International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 8, October - 2012

Projected Clustering Algorithms – A Review

Ilango Murugappan[1]
[1]*Professor, Department of Computer Applications,*
K L N College of Engineering,
Pottapalaym- 630611.
Sivagangai District, Tamilnadu, India

Dr Mohan Vasudev[2]
[2]*Professor and Head, Department of Mathematics,*
Thiagarajar College of Engineering,
Madurai, Tamilnadu, India-625015

Abstract

Clustering is an important task in Data Mining. It is an unsupervised procedure. Clustering is a process which partitions a collection of objects S into a set of groups (Clusters) so that the similarity of the objects in the same group is high and objects from different groups are dissimilar. Finding clusters and their relevant attributes from a data set is known as projected clustering. Large number of Projected Clustering techniques have emerged whose task is to find the i) set of clusters C and ii) for each cluster C_i, the set of dimensions D_i that are relevant to C_i . In this paper we have analyzed some of the projected clustering algorithms. The purpose of this research paper is to analyze the performance issues of projected clustering algorithms such as PCKA (Projected Clustering based on K-Means Algorithm), EPCH (Efficient Projective Clustering by Histograms), EPPC (Emerging Pattern Based Projected Clustering for Gene Expression data).

Key Terms: Data Mining, Projected Clustering, outliers

IJERT

1. Introduction

Partitioning a set of objects into homogeneous clusters is a fundamental operation in Data Mining. The clustering problem has been discussed extensively in the database literature as a tool for similarity search, customer segmentation, pattern recognition, trend analysis and classification. Clustering plays an outstanding role in various data mining applications such as spatial database applications, information retrieval and text mining, medical diagnostics, computational biology and many others.

Clustering data is required in many disciplines and has many applications. Important goal of cluster analysis is the identification of a finite set of categories, classes, groups (clusters) in the data set. We can mention the following basic requirements for clustering techniques for large data files: *scalability* (clustering techniques must be scalable, both in terms of computing time and memory requirements), *independence of the order of input* (i.e. order of objects which enter into analysis) and *ability to evaluate the validity of produced clusters*. Clustering algorithms usually employ a distance metric (e.g Euclidean) or a similarity measure in order to partition the database so that the data points in each partition are more similar than the points in other partitions. Clusters may exist in different sub spaces comprised of different combinations of attributes.

The main purpose of projected clustering is to identify a set of clusters and their relevant dimensions [1]. For each cluster, a projected clustering algorithm determines a set of attributes that it assumes to be most relevant to the cluster. Such attributes are called as relevant attributes and all other attributed are called as irrelevant attributes of the data set. Let us now consider some of the terms and notations used in projected clustering: Given a data set with N objects and a set V of d dimensions, a projected cluster C_i contains N_i member objects, and is defined by a set V_i of d_i dimensions. We will call the dimensions in V_i the relevant dimensions of cluster C_i, and the ones in $V - V_i$ the irrelevant dimensions

of it. The sub space formed by two set of dimensions will be called the relevant sub space and irrelevant sub space of C_i respectively [2]

2. PCKA (Projected Clustering based on K-Means algorithm)

Clustering algorithms uses a distance metric such as Euclidean to measure the similarity between objects. In the case of high dimensional data, the concept of similarity between the objects will not helpful. It is not effective to differentiate the data objects based on a distance or a similarity measure computed using all dimensions.

Feature selection techniques such as Principal Component Analysis are used as preprocessing step for clustering. Sometimes traditional feature selection techniques may lead to substantial loss of information [3].

A projected cluster is a subset SP of data points, together with sub space SD of dimensions, such that the points in SP are closely clustered in SD [3]. PCKA is able to detect clusters of low dimensionality which are embedded in high dimensional data. PCKA avoids distance calculation in full dimensional data. PCKA has 3 phases 1) Attribute Relevance Analysis 2) Outlier Handling 3) Discovery of Projected Clusters.

Attribute Relevance Analysis: The aim of this phase is to identify the relevant dimensions which exhibit some cluster structure by discovering dense regions. Irrelevant dimension may have noise / outliers and sparse data points. So, these irrelevant dimensions must be identified and removed. A binary matrix is obtained which contains the information whether each data point falls into a dense region of an attribute. By detecting dense regions in each dimension, PCKA identifies relevant dimensions.

