Subspace clustering for high dimensional datasets

G.N.V.G. Sirisha^{1*} and M. Shashi²

S.R.K.R. Engineering College, Chinna Amiram, Bhimavaram, Andhra Pradesh, India¹ A.U. College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India²

Received: 04-June-2016; Revised: 10-August-2016; Accepted: 12-August-2016 ©2016 ACCENTS

Abstract

Clustering high dimensional data is a challenging problem because of the existence of many irrelevant and redundant attributes. Conventional clustering algorithms identify a global set of relevant attributes prior to clustering using attribute selection and feature extraction techniques. All the globally relevant attributes are used in the similarity calculation while clustering. These algorithms fail to identify true clusters that are present in a subset of attributes. So, subspace clustering has become the thrust area of research in the recent past. Subspace clustering detects the clusters that exist in subsets of dimensions. Different types of subspace clustering algorithms are proposed in the literature. This paper discusses the different types of subspace clustering algorithms with main emphasis on 2D subspace clustering. Availability of new and huge datasets like spatiotemporal datasets, temporal datasets, spatial datasets and genomic data has necessitated the development of 3D subspace clustering. This paper presents an overview of subspace clustering for the research community who is interested in subspace clustering.

Keywords

Subspace clustering, Curse of dimensionality, Density divergence, 3D subspace clustering.

1.Introduction

Development of communication technology and advancements made in data collection equipment has enabled collection of different types of data at high space and time granularities. The devices like bar code scanners, radio frequency identification devices and smart card readers are being used by many businesses and organizations nowadays for collecting data. Different types of sensors and video surveillance equipment are used frequently in almost all fields to ensure safety and security. Remote sensing satellites are extensively used for collecting data in applications including meteorological data collection, forest and land usage monitoring, oceanography, global positioning systems, etc. Some of these devices generate a continuous flow of data called data streams. With the availability of high capacity storage devices every organization is accumulating large volumes of data over the years. Extracting useful information and knowledge from the accumulated data is essential for gaining competitive advantage. It is impossible to extract knowledge from large volumes of data manually.

So there is a need for automated tools that incorporate intelligent techniques to aid in information extraction and data analysis. Data mining, a confluence of multiple fields provides users with such tools and techniques. Data mining extracts the hidden knowledge in the form of interesting patterns.

A wide variety of data mining techniques are developed specifically to handle different types of datasets which may range from relational databases to data warehouses, text databases, time series data, temporal data, multimedia data, sequence data, spatial data, spatiotemporal data, world wide web etc. Appropriate data mining techniques have to be selected depending on the type of data to be processed as well as the type of knowledge to be extracted.

Based on the type of knowledge extracted data mining techniques can be classified into various functionalities like class/concept description, association rule mining, classification, regression, cluster analysis, outlier analysis and evolution analysis.

^{*}Author for correspondence

Class/Concept description data mining functionality derives general characteristics of a class or specific group of data. Association rule mining derives rules that describe the association between frequently cooccurring variable-value pairs in a given dataset. Classification and prediction are predictive data mining tasks that build models from labeled data (training data). The models are then used for prediction. Classification is used for predicting the class or label of a new or unseen object and regression is used for predicting the value of numerical attribute. Classification and regression are also called as supervised learning as they require training data for model construction.

Outlier analysis identifies the data objects or values which are significantly different from majority of objects/values in the dataset. Evolution analysis is analysis of data whose state or behavior changes with time.

Clustering is the process of grouping objects such that all the objects in each group share similar characteristics. Clustering can be used as a standalone tool for applications like hierarchy generation. It can also be used as a pre-processing step for other algorithms such as characterization, attribute subset selection, classification etc. [1]. Irrespective of the criteria used for clustering all the clustering algorithms try to maximize the intra cluster similarity and minimize inter cluster similarity. Clustering is also called as unsupervised learning because the data that is to be clustered in not labeled.

A wide variety of clustering algorithms are proposed in the literature. They are categorized as partitional clustering, hierarchical clustering, density based clustering, grid based by clustering, model based clustering. The choice of the clustering algorithm depends on the type of data, the type of clusters that are to be mined and the purpose for which the derived clusters are used.

