International Journal of Management, IT & Engineering

CONTENTS

Chief Patron

Editorial Board

Reviewers

Title

A STUDY ON MBC ALGORITHM WITH GOODNESS FUNCTION

Author(s)

P. Usha Madhuri
Research Scholar,
Dr.M.G.R University,
Chennai, India

Dr.S.P. Rajagopalan
Emeritus Professor,
Dr.M.G.R University,
Chennai, India

Introduction:

In Data Mining, clustering is one of the efficient techniques used to extract useful information from large quantities of data. A cluster is a collection of data objects relatively similar to one another in some respect and relatively dissimilar to the objects in other clusters. Clustering analysis is an important technique in data mining. It is a process of grouping a set of physical or abstract objects into classes of similar objects. Clustering can be viewed as unsupervised classification.

Matrix Based Clustering (MBC) is a hierarchical clustering method with a goodness function based on notions of bond and inner bond that in turn involve direct and indirect link measures. MBC employs an operation close to matrix multiplication, but with a little modification. The matrix base is thought helpful for calculations on advanced parallel machinery. The goal is to achieve good clustering performance relative to other clustering methods. Real data is tested on MBC to provide useful classification information.

Matrix Based Clustering algorithm (MBC) measures the “bond" of two clusters based on a goodness function which is computed via matrix manipulation that utilizes not only direct links but also indirect links between two clusters. The effectiveness of MBC is demonstrated with several data sets that contain points in 2D space, a couple of which cannot be captured by other methods such as OPTICS, CHAMELEON, or Matlab Fuzzy Clustering.

Cluster’s Quality:

Good clusters show high similarity within a group and low similarity between clusters. The quality of a clustering result typically is assessed not only by some mathematical measures but the ability to discover hidden patterns. Evaluation of clustering algorithms typically uses criteria such as efficiency and effectiveness. With increasing data size efficiency is very important. The quality of the clustering result is even more important and it is generally more difficult to improve.

Classical approaches to clustering include partitioning methods such as k-means, hierarchical clustering, density-based approaches and graph-based algorithms. There is also a variety of soft clustering techniques, such as those based on fuzzy logic or statistical mechanics. In these cases,

a data point may belong to multiple clusters with different degrees of membership. Hierarchical agglomerative clustering methods are very popular with prominent versions such as single-link (SLINK)], CHAMELEON, BIRCH, CURE.

MBC: Matrix Based Clustering:

Matrix Based Clustering Method is a novel hierarchical agglomerative clustering algorithm that measures the similarity of two clusters using a new bond" concept "Bond" is computed according to operations on the similarity matrix and counts not only direct link between two clusters but also an indirect link between two clusters.

The effectiveness of MBC is analysed with several data sets containing points in 2-D space and clusters of different shape, density, size, noise and artifacts. A principal view adopted in MBC is that improved clustering quality can be achieved through exploiting commonalities among existing clustering methods, e.g., considerations relating to merging clusters and criteria for it. Several commonalties discussed in this chapter includes single link merging (SLINK, OPTICS), edge cut merging (CHAMELEON, ROCK), and criteria based on the square of the adjacency matrix (OPTICS, ROCK). In response, some comparative information is analyzed to uncover related clustering methods. This serves as background for MBC, a new hierarchical clustering algorithm.

