(31) clusters are defined by the density-based approach as crowded areas that are divided by thin regions. Similarly, PROCLUS,(22) which is an approach that is oriented toward clustering, defines properties of the complete set of clusters. A few examples of these characteristics are the typical number of the clusters, the total number of clusters that make up the dataset, or properties that have a statistical orientation.

Important subspace clustering algorithms

In this section, we take a look at a few of the most important subspace clustering algorithms that were discovered in published research for a limited amount of time. These algorithms were a significant contributor to the development of novel methodologies. It is easy to see that most of the algorithms are merely adjustments and modifications of one of the pre-existing methods.

CLIQUE[15], in order to identify clusters that are contained within the subdomains of the dataset, one of the earliest algorithms makes use of a method known as APRIORI. In order to locate clusters, this algorithm utilizes a hybrid approach that incorporates density-based clustering and grid-based clustering. This technique makes use of coverage, which can be understood as a percentage of the raw data that is comprised of small quantities that are located within the subdomain, in order to determine which clusters actually exist. After the dense subspaces have been identified, they are arranged according to their covering. Only the subspaces that have the greatest amount of covering are retained, while the others are deleted. After completing this stage of the process, the algorithm then utilizes a depth-first search strategy in order to find neighboring grid elements packed closely together in all of the subdomains that were selected in the previous phase. After that, clusters are produced by integrating these units in order to achieve a greedy development strategy. The process begins with a randomly selected dense block, and then it from there, it expands a maximal region in an increasingly wasteful manner until it achieves the largest size that is feasible for it. This development is carried out in every conceivable facet up until the point at which the totality of the cluster is encompassed by the combined efforts of all of its parts. After repeating the process, the duplicate parts of the smallest size are eliminated one by one until there are no more maximal regions left to eliminate. An expression known as the Disjunctive Normal Form (DNF) is then used to describe the hyper-rectangular clusters.

CLIQUE is able to locate a wide variety of clusters, regardless of their shape, and can display the results in a manner that is simple to understand. CLIQUE has a technique for generating clusters that is based on the increasing density of regions, which enables it to locate clusters of arbitrary shapes. This gives CLIQUE the capacity to identify groupings of any form. Additionally, CLIQUE has a technique for generating clusters that is based on the increasing density of regions, which enables it to locate clusters of arbitrary shapes. This gives CLIQUE the skill of locating concentrations of any form. Clusters can be located in the same subspace as one another, in overlapping subspaces, or in subspaces that are not connected. The DNF formulations that are utilized to symbolically represent clusters are typically very simple and straightforward to comprehend. This is useful information for the subspace clustering process because clusters frequently exist across various subspaces and, as a result, symbolize a variety of different relationships. CLIQUE, like other bottom-up algorithms, has high levels of scalability in proportion to the amount of examples and variables that are present in the collection. CLIQUE, along with other methods of a similar nature, does not, however, increase very well with the total amount of variables in its result clusters. This is due to the fact that CLIQUE was designed to handle a much smaller number of dimensions. Due to the fact that subspace clustering is typically utilized in order to locate low dimensional groupings among high dimensional data, this is not typically a significant problem.

A subspace clustering technique known as ENCLUS[33] is strongly influenced by the algorithm known as CLIQUE. One of the most significant distinctions between ENCLUS and CLIQUE is that ENCLUS doesn't directly measure abundance or cover. Instead, it counts the data's entropy. CLIQUE does so. The fundamental idea that underlies the functioning of ENCLUS is that, on average, a subspace that includes clusters will have less entropy than one that does not compared to a subspace and that does not contain clusters. The cluster ability of a subspace can be described with the help of these three criteria: coverage, density, and correlation. Entropy is a useful tool for evaluating all three of these parameters simultaneously. It is also essential to keep in mind that the entropy of a system diminishes as the density of its constituent cells rises. However, under particular circumstances, entropy can also diminish in response to an increase in coverage in a similar vein, interest is a definition of association that can be characterized as the variation in a number of variance values for a collection of variables as well as the overall entropy of the complex distribution. This difference can be calculated using the formula. To put it another way, interest is a measurement of the strength of the correlation between two different sets of characteristics. Values closer to one another indicate a higher degree of correlation between the parameters. Whereas a value of zero indicates that the dimensions are independent. In order to mine significant subspaces, ENCLUS takes the same bottom-up strategy as CLIQUE, which is inspired by the APRIORI style. Finding subspaces with the lowest possible correlation is the goal of the pruning process, which is accomplished by combining the entropy's downward closure property with the correlation's upward closure property of interest.