2.Subspace clustering

Conventional cluster algorithms use all the dimensions that are selected prior to clustering in the similarity calculation while clustering objects. Different types of distance measures like Euclidean distance, Manhattan distance are used for this purpose. With the increase in the number of dimensions that are used to describe the data, these algorithms face a problem called as Curse of Dimensionality. It refers to the situation that the ratio of the distances of the nearest and farthest neighbors

to the given object/target tends to 1 as the number of dimensions used to describe the data increases irrespective of distance estimates and data distributions [2-4].

Also, presence of irrelevant and redundant attributes masks the clusters that are present in subsets of dimensions. Hence conventional clustering algorithms fail to detect true clusters in such domains. Feature extraction and attribute selection are two popularly used methods for dimensionality reduction. Feature extraction methods like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) create new attributes which are linear combinations of the original attributes. The most significant of the new attributes are then used for clustering. The disadvantage with this approach is that new attributes are difficult to interpret, making the clustering results less useful. Attribute selection methods aim at selecting a subset of dimensions by removing irrelevant and redundant dimensions in the global perspective.

Even after relevant attribute selection a cluster that is present in a subset of attributes (subspace) may not be discovered when seen in full dimensional space defined by the global set of relevant attributes. A subspace is a subset of relevant attributes/dimensions that are shared by the members of a cluster. The subset of relevant attributes shared by members of one cluster may be different from a subset of relevant attributes shared by members of another cluster. Attribute subset selection, retrieves only a single set of relevant attributes in global perspective prior to clustering. Both attributes subset selection and feature extraction are filter approaches to relevant attribute selection. They thus does not aid in finding clusters which are meaningful in subspaces only. Subspace clustering is a wrapper approach where interesting subspaces are detected and clusters that exist in those subspaces are detected by the cluster algorithm.

Subspace clustering detects the clusters that are existing in subspaces of the multidimensional data space. Let $\mathbf{D} = \mathbf{O} \times \mathbf{A}$ be a dataset represented in the form matrix, where \mathbf{O} is the set of objects and \mathbf{A} is the set of attributes. A subspace cluster C is a submatrix of \mathbf{D} defined as $O \land A$, where the set of objects $O \subseteq \mathbf{O}$ is homogeneous in the set of attributes defined by the subspace $A \subseteq \mathbf{A}$ [3]. A number of subspace clustering algorithms have been proposed in the literature. Subspace clusters may overlap, i.e. an

object O_1 may cluster with object O_2 in subspace S_1 and O_1 may cluster with O_3 in subspace S_2 . Similarly an attribute may be a member of different subspaces.

2.1Applications of subspace clustering 2.1.1Gene expression analysis

Microarray chip technology helps us to measure expression level of different genes under different conditions like different environments, at different times and for different persons. The data usually contain expression levels of thousands of genes under thousands of conditions. The data is stored in the form a matrix with rows corresponding to genes and columns corresponding to different conditions.

Genes having a homogeneous expression level indicate common function. When we try to cluster genes having a homogeneous expression level they cluster differently in different subsets of conditions or time slots. That is, for example, under different cellular environmental conditions the genes may be clustered into different groups. That is a gene X may be clustered with gene Y but not with gene Z under certain conditions. The same gene X may be clustered with gene Z but not with Y under different set of conditions.

2.1.2Text documents

Text documents are clustered based on term frequency of feature vectors. In this application the data can be viewed as a matrix where rows represent the different text documents and columns represent the terms occurring in one or more documents under consideration, which can be hundreds or thousands or more. For each pair of document and term combination the number of times the term occurred in that document is recorded in corresponding matrix cell. This value can be zero or more. When these documents are clustered the documents in each cluster may have similar frequency only in a subset of terms. The terms shared by documents of one cluster may be different from the terms shared by documents of another cluster. Also the same document may be assigned to different clusters based on different sets of terms.

2.1.3Census data

Governments of all the countries periodically collect the data regarding the demographic, educational, economic, social conditions etc. of their people. Government can cluster that data to identify different sections of people who are eligible for different government schemes. In this dataset, the rows correspond to different persons and columns correspond to the different attributes describing persons. The same person may belong to different groups based on different attributes. For example, the same person may put into different clusters based on the purpose for which the data are analyzed; which would be found eligible for different government schemes based on known or unknown criteria. The same data can be used for better town planning, for identifying the best places for community/utility centre, etc. by clustering the same data in different subspaces. Subspace clustering has the capacity to automatically identify interesting subspaces to cluster data [5].

