ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

Exploring Constraints Inconsistence for Value Decomposition and Dimension Selection Using Subspace Clustering

K.Prema[1], K.Sangeetha[2], Dr.S.Karthik[3]

SNS College of Technology, Anna University Chennai, India[1, 2, 3]

Abstract: The datasets which are in the form of object-attribute-time is referred to as three-dimensional (3D) data sets. As there are many timestamps in 3D datasets, it is very difficult to cluster. So a subspace clustering method is applied to cluster 3D data sets. Existing algorithms are inadequate to solve this clustering problem. Most of them are not actionable (ability to suggest profitable or beneficial action), and its 3D structure complicates clustering process. To cluster these three-dimensional (3D) data sets a new centroid based concept is introduced in the proposed system called PCA. This PCA framework is introduced to provide excellent performance on financial and stock domain datasets through the unique combination of Singular Value Decomposition, Principle Component Analysis and 3D frequent item set mining.PCA framework prunes the entire search space to identify the significant subspaces and clusters the datasets based on optimal centroid value. This framework acts as the parallelization technique to tackle the space and time complexities.

Keywords— 3D supspace clustering, singular value decomposition, numerical optimization, financial data mining

I. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. The term is a buzzword, and is frequently misused to mean any form of large-scale data or information processing but is also generalized to any kind of computer decision support system, including artificial intelligence, machine learning, and business intelligence.

Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise. Data mining works to analyze data stored in data warehouses that are used to store that data that is being analyzed. Managers also use data mining to decide upon marketing strategies for their product.

Data mining interprets its data into real time analysis that can be used to increase sales, promote new product, or delete product that is not value-added to the company. Data mining interprets its data into real time analysis that can be used to increase sales and promote new product. Data mining mostly is used in decision making process which is also called Business Intelligence. Business related decision making is made using data mining techniques.

Calculating and Pruning homogeneous tensor using SVD

Mining CATs from high dimensional data and continuous valued tensor is a difficult and time consuming process. The main role is to remove regions that donot contain CATS.A simple solution to remove the values that are less than the threshold but it is difficult to know the right threshold. The proposed method is to efficiently prune the tensor in a parameter free way.

Calculating the probability value using augmented Lagrangian Multiplier Method

We use the homogeneous tensor value to calculate the probability of each value of the each data to be clustered with centroid.

Mining CATS using 3D closed pattern mining

After calculating the probability values then the high probability value is high as1.we efficiently mine the closed pattern to mine the CATS.

II. CONSTRAINT-BASED SUBSPACE CLUSTERING

An approach to deal with the high dimensional data. Clustering techniques are widely used unsupervised classification techniques to discover groupings of similar objects in data. However, when the dimensionality of the data become too high, usual criteria to define similarity between objects based on distance or density become irrelevant. Besides, some dimensions may be too noisy to clearly identify clusters in the original data. Subspace clustering techniques have been developed to overcome these problems. The idea, related to feature selection or dimension reduction, is to look for clusters in subsets of dimensions.

There is exist no constraint-based subspace clustering algorithms, in particular semi- supervised subspace algorithms, even though the integration of instance-level constraints in traditional clustering algorithms has proven to be successful and for another type of semi-supervised clustering. In this context, the aim of this work is to investigate how instance- level constraints can influence the subspace clustering process, making it not only more efficient but also more accurate. Towards this goal, the proposed method is to extend the common framework of bottom-up subspace clustering algorithms by integrating instance-level constraints into the mining process. The extended framework is able to consider several evaluation criteria (e.g. density, distance) in order to identify meaningful clusters in the data.

In high dimensional data, the general performance of traditional clustering algorithms decreases. This is partly because the similarity criterion used by these algorithms becomes inadequate in high dimensional space. Another reason is that some dimensions are likely to be irrelevant or contain noisy data, thus hiding a possible clustering. To overcome these problems, subspace clustering techniques, which can automatically find clusters in relevant subsets of dimensions, have been developed. However, due to the huge number of subspaces to consider, these techniques often lack efficiency.

It shows how this new framework can be applied to both density and distance- based bottom-up subspace clustering techniques. The experiments on real datasets show that instance-level constraints cannot only increase the efficiency of the clustering process but also the accuracy of the resultant clustering.

The subspace clusters (O,D) where O is a set of objects and D a subspace, i.e. a set of dimensions (attributes), are then iteratively merged to form higher-dimensional subspace clusters. The bottleneck of these algorithms is the NP-completeness of the enumeration of the subspace clusters. To make the bottom-up algorithms more efficient, strong constraints have to be pushed in the enumeration process to prune large parts of the search space.

III. DENSITY-CONNECTED SUBSPACE CLUSTERING FOR HIGH-DIMENSIONAL DATA

The monotonicity of density-connectivity is used to efficiently prune subspaces in the process of generating all clusters in a bottom up way. And then all small clusters are Automated analysis tools, there is an ever increasing need for efficient and effective data mining methods to make use of the information contained implicitly in that data. One of the primary data mining tasks is clustering which is intended to help a user discovering and understanding the natural structure or grouping in a data set. In particular, clustering is the task of partitioning objects of a data set into distinct groups (clusters) such that two objects from one cluster are similar to each other, whereas two objects from distinct clusters are not.