Outlier Detection: Outliers can be defined as set of data points that are considerably dissimilar, exceptional or inconsistent with respect to the remaining data [4]. PCKA makes use of Jaccard coefficient in order to find the outliers from the binary matrix and the corresponding data object is removed from the data set.

IJERT

Discovery of Projected Clusters: It contains 2 steps. In the first step clusters are discovered. In the next step, relevant dimensions of the clustered are selected. 1) In the first step, K-Means algorithm is used to find the clusters and distance is calculated based on sub set of dimensions where object values are dense. 2) Relevant dimensions are selected for the obtained clusters.

Performance Issues

i) PCKA scales quadratically with increase in data size.
ii) PCKA scales linearly with the increase in the data dimensions
iii) PCKA is able to achieve highly accurate results in different situations involving data sets with different characteristics
iv) In terms of clustering quality for the data sets SCGE and MF, accuracy of clusters produced by PCKA is far better than HARP, PROCLUS, FASTDOC and SSPC
v) When the average cluster dimensionality is very low, only PCKA yields acceptable results.
vi) In the presence of outliers, PCKA performs well compared to other projected clustering algorithms

3. EPCH (Efficient Projective Clustering by Histograms)

EPCH, proposed by Eric Ka Ka Ng et al [5] is able to detect projected clusters in high dimensional data. It does not require the users to input the average dimensionality of associated sub spaces as well as the number of clusters. It requires only one input from the user max-no-cluster. This input max-no-cluster represents the maximum number of clusters that the user wants to uncover from the given data set. Clusters of varying densities and/or varying dimensionalities can be easily handled by EPCH.

There are some tuning parameters which are used to improve the clustering quality

i) for each projection domain, an upper bound f of the spread of a cluster.
ii) To speed the final clustering process, a value k.
iii) Expected noise / outlier percentage

The sub space corresponding to a projected cluster is called associated sub space and the dimension that are being included in the associated sub spaces are called bounded dimensions and the other dimensions are called as unbounded dimensions.

Each histogram is related to a sub space. For example, if there are three attributes A, B, and C in our data set then 2 dimensional histograms can be generated. The corresponding sub spaces will be AB, BC and AC and these sub spaces will be denoted as S_1, S_2 and S_3.

In EPCH, d-dimensional histograms are constructed. This d may vary from 1 to n which depends on cluster quality and running time. If d is set to 1 then the implementation is called EPC1 which uses 1 dimensional histogram. If d is set to 2 then the implementation is called EPC2 which uses 2 dimensional histograms.

There are five phases in EPCH algorithm

i) **Histogram Building Phase**: In this phase, histograms are built. Each histogram corresponds to one d-dimensional space. ii) **Dense Region Detection Phase**: This phase identifies all dense regions iteratively. iii) **Signature List construction phase:** Based on the id of the dense regions for each histogram, a signature is derived for each data object. Derived sub space is found by the union of all the bounded sub spaces iv) **Merging similar sub spaces phase**: All similar sub spaces are merged in this phase. v) **Membership Degree assigning phase:** Clusters are discovered and data objects are associated with degree of membership.

Performance Issues

The total time complexity is $O(ND^d + mHD^d + N(l^d + D^d) + N \log(km) + (km)^2 l^2 + Nml^d)$ where N be the number of data points, m = max-no-cluster, D be the number of dimensions, d be the dimensionality of histograms, H be the number of bins in each histogram, l be the average dimensionality of the associated sub spaces of clusters, and km be the number of signatures kept in the merging step.

IJERT

Comparison of EPC1 (EPCH with 1-d histograms), EPC2 (EPCH with 2-d histograms) with PROCLUS and ORCLUS

i) EPC1 produces more accurate results than PROCLUS
ii) EPC1 is much faster than PROCLUS
iii) Accuracy of EPC2 does not vary much with different dimensionalities of the original space and different dimensionalities of associated spaces.
iv) When the number of data points scales linearly, EPC1 is extremely fast compared to ORCLUS and EPC2

4. EPPC (Emerging Pattern Based Projected Clustering for Gene Expression Data)

Gene Expression data is usually high dimensional in nature. Gene expression data can't be manipulated easily. It is also not easy to understand gene expression data. Larry T H Yu et al [6] in their algorithm EPPC said that the emerging patterns and projected clustering techniques can be integrated to improve quality of projected clusters. The main aim of EPPC algorithm is to make use of strong discriminatory of powers of emerging patterns in the dimension projection process.

