Data Mining with Sparse Grids

Seminar: Computational Aspects of Machine Learning

Sebastian Kreisel

Datasets with large size and high-dimensional data pose a challenge even with the steadily growing computational power. Data mining algorithms often scale quadratic or worse in the number of data points and the computational effort usually grows exponentially with respect to the dimensionality of the data.

To tackle these problems *discretization methods* can be employed. Discretizing the feature space (given by the dataset) allows to handle a large amount of data by operating on carefully chosen *grid points* instead of the data points. Making the grid *sparse* mitigates the exponential dependency between computational effort and dimensionality without major drawbacks in accuracy.

This paper explores the sparse grid technique in the context of data mining. First, the core concepts of discretization will be explained by discussing full grid discretization of functions using equidistant grid points. However, this approach exhibits the *Curse of Dimensionality* and is thus not applicable to high-dimensional problems. To address that, a hierarchical structured grid is then examined and *sparse grids* are introduced which are able to deal with high dimensionality.

Through applying the sparse grids, it is possible to find a optimal structured grid for a general function (*a-priori*). Although this might suffice to accurately discretize well-behaved functions, machine learning tasks often require additional care. By introducing *spatial adaptivity* the grid can be modeled to the specific function at hand and the accuracy improves even in difficult scenarios.

After these basic notions, sparse grids are applied to the data mining tasks classification and regression. The commonly used *least squares estimation* gets review and then modified to conform with the previously established formulation of sparse grids. In order to examine the performance of sparse grids, results of artificial and real-life data mining scenarios are presented.

In the last section the implementation of sparse grids on modern systems is discussed briefly. Due to the hierarchical structure some consideration has to go into an computationally efficient implementation allowing parallelization and architecture dependent optimizations. One approach presented, disregards the nested structure of a hierarchical grid and trades unnecessary computations for better parallelization with good results.

To summarize, sparse grids are a viable option when confronted with high dimensionality and a large amount of data. This is confirmed by difficult test datasets and existing applications with real scenarios. Although sparse grids are challenging to implement efficiently there exist approaches exploiting the capabilities of modern hardware.