The top-down subspace clustering algorithm known as PROCLUS[22] was the first of its kind. In a manner analogous to that of CLARANS[34], PROCLUS takes a random selection of the data, after which it chooses a group of k medoids and continually enhances the clustering. Initialization, repetition, and cluster refinement are the three components that make up the algorithm's distinct phases of operation. When it comes time to begin the process of initialization, In order to make room for additional ones, a greedy algorithm is used to select a collection of potential medoids that are situated at a considerable distance from one another in order to clear some space. This is done in order to guarantee that each cluster will have at least a single instance to symbolize it within the set that is selected. During the repetition phase, a random collection of k medoids is chosen from the smaller dataset that was utilized in the previous step. During the cluster refinement phase, poor medoids are exchanged for new, arbitrarily selected medoids, and it is determined whether or not the clustering has been improved. The average distance that separates each instance from the medoid that is geographically nearest to it can be used to evaluate the quality of a cluster.

A collection of measurements is chosen for each medoid whose average distances are significantly closer together than what a statistical study would lead one to anticipate they should be. It is necessary for there to be k*l total dimensions associated with medoids, where l is an input value that establishes the usual number of lengths that cluster subspaces have. After the subspaces for each medoid have been selected, the average Manhattan segmental distance will be utilized in order to allocate points to the medoids, which will ultimately result in the formation of clusters. The medoid of the cluster that has the fewest number of points is discarded, along with any other medoids associated with fewer points than (N/k) *min Deviation. The term "min Deviation" refers to the name of the statistic that was utilized in the process of inputting the information. This occurs after determining which cluster has the fewest number of points. During the phase known as "refinement," in order to calculate more dimensions for each medoid, the clusters that have already been generated are utilized as a starting point. Additionally, during this phase, outliers are eliminated, and their scores are redistributed to the medoids. PROCLUS, like many other top-down techniques, has a strong preference for clusters that have a hyper-spherical shape. In addition, even though clusters can be discovered in various subspaces, those subspaces have to have sizes that are comparable to one another mainly due to the fact that the individual is required to input the typical amount of dimensions for the clusters. Clusters can be thought of as collections of instances, with each individual instance having its own unique medoids and subspaces. These medoids and subspaces generate distinct subdivisions of the dataset, which may also include outliers in their composition.

DBSCAN(35), density-based clustering algorithms look for clusters in a location based on the number of data points that are concentrated there, the core idea that underpins density-based clustering is that the area surrounding a given radius (Eps) must have at least the threshold number that was specified of instances for each individual cluster instance. This is the foundational concept that underpins density-based clustering (MinPts). The DBSCAN is an example of several of the most common approaches to density-based clustering that are currently in use,(35) which can be found here.

The data points are arranged in DBSCAN according to one of three different categories.

Core points: these are areas that are physically located in greater proximity to the core of a cluster. When there are sufficient points surrounding a point, we refer to that point as an internal point.

Border points: the term "boundary point" refers to any location on the map that is not a central point. This indicates that there are insufficient points in the neighborhood of the boundary point; however, the boundary point is still in close proximity to the central location where the action takes place.

Noise points: a noise point is any location in an area that does not qualify as either the center or a boundary point in the area.

