Assessing Performances of Learning Algorithms when Trained using Incomplete Data

M. Kashif Gill[1], Tirusew Asefa[2], and Mac McKee[1]
Utah Water Research Laboratory and Department of Civil and Environmental Engineering, Utah State University

Abstract. A common practice in pre-processing step of hydrological modeling is to ignore observations with any missing variable values at any given time step even if it is only one of the independent variable that is missing. These rows of data are labeled incomplete and would not be used in either model building or subsequent testing and verification steps. This is not necessarily the best way of doing it as information is lost when incomplete rows of data are thrown out. Learning algorithms are affected by such problems more than physically-based models as they rely heavily on the data to learn the underlying input/output relationships. In this study, the extent of damage to the performance of the learning algorithm due to missing data is explored in a field-scale application. We have tested and compared the performance of two well-known learning algorithms, namely Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) for short-term prediction of groundwater levels in a well field. The comparison of these two algorithms is made using various percentages of missing data. In addition to understanding the relative performance of these algorithms in dealing with missing data, a solution in the form of an imputation methodology is proposed for filling the data gaps. The proposed imputation methodology is tested against observed data.

[1] Utah Water Research Laboratory and Department of Civil and Environmental Engineering, Utah State University, Logan, 84321 UT
[2] Source Rotation and Environmental Protection Department, Tampa Bay Water, Clearwater 33761 FL