Approximation Clustering as a Framework for Solving Challenging Problems in Processing Massive Data Sets*

Boris Mirkin
DIMACS, Rutgers Univ., P.O.Box 1179, Piscataway, NJ 08854-1179, USA
mirkin@dimacs.rutgers.edu

October 20, 1998

1 Introduction

Approximation clustering is an extension of what is usually called "square-error clustering" onto various cluster structures and criteria. This may be considered a mathematical device to support, specify and extend K-Means and other fast clustering techniques. Also it appears a framework for developing workable concepts and fast procedures for such challenging problems in machine learning and data processing as approximate concept learning,feature selection, data mining, data compression/decompression, and design and use of multiresolution hierarchies. In the remainder, we (1) overview how three clustering structures, single cluster, partition, and binary hierarchy, are represented in this framework, (2) discuss the least-squares based criteria and decompositions, and (3) show some theoretical results and methods found with the model in the areas listed above. Move on approximation clustering and corresponding references can be found in Mirkin (1996).

2 Approximation Clustering

2.1 Cluster Structures and Their Linear Representation

Let I be an N-element set and $R(I)$ the N-dimensional Euclidean space of vectors $x = (x_1, \ldots, x_N)$. A subset $S \subset I$ corresponds to a basis of a unidimensional subspace presented by its Boolean N-dimensional indicator vector $s = (s_i)$ where $s_i = 1$ if $i \in S$ and $s_i = 0$ otherwise. A partition, $S = \{S_1, \ldots, S_m\}$ on I, can be associated with an m-dimensional orthogonal basis presented by an $N \times m$ Boolean matrix $s = (s_{it})$ where $s_{it} = 1$ if $i \subset S_t$ and $= 0$ if $i \notin S_t$. A binary hierarchy,

*The reach was supported by the Office of Naval Research under grants number N00014-93-1-0222 and N00014-96-1-0208 to Rutgers University

$S_W \subseteq 2^I$, is s nested set of subsets $S_w \subseteq I$ called clusters, containing all singletons $i \subset I$ and I itself, such that every non-singleton $S_w \in S_W$ is split in two clusters,S_{w1} and S_{w2}, from S_W. A binary hierarchy is associated with $(N-1)$ dimensional orthonormal basis presented by three-valued N-vectors ϕ_w defined for every nonsingleton $S_w \in S_W : \phi(i)$ is a a_w for $i \in S_{w1}$ and b_w for $i \in S_{w2}$ and 0 for $i \notin S_w$, where $a_w$and b_w are defined so that $\sum_i \phi_w(i) = 0$ and $\sum_i \phi^2{}_w(i) = 1$.

When an $N \times n$ data matrix Y is given, the vases are employed to form matrices matching the data table. Given a cluster center vector, c , the single cluster matrix is $sc^T = (s(i)c_k)$ whose rows are either c_t or 0. given cluster vectors, $c_1, \ldots, c_m$, the partition cluster matrix is sc, whose i-th row is c_t for $i \in S_t (t = 1, \ldots, m)$. Given cluster center vectors c_w, a (partial) binary hierarchy leads to matrix ΦC where Φ is composed with ϕ_w as its columns and C with c_w as its columns.

2.2 Bilinear Clustering Model

Let us take $Z = (z_{it})(i \in I, t \in T)$ as a common denotation of either kind of matrices related to the three clustering structures above: single cluster $(|T| = 1)$, partition $((|T| = m)$, and binary hierarchy $(T = W)$. The approximation clustering model under consideration is presented with equations

$$y_{ik} = \sum_t c_{tk} z_{it} + e_{ik} \tag{1}$$

or, in matrix form,

$$Y = ZC + E,$$

where $E + (e_{ik})$ is matrix of residuals to be minimized. The least-squares optimalc_{tk} satisfy equation

$$C = (Z^T Z)^{-1} ZY$$

The model is an extension of the well-known properties of SVD-decomposition of matrices, to the case when z_{ik} are under constraints (to be binary for single cluster and partition clustering or to be ternary for hierarchical clustering).

2.3 Decompositions and Contributions

When c_{tk} are least-squares optimal, the data scatter can be decomposed according to the approximation model as

$$\sum_{i,k} y^2{}_{ik} = \sum_t \mu_t^2 + \sum_{i,k} e_{ik}^2 \tag{2}$$

where

$$\mu_t^2 = \sum_{k=1}^{n} c_{tk}^2 \sum_I z_{it}^2 \tag{3}$$

which is $\sum_{k=1}^{n} c_{tk}^2 |S_t| = d^2(0, c_t)|S_t|$, for partition or single cluster clustering, and $\frac{n_{w1}n_{w2}}{n_w} d^2(c_{w1}, c_{w2})$ for hierarchical clustering. Here d is Euclidean distance and c with corresponding index is the gravity center corresponding cluster.

This can be used for making theoretical inferences, developing clustering procedures, and for deriving interpretation aids.

3 Some Applications of the Approximation Model

3.1 Traditional and Conceptual Clustering

Some inferential results concern interconnections between the traditional and conceptual clustering. It appears, the least-error clustering criterion, On the other hand, in the case of categorical data presented by non-normalized (though centered) binary vectors, the least-squares criterion is equivalent to what is called the Category Utility Function in conceptual clustering. This allows us for considering both techniques as just different search procedures for the same criterion.

