Classification of rare diseases for medical decision support systems using RWeka

Tim Van den Bulcke[1,2,*], Kristien Wouters[1,3], Paul Vanden Broucke[1,2], Vanessa De Wit[1,2], François Eyskens[4,5]

1. biomina (biomedical informatics research center Antwerp), CRC Antwerp, Antwerp University Hospital - University Antwerp, Wilrijkstraat 10, Edegem, Belgium.
2. i-ICT, Antwerp University Hospital, Wilrijkstraat 10, Edegem, Belgium.
3. Dept. of Scientific Coordination, Antwerp University Hospital, Wilrijkstraat 10, Edegem, Belgium.
4. Provinciaal Centrum voor de Opsporing van Metabole Aandoeningen (PCMA), Antwerpen, Belgium.
5. Dept. of Paediatrics/Metabolic Diseases, University Hospital Antwerp, Edegem, Belgium.
*Contact author: tim.van.den.bulcke@uza.be

Keywords: machine learning, rare diseases, RWeka

Newborn screening tests for treatable rare diseases, do not only improve the infant life expectancy and quality of life, but also cut down on health-care spending. Several metabolic disorders can be detected in infants via a blood sample which is taken within a few days after birth using a standard heel prick test. Currently, the diagnosis is determined by a medical expert using previously published cutoff values. However, these cutoffs are chosen conservatively and provide only a univariate approach to the rare disease classification. This results in a lower accuracy and a higher number of false positives to what is technically achievable with the available data.

In order to address these issues, we compared and assessed the diagnostic performance of a number of machine learning methods for the classification of MCADD (a specific metabolic disorder). The use of the **RWeka** package allowed us to setup an elegant analysis pipeline by providing a fast and uniform interface to various machine learning methods in *Weka* and by integrating this with the statistical capabilities of *R*. The best performing model achieved a sensitivity of 100% and a specificity of 99.987%, obtained in a stratified cross-validation setting. This resulted in a significant performance improvement compared to the current state-of-the-art and indicates the potential value of machine learning methods as a decision support tool for diagnosis of rare diseases.

References

Wilcken B. (2009). Fatty acid oxidation disorders: outcome and long-term prognosis. *J Inherit Metab Dis 1-6.*

Eyskens FJM, Philips E. (2007). Newborn mass screening using tandem mass spectrometry: results of the validation and comparison of two methods (derivatized/non-derivatized). *J Inherit Metab Dis 30 (Suppl. 1).*

Chace D, Kalas T, Naylor E. (2003). Use of tandem mass spectrometry for multianalyte screening of dried blood specimens from newborns. *Clin Chem;49(11):1797–817.*

Baumgartner C, Böhm C, Baumgartner D. (2005). Modelling of classification rules on metabolic patterns including machine learning and expert knowledge. *J Biomed Inform 38(2):89–98.*

Ho S, Lukacs Z, Hoffmann G, Lindner M, Wetter T. (2007). Feature construction can improve diagnostic criteria for high-dimensional metabolic data in newborn screening for medium-chain acyl-CoA dehydrogenase deficiency. *Clin Chem 53(7):1330–7.*