Mahalanobis metrics for k-means algorithms

Metriche di Mahalanobis per l'algoritmo delle k-medie

Agostino Tarsitano
Dipartimento di Economia e Statistica - Università degli Studi della Calabria
agotar@unical.it

Summary. Solutions obtained for a k-means algorithm depends heavily on the starting partition. In this article an aid is developed for implementing initialization methods particularly suited when all the cluster covariances matrices are presumed to be identical.

Keywords: covariance estimates, determinant minimization, multiple quickselect

1. Introduction

The efficacy of a k-means algorithm is influenced by many factors. Most obvious is the starting partition. In fact, k-means algorithms have differential recovery rates depending on the quality of the starting configuration. As the cluster analysis has evolved, a wide variety of techniques has emerged for choosing the first *k* centroids (or, alternatively, for specifying an appropriate starting partition). So far no attempt has been made to set up a specific procedure by which an iterative partitioning based on the Friedman-Rubin approach could be triggered. This is partly due to a vicious circle: to determine a preliminary classification of the entities a good estimate of the within-group scatter matrix $\boldsymbol{W}$ is required, but to estimate $\boldsymbol{W}$ a plausible classification should be known in advance. The primary purpose of this paper is to refine the algorithm proposed by Art *et al.* (1982) which offers an interesting solution to this problem.

2. Preliminary estimation of the within-clusters matrix

Art *et al.* (1982) proposed an algorithm to compute an estimate of $\mathbf{W}$ without knowing the cluster structure, but assuming that the clusters have different means and a common covariance matrix. The standard multivariate analysis decomposition $\boldsymbol{T}=\boldsymbol{W}+\boldsymbol{B}$ can also be made in terms of pairwise differences:

$$\boldsymbol{T} = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\boldsymbol{X}_i - \boldsymbol{X}_j\right)\left(\boldsymbol{X}_i - \boldsymbol{X}_j\right)^t =$$

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)^t}_{Within} + \underbrace{\frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)^t}_{Between} = W^* + B^* \qquad (1)$$

where, $\boldsymbol{X}_i$ for $i=1,2,\ldots,n$ gives the values of m real-valued features on n distinct entities. The first term on the right side involves all the pairs which belongs to the same clusters, and the second term involves all the distance measurement occurring between those pairs

where one entity comes from cluster i and the other entity comes from cluster j with $i \neq j$. No explicit indication is made to a cluster membership or to a fixed number of clusters. Moreover, $\boldsymbol{W}^*$ avoids the estimation of the centroids. Under normal sampling assumptions, with $X_{ij} \tilde{} N(\mu_i, \boldsymbol{\Omega})$ we have $E(\boldsymbol{W}) = (n-k)\boldsymbol{\Omega}$ and $E(\boldsymbol{W}^*) = c\boldsymbol{\Omega}$ where c is a constant depending only on the cardinalities of the clusters. Hence $\boldsymbol{W}$ and $\boldsymbol{W}^*$ can be used to construct an unbiased estimate for $\boldsymbol{\Omega}$, but $\boldsymbol{W}^*$ gives relatively more weight to large clusters than does $\boldsymbol{W}$. The fact that $\boldsymbol{W}^*$ needs to be scaled by a factor to remove the bias is not relevant since a clustering based on $\boldsymbol{W}^*$ is invariant with respect of the transformation $a\boldsymbol{W}^*$ with $a > 0$. Naturally, since the cluster structure is unknown, neither $\boldsymbol{W}$ nor $\boldsymbol{W}^*$ can be computed. The initialization of the k-means, however, requires something less demanding. In this sense, I extend an earlier work of Gnanadesikan *et al.* (1993) based on the work of Art *et al.* (1982). A first approximation $\boldsymbol{W}_1$ to $\boldsymbol{W}^*$ can be obtained by a Winsorizing-type scheme

$$\boldsymbol{W}_1 = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \delta_h \left(\boldsymbol{X}_i - \boldsymbol{X}_j \right) \left(\boldsymbol{X}_i - \boldsymbol{X}_j \right)^t; \quad h = j + (i-1)n - \frac{i(i+1)}{2} \tag{2}$$

$$\delta_h = \begin{cases} f\left[M_{ij}(\boldsymbol{W}_0) \right] > 0 & if \quad h \leq q; \\ 0 & if \quad h > q \end{cases} \qquad M_{ij}(\boldsymbol{W}_0) = \left(\boldsymbol{X}_i - \boldsymbol{X}_j \right) W_0^{-1} \left(\boldsymbol{X}_i - \boldsymbol{X}_j \right); \quad \sum \delta_h = 1 \tag{3}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are among the closest q pairs in terms of the given metric $\boldsymbol{W}_0$. The weight or score δ_h is a non increasing function of the rank position of the Mahalanobis distance M_{ij}. This relationship reflects the fact that entities belonging to the same cluster are closed together than entities belonging to different clusters. The weighting scheme has the effect of decreasing the impact of the larger distances and increasing the smaller ones, thereby counteracting the biasing inherent in sample-based estimation of a covariance matrix.