DBSCAN begins with an arbitrary case in the data collection (D), and then obtains all instances of D with respect to Eps and MinPts in order to identify a cluster. This process is repeated until a cluster is found. This process allows DBSCAN to locate the cluster. In order to determine which locations are inside Eps separation from the central nodes of the clusters, the method employs a spatial data arrangement known as R*tree(36), which stores information in a tree-like format. It was demonstrated that a modified variant of DBSCAN called an incremental DBSCAN could achieve the same level of success as the original DBSCAN algorithm. In addition, another clustering algorithm known as GDBSCAN is described in Ester et al.(37), which generalizes the density-based algorithm known as DBSCAN.

SUBCLU (density-connected SUBspace CLUstering)(21): this is the initial approach to make use of density-based subspace clustering, and it expands the concept behind DBSCAN so that it can function with data that is extremely dimensional. The method makes use of an algorithm to discover highly linked clusters across every one of the subspaces of high-dimensional information, and it makes use of the monotonicity property to get rid of higher-dimensional projections. This guarantees that the search space is reduced to a large extent, which makes it easier to find what you're looking for. It is oblivious to the problems that are inherent in grid-based approaches, like being dependent in the location of the arrays or having a specified form for the clusters, both of which are required by grid-based approaches. This is because it is independent of both of these factors. In order to generate all 1-dimensional clusters, the technique starts by employing DBSCAN(35) to every 1-dimensional region. This is done in order to generate the clusters. The input parameters used to generate these clusters are the density threshold and distance (radius). The next step is for it to ascertain, for each k-dim cluster, that it is present in any of the (k-1)-dim clusters; when it does not, the cluster in question is eliminated from the database. In the final step, clusters are produced by employing DBSCAN on each potential subspace that possesses (k+1) dimensions. These stages will continue to be carried out in a recursive manner as long as the collection of k-dimensional subdomains that contain groups remains infinite.

INSCY (INdexing Subspace Clusters with in-process-removal of redundancY)(38) has yet another effective sector clustering algorithm, and it was founded on the idea of subspace clustering that was described in Assent et al.(39) In order to mine in an iterative manner in a region that contains all clusters across each subspace projection, the following steps must be taken, it uses a depth-first strategy, and then it continues with the next region. It begins by quickly trimming away all of its redundant low-dimensional projections before moving on to evaluating the maximum high-dimensional projection. This strategy improves efficiency because it eliminates a number of the disadvantages associated with breadth-first subspace clustering, and it significantly cuts down on the amount of time required for runtimes. In addition to that, it enables the tracking of potentially useful sub domain cluster regions. INSCY suggests a brand-new index structure that will be called the SCY-tree when it is implemented. This structure provides an organized version of the data and permits unrestricted access to the various subspaces SCY-tree is able to combine during duplication cutting, which enables extremely effective subspace clustering and also enables INSCY rapid and concise. These benefits are all due to SCY-tree's ability to combine these two features.

Scalable Density Based Subspace Clustering(40) reduces the amount of time spent processing subspaces by locating and grouping together potentially useful subspaces; only mines a small number of carefully chosen subspace groupings. It as well as their combinations directly, which helps to narrow down the search universe while maintaining the accuracy of the results.(41) It operates under the premise that any high-dimensional subspace grouping can be seen to exist in a variety projection of a limited dimensions. The process is capable of acquiring data by mining a subset of them, which is only some of them, enough information to skip processing the intermediate subspaces and go straight to the high-dimensional subspace clusters that contain the most fascinating data. When utilizing this method, it is possible to steer the process of subspace clustering without having to conduct any database searches at all for the many intermediate and redundant subspace projections. It accomplishes this goal by employing a priority queue in order to begin the process of initializing the data of density projections. It lays the groundwork for selecting the perfect applicant from the pool of candidates who have been given their priorities in order of importance. In order to make room for a very diverse range of density granularities, the priority queue has been sectioned off into three distinct categories. It does not proceed through the intermediary subspaces in a best-first fashion but rather goes straight to the higher dimensional subspaces.