3.Types of subspace clustering techniques

Subspace clustering techniques are classified based on three criteria. The first criterion is the type of data they handle, the second criterion is the dimensionality of cluster solutions, and the third criterion is approaches used for clustering data. Different types of subspace clustering algorithms are devised for handling different types of data, such as continuously valued data, categorical data, sequence data, stream data. Subspace clustering techniques produce either 2D or 3D cluster solutions. A 2D cluster solution defines each cluster in 2 dimensions, with the first dimension representing the objects of the cluster while the second dimension representing the set of attributes shared by the members of a cluster. A 2D cluster solution is a set of 2D clusters. A 2D cluster is a set of objects that are homogenous in a subspace defined by the set of attributes.

All conventional clustering algorithms produce a one-dimensional cluster solution. A 1D cluster solution is a set of 1D cluster. A 1D cluster is a group of objects that are homogenous in a predefined problem space specified in terms of the relevant set of attributes [6]. In 3D cluster solution, first dimension represents objects, second dimension represents attributes describing the objects and the third dimension represents an attribute that has to be handled differently like time or location.

3.1Types of 2-D subspace clustering algorithms

Based on the approach taken 2-D subspace clustering algorithms can be classified as grid based, density based and window based algorithms. All these algorithms handle quantitative 2-D data. Grid based subspace clustering partitions the data space into grids. Dense grid cells are used to form subspace clusters. Grid based subspace clustering algorithms take two input parameters τ and ξ . τ is density threshold that is used in identifying dense grid cells. ξ specifies the number of intervals into which the domain of each attribute is to be split. Window based subspace clustering generates overlapping intervals for each attribute using sliding windows. The size of the sliding window is determined by a parameter ' δ '. The intervals are then used for finding maximal subspace clusters. The algorithm MaxNCluster based on nCluster model is a window based subspace clustering algorithm. Grid and Window based subspace clusters.

In Density based subspace clustering a cluster is defined as a maximal set of density connected points. It takes two input parameters \mathcal{E} and m. An object with at least m points in its \mathcal{E} neighborhood is called a core object. The density, connected core objects along with border objects form a cluster. All the subspace clusters at all possible subspaces of all dimensionalities satisfying this property are mined. The monotonicity property of density connectivity is used to efficiently discover core objects in all subspaces at all dimensionalities. Density based subspace clustering algorithms can detect arbitrary shaped clusters.

3.2Grid based subspace clustering algorithms

Grid based subspace clustering algorithms partition the domain of each dimension into ξ number of equal size intervals. This converts a *n*-*D* data space into a *n* dimensional grid. Grid cells which have high density are used to form subspace clusters. A subspace corresponds to a set of variable-value pairs. A *1-D* subspace is thus one variable-value pair or an interval corresponding to 1 dimension. A *K-D* subspace is a set of *K* variable-value pairs or *K* intervals corresponding to *K* different dimensions. A *K-D* subspace is interesting if its density exceeds the density threshold where the density of a subspace is the ratio of the number of data points (database objects) it holds to the total number of data points(database objects) in the dataset.

Clustering in QUEst (CLIQUE) is a pioneering algorithm in this category [7]. It discretizes the domain of each dimension into a user specified ξ number of equal sized intervals. Each discretized interval representing a dimension-value pair is a *1-D* subspace. *K-D* subspace has *K* intervals corresponding to *K* dimensions. It first identifies interesting subspaces at all dimensionalities. At each dimensionality, each dense grid cell is called an interesting subspace.

The interesting subspaces are identified using Apriori [8] style iterative level wise approach. Next at each dimensionality, connected dense grid cells in the multidimensional grid is grouped to form subspace clusters. To prune their search at a faster rate, they use the minimum description length (MDL) principle as a heuristic, thereby making it an approximate search. The 'dense' subspaces sharing faces are merged, and covering algorithms are used to mine the minimal descriptions of the subspaces.

