Supervised machine learning techniques for the classification of metabolic disorders in newborns

C. Baumgartner, C. Böhm2, D. Baumgartner, G. Marini,

K. Weinberger, B. Olgemöller, B. Libel and A. A. Roscher

Background

- Usually blood sample that is collected during the first few days of life to screen for metabolic disorders.

- Test now simultaneously screens the concentrations of up to 50 metabolites to detect more than 20 inherited metabolic disorders .
- **The amount and complexity of the experimental data is quickly becoming unmanageable to be evaluated manually.**

Objective

- Focusing on two representative inborn errors of metabolism—
- phenylketonuria (PKU), an amino acid disorder,
- and mediumchain acyl-CoA dehydrogenase deficiency (MCADD), a fatty acid oxidation defect
- six well-established supervised machine learning techniques were evaluated to determine the ‘best’ screening model

Criteria

- discriminatory performance of the learning algorithm based on pre-classified, selected and clinically validated sub-databases of PKU and MCADD newborns.
- diagnostic prediction of constructed classifiers with optimizing sensitivity and minimizing the number of false positive results considering a large database.

Methods

Used Tandem mass spectrometry (MS/MS)

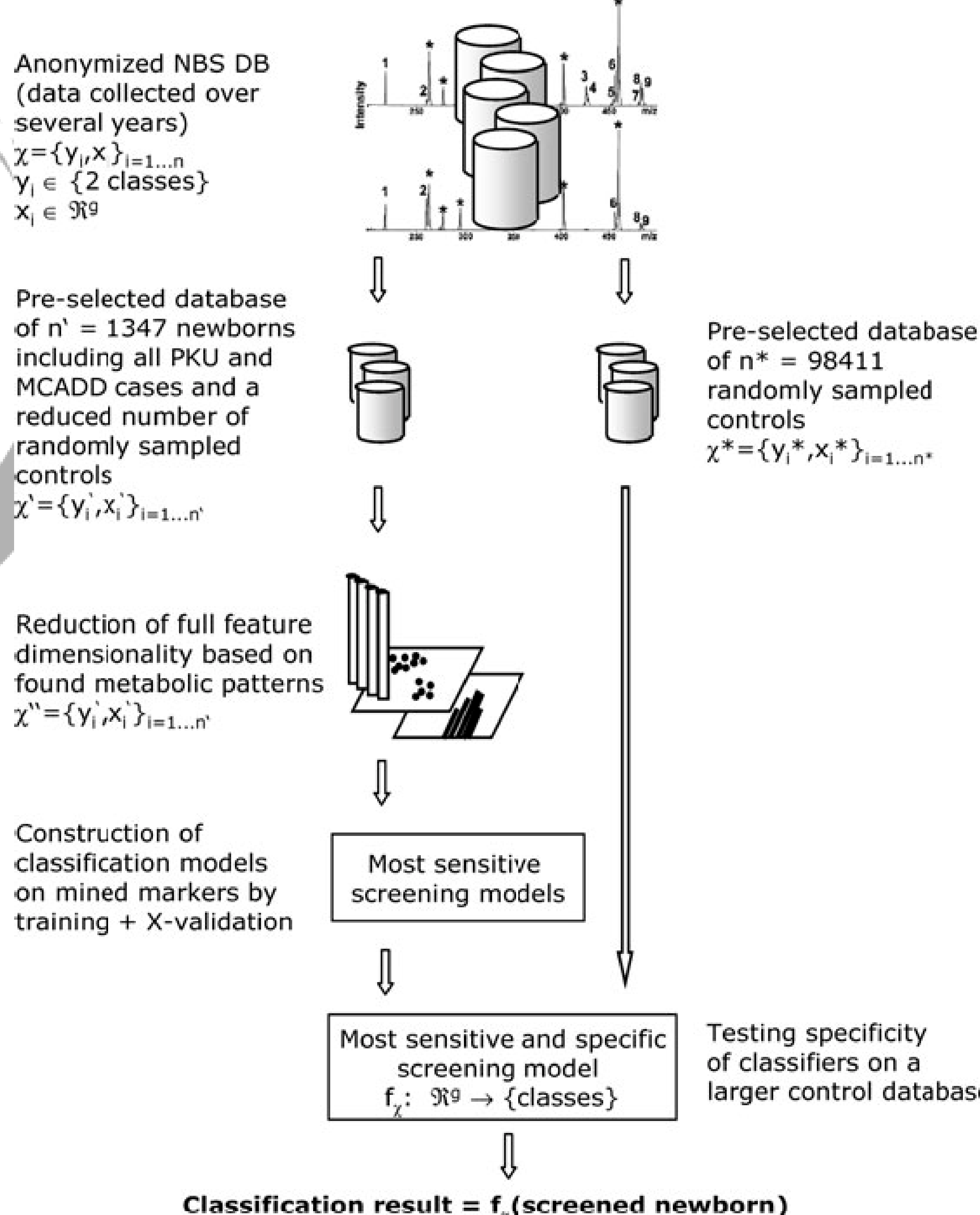
To find

- Phenylketonuria is an amino acid disorder which is caused primarily by a deficiency of phenylalanine hydroxylase
- Medium-chain acyl-CoA dehydrogenase deficieny is a fatty acid oxidation defect which leads to an accumulation of fatty acids and a decrease in cell energy metabolism.

experimental datasets were anonymously provided from the newborn screening program in Bavaria, Germany

- *Discriminate analysis (DA)*
- *Logistic regression analysis (LRA)*
- *Decision trees (DT)*
- *K-nearest neighbor classifier (k-NN)*
- *Artificial neural networks (ANN)*
- *Support vector machines (SVM)*

Fig 1



Anonymized NBS DB (data collected over several years)
$\chi=\{y_i,x\}_{i=1\ldots n}$
$y_i \in \{2 \text{ classes}\}$
$x_i \in \Re^g$
Intensity
m/z
Pre-selected database of n‘ = 1347 newborns including all PKU and MCADD cases and a reduced number of randomly sampled controls
$\chi`=\{y_i`,x_i`\}_{i=1\ldots n`}$
Pre-selected database of n* = 98411 randomly sampled controls
$\chi^*=\{y_i^*,x_i^*\}_{i=1\ldots n^*}$
Reduction of full feature dimensionality based on found metabolic patterns
$\chi``=\{y_i`,x_i`\}_{i=1\ldots n`}$
Construction of classification models on mined markers by training + X-validation
Most sensitive screening models
Most sensitive and specific screening model
$f_\chi: \Re^g \rightarrow \{\text{classes}\}$
Testing specificity of classifiers on a larger control database
Classification result = f_χ(screened newborn)

Use WEKA

- Weka tool set and ADE-4 were used to evaluate results and perform DA and statistical analysis

Winner

- Logistic regression analysis led to superior classification rules (sensitivity>96.8%, specificity >99.98%) compared to all investigated algorithms.
- For the routine clinical screening LRA models proved particularly feasible because of their highly significant prognostic accuracy.

results

- To sum it up, the top three machine learning techniques, LRA SVM and ANN, delivered results of high predictive power
- the DA classifier discriminated worse for both disorders

points

- Machine learning works well in defined tasks.
- LRA and DA work is a similar manner yet one is the best one is the worst. Lesson?
- ANN as labeled the future by many, do you think ANN will become the best