A Fast Algorithm for Lossless Compression of Data Tables by Reordering

Slobodan Vucetic
Center for Information Science and Technology, Temple University, USA
Email: vucetic@ist.temple.edu

A data table is represented as a matrix $D = \{x_{ij}, i = 1\ldots N, j = 1\ldots M\}$, where N is number of rows, or data points, and M number of columns, or attributes. A major property of data tables is that they allow arbitrary permutations of rows and columns, without loss of information. This property can be used to improve compressibility of data tables by ordering of rows or columns in a way that that allows grouping of similar table elements. Here, an algorithm for lossless compression of tables with numeric attributes based on row ordering is proposed. The optimization problem is defined as finding the permutation of table rows π that maximizes compression of the permuted table $\pi(D)$ when differential predictive coding is used for compression. The optimal permutation can be expressed as $\pi^* = arg\,min_\pi L(\pi(D))$, where $L(D) = -\sum_i \sum_j log(p_j(\varepsilon_{ij}))$ is the minimal achievable compressed length of differentially encoded table D, ε_{ij} is difference $\varepsilon_{ij} = x_{ij} - x_{i\_1,j}$, and $p_j(\varepsilon)$ is distribution of the differences for the j-th attribute.

Good table ordering will result in $p_j(\varepsilon)$ distributions whose mass is concentrated around zero. Empirical results show that $p_j(\varepsilon)$ of good orderings can often be approximated with Laplace or Gaussian distributions. In the Gaussian case, it can be shown that the optimal permutation can be expressed as the one that minimizes the sum of weighted squared distances,

$$\pi^* = arg\,min_\pi \sum_{j=1}^{M} \sum_{i=2}^{N} (x_{\pi(i),j} - x_{\pi(i-1),j})^2 / \sigma_j(\pi)^2 , \qquad (1)$$

where $\sigma_j(\pi)$ is the standard deviation of distribution $p_j(\varepsilon;\pi)$ and where argument π reflects the fact that these quantities are dependent on the permutation π. In the Laplace case, a similar expression that uses weighted city block distances can be derived.

The stated optimization problem is challenging since π and distributions $p_j(\varepsilon;\pi)$ are mutually dependent. The problem can be solved in an iterative manner by starting from an initial guess $\pi(i) = i$, $i = 1\ldots N$, and repeating until convergence: estimate $\sigma_j(\pi)$, fix their values, and find new permutation π that solves (1). Note that solving (1) for fixed $\sigma_j(\pi)$ is equivalent to the well-known traveling salesman problem (TSP). By assuming that the optimal TSP algorithm is used, each iteration of the iterative ordering procedure leads to decrease of $L(\pi(D))$, thus ensuring convergence to a local minimum.

To allow application of the ordering algorithm on large data tables, a fast recursive partitioning TSP heuristics was used that has O($NlogN$) time and O(N) space complexity. It starts by building a binary tree whose leafs are individual data points by recursively applying k-means ($k = 2$) clustering algorithm at each node. It finishes by reordering the tree nodes by a dynamic programming algorithm.

Extensive experiments were performed on randomly generated and scientific multidimensional tables with numerical attributes. The results showed that ordering is useful for compression of moderately large to large tables with intrinsic dimensionality below 20 and with attributes represented with low to moderate precision. The benefits of the iterative ordering procedure are the largest on data tables with correlated attributes and heterogeneous attribute types.