CONCLUSION

From social media promotion to genome mapping, big data study has tremendous breadth and a lot of room for improvement. This is because big data has become the primary focus of many businesses and industries. The use of subscape clustering is one of the essential techniques that can be of assistance in the examination of big data. The purpose of this document is to accomplish the general introduction to Big Data and the challenges. A literature review is presented about the various types of Subspace clustering approaches, along with their categorization, in order to facilitate straightforward comprehension of the various classifications of Subspace algorithms and their classifications. A concise presentation on essential approaches to subspace clustering is also discussed here. This paper can help researchers comprehend big data, its challenges, and existing approaches, and it can also contribute to the design of new clustering methods for big data.

REFERENCES

1. David JM, Balakrishnan K. Prediction of Key Symptoms of Learning Disabilities in School-Age Children using Rough Sets. Int J Comput Electr Eng. 2011;3(1):163-169.

2. Gupta R. Journey from data mining to Web Mining to Big Data. IJCTT. 2014;10(1):18-20.

3. Sharma PP, Navdeti CP. Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution. IJCSIT. 2014;5(2):2126-2131.

4. Gupta R, Gupta S, Singhal A. Big Data: Overview. IJCTT. 2014;9(5).

5. Jain A. Data clustering: 50 years beyond k-means. Pattern Recognition Letters. 2010;31(8):651-666.

6. Jain A, Dubes R. Algorithms for Clustering Data. Prentice Hall; 1988.

7. Karger DR. Random sampling in cut, flow, and network design problems. STOC. 1994;648-657.

8. Laney D. 3-D data management: Controlling data volume, velocity and variety. Application Delivery Strategies by META Group Inc. [Internet]. 2001 [cited 2023 Jun 10]. Available from: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

9. Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: From big data to big impact. MIS Quarterly. 2012;36(4):1165-1188.

10. Kwon O, Lee N, Shin B. Data quality management, data usage experience and acquisition intention of big data analytics. Int J Inf Manage. 2014;34(3):387-394.

11. TechAmerica Foundation’s Federal Big Data Commission. Demystifying big data: A practical guide to transforming the business of Government. [Internet]. 2012 [cited 2023 Jun 10]. Available from: http://www.techamerica.org/Docs/fileManager.cfm?f=techamerica-bigdatareport-final.pdf

12. Gartner IT Glossary. [Internet]. n.d. [cited 2023 Jun 10]. Available from: http://www.gartner.com/it-glossary/big-data/

13. Cukier K. The Economist, Data, data everywhere: A special report on managing information. February 25, 2010. [Internet]. [cited 2023 Jun 10]. Available from: http://www.economist.com/node/15557443

14. Chen L. Curse of Dimensionality. In: Liu L, Özsu MT, editors. Encyclopedia of Database Systems. Springer; 2009. p. 133.

15. Agrawal R, Gehrke J, Gunopulos D, Raghavan (1998) Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. SIGMOD. 1998;27(2):94-105.

16. Berchtold S, Bohm C, Keim D, Kriegel H-P. A Cost Model for Nearest Neighbour Search in High Dimensional Data Space. PODS. 1997;78-86.

17. Lance P, Haque E, Liu H. Subspace Clustering for High Dimensional Data: A Review. ACM SIGKDD Explorations Newsletter. 2004;6(1):90-105.

18. Ilango MR, Mohan V. A survey of Grid Based Clustering Algorithms. Int J Eng Sci Technol. 2010;2(8):3441-3446.

19. Karlton S, Zaki M. SCHISM: A New Approach to Interesting Subspace Mining. Int J Bus Intell Data Min. 2005;1(2):137-160.

20. Goil S, Nagesh H, Choudhary A. MAFIA: Efficient and Scalable Subspace Clustering for Very Large Data Sets. Technical Report CPDC-TR-9906-010. Northwestern University; 1999.

