International Journal of Computer and Communication Technology

Volume 3 | Issue 2 Article 5

April 2012

Partitional Clustering - An Iterative Algorithm

S. S. Patra
1School of Computer Application, KIIT University Bhubaneswar-751024, India,
sudhanshupatra@gmail.com

S. R. .Dash
School of Computer Application, KIIT University Bhubaneswar-751024, India,
satyaranjan.dash@gmail.com

Partitional Clustering - An Iterative Algorithm

S.S. Patra[1] , S.R.Dash[2]
[1]*School of Computer Application, KIIT University*
Bhubaneswar-751024, India
Email id:sudhanshupatra@gmail.com
[2]*School of Computer Application, KIIT University*
Bhubaneswar-751024, India
Email id:satyaranjan.dash@gmail.com

Abstract –*Cluster analysis is the term applied to a group of analyses that seek to divide a set of objects into a number of homogeneous groups or clusters, when there no a priori information about the group structure of the data. Clustering is an active research topic in data mining and different methods have been proposed in the literature. Most of these methods are based on the use of a distance measure defined either on numerical attributes or on categorical attributes. There are three basic categories of clustering methods: partitional methods, hierarchical methods and density-based methods. This paper proposes an iterative algorithm for partitional clustering.*

1 Introduction

Data analysis underlies many computing applications, either in a design phase or as part of their online operations. Data analysis procedures can be dichotomized as either exploratory or confirmatory, based on the availability of appropriate models for the data source, but a key element in both types of procedures (whether for hypothesis formation or decisionmaking) is the grouping, or classification of measurements based on either (i) goodnessoffit to a postulated model, or (ii) natural groupings (clustering) revealed through analysis. Cluster analysis[2] is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. The variety of techniques for representing data, measuring proximity (similarity)between data elements, and grouping data elements has produced a rich and often confusing assortment of clustering methods[9].

It is important to understand the difference between clustering (unsupervised classification) and discriminant analysis (supervised classifica-

tion). In supervised classification, we are provided with a collection of labeled (preclassified) patterns; the problem is to label a newly encountered, yet unlabeled, pattern. Typically, the given labeled (training) patterns are used to learn the descriptions of classes which in turn are used to label a new pattern.In the case of clustering, the problem is to group a given collection of unlabeled patterns into meaningful clusters. In a sense, labels are associated with clusters also, but these category labels are data driven; that is, they are obtained solely from the data. Clustering is useful in several exploratory patternanalysis, grouping, decisionmaking, and machine-learning situations, including data mining, document retrieval, image segmentation, and pattern classification. However, in many such problems, there is little prior information (e.g., statistical models) available about the data, and the decisionmaker must make as few assumptions about the data as possible.It is under these restrictions that clustering methodology is particularly appropriate for the exploration of interrelationships among the data points to make an assessment (perhaps preliminary) of their structure.

The problem of partitional clustering[7] is stated as follows: "given n patterns in a d-dimensional metric space ,determine a partition of the patterns into K groups or clusters such that the patterns in a cluster are more Similar to each other than patterns in different clusters." Global criteia require some prototypes to assign each pattern to a cluster but often they are not available.Then local criteria ,which form clusters by means of local structure such as high density areas or by assigning a particular pattern and its k-nearest neighbours to the same cluster,are the most appropriate.

2 Model Description and Analysis

The basic idea of partitional clustering method

- Seed Points
- Initial Partition

Seed Points:

1. Choose the first k objects in the data set.
2. Label the objects from 1 to n and choose those labeled $n/k, 2n/k, \ldots, (k-1)n/k$, and n.
3. Subjectively choose any k objects from the data set.
4. Label the objects from 1 to n and choose the objects corresponding to k different random numbers in the range $[1, n]$.
5. Take any desired partition of the objects into k mutually exclusive clusters and compute the cluster centroids as seed points.

