Accuracy-Optimized Quantization for High-Dimensional Data Fusion

Slobodan Vucetic

Center for IST, Temple University, Philadelphia, PA 19122, USA
Email: vucetic@ist.temple.edu

Decentralized estimation is an essential problem for a number of data fusion applications. Given a system of n distributed data sources $S_1 \ldots S_n$ producing random vectors $X_1 \ldots X_n$ the task of decentralized estimation is to determine the optimal coding of each source such that the unknown continuous variable Y can be accurately estimated at the centralized site from the compressed vectors $\hat{X}_1,\ldots,\hat{X}_n$. The accuracy can be defined in terms of the Mean Square Quantization Error, MSQE = $E[(f(X_1,\ldots, X_n) - h(\hat{X}_1,\ldots,\hat{X}_n))^2]$, where f is the best approximation of the conditional expectation $E(Y|x_1,\ldots, x_n)$ obtained by the least squares estimation from data set $D = \{(x_1^k \ldots x_n^k, y^k), k = 1 \ldots N\}$, and h is an appropriate fusion function. While the necessary conditions for the optimal source coding and fusion function h have been derived [1], their practical implementation is constrained to low-dimensional and low-rate applications.

In this work, a computationally efficient and robust algorithm was developed for high-dimensional and high-rate decentralized estimation scenarios. The algorithm was developed based on the constraint that $h = f$. Starting from this constraint, it can be shown that for high-rate coding MSQE can be closely approximated by $E[\sum_i \sum_j d(x_{ij}, \hat{x}_{ij})]$, where x_{ij} is the j-th element of vector x_i generated by source S_i, and distortion measure d is defined as $d(x_{ij}, \hat{x}_{ij}) = (\int_{x_{ij}}^{\hat{x}_{ij}} g_{ij}(x)dx)^2 + \varepsilon$, where ε is a small positive constant. g_{ij} is the trend function defined as $g_{ij}(x) = \sqrt{E_{X_R}[f_{x_{ij}}(x, X_R)^2]}$, where $f_{x_{ij}}$ is the partial derivative of f with respect to x_{ij} and X_R is the difference vector $(X_1 \ldots X_n) \backslash X_{ij}$. Since the trend function should be estimated from the data set D, constant ε is introduced as a regularization factor that reduces the effects of the estimation error.

Using the companding function $c_{ij}(x) = \int_0^x g_{ij}(x)dx$, x_{ij} can be transformed into a new variable z_{ij} as $z_{ij} = c_{ij}(x_{ij})$. Therefore, MSQE can be approximated with $E[\sum_i \sum_j d_E(z_{ij}, \hat{z}_{ij})]$, where d_E is the Euclidean distance. As a consequence, determining the optimal coding for source S_i can be solved by applying standard Vector Quantization algorithms on data set $D_i = \{(z_{i1}^k \ldots z_{iI}^k), k = 1 \ldots N\}$, where z_{ij}^k is obtained by companding of x_{ij}^k and I is the dimensionality of vector X_i.

Experiments were performed on a 2-source 21-dimensional problem. The proposed algorithm is compared with standard vector quantization, due to lack of alternative high-rate and high-dimensional decentralized estimation algorithms. The results showed that the proposed algorithm was consistently more accurate than standard VQ and that the difference increased with increase in number of codewords and size of the data set.

[1] Lam W.M., Reibman A.R., Design of Quantizers for Decentralized Estimation Systems, IEEE Trans. on Communications, 41(11), 1602–1605, 1993