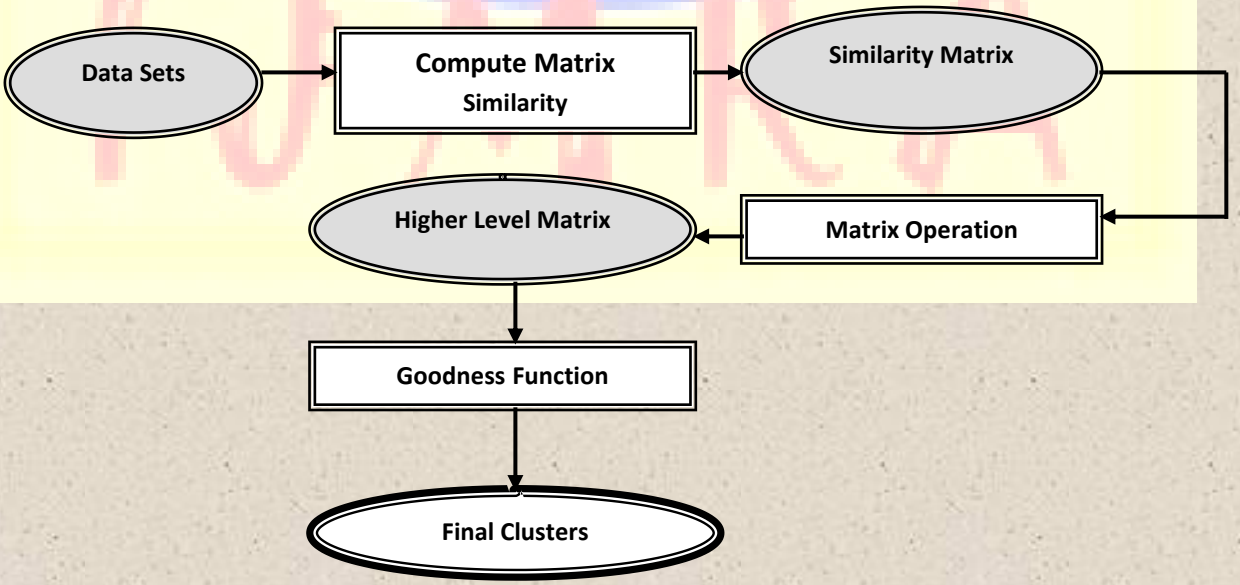


Figure 3.1 Overview of MBC

Figure3.1 provides the key stages (phases) of MBC algorithm. MBC operates on a matrix *M* in which *M*[*i; j*] represents the similarity between two data points *i* and *j*.

There are three phases in the algorithm:

First a similarity matrix is computed from input data. Different similarity functions can be employed. For two-dimensional data the best choice is to use the reciprocal of Euclidean distance as a similarity measure.

Second, matrix operations are applied to the similarity matrix. The most common operation is "*multiple*," which is somewhat different from standard matrix multiplication. Another operation is "*pow*", i.e., the exponential of the similarity matrix.

Third, a general hierarchical clustering algorithm is applied with a specified goodness function. The goodness function is essential for the clustering quality. Goodness function measures both between-group and within-group properties.

A key feature of MBC is that it determines the pair of most similar sub clusters by taking account both direct link and indirect link. Alternately, it merges two clusters according to both within-cluster and between-cluster properties.

Similarity Matrix Construction:

Many methods can be applied to build a similarity matrix. Since experimental data sets are two-dimensional, a function to convert the Euclidian distance to similarity values is used. The similarity value should lie between 0 and 1 where 0 means no similarity at all and 1 means identity.

Two parameters are used here: a normalized coefficient **α** and a cutoff value **γ**. Both **α** and **γ** are values between 0 and 1. Link of a data point to itself is set at 1,

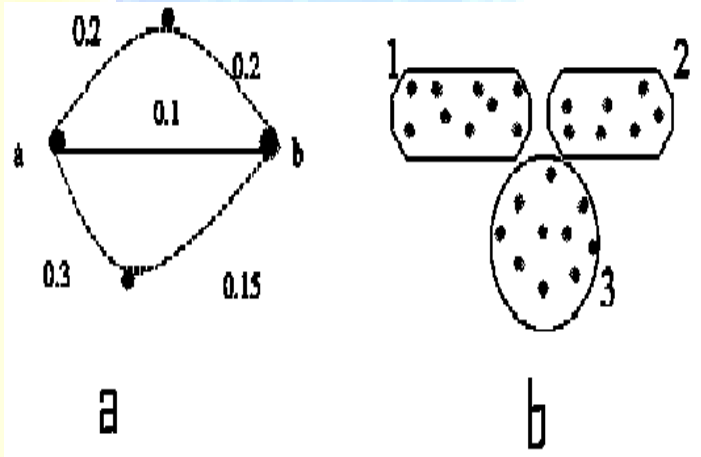
Link[i][j] = Link of a point to another point is defined as

*link[i][j] = **α** * **(min / distance[i][j])***

where *min* is the minimum distance determined over the entire distance matrix. If distance between two points *i* and *j* is larger than **γ** * *max*, then *link*[*i*][*j*] = 0, where *max* is the maximum distance.

The higher the cutoff value *γ*, the more global the information becomes. In a similar fashion, lower cutoff values lead to more local clustering information. A commonly used cutoff value is 0.3 but depends on properties of data.