Emerging pattern (EPs) are specific patterns that has got significant values in different partition of the data set. In gene expression data, emerging pattern may exist only in cancerous tissues and may not exist in normal tissues. EPs are easy to understand and have strong discriminatory powers.

Emerging patterns (EPs) were introduced initially by Dong Li [7]. They are defined as item sets whose values increases from one data set to another data set more than the threshold value called the

growth rate (ρ). Jumping emerging patterns (JEPs) and plateau EPs are the classification of emerging patterns. Each category has different properties and they can be applied to different problems.

Three phases of EPPC algorithm

i) **Initialization Phase.** The user has to give the input for number of clusters to be formed. Initially some random seeds K_0 are picked up from the data set where K_0 is larger than K (number of clusters). All dimensions are selected as projected dimensions for the initial seeds.

ii) **Iterative Phase:** In order to detect the best clusters, the main aim of this phase is to improve the quality of the cluster seeds iteratively. Three operations are performed in this iterative phase. a) **Assignment Operation**. As the name implies, the data points are assigned to the closest seed. City block distance or Euclidean distance metric is used to determine the distance between the data point and the cluster seed under the projected dimensions. The centroid of each partition formed is evaluated and they are used as the new seeds in the next iteration. b) **Dimension Projection Operation**: For each projected clusters, the projected dimensions are evaluated by the own data points. For each partition, the embedded emerging patterns are found. The emerging patterns with most frequent occurrences are selected as projected dimensions. C) **Merging Operation**: The closest pair of clusters are merged together to form a new cluster.

iii) **Refinement Phase:** Final clusters are formed by refining the cluster seeds obtained from the iterative phase.

Performance Issues

i) EPPC is suitable and efficient algorithm to find projected clusters in Gene Expression data which always consists of large number of numerical attributes, i.e. gene expression values, but limited records.

ii) It makes use of domain knowledge and patterns generated from different classes in the gene expression data to form clusters.

IJERT

5. Conclusion

Large number of projected clustering algorithms has been evolved. We have analyzed Projected Clustering algorithms such as PCKA, EPCH and EPPC and also their performance issues are also analyzed.

6. Acknowledgements

I would like to thank my colleague Prof. Ashok Baburaj for his help and support in this work.

7. References

[1] Kevin Y Yip, David W Cheung, Micheal K Ng, "A highly-usable Projected Clustering Algorithms for Gene Expression profiles", BIOKDD03: 3rd ACM SIGKDD workshop on Data Mining in Bio Informatics, 2003.
[2] Kevin Y Yip, David w Cheung, Micheal K Ng, "HARP: A Practical Projected Clustering Algorithm", IEEE Transactions on Knowledge and Data Engineering, Vol 16, No.11, November 2004.
[3] Aggarwal C. C. Procopiuc, C. Wolf J.L.. Yu P.S. and Jong S.P. (1999), " Fast algorithms for projected clustering", ACM SIGMOD International Conference on Management of Data, 61-72.
[4] J.Han, M.Kamber, "Data Mining, Concepts and Techniques", Morgan Kaufman, 2001.
[5] Eric Ka Ka Ng. Ada Wai-chee Fu. and Raymond Ch-Wing Wong, "Projective Clustering by Histograms"
[6] Larry T H Yu, Fu-Lai Chung, Stephan C F Chan, "Proceedings of European workshop on Data Mining and Text Mining for Bio Informatics, held in conjunction with ECML/PKDD", Dubrovnik, Croatia, 2003, "Emerging Pattern Based Projected Clustering for Gene Expression Data"
[7] G Dong, J Li, "Efficient Mining of emerging patterns: Discovering trends and differences. Proceedings of the fifth ACM SIGKDD International conference on knowledge discovery and Data Mining", San Diego, California, United Sates, 1999. 43-53.