576

DATA MINING METHODS FOR CLASSIFICATION OF MEDIUM-CHAIN ACYL-COA DEHYDROGENASE DEFICIENCY (MCADD) USING NON-DERIVATIZED TANDEM MS NEONATAL SCREENING DATA

F.J. Eyskens

Paediatrics, Antwerp University Hospital, Antwerpen, Belgium

Newborn screening programs for metabolic disorders using tandem mass spectrometry are widely used. Medium Chain Acyl coA dehydrogenase deficiency (MCADD) is the most prevalent mitochondrial fatty acid oxidation defect and it has been proven that early detection of this metabolic disease decreases mortality and improves the outcome. In previous studies, data mining methods on derivatized tandem MS datasets, have shown high classification accuracies. However, no machine learning methods currently have been applied to datasets based on non-derivatized screening methods.

A dataset with 31,924 blood samples was collected using a non-derivatized screening method as part of a systematic newborn screening by the PCMA screening center (Belgium). Nine MCADD cases were present in this partially MCADD-enriched dataset. We compared three data mining methods, namely C4.5 decision trees, logistic regression and ridge logistic regression and evaluated their applicability as a diagnostic support tool. Within a 9-fold stratified crossvalidation setting, a grid search was performed for each model for a wide range of model parameters, included variables and classification thresholds.

The best performing model used ridge logistic regression and achieved a sensitivity of 100% and specificity of 99.987% in a 9-fold stratified crossvalidation setting. Our results on non-derivatized tandem MS neonatal data show a slightly better performance to the current state-of-the-art for derivatized MS data while retaining more interpretability and requiring less variables. The results indicate the potential value of data mining methods as a diagnostic support tool and show that comparable classification performances are achieved for non-derivatized data compared to derivatized datasets.