The Mining of the Real Valued Datasets using Clustering Algorithms: Review Paper

Preshita S. Mahiskar  
M.Tech,  
Department Computer Science & Engg  
Govt. College of Engg, Amavati(M.S.) India-444604.  
preshita1808@gmail.com

Prof. A. W. Bhade  
Assistant Professor,  
Department Information Technology, Govt. College of Engg, Amavati(M.S.) India - 444604.

Dr. P. N. Chatur  
Head of Department  
Computer Science & Engg, Govt. College of Engg, Amavati(M.S.) India - 444604.

Abstract

Cluster analysis has been widely used in several disciplines, such as statistics, software engineering, biology, psychology and other social sciences, in order to identify natural groups in large amounts of data. These data sets are constantly becoming larger, and their dimensionality prevents easy analysis and validation of the results. The subspace pattern mining has been tailored to microarray data clustering to find biclusters and triclusters. We focus on deterministic clustering algorithm: Triclusters, which can mine arbitrarily positioned and overlapping biclusters/triclusters. Depending on different parameter values, they can mine different types of clusters, including those with constant or similar row/column values, as well as scaling and shifting expression patterns. We also give a useful set of metrics to evaluate the clustering quality, and show their effectiveness on real data.

1. Introduction

Data Mining, which is a process of nontrivial extraction of implicit, previously unknown and potentially useful information [1], can be employed to find out the valuable information. Currently, the data in manufacturing industry is really huge and consists of various stages. However, because of the insufficient information many problems especially in the operation management could not be solved efficiently. Data mining is an analytic process designed to explore data in search for consistent pattern and relationship [1]. Typically, data mining techniques are already used in business, medical and various domains.

As a result of modern methods for scientific data collection, huge quantities of data are getting accumulated at various databases. Such data banks are growing so rapidly that it is practically difficult to extract useful information from them by using conventional database techniques. Effective and efficient algorithms for data mining are necessary to unravel implicit information from huge databases.

Clustering is the process of grouping a set of objects into classes with maximum intra-class similarity and minimum inter-class similarity [2]. This similarity criterion is based upon the entire set of attributes. For example two objects will be considered similar to each other if they exhibit strongly co-related values over the entire set of attributes. This method restricts us to finding global patterns only and leaves out any chance of finding local patterns where two objects may be similar to each other based upon only a subset of attributes. This subset similarity criterion has been named as biclustering, co-clustering or block clustering [5].

Most of the biclustering research has been focused on model based solutions. Biclustering is used for discovering correlations among subsets of attributes with subsets of transactions in a transaction database. It has an extensive set of applications ranging from Gene co-regulation analysis [4], document keyword clustering and collaborative filtering for online recommendation systems [3].

The concept of Triclusters has been investigated recently in the context of two relational datasets that share labels along one of the dimensions [6]. By simultaneously processing two datasets to unveil triclusters, new useful knowledge and insights can be obtained.

2. Clustering Techniques

Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning. Clustering is one of the most popular approaches for analyzing data and has proved to be successful in many applications such as classification and prediction. Traditional clustering method are hierarchical clustering, k-means algorithm, and self organizing map [2].
Clustering is distinguished from pattern recognition or the areas of statistics known as discriminant analysis and decision analysis, which seek to find rules for classifying objects from a given set of preclassified objects.

2.1 Biclustering

There are many biclustering techniques, based on Graph theory for example, SAMBA, on Fuzzy and Possibilistic approaches [5], and many other algorithms. The aim of any biclustering algorithm for a given data matrix is to search for sub-matrices that are tightly co-related based upon some scoring criterion. Note that the identified sub-matrices need not be disjoint or cover the entire row or column. Instead a biclustering algorithm builds a wide variety of sub-matrices that will capture all the significant information in the data. Let us consider a data matrix, , with set of rows and set of Columns . For this matrix, a cluster of rows is a subset of columns that show similar behavior across the set of all rows. A cluster of rows is a subset of rows that exhibit similar behaviors across all columns of the matrix. On the other hand, a bicluster is a subset of rows that exhibit a similar behavior across a subset of columns and vice versa.

We use subspace clustering, biclustering (a clustering process of simultaneously mining column and row for a 2D real-valued data matrix) to mine the subspace patterns (biclusters) from real-valued datasets [6]. Traditional clustering algorithms work in the full dimensional space, which consider the value of each point in all the dimensions and try to group the similar points together.

![Figure 1. Difference between (a) Fullspace Clustering and (b) Biclustering](image)

For example, in Figure 1 (a), if points , and are similar in the whole space, they can be clustered together. Biclustering, however, does not have such a strict requirement. If some points are similar in several dimensions (a subspace), they will be clustered together in that subspace. For example, in Figure 1 (b), if points , and are similar in the subspace composed of dimensions , , and , they form a bicluster. Biclustering is very useful, especially for clustering in a high dimensional space where often only some dimensions are meaningful for some subset of points [6].

A. Bi-clustering: definition and notations

A matrix can be defined as a subset of elements that represent similar activity patterns with regard to a subset of features. Let be a matrix denoted by having rows and columns . In this notation, we denote as representing the sub-matrix of that contains only the elements that belongs to the subsets of rows and columns .

Definition: A bi-cluster of elements of matrix is an where is the subset of rows and is the subset of columns.

