

Complementing real datasets with simulated data: a regression-based approach

M. A. Ortiz-Barrios · J. Lundstrom · J. Synnott · E. Jarpe · A. Sant'Anna

Abstract

Activity recognition in smart environments is essential for ensuring the wellbeing of older residents. By tracking activities of daily living (ADLs), a person's health status can be monitored over time.

Nonetheless, accurate activity classification must overcome the fact that each person performs ADLs in different ways and in homes with different layouts. One possible solution is to obtain large amounts of data to train a supervised classifier. Data collection in real environments, however, is very expensive and cannot contain every possible variation of how different ADLs are performed. A more cost-effective solution is to generate a variety of simulated scenarios and synthesize large amounts of data. Nonetheless, simulated data can be considerably different from real data. Therefore, this paper proposes the use of regression models to better approximate real observations based on simulated data. To achieve this, ADL data from a smart home were first compared with equivalent ADLs performed in a simulator. Such comparison was undertaken considering the number of events per activity, number of events per type of sensor per activity, and activity duration. Then, different regression models were assessed for calculating real data based on simulated data. The results evidenced that simulated data can be transformed with a prediction accuracy $R^2 = 97.03\%$.

Keywords Activity recognition · Activity duration · Regression analysis · Non-linear models · Determination coefficient · Quantile-quantile plots