The normalized coefficient **α** is often set in the range 0:5 ~ 0:8. It affects results very much. For high dimensional data or protein sequence data, different methods may be applied to do the transformation. For example, the linear correlation coefficient may be used for high dimensional data such as micro array data set. A normalized Smith-Waterman score is used for protein sequence clustering as in *Matrix operation.*



A very common operation is the multiplication of two matrices. But MBC's matrix multiplication is a little different from general multiplication. The following psuedocode illustrates what "*multiple*" does. The algorithm does not sum up the link between an object and itself. This removes the duplication of the direct link from a data point to itself.

Pseudocode 1

for i = 0 to m1 Row

for j = 0 to m2 Column

for k = 0 to m1 Column

If(k!=j)

*Mult[i][j] = mult[i][j] + m1[i][k] * m2[k][j];*

Computation of Goodness function:

The main difference among hierarchical agglomerative clustering methods is the distance (or similarity) measure. Goodness function is a unified function that counts both within cluster information and between cluster information. Two definitions are used:

1. $Bond(c1,c2) = \frac{link(c1,c2)}{|c1| * |c2|}$,

Where $link(c1,c2) = \sum link(p1,p2), p1 \in c1, p2 \in c2,$ the average link between two clusters;

2. *innerBond*(*c*) = *bond*(*c1; c2*).

The goodness function is defined as the following:

Goodness(c1,c2)=

$$\frac{Bond(c1,c2)}{innerBond(c1) \times \frac{|c1|}{|c1|+|c2|} + innerBond(c2) \times \frac{|c2|}{|c1|+|c2|}}$$

Remarks on MBC's goodness function

The goodness function defines two basic ideas of MBC:

1. The bond between two clusters depends not only on the direct link but also on indirect links.
2. The merging criterion depends not only on the bond between two clusters but also on the inner bond of each cluster.

MBC Performance Analysis:

The overall computational complexity of MBC depends on the time to carry out the matrix operation. If it is a full matrix, then it requires $O(n3)$ to do the matrix multiplication. If it is a sparse matrix data structure, the average number of non-zero neighbors of a vertex is m, then the time complexity reduces to $O(mn2)$.

The amount of time required by the first phase is $O(n2)$. Basically in this step one just needs to convert the distance to similarity. For high dimensional data the computation per distance is $O(d)$ where d is the number of dimensions. So the total complexity for this step would be $O(dn2)$. The amount of time required by the third phase is $O(n3)$. If a priority queue is used, each merge is $O(n \log n)$. Thus this phase uses $O(n2 \log n)$ time.

Overall the time complexity is $O(mn2)$. This is an expensive algorithm but the clustering quality is deemed good. One advantage of this algorithm is this can be used for outliers detection by checking the density of each of the data points. Low-density data points tend to be outliers. Another advantage is that it is possible to create a termination criterion according to the goodness function.

Conclusions and Future Work:

It involves not only direct link but also an indirect link, not only between-cluster information but also within-cluster information. MBC, was able to discover clusters with different shape and density

The possible future work about MBC, includes:

- By optimizing the algorithm with sparse matrix and/or graph theory the speed can be increased.
- Matrix Based Clustering Algorithm can be applied in protein/DNA sequences clustering and extending it to create a multiple sequences alignment method.
- MBC can be applied to much larger datasets using Parallel computing technique.

References:

- Jiawei Han and Micheline Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, San Francisco, 2001.
- J. Hartigan. *Clustering Algorithms*. Wiley, 1975.
- D. fasulo. An analisys of recent work on clustering algortihms. Technical report, 1999.
- C. Olson. Parallel algorithms for hierarchical clustering.
- Parsons,L., Haque,E., Liu, H.: Evaluating subspace clustering algorithms. Workshop on Clustering High Dimensional Data and its Applications, SIAM Int. Conf. on Data Mining. (2004) 48–56
- Candillier,L.,Tellier, I., Torre, F., Bousquet, O.: SSC : Statistical Subspace Clustering. In Perner,P.,ed.: Machine Learning and Data Mining in Pattern Recognition (MLDM). LNCS, Leipzig, Germany, Springer Verlag (2005) 100–109
- BANERJEE, A. and GHOSH, J. 2002. On scaling up balanced clustering algorithms. In Proceedings of the 2nd SIAM ICDM, 333-349, Arlington, VA.
- YONGHUI CHEN, ALAN P. SPRAGUE, AND KEVIN REILLY. MABAC - Matrix Based Clustering Algorithm. In *MSV/AMCS*, pages