Subspace entropy had used for selecting interesting subspaces [9]. Three criteria, namely coverage, density and correlation are used for identifying interesting subspaces. Entropy is used to measure all the three criteria. As density increases entropy decreases, under some conditions as coverage increases entropy decreases. 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. The subspaces that are found to be interesting based on all these three criteria are used for discovering the clusters. The same clustering model of CLIQUE with fixed grids is used for cluster discovery.

The algorithms that use uniform grids incur high computational costs if fine grained grids are used and poor cluster quality if coarse grain grids are used. So the maximal frequent itemset algorithm (MAFIA) [10] which is a major extension of CLIQUE uses adaptive grids for cluster discovery. Adaptive interval size is used to partition the dimensions depending on the distribution of data in the dimension. The bins and grid cells, which have low density will be pruned thereby reducing the computational cost. All the above algorithms use the same density threshold at all subspace dimensionalities and suffer from the density divergence problem.

3.3Density divergence problem

All the grid and density based subspace-clustering algorithms that use the same density threshold at all subspace dimensionalities suffer from the density divergence problem [4] [11]. As the dimensionality of the subspace increases, the number of constraints imposed by the subspace increases. Hence, as the dimensionality of subspace increases, the number of points/objects satisfying the constraints imposed by the subspace decreases.

Row id	Age	Salary	Experience	
1	33	15000	7	
2	35	35000	5	
3	35	16000	13	
4	34	32000	9	
5	36	12000	6	
6	37	38000	17	
7	34	17000	12	
8	32	33000	18	
9	31	14000	8	
10	48	37000	13	
11	45	17000	14	
12	48	38500	27	
13	43	17000	16	
14	46	39000	16	
15	44	13000	26	
16	47	37000	15	
17	45	22500	24	
18	50	38000	27	

 Table 1 Sample dataset with clusterable tuples

Lower dimensional subspaces have high density compared to the higher dimensional subspaces. This phenomenon is called as density divergence. If we use same global density threshold to identify the dense subspaces at all dimensionalities, we may not find high-quality subspaces at all dimensionalities i.e. if the density threshold is set too high, many subspace clusters are missed at high subspace dimensionalities. If the density threshold is set too low, then many unnecessary clusters are mined at low subspace dimensionalities. To overcome this problem density threshold should be a function of subspace dimensionality. The value of the density threshold should be lowered as the dimensionality of the subspace increases.

For example, consider the dataset given in Table 1. From Figures 1, 2 and 3 shows the data clustering based on different attributes, i.e. age, salary and experience. Then the two clusters have been formed having 9 tuples/objects each. When data are clustered based on age and salary the two clusters in Figure 1 are split into 4 clusters. The number of objects in each cluster has reduced to either 4 or 5. When the data are clustered based on 3 attributes, i.e. age, salary and experience the database tuples/objects are split into 8 clusters with 2 to 3 objects each. So as the dimensionality of subspace increases the number of points/tuple/objects satisfying the constraints imposed by the subspace decreases and hence as the dimensionality of subspace increases the number of objects in subspace clusters decrease. So as subspace dimensionality increases density threshold should be lowered. By setting high density thresholds some of the interesting clusters are missed in high dimensionalities. Suppose if the density threshold is to 8 for sample data many interesting clusters will be missed in two and three dimensionality.



Figure 1 Clusters formed when the data is clustered based on age (1 dimension)



Figure 2 Clusters formed when the data in table 1 are clustered based on age and salary



Figure 3 Clusters formed when the data is table 1 is clustered based on age, salary and experience

Figure 4 shows the R code used to generate the figures 1.

path<-"D:\\Sample_Dataset.csv"; mydata<-read.csv(path,as.is=TRUE) stripchart(mydata[,2], xlab="age", method="jitter", col="blue", pch=16)

Figure 4 R code used to generate figure 1

Figure 5 shows the R code used to generate figure 2.

path<-"D:\\Sample_Dataset.csv"; mydata<-read.csv(path, as.is=TRUE) plot(mydata[,2],mydata[,4],pch=16,col="blue",xlab=" age",ylab="salary")

Figure 5 R code used to generate figure 2

path<-"D:\\Sample_Dataset.csv";		
mydata<-read.csv(path, as.is=TRUE)		
install.packages("scatterplot3d");		
library(scatterplot3d);		
<pre>scatterplot3d(mydata[,2],mydata[,3],mydata[,4],pch=</pre>		
16,xlab="age",ylab="salary",zlab="Experience",colo		
r="blue");		

Figure 6 R code used to generate figure 3

Figure 6 gives the R code used to generate figure 3.