A lot of work has been done in the area of clustering. Nevertheless, clustering real-world data sets is often hampered by the so called curse of dimensionality since many real-world data sets consist of a very high dimensional feature space. In general, most of the common algorithms fail to generate meaningful results because of the inherent sparsity of the objects. In such high dimensional feature spaces, data does not cluster anymore. But usually, there are clusters embedded in lower dimensional subspaces. In addition, objects can often be clustered differently in varying subspaces.

Gene expression data is a prominent example: Microarray chip technologies enable a user to measure the expression level of thousands of genes simultaneously. Roughly speaking, the expression level of a gene is a measurement for the frequency the gene is expressed (i.e. transcribed into its mRNA product).The expression level of a gene allows conclusions about the current amount of the protein in a cell the gene codes for. Usually, gene expression data appears as a matrix where the rows represent genes, and the columns represent samples (e.g. different experiments, time slots, test persons, etc.). The value of the i-th feature of a particular gene is the expression level of this gene in the i-th sample.

IV. MINING ACTIONABLE SUBSPACE CLUSTERING

The Subspace clusters represent useful information in high-dimensional data. However, mining significant subspace clusters in continuous-valued 3D data such as stock-financial ratio-year data, is difficult. Firstly, typical metrics either find subspaces with very few objects, or they find too many insignificant subspaces those which exist by chance. Besides, typical 3D subspace clustering approaches abound with parameters, which are usually set under biased assumptions, making the mining process a 'guessing game'.

Three-dimensional (3D) data, in the general form of object-attribute-time/location has become increasingly popular in data analysis. Many real-world applications, such as stock analysis based on stock-financial ratio-year data, basically cluster the continuous 3D data to perform the task. However, because these data are essentially high dimensional, traditional clustering approaches operating on the full data space become ineffective.

The problem to cluster subspaces in the 3D data is solved easily. The objects are grouped based upon their similarity in some subset of attributes and time. In such formulations, a 3D subspace cluster can be considered as a cuboid spanned by a group of objects, a group of attributes and a group of timestamps. This cuboid is inherently axis- parallel, which is important for the user to easily interpret and understand the cluster.

V. DISCOVERING CORRELATED SUBSPACE CLUSTERS IN 3D CONTINUOUS-VALUED DATA

The Subspace clusters represent useful information in high-dimensional data. However, mining significant subspace clusters in continuous-valued 3D data such as stock-financial ratio-year data, is difficult. Firstly, typical metrics either find subspaces with very few objects, or they find too many insignificant subspaces those which exist by chance.

Besides, typical 3D subspace clustering approaches abound with parameters, which are usually set under biased assumptions, making the mining process a 'guessing game'. Information theoretic measure is introduced to group the datasets, which allows us to identify 3D subspace clusters that stand out from the data. And a highly effective, efficient and parameter-robust algorithm, which is a hybrid of information theoretical and statistical techniques, to mine these clusters is introduced here.

Three-dimensional (3D) data, in the general form of object-attribute-time/location has become increasingly popular in data analysis. Many real-world applications, such as microarray analysis based on gene-sample-time or gene- sample-region data, and stock analysis based on stock- financial ratio-year data, basically cluster the continuous 3D data to perform their task. However, because these data are essentially high dimensional, traditional clustering approaches operating on the full data space become ineffective.

Hence, it is the user who determines the results, based upon his/her biased assumptions. Besides, these algorithms typically abound with parameters, thereby increasing the burden on the user. For example, due to the complex nature of 3D continuous- valued data, the pioneering work in 3D subspace clustering requires a total of 7 parameter settings.

VI. DISTANCE BASED SUBSPACE CLUSTERING WITH FLEXIBLE DIMENSION PARTITIONING

Clustering seeks to find groups of similar objects based on the values of their attributes. Traditional clustering algorithms use distance on the whole data space to measure similarity between objects. As the number of dimensions in a dataset increases, distance measures become increasingly meaningless. In very high dimensional datasets, the objects are almost equidistant from each other. This is known as the curse of high dimensionality.

The concept of subspace clustering has been proposed to cope with the problems caused by high dimensionality by discovering clusters embedded in subspaces of high dimensional datasets. Many subspace clustering algorithms use a grid and density based approach. They partition the data space into non-overlapping rectangular cells by discretizing each dimension into a number of bins. A cell is dense if the fraction of total objects contained in the cell is greater than a threshold. Clusters are formed by merging connected dense cells in the same subspace.

VII. MINING ACTIONABLE SUBSPACE CLUSTERS

The Subspace clusters represent useful information in high-dimensional data. However, mining significant subspace clusters in continuous-valued 3D data such as stock-financial ratio-year data, is difficult. Firstly, typical metrics either find subspaces with very few objects, or they find too many insignificant subspaces those which exist by chance. Besides, typical 3D subspace clustering approaches abound with parameters, which are usually set under biased assumptions, making the mining process a 'guessing game'.