Initial Partition:

1. Assign each object to the cluster built around the nearest seed point.This point remains stationary throughout one full pass over all objects.
2. Let each seed point to form a cluster of one member. Then assign objects one at a time to the cluster with the nearest centroid; after an object is assigned to a cluster, update the centroid so that it is the true mean vector for all the objects currently in that cluster. 3. Use

hierarchical clustering to obtain an initial partition. 4. The analyst could use his judgment to sort the set of objects into an initial partition. 5. The analyst could rely on some random allocation Schemes.

Criteria for Partitional Clustering

Let $X = [X_{*1}, X_{*2}, \ldots, X_{*n}]_{m\times n}$ The problem is to partition X into k clusters, such that $X = G1 \bigcup G2 \bigcup \ldots \bigcup G_c, and G_i \bigcap G_j = \emptyset, 1 \leq i, j \leq k$ and $i \neq j$ samples in a group are more similar than those in different groups.

Let $|Gi| = n$ and $\sum_{i=1}^{c} n_i = n$

2.1 Sum-of-squared-error criterion

The centroid of cluster G_i,

$$m_i = \frac{1}{n_i} \sum_{X_{*j}\in G_i} X_{*j}$$

= mean of the ith sample

The square error J_i for cluster G_i is the sum of the squared Euclidean distance between each object in G_i and its cluster centroid m_i,

$$\begin{aligned} J_i &= \sum_{X_{*j}\in G_i} \parallel X_{*j} - m_i \parallel^2 \\ &= \sum_{X_{*j}\in G_i} \Big(X_{*j} - m_i\Big)^T \Big(X_{*j} - m_i\Big) \end{aligned}$$

The square error,J_e for the entire clustering containing c clusters is the sum of square-error of the individual clusters,

$$\begin{aligned} J_e &= \sum_{i=1}^{c} J_i \\ &= \sum_{i=1}^{c} \sum_{X_{*j}\in G_i} (X_{*j} - m_i)^T (X_{*j} - m_i) \quad (1) \end{aligned}$$

The objective of a partitional clustering algorithm based on the square error criterion is to find a partition that minimizes J_e.

The resulting clusters is called Minimum variance partition.

Related Minimum Variance Criterion

Substitute for m_i in equation (1)

$$\begin{aligned} J_e &= \sum_{i=1}^{c} \sum_{X_{*j}\in G_i} (X_{*j} - \frac{1}{n_i} \sum_{X_{*j}\in G_i} X_{*j})^T * \\ &\qquad (X_{*j} - \frac{1}{n_i} \sum_{X_{*j}\in G_i} X_{*j}) \\ &= \sum_{i=1}^{c} \sum_{X_{*j}\in G_i} \frac{1}{n_i^2} (n_i X_{*j} - \sum_{X_{*j}\in G_i} X_{*j})^T * \\ &\qquad (n_i X_{*j} - \sum_{X_{*j}\in G_i} X_{*j}) \\ &= \frac{1}{2} \sum_{i=1}^{c} n_i \overline{S_i}, \quad where \\ \overline{S_i} &= \frac{1}{n_i^2} \sum_{X\in G_i} \sum_{X'\in G_i} (X - X')^T (X - X') \end{aligned}$$

= average squared distance between points in G_i

• This uses Euclidean distance for the similarity measure.

• If $S(X, X')$ is any other similarity measure we can define

$$\begin{aligned} \overline{S_i} &= \frac{1}{n_i^2} \sum_{X} \sum_{X'} S(X, X') \\ or & \\ \overline{S_i} &= min \sum_{X,X'} \in G_i S(X, X') \\ etc & \end{aligned}$$

2.2 Scatter Matrix Criterion

Consider C Clusters $G_1, G_2, \ldots, G_c$. Let m_i be the mean of the ith cluster G_i , and let m be the pooled mean of all objects in X.

$$m_i = \frac{1}{n_i} \sum_{X*j \in C_i} X_{*j} \ldots mean\ of\ ith\ group$$

$$m = \frac{1}{n} \sum_{j=1}^{n} X_{*j} = \frac{1}{n} \sum_{j=1}^{n} n_i m_i \ldots Pooled\ Mean$$