3.4Grid based subspace clustering algorithms which use adaptive density thresholds

Support and Chernoff-Hoeffding bound-based interesting subspace mine (SCHISM) and Density conscious subspace clustering (DENCOS) are grid based subspace clustering algorithms that use adaptive density thresholds. These algorithms, lower the density threshold as the dimensionality of subspace increases. All subspaces of same dimensionality have same density threshold. SCHISM uses Chernoff-Hoeffdings bound for setting the density thresholds at different dimensionalities. SCHISM mines maximal interesting subspaces.

DENCOS uses adaptive density thresholds and uses a divide and conquer strategy for mining all dense subspace clusters at all dimensionalities. In the context of density divergence, apriori property does not hold and a lower dimensional subspace may not satisfy its density threshold, though it is a part of higher dimensional dense subspace. Thus, unlike maximal frequent pattern mining all of whose subpatterns are also frequent, a dense higher dimensional subspace does not imply that all subsets of it are dense. Since SCHISM mines only the maximal interesting subspaces, it may miss many interesting lower dimensional subspaces. DENCOS mines all interesting subspaces and the number of dense subspaces mined is very huge and requires post pruning. To overcome the drawbacks of both these algorithms the concept of closed interesting subspaces is proposed [6].

3.5Density based subspace clustering algorithms

Density-connected Subspace Clustering (SUBCLU) [12] is the pioneering density based subspace clustering algorithm. It uses density-based spatial clustering of applications with noise (DBSCAN) [13] cluster model at every possible subspace to identify subspace clusters. At any subspace 'S' a cluster is defined as a maximal set of density connected points. It takes two input parameters \mathcal{E} and m. An object with at least m points in its \mathcal{E} neighborhood is called a core 182

object. The density, connected core objects along with border objects form a cluster. Only the attributes/variables defining the subspace 'S' are used in calculating the distance between objects when finding clusters in subspace 'S'. The procedure is repeated for all possible subspaces of a multidimensional data space. SUBCLU uses the same density threshold at all subspace dimensionalities. So it may miss interesting subspace clusters at different subspace dimensionalities.

Indexing Subspace Clusters with in-process-removal of redundancY (INSCY) [14] and scalable densitybased subspace clustering [15] are two other efficient density-based subspace clustering algorithms. INSCY algorithm uses a novel index structure called SCYtree for efficient pruning of sparse and redundant subspace clusters during depth first enumeration of subspace clusters. The algorithm first mines higher dimensional subspace clusters before mine their lower dimensional counterparts. After mining a higher dimensional subspace cluster the algorithm steps back to mine lower dimensional subspace clusters in the same subspace region, at which time the redundant clusters are pruned. This is made possible by using SCY-tree which is a compact representation of potential subspace cluster regions.

Scalable density-based subspace clustering [15] uses a better first approach and a steering technique to directly identify interesting subspaces and their combinations. It thus identifies subspace clusters efficiently. Both INSCY and scalable density-based subspace clustering do not deal with the density divergence problem.

Filter-refinement based subspace clustering algorithm (FIRES) [16] is proposed to overcome the scalability and density divergence problems faced by most density based subspace clustering algorithms. It is a generic framework in that it works with all kinds of clustering notions. It starts with *1-D* clusters which can mined by using any clustering algorithm of choice. Then the *1-D* clusters are merged to generate approximations of subspace clusters. An optional refinement step is used to refine the cluster approximations obtained in the second step.

Dimensionality Unbiased Subspace Clustering (DUSC) [17] is a density-based subspace clustering algorithm which uses DBSCAN model for cluster discovery. It uses a density measure that is adaptive to the dimensionality. Due to the pruning strategies applied it may produce incomplete clustering results.

DerIving Subspace Hierarchies (DiSH) [18] is a density based subspace clustering algorithm that is used for constructing subspace cluster hierarchies.

