2009-11-01

Dataset threshold for the Performance Estimators in Supervised Machine Learning Experiments

Zanifa Omary
Technological University Dublin, zanifa.omary@student.dit.ie

Fredrick Mtenzi
Technological University Dublin, Fredrick.Mtenzi@tudublin.ie

Recommended Citation

Omary, Z. & Mtenzi, F. (2009). Dataset threshold for the Performance Estimators in Supervised Machine Learning Experiments. *Proceedings of the 4th International Conference for Internet Technology and Secured Transactions*, November 9-12, London, UK. doi:10.1109/ICITST.2009.5402500

Dataset threshold for the Performance Estimators in Supervised Machine Learning Experiments

Zanifa Omary, Fredrick Mtenzi
School of Computing
Dublin Institute of Technology
zanifa.omary@student.dit.ie, fredrick.mtenzi@dit.ie

Abstract

The establishment of dataset threshold is one among the first steps when comparing the performance of machine learning algorithms. It involves the use of different datasets with different sample sizes in relation to the number of attributes and the number of instances available in the dataset. Currently, there is no limit which has been set for those who are unfamiliar with machine learning experiments on the categorisation of these datasets, as either small or large, based on the two factors. In this paper we perform experiments in order to establish dataset threshold. The established dataset threshold will help unfamiliar supervised machine learning experimenters to categorize datasets based on the number of instances and attributes and then choose the appropriate performance estimation method. The experiments will involve the use of four different datasets from UCI machine learning repository and two performance estimators. The performance of the methods will be measured using f1-score.

Publication details: In the conference proceeding of the 4th International Conference for Internet Technology and Secured Transactions, November 9-12, 2009, London, UK.