Define S_i to be the scatter matrix for the ith cluster,

$$S_i = \sum_{X*j \in G_i} (X_{*j} - m_i)(X_{*j} - m_i)^T$$

The within-cluster,S_w, is the sum of scatter matrices of the individual clusters,

$$S_w = \sum_{i=1}^{c} S_i$$

The between cluster scatter matrix,S_B , is defined as

$$S_B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T$$

The total clusters scatter matrix,S_T , is defined as

$$S_T = \sum_{X*j \in C} (X*j - m)(X*j - m)^T$$

It can be verified that $S_T = S_B + S_W$ i.e.,there is exchange between S_B and S_W matrices: S_B goes up as S_W goes down.By minimizing S_W we will tend to maximize S_B. We need for a criterion, a scalar measure. Two useful measures are : Trace and Determinant.

2.2.1 Trace criterion

A good partition can be obtained by minimizing the trace of S_W.

$t_r(S_W) = \sum_{i=1}^{c} t_r(S_i)$

By expansion

$$t_r(S_W) = \sum_{i=1}^{c} \sum_{X*j \in G_i} (X_{*j} - m_i)(X_{*j} - m_i)^T$$

$$= J_e$$

Therefore minimizing $t_r(S_W)$ immediately implies that

$$t_r(S_B) = \sum_{i=1}^{c} n_i (m_i - m)^T (m_i - m)$$

is maximized and hence ,the resulting partition is optimal.

2.2.2 Determinant criterion

Let $x = (x_1, x_2 \ldots, x_n)^T$ be a vertex .X is said to be normally distributed with mean vector μ and covariance matrix Σ denoted by $x \approx N(\mu, \Sigma)$ if $f(x)$ is given by

$f(x) = \frac{1}{(2\Pi)^{n/2}} |\Sigma|^{\frac{1}{2}} \exp\lfloor -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \rfloor$ where $\mu \in R^n, \Sigma \in R^{nXn}$ is a symmetric and positive definite matrix and $|\Sigma|$ is the determinant of Σ. It can be shown that $E(x) = \mu$ and

$\Sigma = E[(x-\mu)(x-\mu)^T]$

• Samples drawn from $N(\mu, \Sigma)$ tend to form a single cluster with center μ and the shape determined by Σ.

• Locus of points of constant density form a hyper ellipsoid for which

$(x-\mu)^T\Sigma^{-1}(x-\mu)$=constant=$r^2$

• The principal axes of this ellipsoid is given by the eigen vectors of Σ. • The eigen values determine the length of the semi-axes. • Volume of the hyper ellipsoid measures the scatter of the samples about the mean. • Let e be an n-d ellipsoid .Intutively, an n-d ellipsoid is simply a sphere that has been stretched along the n orthogonal semiaxes of the ellipsoid. This indicates that the volume of the ellipsoid is simply the volume of a unit hyper-sphere multiplied by the length of each semi axes. The volume of n-d ellipsoid is

$$V = V_n|\Sigma|^{\frac{1}{2}}\lambda_1\lambda_2\ldots\lambda_n$$

where V_n the volume of an n-d hyper sphere and $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the lengths of the n semi-axes of the ellipsoid. The volume V_n of an n-d unit hemisphere is

$V_n = \frac{\Pi^{\frac{n}{2}}}{\Gamma(1+\frac{1}{2}n)}$ where Γ function is a mathematical extension of factorial frunction from positive integers to real numbers.

Equivalently,

$$\mathbf{V}_n = \begin{cases} \frac{\pi^{n/2}}{\left(\frac{n}{2}\right)!}, & n \text{ even} \\ \frac{2^n\pi^{\frac{n-1}{2}}\left(\frac{n-1}{2}\right)!}{n!}, & n \text{ odd} \end{cases}$$

3 Square-Error Clustering Algorithm

Let X be a data matrix, and let there be C clusters.We used the square-error criterion J_e

$$J_e = \sum_{i=1}^{c}\left(J_i\right)$$

$$J_i = \sum_{X_{*j}\in G_i}(||X_{*j}-m_i||)^2$$

Start with an initial partition and move samples from one group to another if such a move improves the criterion.