4.3-D subspace clustering

3-D subspace clustering is used in applications when the data objects are homogeneous not only in a set of attributes, but also in a third dimension like time or location. A 3-D subspace cluster is either a set of objects that are homogenous in a set of attributes or a set of attributes that are homogeneous in a set of objects along a third dimension like time or location. The three dimensions constituting the 3-D subspace cluster are the set of objects, the set of attributes and the third dimension being time or location. TriCluster [19] is a 3-D subspace clustering algorithm designed to mine tri-clusters of 3D Microarray data. It mines coherent clusters along gene-sample-time (temporal) or gene-sample-region (spatial) dimensions. Crossgraph quasi-biClique (CGOB) algorithm [20] is proposed to mine stocks with similar financial ratio values across the years. CARSeeker [21] is a 3-D centroid based, actionable subspace clustering algorithm which is parameter insensitive and uses domain knowledge to mine actionable 3-D subspace clusters.

5.Enhanced subspace clustering

All of the subspace clustering algorithms discussed in the previous sections can be applied to quantitative data. Specialized subspace clustering algorithms are developed to handle complex data types and to improve cluster solutions. Specialized algorithms are designed to handle categorical data, noisy data, etc. A few subspace clustering algorithms incorporate domain knowledge and hence are semi supervised.

6.Subspace clustering versus projected pattern based and correlation clustering

Projected clustering also finds clusters which are sets of objects that are homogeneous in a set of attributes. The difference between subspace clustering and projected clustering is that projected clustering produces non-overlapping clusters. Pattern based clusters are similar to subspace clusters. The objects and attributes of a pattern based cluster exhibit a pattern. Objects and attributes are treated equally in a pattern based cluster, i.e. for example, objects having the same values for different attributes or attributes having the same values for all the objects can form a pattern based cluster. Correlation cluster is a set of objects whose values are positive and/or negatively correlated with a set of attributes.

7.Discussion

Different types of 2D subspace cluster algorithms are discussed in this paper. An overview of 3D subspace clustering, enhanced subspace clustering, projected clustering, pattern based clustering and correlation clustering is presented. Most of the 2D subspace clustering algorithms are designed for quantitative data. Grid based subspace clustering algorithms is more efficient and scalable compared to the density and window based algorithms. Grid based subspace clustering algorithms produces axis parallel clusters; if arbitrary shaped clusters are to be detected we have to use density based subspace clustering. Density divergence is inevitable in grid and density based subspace clustering so algorithms that set different density thresholds different at subspace dimensionalities can mine interesting subspace clusters at all subspace dimensionalities with little or no loss of subspace clusters. When data are in the form of a n-dimensional matrix, for example, geneexpression micro-array data, and consists of patterns or correlation relationships among object then pattern based and correlation clustering should be used.

8.Conclusions

This paper has discussed the concept of subspace clustering and different types of subspace clustering algorithms. Subspace clustering is essential to cluster high dimensional data sets to overcome the curse of dimensionality. Subspace clustering algorithms can be classified based on the type of data they handle, dimensionality of the cluster solution derived and approach taken. Different 2-D subspace clustering algorithms were discussed in this paper. 2-D subspace clustering algorithms that use the same global density thresholds face density divergence problem. To handle it density threshold should be lowered to increase in subspace dimensionality. In certain applications like Microarray data analysis and stock market data analysis 3-D subspace clustering has to be applied to find the objects that are homogeneous in time along with a subset of attributes. To generate a concise set of subspace clusters it is necessary to generate closed subspaces clusters. Though closed subspace clustering is developed in the context of density divergence for grid based subspace clustering algorithms, it has to be extended to density based and window based algorithms also. Subspace clustering can be used as a preprocessing step to reduce the dimensionality of datasets like temporal, spatiotemporal and multivariate time series for sequential pattern extraction from these datasets.

