Efficient Regression for Big Data Problems using Adaptive Sparse Grids

Michael Lettrich
m.lettrich@mytum.de
Fakutät für Informatik, Technische Universität München

The amount of available data increases rapidly. This trend, often related to as *Big Data* challenges modern data mining algorithms, requiring new methods that can cope with very large, multi-variate regression problems. A promising approach that can tackle non-linear, higher-dimensional problems is regression using sparse grids. Sparse grids use a multi-scale system of grids with basis functions φ with local support to circumvent the curse of dimensionality [BG04].

Sparse grid based regression [Pfl10] iteratively constructs a function $\hat{f}(x) = \sum_i \alpha_i \varphi_i(x)$ that approximates the solution of the regression problem f finding optimal weights α_i for a fixed grid using least squares optimization. For further error reduction, refinement algorithms then add new basis functions and the optimization is repeated.

Current refinement algorithms add hierarchical child functions in all d input dimensions if a neighborhood with a high local error is discovered. This results in the addition of less beneficial basis functions if not all input dimensions are equally important. Instead, we propose a heuristic method that identifies and introduces only new basis functions with the potentially highest error reduction. We found, that the average of the local error r at data points x_i weighted with the evaluation of the basis function $\varphi(x_i)$ in question is suited as a refinement indicator $\alpha \approx n^{-1} \sum_{i=1}^{n} \varphi(x_i) r(x_i)$, as we then can use the heuristics from [Pfl10] to estimate potential error reduction. The number of data points n required for a significant approximation is in $\mathcal{O}(d)$.

The algorithm has been tested with both synthetic and real world datasets from the UCI repository. We have found that our algorithm offers a 30 to 50 percent improvement in error decay over current methods while reducing the grid size between 40 to 75 percent, allowing us to tackle large, high dimensional datasets more efficiently. The accuracy remains comparable with current approaches.

References

[BG04] Hans-Joachim Bungartz and Michael Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004.

[Pfl10] Dirk Pflüger. *Spatially Adaptive Sparse Grids for High Dimensional Problems*. PhD thesis, Technische Universität München, 02 2010.