Information theoretic measure is introduced to group the datasets, which allows us to identify 3D subspace clusters that stand out from the data. And a highly effective, efficient and parameter-robust algorithm, which is a hybrid of information theoretical and statistical techniques, to mine these clusters is introduced here.

Three-dimensional (3D) data, in the general form of object-attribute-time/location has become increasingly popular in data analysis. Many real-world applications, such as stock analysis based on stock-financial ratio-year data, basically cluster the continuous 3D data to perform the task. However, because these data are essentially high dimensional, traditional clustering approaches operating on the full data space become ineffective.

The problem to cluster subspaces in the 3D data is solved easily. The objects are grouped based upon their similarity in some subset of attributes and time. In such formulations, a 3D subspace cluster can be considered as a cuboid spanned by a group of objects, a group of attributes and a group of timestamps. This cuboid is inherently axis- parallel, which is important for the user to easily interpret and understand the cluster.

VIII. MINING FREQUENT CLOSED CUBES IN 3D DATASETS

The concept of frequent closed cube (FCC), which generalizes the notion of 2D frequent closed pattern to 3D context. Two novel algorithms to mine FCCs from 3D datasets is introduced.The first scheme is a Representative Slice Mining (RSM) framework that can be used to extend existing 2D FCP mining algorithms for FCC mining. The second technique, called CubeMiner, is a novel algorithm that operates on the 3D space directly. In 3D context the frequent closed pattern is referred as frequent closed cube (FCC). Even in the traditional 'market-basket' analysis, it is not uncommon to have consumer information on a number of dimensions.

For example a number of dimensions, in the region-time-items data simply stores the sales of itemsets in certain locations over certain time periods. This trend motivates us to extend existing 2D frequent closed pattern analysis to 3D context. In 3D context the frequent closed pattern is referred as frequent closed cube (FCC).

The problem of mining FCC from 3D datasets is solved by RSM. First, the notion of FCC is introduced and formally it is defined. Second, two approaches to mine FCCs is proposed. The first approach is a three-phase framework, called Representative Slice Mining algorithm (RSM) that exploits 2D FCP mining algorithms to mine FCCs.

IX. CONCLUSION

The proposed method called PCA is applied to the large amount of datasets. The Principle Component Analysis is the optimization and parallel methodology which is used to obtain the best clustering results. In PCA the clustering is made based on the centroid value. Since PCA is the optimization technique it is used to find the optimal centroids based on the velocity of the particle. The centroid value for each iteration is updated using particle's velocity. PCA is the parallel methodology it is used to reduce the time and space complexities. This framework can be applied to both real-world and synthetic datasets. The PCA framework can work well with the increasing data sizes with increased cluster quality and with minimal time and space requirement.

REFERENCES

1. G.Liu, J. Li, K. Sim, and L. Wong, "Efficient Mining of Distance-Based Subspace Clusters," Statistical Analysis Data Mining, vol. 2,nos. 5/6, pp. 427-444, 2009.
2. G. Moise and J. Sander, "Finding Non-Redundant, Statistically Significant Regions in High Dimensional Data: A Novel Approach to Projected and Subspace Clustering," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 533-541, 2008.
3. H.-P. Kriegel, P. Kroger, and A. Zimek, "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering," ACM Trans Knowledge Discovery from Data, vol. 3, no. 1, pp. 1-58, 2009.
4. H.P.Kriegel et al.,"Future Trends in Data Mining," Data Mining Knowledge Discovery, vol. 15, no. 1, pp. 87-97, 2007.
5. K. Sim, A.K. Poernomo, and V. Gopalkrishnan, "Mining Actionable Subspace Clusters in Sequential Data," Proc. SIAM Int'l Conf. Data Mining (SDM), pp. 442-453. 2010.
6. Kailing K., Kriegel H.P., Kroger P., and Wanka S. (2003) "Ranking Interesting Subspaces for Clustering High Dimensional

Data,"Proc. Practice of Knowledge Discovery in Databases (PKDD), pp. 241- 252.

7. L. Ji, K.L. Tan, and A.K.H. Tung, "Mining Frequent Closed Cubes in 3D Data Sets," Proc. 32nd Int'l Conf. Very Large Databases (VLDB), pp. 811-822, 2006.
8. E. Fromont, A. Prado, and C. Robardet,"Constraint-Based Subspace Clustering," Proc. SIAM Int'l Conf. Data Mining (SDM),pp.26-37, 2009.
9. P. Kroger, H.P. Kriegel, and K. Kailing, "Density-Connected Subspace Clustering for High-Dimensional Data," Proc. SIAM Int'l Conf. Data Mining (SDM), pp. 246-257, 2004.
10. Fu Q. and Banerjee A. (2009) "Bayesian Overlapping Subspace Clustering," Proc. IEEE Ninth Int'l Conf. Data Mining (ICDM), pp. 776-781.