G.N.V.G. Sirisha et al.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

- [1] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.
- [2] Parsons L, Haque E, Liu H. Subspace clustering for high dimensional data: a review. ACM SIGKDD Explorations Newsletter. 2004; 6(1):90-105.
- [3] Sim K, Gopalkrishnan V, Zimek A, Cong G. A survey on enhanced subspace clustering. Data Mining and Knowledge Discovery. 2013; 26(2):332-97.
- [4] Sequeira K, Zaki M. SCHISM: a new approach to interesting subspace mining. International Journal of Business Intelligence and Data Mining. 2005; 1(2):137-60.
- [5] Dharmavaram VG, Mogalla S. A framework for context-aware semi supervised learning. Global Journal of Computer Science and Technology. 2014; 14(1):61-70.
- [6] Sirisha GNVG, Shashi M. Mining closed interesting subspaces to discover conducive living environment of migratory animals. In proceedings of the 4th international conference on frontiers in intelligent computing: theory and applications (FICTA) 2015 (pp. 153-66). Springer India.
- [7] Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications. ACM. 1998; 27(2):94-105.
- [8] Agrawal R, Srikant R. Fast algorithms for mining association rules. In proceedings of 14th international conference on VLDB 1994 (pp. 487-99).
- [9] Cheng CH, Fu AW, Zhang Y. Entropy-based subspace clustering for mining numerical data. In proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining 1999 (pp. 84-93). ACM.
- [10] Goil S, Nagesh H, Choudhary A. MAFIA: efficient and scalable subspace clustering for very large data sets. In proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining 1999 (pp. 443-52). ACM.
- [11] Chu YH, Huang JW, Chuang KT, Yang DN, Chen MS. Density conscious subspace clustering for highdimensional data. IEEE Transactions on Knowledge and Data Engineering. 2010; 22(1):16-30.
- [12] Kailing K, Kriegel HP, Kröger P. Density-connected subspace clustering for high-dimensional data. In proceedings of 4th international conference on data mining SDM 2004 (pp. 246-56).
- [13] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD 1996 (pp. 226-31).
- [14] Assent I, Krieger R, Müller E, Seidl T. INSCY: Indexing subspace clusters with in-process-removal of

redundancy. In eighth IEEE international conference on data mining 2008 (pp. 719-24). IEEE.

- [15] Müller E, Assent I, Günnemann S, Seidl T. Scalable density-based subspace clustering. In proceedings of the 20th ACM international conference on information and knowledge management 2011 (pp. 1077-86). ACM.
- [16] Kriegel HP, Kroger P, Renz M, Wurst S. A generic framework for efficient subspace clustering of highdimensional data. In fifth IEEE international conference on data mining (ICDM'05) 2005 (pp. 1-8). IEEE.
- [17] Assent I, Krieger R, Müller E, Seidl T. DUSC: dimensionality unbiased subspace clustering. In seventh IEEE international conference on data mining (ICDM 2007) 2007 (pp. 409-14). IEEE.
- [18] Achtert E, Böhm C, Kriegel HP, Kröger P, Müller-Gorman I, Zimek A. Detection and visualization of subspace cluster hierarchies. In international conference on database systems for advanced applications 2007 (pp. 152-63). Springer Berlin Heidelberg.
- [19] Zhao L, Zaki MJ. Tricluster: an effective algorithm for mining coherent clusters in 3d microarray data. In proceedings of the international conference on management of data 2005 (pp. 694-705). ACM.
- [20] Sim K, Liu G, Gopalkrishnan V, Li J. A case study on financial ratios via cross-graph quasi-bicliques. Information Sciences. 2011;181(1):201-16.
- [21] Sim K, Yap GE, Hardoon DR, Gopalkrishnan V, Cong G, Lukman S. Centroid-based actionable 3D subspace clustering. IEEE Transactions on Knowledge and Data Engineering. 2013; 25(6):1213-26.



G.N.V.G. Sirisha is an Assistant Professor in the department of CSE in S.R.K.R. Engineering College, India. She did her B.Tech.(CSE) and M.Tech.(CST) from the same college. She stood first in university level in M.Tech.(CST). She has recently submitted her Ph.D. thesis to Andhra

University. Her research interests include data mining, information retrieval, big data analytics. Email: sirishagadiraju@gmail.com



M. Shashi is a Professor and Chairperson of Board of Studies of the department of CS&SE, A.U. College of Engineering, Andhra University. She received the AICTE Career Award and best thesis prize from Andhra University in the year 1994. 12 Ph.D.'s were awarded under her guidance. She

has co-authored nearly 60 technical research papers in International Journals and 50 International Conferences and delivered many invited talks in such academic events. Her current research interests include Data warehousing and Mining, Artificial Intelligence, Soft Computing and Machine Learning.