21. Kailing K, Kriegel H-P, Kroger P. Density-Connected Subspace Clustering for High Dimensional Data. SIAM International Conference on Data Mining. 2004;46-257.

22. Aggarwal CC, Wolf JL, Yu PS, Procopiuc C, Park JS. Fast Algorithms for Projected Clustering. ACM SIGMOD International Conference on Management of Data. 1999;61-72.

23. Ng RT, Han J. CLARANS: A Method for Clustering.

24. Aggarwal C, Yu P. Finding Generalized Projected Clusters in High Dimensional Spaces. ACM SIGMOD International Conference on Management of Data. 2000;70–81.

25. Procopiuc C, Jones M, Agarwal PK, Murali TM. A Monte Carlo Algorithm for Fast Projective Clustering. ACM SIGMOD International Conference on Management of Data. 2002;418-427.

26. Kriegel HP, Kroger P, Zimek A. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, & Correlation Clustering. ACM TKDD. 2009;3(1):1.

27. Wang H, Wang W, Yang J, Yu P. Clustering by Pattern Similarity in Large Data Sets. ACM SIGMOD International Conference on Management of Data. 2002;394-405.

28. Bohm C, Kailing K, Kriegel H-P, Kroger P. Density Connected Clustering with Local Subspace Preferences. IEEE International Conference on Data Mining. 2004;27-34.

29. Friedman J, Meulman J. Clustering objects on subsets of attributes. J R Stat Soc Ser B. 2004;66:815-849.

30. Kriegel HP, Kroger P, Renz M, Wurst S. A Generic Framework for Efficient Subspace Clustering of High Dimensional Data. IEEE International Conference on Data Mining. 2005;250-257.

31. Blum A, Langley P. Selection of Relevant Features and Examples in Machine Learning. Artif Intell. 1997;97:245–271.

32. Müller E, Günnemann S, Assent I, Seidl T. Evaluating Clustering in Subspace Projections of High Dimensional Data. VLDB Endowment. 2009;2(1):1270-1281.

33. Cheng CH, Fu AW, Zhang Y. Entropy-based subspace clustering for mining numerical data. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1999;84-93.

34. Ng R, Han J. Efficient and effective clustering methods for spatial data mining. VLDB Conference. 1994;144-155.

35. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial data sets with noise. Proc Int Conf Knowl Discov Data Min. 1996;226–231.

36. Katayama N, Satoh S. The SR-tree: An Index Structure for High-Dimensional Nearest Neighbor Queries. ACM SIGMOD International Conference on Management of Data. 1997.

37. Ester M, Kriegel H-P, Sander J, Wimmer M, Xu X. Incremental Clustering for Mining in a Data Warehousing Environment. VLDB Conference. 1998.

38. Assent I, Krieger R, Müller E, Seidl T. INSCY: Indexing Subspace Clusters with In Process-Removal of Redundancy. IEEE International Conference on Data Mining. 2008;414–425.

39. Assent I, Krieger R, Muller E, Seidl T. DUSC: Dimensionality Unbiased Subspace Clustering. IEEE Intl. Conf. on Data Mining (ICDM). 2007;409-414.

40. Müller E, Assesnt I, Gunnemann S, Seidl T. Scalable Density based Subspace Clustering. ACM Conference on Information and Knowledge Management (CIKM'11). 2011;1076-1086.

41. Sangapu SC, Prasad KSN, Kannan RJ, et al. Impact of class imbalance in VeReMi dataset for misbehavior detection in autonomous vehicles. Soft Comput. 2023. https://doi.org/10.1007/s00500-023-08003-4.

FINANCING

No financing.

CONFLICT OF INTEREST

None.

AUTHORSHIP CONTRIBUTION

Conceptualization: Vijaya Kishore Veparala, Vattikunta Kalpana.
Research: Vijaya Kishore Veparala, Vattikunta Kalpana.
Writing - proofreading and editing: Vijaya Kishore Veparala, Vattikunta Kalpana.