Next $\boldsymbol{W}_2$ is formed in the same manner except that a new Mahalanobis distance is used to define the weights: $\delta_h = f[M(\boldsymbol{W}_1)]$. The algorithm continues in a like manner until the process stabilizes. The procedure is controlled by the following parameters:

1) The first metric. Art *et al.* (1982) used $\boldsymbol{W}_0 = \boldsymbol{I}$, that is the first allocation is made by using Euclidean distance, although $\boldsymbol{W}_0 = \boldsymbol{T}$ seems a more plausible choice when the data consist of a number of variables measured in different scales and $\boldsymbol{T}$ is well-conditioned. Another plausible choice is $\boldsymbol{W}_0 = diag(v_1, v_2, \ldots, v_m)$. It should be noted that choosing a diagonal is justified only when the variables are uncorrelated or weakly correlated. If this fact is not taken into account, the measure of closeness of the entities, suffers. Gnanadesikan *et al*. (1993) consider a sequence $q_0, q_1, q_2, \ldots$, and revise iteratively the starting metrics: $\boldsymbol{W}_{0,1} = I$, $\boldsymbol{W}_{0,2} = \boldsymbol{W}_{q0}, \boldsymbol{W}_{0,3} = \boldsymbol{W}_{q1}, \ldots$ In practice, this can be done if $\boldsymbol{W}_{0,i}$ is not ill-conditioned and there is enough evidence that $\boldsymbol{W}_{0,i}$ is a better choice than $\boldsymbol{W}_0$ which cannot be guaranteed because of the oscillating behavior of $\boldsymbol{W}_q$.

2) The number of pairs. The integer q is chosen conservatively small to avoid contamination by between-cluster pairs. However, $q > n-m$ to ensure a regular positive definite matrix. It should be noted that if the number of within-group pairs is small compared with the true number of pairs, the covariance matrix estimates become highly variable. The reduction of q has, undoubtedly, some advantages from a computational point of view. In fact, instead of sorting the $n(n-1)/2$ vector of all possible distances, it is sufficient to determine the smallest q elements. The Hoare's Quicksort can be adjusted to this purpose avoiding many of the computation required to do a complete sort. Hoare's algorithm takes advantage

of the remarkable power of the recursive function (a function that calls itself). Lent and Mahmoud (1996) analyze multiple Quickselect (MQS), a variant of Quicksort designed to search for several order statistics simultaneously. I have adapted MQS to select the first q ordered Mahalanobis distances.

3) The weights. Art *et al.* used $\delta_h = 1/q$. After some experiments the following formula has given better results and more robustness against non-singularity

$$\delta_h(\alpha) = \alpha(1-\alpha)^h\left[1-(1-\alpha)^{q+1}\right]^{-1}, \quad h = 1,2,\ldots,q, \quad 0 < \alpha < 1 \tag{4}$$

where h indicates the h-th closest pairs. The lower α is, the slower the decrease of the weights and the larger the number of distances that receive a score significantly different from zero. Of course, $\alpha \to 0$ implies $\delta_h(\alpha) = constant$. As $\alpha \to 1$ $\boldsymbol{W}^*$ becomes insensitive to q and ill-conditioned which, in turn, leads to lowered classification accuracy.

4) The measure of closeness. Art *et al.* defined $\varepsilon_{i+1} = \mathrm{tr}[\mathbf{W}_i(\mathbf{W}_{i+1})^{-1} - \mathbf{I}]^2$ and convergence is considered satisfactory if $\varepsilon_{i+1} \le \varepsilon$. The method outlined by Art *et al.* is available in the *Aceclus* procedure of the SAS routine *Fastclus***.**

