Improving Data Imputation for High Dimensional Datasets

Neta Rabin
Afeka - Tel-Aviv Academic College of Engineering
neta.rabin@gmail.com

Dalia Fishelov
Afeka - Tel-Aviv Academic College of Engineering and Tel-Aviv University
fishelov@gmail.com

Abstract

A common pre-possessing task in machine learning is to complete missing data entries in order to form a full dataset. Known techniques provide simple solutions to this problem by replacing the missing data entries with the mean or median value calculated for the known data of the same measurement or type. Alternatively, missing entries may be replaced by random values that are drawn from a distribution that fits to the known data values. More sophisticated methods use regression to complete missing data. For a given column with missing values, the column is regressed against other columns for which the values are known. In case the dimension of the input data is high, it is often the case that the data columns lay on a low-dimensional space. Constriction of low-dimensional embedding of the subset of the complete data produces a loyal representation of it. This new representation can now be used to construct regression or regression-type models for imputing the missing values. In previous work [1], we proposed a two-step algorithm for data completion. The first step utilizes a non-linear manifold learning technique, named diffusion maps [2], for reducing the dimension of the data. This method faithfully embeds complex data while preserving its geometric structure. The second step is the Laplacian pyramids [3] multi-scale method, which is applied for regression. Laplacian pyramids construct kernels of decreasing scales to capture finer modes of the data and the scale is automatically fit to the data density and noise. In this work, we improve the previous method by considering the dual geometry of the dataset. We construct a model that learns the geometry of the rows and of the columns of the full subset alternately. Experimental results demonstrate the efficiency of our approach on a publicly available dataset.

References

1. N. Rabin and D. Fishelov. Missing Data Completion Using Diffusion Maps and Laplacian Pyramids. International Conference on Computational Science and Its Applications ICCSA. (2017) 284-297.
2. R. Coifman and S. Lafon. Diffusion Maps. Appl. Comput. Harmon. Anal. 21 (2006) 5-30.
3. N. Rabin and R. Coifman. Heterogeneous Datasets Representation and Learning Using Diffusion Maps And Laplacian Pyramids. 12th SIAM International Conference on Data Mining. (2012) 189-199.