Details of the Algorithm

Let $y \in C_i$. Decide to move object y from C_i to C_j.As a result of this move the quantities m_j, J_j, m_i and j_i will change. Let m_j^*, J_j^*, m_i^* and j_i^* be the value of these quantities after the move.Then

$$\begin{aligned} m_j^* &= m_j + \frac{y-m_j}{n_j+1} \\ J_j^* &= J_j + \frac{n_j}{n_j+1} \parallel y-m_j \parallel^2 \\ m_i^* &= m_i - \frac{y-m_i}{n_i-1} \\ J_i^* &= J_i - \frac{n_i}{n_i-1} \parallel y-m_i \parallel^2 \end{aligned}$$

Therefore the transfer of y from C_i to C_j is welcome only if

$\mid J_i^* - J_i \mid > \mid J_j^* - J_j \mid$

which is same as

$\frac{n_i}{n_i-1} \parallel y-m_i \parallel^2 > \frac{n_j}{n_j+1} \parallel y-m_j \parallel^2$

An iterative algorithm using this method can be described as follows.

1. Select an initial partition of the n objects into k clusters and compute m_i and J_e.

LOOP:

2. Select a candidate for move $y \in C_i$

3. If $n_i = 1$, go to NEXT

ELSE compute

$$\mathbf{R}_j = \begin{cases} \frac{n_j}{n_j+1} \parallel y-m_j \parallel^2, & j \neq 1 \\ \frac{n_i}{n_i-1} \parallel y-m_i \parallel^2, & j = 1 \end{cases}$$

4.Transfer y to C_k , if $R_k \leq R_j$ for all j
5.Update m_i, m_k, J_e
NEXT
6.If J_e has not changed in n steps then STOP
ELSE goto LOOP.

4 Conclusion

(1) Its time complexity is $O(nkl)$,where n is the number of patterns,k is the number of clusters, and l is the number of iterations taken by the algorithm to converge. Typically, k and l are fixed in advance and so the algorithm has linear time complexity in the size of the data set
(2) Its space complexity is O(k+n). It requires additional space to store the data matrix. It is possible to store the data matrix in a secondary memory and access each pattern based on need. However, this scheme requires a huge access time because of the iterative nature of the algorithm, and as a consequence processing time increases enormously.

However, the this algorithm is sensitive to initial seed selection and even in the best case, it can produce only hyperspherical clusters.

REFERENCES

[1]K.J. Cios, W. Pedrycz, and R.W. Swiniarski, Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers, 1998.
[2]R. J. Kuo, H. S. WANG, TUNG-LAI HU AND S. H. CHOU Application of Ant K-Means on Clustering Analysis Computers and Mathematics with Applications 50 (2005) 1709-1724
[3]Raymond T. Ng and Jiawei Han, Member, IEEE Computer Society CLARANS: A Method for Clustering Objects for Spatial Data Mining IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 14, NO. 5, SEPTEMBER/OCTOBER 2002
[4]An Efficient Concept-Based Mining Model for Enhancing Text Clustering Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, OCTOBER 2010
[5]H. Jin, M.-L. Wong, and K.S. Leung, Scalable Model-Based Clustering for Large Databases Based on Data Summarization, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 11, pp. 1710-1719, Nov. 2005.
[6]S. Shehata, F. Karray, and M. Kamel, Enhancing Text Clustering Using Concept-Based Mining Model, Proc. Sixth IEEE Intl Conf. Data Mining (ICDM), 2006.
[7]A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Prentice Hall, 1988.
[8]D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. IEEE Transactions on Knowledge and Data Engineering, 17(12):16241637, December 2005.
[9]J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001 pp. 346389.