3. Empirical investigation

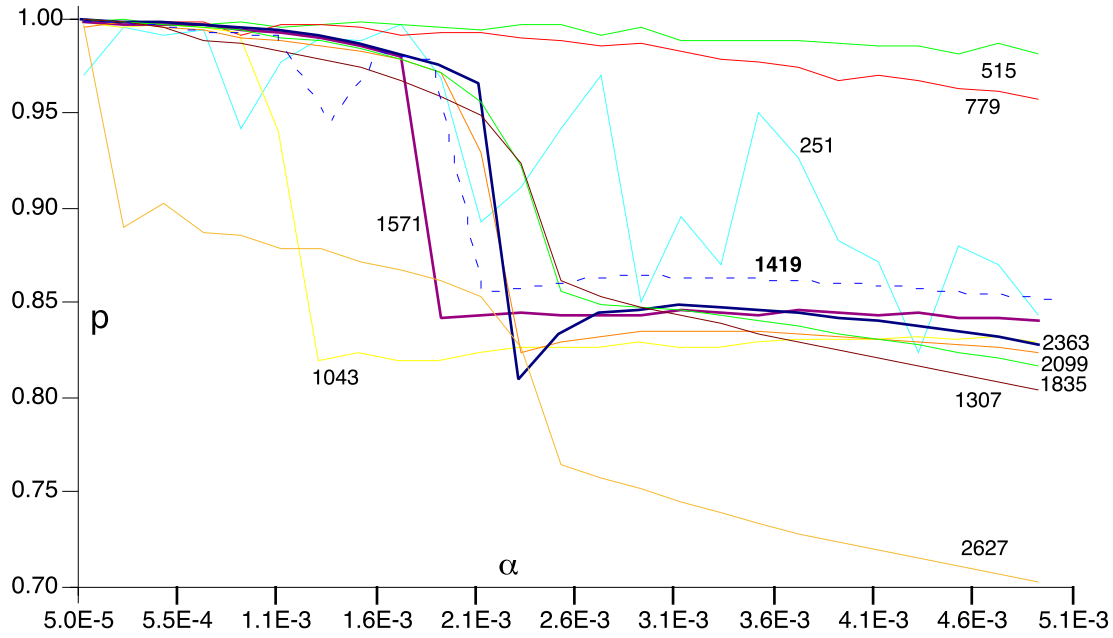
The performance of the method discussed in the previous section has been evaluated in terms of quality for several test data sets. In particular, the configuration with $\boldsymbol{W}_0 = \boldsymbol{I}$, $q_0 = max\{(n/3)[n/(k*3)-1], 2*(n-m)\}$ has been applied (k is the number of clusters suspected to be present in the data set). The procedure is stopped after $\varepsilon_{i+1} \le \varepsilon = 0.001$; the process is also interrupted after 30 iterations. The crucial elements of the algorithm are the number of pairs q and the score system $\delta_h(\alpha)$. A good couple of values (q,α) is not likely to be known in advance. My approach to covariance estimation is to explore a grid of points on the (q,α) plane and then determine the point with the greatest value of

$$p = \left[m^{-1}\sum_{i=1}^{m}\lambda_i\right]^2\left[\sum_{i=1}^{m}\lambda_i^2\right]^{-1}; \quad \lambda_i = eigenvalue \ \ of \ \ \boldsymbol{\Omega}^{-1}\boldsymbol{W}_q \tag{5}$$

Note that $(1/m) \le p \le 1$ and $p = 1$ if and only if $\boldsymbol{W}_q \propto \boldsymbol{\Omega}$. For the distance mixing parameter the following we used $\alpha \in \{0.005*(i-1), i = 1,2,\ldots,21\}$; for each α the procedure has been repeated for $q \in \{q_0 + gi, i = 0,1,2,..,9\}$ where $g = [(\nu - q_0)/10]$, $\nu = n(n-1)/(2\sqrt{2})$. One expects that p increases with q (as a consequence of the growing number of within-group pairs which is included in the estimate) and reaches a maximum, hopefully in the vicinity of the upper bound. As q continues to increase, the value of (5) should decline because of the contamination due to the between-group pairs which are progressively included in the estimate. The value of α should anticipate the achievement of the peak. If the optimal p is far from unity than the hypothesis of homogeneous clusters in the data is not defensible.

Usually we do not known $\boldsymbol{\Omega}$ or even $\mathbf{W}$. A reasonable strategy is to replace $\boldsymbol{\Omega}$ with the metric $\boldsymbol{W}_q$ obtained at the last iteration of the algorithm and evaluate (5) for a variety of different choices of q and α. To this end, a graphical aid consisting of a plot of p against α for a range of q-values would be useful. The idea is illustrated in Figure 1.

Figure 1: *Plot of p vs α for the Chernoff data set.*



The real data example from Chernoff (1973) involves $m=6$ variables measured on each of nummulited specimens from Eocene Yellow Limestone formation of Northwestern Jamaica. According to Chernoff the entities divide into $k=3$ distinct clusters. The cardinalities are $\{40, 34, 13\}$. The choice of $q=251$ is too small because the value of p for consecutive estimates of $\mathbf{\Omega}$ are too different. The dashed line uses the true number of within-cluster pairs. The curves corresponding to the largest levels of q seem instable. Instead, there is a flat region between $m=515$ and $m=779$ where adjacent nearby estimates are similar enough to keep p at a stationary level. In fact, the optimal value of q when $\mathbf{\Omega}$ is replaced by the $\mathbf{W}$ (the scatter matrix based on the actual cluster membership) is 709. The best choice for α is from 0.00005 to 0.00010.

When n is large, computation of exact quantiles is impractical due to the large requirement of memory storage and execution times. In this case, the closest q pairs can be selected in a random sample of pairs.

Even if the true covariance matrix of each cluster differs greatly, the common covariance matrix estimate can be useful, especially in small-sample settings.

References

Art D. , Gnanadesikan R., Kettenring J.R. (1982). Data-based metrics for cluster analysis. *Utilitas Mathematica*, 21A, 75-99.

Chernoff H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 73, 361-368.

Gnanadesikan R., Harvey J.W., Kettenring J.R. (1993). Mahalanobis metrics for cluster analysis. *Sankhya, A,* 55, 494-505.

Lent J., Mahmoud H.M. (1996). Average-case analysis of multiple Quickselect: an algorithm for finding order statistics. *Statistics and Probability Letters,* 28, 299-310.