Bi-clusters can thus be seen as sub-matrices of a matrix representing features of elements. Using the notation above, the bi-clustering problem can be defined as follows: given a matrix , compute a family of bi-clusters such that each bi-cluster satisfies certain “homogenous” properties. Similarly, as any clustering technique, the objective of bi-clustering is to compute/identify clusters of data grouped according to their similarities [7]. In the case of bi-clustering the data is represented in a matrix in which rows are features of the elements of the sample under study represented by the columns of the matrix.

2.2 Triclustering

The biclustering itself is known to be a NP-hard problem, and thus many proposed algorithms of mining biclusters use heuristic methods or probabilistic approximations, which as a tradeoff decrease the accuracy of the final clustering results [7]. Tricluster can mine arbitrarily positioned and overlapping clusters, and depending on different parameter values, it can mine different types of clusters, including those with constant or similar values along each dimension.

Another difficulty arises from the fact that individual biclusters, if found in local contexts only, would be maximal in the number of attributes and rows; but a Low-Variance 3-Cluster may be formed by using non-maximal local biclusters which contain only subsets of attributes forming the locally optimal biclusters, and also consequently, larger number of rows, increasing chances of larger sets of shared rows [8].

Here present a novel, efficient, deterministic, triclustering method called triclusters [9], which having the key features of approach include: 1) It mines only the maximal triclusters satisfying certain homogeneity criteria.
2) The clusters can be arbitrarily positioned anywhere in the input data matrix and they can have arbitrary overlapping regions.
3) The use of a flexible definition of a cluster which can mine several types of triclusters, such as triclusters having identical or approximately identical values for all dimensions or a subset of the dimensions, and triclusters that exhibit a scaling or shifting expression values.
4) Triclusters is a deterministic and complete algorithm, which utilizes the inherent unbalanced property in real valued datasets, for efficient mining.
5) Triclusters can optionally merge/delete triclusters that have large overlaps, and can also automatically relax the similarity criteria. It can thus tolerate some noise in the dataset, and lets the user focus on the most important clusters.
6) It present a useful set of metrics to evaluate the clustering quality, and we show that tricluster can find substantially significant triclusters in the real valued datasets.

Thus the Triclustering algorithm introduces ideas underlying the design of standard deviation threshold and the accompanying search algorithm which accommodates datasets with different sizes and/or different distributions of values [9].

A 3-Cluster T, represented as T = \{B1\|B2\} is a pair of Biclusters B1, B2 taken from two datasets D1 and D2 respectively; the two component biclusters are such that |r1\cap r2| is the number of data objects shared between the two biclusters and it is greater than zero.

A Low-Variance 3-Cluster is a 3-Cluster in which the standard deviation of each component bicluster is less than or equal to some pre-defined threshold for the dataset from which the bicluster is taken. That is, sB1 ≤ s1 and sB2 ≤ s2. And the number of rows and columns contained in each component bicluster are equal to or more than the pre-defined thresholds.

A bicluster B = \{r, c\} contains minimum number of data objects in a data table D = \{R, C\} if:

\[|r| ≥ α * |R| \text{ and } |c| ≥ β * |C| \text{ (α, β ∈ (0, 1))}\]

Where \(a\) and \(b\) are pre-defined size ratio.

In algorithm, for each candidate bicluster, we compare their Range values (difference between the maximum and minimum value of the biclusters) with a pre-specified Range threshold. The Range of a component biclusters is:

\[ΔB = \max_{i,j∈B} d_{ij} - \min_{i,j∈B} d_{ij}\]

Where \(\max_{j\neq j} d_{ij}\) is the maximum value of the bicluster and \(\min_{j\neq j} d_{ij}\) is the minimum value.

A Low-Variance 3-Cluster is a 3-Cluster when each of its component bicluster’s range is less than or equal to the background Range threshold (uniform distribution Range) and contains minimum required number of rows and columns [10]. That is:

\[Δ_{B1} ≤ C_{D1} * α * |R| * β * |C|\]
\[Δ_{B2} ≤ C_{D2} * α * |R| * β * |C|\]
\[|r1| ≥ α * |R| |c1| ≥ β * |C|\]
\[|r2| ≥ α * |R| |c2| ≥ β * |C|\]

Where \( α \) and \( β \) are pre-defined size ratios.

The search algorithm starts to examine bicluster candidates for Low-Variance 3-Clusters from one dataset and then finds corresponding biclusters in the second dataset [10]. The algorithm employs a number of pruning strategies to make the algorithms more efficient. Generally, our algorithm conducts heuristic beam search in two prefix-tree defined search spaces, which are the same as the lattices of all possible biclusters in the two datasets.

3. Conclusion

The results of the algorithm from both datasets are very promising and demonstrate the value of our algorithm and its design. There are no other known algorithms that find the Low-Variance 3-Clusters in real-valued datasets. Therefore, compare the results to some of the biclustering algorithms and demonstrate the overall results.

Algorithms for generating triclusters whose cell-values demonstrate simple well known statistical properties, such as upper bounds on standard deviations, are needed for many applications.

The review show the 3-clusters in which the variance is kept under control while the overlap between the rows shared between the two components of a 3-Clusters is maximized reviewing the various biclustering algorithm such as CC, SAMBA, Co-Clustering.

References

[1] Jiawei Han and M. kamber “,” Data mining: Concepts and Techniques,” Morgan Kaufmann, 2004
18th IEEE International Conference on Data Engineering, pages 517–528, 2002.


