Abstract

Sherine . Nagy

Handling Varying Amounts of Missing Data when Classifying Mental-Health Risk Levels

One of the main challenges of classifying clinical data is determining how to handle missing features. Most research favours imputing of missing values neglecting records that include missing data, both of which can degrade accuracy when missing values exceed a certain level. In this research we propose a methodology to handle data sets with a large percentage of missing values and with high variability in which particular data are missing. Feature Selection is effected by picking variables sequentially in order of maximum correlation with the dependent variable and minimum correlation with variables already Selected. Classification models are generated individually for each test case based on its particular feature set and the matching data values available in the training population. The method was applied to real patients' anonymous mental-health data where the task was to predict the suicide risk judgement clinicians would give for each patient's data, with eleven possible outcome classes: zero to ten, representing no risk to maximum risk. The results compare favourably with alternative methods and have the advantage of ensuring explanations of risk are based only on the data given, not imputed data. This is important for clinical decision support systems using human expertise for modelling and explaining predictions.