Similarly, the hierarchical clustering criterion (for categorical data) can be presented as

$$\mu_w^2 = \frac{n_{w1}n_{w2}}{n_w} \sum_{\mu} |p(u/S_1) - p(u/S_2)|^2$$

which must resembles the so-called Twoing rule criterion in a version if conceptual clustering, CART techniques (Breiman et al. 1984).

3.2 Clustering Algorithms as Local Search

It can be easily proven that the least-squares fitting criterion generates some well-known clustering techniques as local search algorithms:

1.K-Means (with the cluster centers adjusted to distance) as the alternating minimization algorithm for partitioning;

2.Agglomerative/Divisive clustering as the sequential fitting of the model(in the hierarchy structure case).

3.Seriation (by the average linkage) as the local search algorithm for the single cluster case.

The model can be utilized for:

- deepening search;

– extending to nonstandard data formats (mixed, flow, spatial, etc.);

– restricting parameters.

3.3 Principal Cluster Analysis: A Data Mining Device

A method based on the criterion above, $|S|d^2(0,c) \rightarrow max_{S,c}$, leads to a method for finding deviant patterns in the data. The method is based on repeatedly finding the most deviant clusters in the residual data obtained by the previous results from the data matrix:

$$y_{ik} \Leftarrow yik - c_k z_i$$

This may be used for finding patterns presented by overlapping clusters.

3.4 Machine Learning

In the case of single cluster or partition clustering, the decompositions above lead to cluster-specific salience weights of the variables:

$$Weight(Class_t, Variable_k) = (c_k - c_{tk})^2 \# Class_t$$

where c_k is grand mean and c_{tk}within-cluster mean of the variable k.

This can be employed for finding distinctive conceptual descriptions of the clusters. For instance, a cluster given, we can select the most contributing variables and their within-cluster ranges to form a conjunctive description of the cluster. This, combined with transformation of the variables when needed, may produce nice and simple conceptual descriptions of clusters either supervised or unsupervised.

3.5 Hierarchies as Multiresolution Analysis Devices

A recently emerged techniques for fast compression/decompression of data, multiresolution approximation and wavelets, has found a parallel in terms of complete hierarchies. Let us denote the difference between the means of splits of a hierarchy cluster, S_w, as $d_w = c_{w1} - c_{w2}$. Then, obviously,

$$c_{w1} = c_w + \frac{n_{w2}}{n_w} d_w, cw2 = c_w - \frac{n_{w1}}{n_w} d_w \tag{4}$$

which can be used for layer-to-layer compression/decompression of the unidimensional binary data. Amazingly, these formulas coincide with those based on the so-called Haar wavelet basis in the theory of multiresolution approximation, in the case when the hierarchy is complete(case B in Figure 1).

More exact analogy involves two types of linear subspaces defined for each layer of a hierarchy, those ternary valued (space V_m) and those binary valued(space D_m

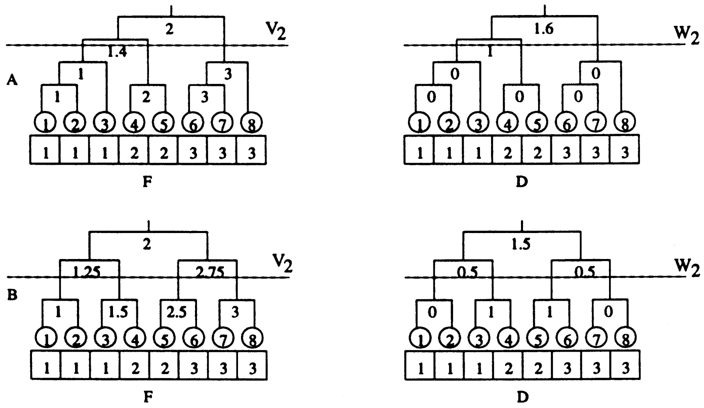

Figure 1: Compression and decompression of the boxed data with two hierarchies, A and B.

which is ortogonal to V_m). The subspaces satisfy equation $V_{m-1} \oplus D_{m-1} = V_m$, an analogue of which is the core of multiresolution approximation.

On the level of plane digitalized image data, the hierarchical representation includes yet another well-known data structure, quad-tree,

The concept of hierarchy can be utilized in data compression/decompression effectively since it allows to drop off the standard requirements of the multiresolution wavelet-based analysis: the classes can be of different cardinalities; they can break the spatial continuity requirement; the layers can be of arbitrary shape. Moreover, the hierarchy classes can be adjusted to the data format(via clustering to minimize the residuals),as it is done in Figure 1 B.

4 Conclusion

Results reported:

1. Presenting some most popular clustering techniques, K-Means, hierarchical clustering, and seriation, as local search approximation algorithms for partitional, hierarchical, or single cluster clustering structures, respectively;

2. Discussing related data decompositions as sources of interpretation aids for cluster analysis and, also, as model-based frameworks for addressing problems in machine learning, data mining and knowledge discovery;

3. Presenting a theory which parallels that of fast wavelet-based compression/decompression of data, but allows to drop off the standard requirements of the multiresolution wavelet-based analysis: the classes can be of different cardinalities; they can break the spatial continuity requirement; the layers can be of arbitrary shape.

References

L.Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone(1984) *Classification and Regression Trees,* Wadsworth: Belmont, Ca.

B. Mirkin(1996) *Mathematical Classification and Clustering,* Kluwer Academic Publishers: Dordrecht,Netherlands.