dHealth 2019 – From eHealth to dHealth
D. Hayn et al. (Eds.)

doi:10.3233/978-1-61499-971-3-89

Ensemble Based Approach for Time Series Classification in Metabolomics

Michael NETZER[a,1], Friedrich HANSER[a], Marc BREIT[a], Klaus M. WEINBERGER[a,c], Christian BAUMGARTNER[b] and Daniel BAUMGARTEN[a]
[a]*Institute of Electrical and Biomedical Engineering, UMIT, Austria*
[b]*Institute of Health Care Engineering, Graz University of Technology, Austria*
[c]*sAnalytiCo Ltd., Belfast, United Kingdom*

Abstract. Background: Machine learning is one important application in the area of health informatics, however classification methods for longitudinal data are still rare. Objectives: The aim of this work is to analyze and classify differences in metabolite time series data between groups of individuals regarding their athletic activity. Methods: We propose a new ensemble-based 2-tier approach to classify metabolite time series data. The first tier uses polynomial fitting to generate a class prediction for each metabolite. An induced classifier (k-nearest-neighbor or naïve bayes) combines the results to produce a final prediction. Metabolite levels of 47 individuals undergoing a cycle ergometry test were measured using mass spectrometry. Results: In accordance with our previous work the statistical results indicate strong changes over time. We found only small but systematic differences between the groups. However, our proposed stacking approach obtained a mean accuracy of 78% using 10-fold cross-validation. Conclusion: Our proposed classification approach allows a considerable classification performance for time series data with small differences between the groups.

Keywords. biomarkers, time series, classification, kinetics

1. Introduction

Machine learning (ML) is amongst the greatest application challenges of Health Informatics (HI), resulting in improved medical diagnoses, disease analyses, and pharmaceutical development [1]. Recently, Rav`ı et al. [2] reviewed the research employing deep learning in health informatics including medical informatics, public health, sensing, bioinformatics and imaging. In the area of biomedical time series analysis, McCoy et al. [3] recently introduced a machine learning method for forecasting hospital discharge volume using a Bayesian forecaster. In general, machine learning methods can be categorized into unsupervised and supervised methods. Unsupervised methods do not require labeled input and search for pattern. A popular unsupervised method is clustering, where groups of instances with similar properties are identified. In contrast, supervised methods require a learning input (outcome variable). Depending on the type of the outcome variable we distinguish between regression (numeric outcome) and classification problems (categoric outcome).

[1] Corresponding Author: Michael Netzer, UMIT Hall, Eduard-Wallnöfer-Zentrum 1, 6060 Hall in Tirol, Austria, E-Mail: michael.netzer@umit.at.

Feature selection is an important step to reduce the number of variables to a smaller set with higher discriminatory ability. The advantages include faster, more cost-effective and better interpretable learning models [4]. In our previous work [5] we introduced a feature selection approach to identify clusters of metabolic biomarker candidates that considerably change over time during physical activity. In particular, biomarker candidates were chosen by using maximum fold changes (MFCs) of metabolite levels and P-values resulting from statistical hypothesis testing. We identified characteristic kinetic patterns using a mathematical modeling approach. Examples for such patterns include early, late, or other forms of kinetic response patterns. The method utilized polynomial fitting and clusters of metabolites with similar kinetics were analyzed applying a cluster analysis. Finally, kinetic shape templates were identified that represent different kinetic response patterns (e.g., sustained, early, late response). The aim of this work is to analyze and classify differences in metabolite time series data between groups of individuals regarding their athletic activity. In particular, we developed a new ensemble-based approach based on the combination of class predictions using polynomial fitting. Our main contribution is the introduction of a new classification method for metabolic time series data. We here exemplarily distinguish athletes from non-athletes; however, this approach can be also applied to other binary classification problems in this context.

2. Material and Methods

2.1. Dataset

A total of 47 individuals underwent a standardized cycle ergometry experiment (starting workload of 50 Watt). The study population can be categorized into two classes: i) average physical activity ($n_{Samples} = 24$) vs. competitive athletes ($n_{Samples} = 23$). Starting the exercise with a workload of 50 W, the workload was increased by 25 W every 3 minutes up to the individual's maximum physical load [5]. Due to different individual maximum workloads, the time series data were normalized and linearly interpolated. Based on dried blood spots of samples taken from the earlobe, levels of 110 metabolites were measured using triple quadrupole tandem mass spectrometry (MS/MS) applying stable isotope dilution for metabolite quantitation. Groups of quantified metabolites include acylcarnitines, amino acids and sugars. Missing values were imputed using a k-nearest-neighbor approach [6]. See [5] for a detailed description of data acquisition and preparation.

2.2. Statistical Analysis

Statistical analysis was performed using non-parametric tests for repeated measures data [7]. The p-value was obtained using an ANOVA-type statistic. In contrast to parametric approaches rank-based methods allow to analyze categorical or heavily skewed data in a systematic way [8].

2.3. Machine Learning Approach

We use a k-nearest neighbor (kNN) and naive bayes (NB) classifier. kNN is a popular non-parametric method that determines the class of a new instance based on the majority

class of its k nearest neighbors. The NB model assumes independence between variables. This assumption is not valid in general and may also be validated in our dataset, however NB is a popular classifier that performs well in many classification tasks [9]. The performance of models is estimated using 10-fold cross validation summarized by micro-average. In particular, the dataset is divided into ten partitions using nine parts for training and the remaining subset for testing. This procedure is repeated ten times. For every iteration the accuracy is calculated and finally summarized by calculating the mean. However, in particular for imbalanced datasets the accuracy is inappropriate. Consequently, we also calculate the parameter $\kappa = \frac{O-E}{1-E}$, where O is the observed and E the expected accuracy to overcome this problem.

2.4. Ensemble Learning Approach

In this work we introduce an ensemble learning approach consisting of the following steps:

1. Training step
   (a) Calculate a representative time course for each metabolite of each group. In particular, we calculate a median curve for each group. The value for each time point is consequently the median value of the group.
   (b) Fit polynomial (degree of 9 as used in [5]) for each class using the median value of each time point ($\hat{y}_t$).
2. Classification step of an unlabeled sample s
   (a) Calculate the residual sum of squares (RSS) for each class $c \in C$ of each metabolite $m \in M$ for all time points $t \in T$

$$RSS(c) = \sum_t (\hat{y}_t - y_t)^2 \quad (1)$$

   , where $\hat{y}_t$ represents the value for each time point from step 1 and y_t is the actual value.
   (b) Determine for each metabolite $m \in M$ the class $c \in C$ by selecting the class with the smallest RSS

$$c_m = \underset{c}{\operatorname{argmin}} RSS(c) \quad (2)$$

   (c) Select the final class using class predictions of the previous steps (class selection step).

For the class selection step, we consider three methods:

- *Majority voting*: Select the majority class based on the class predictions for all metabolites. For instances, having 5 metabolites where four metabolites select class 1 and one metabolite selects class 2, we use class 1.
- *Majority voting and feature selection*: Use only a subset of n_{top} ranked features for the voting step. The ranking is calculated using the area between the time

curves (ABC). The idea is that discriminatory metabolites are represented by higher ABC values. Additionally, we weight the ABC scores by adjusted R^2 (i.e., $s = ABC \times R^2_{adjusted}$).

- *Stacking approach*: We propose a 2-tier approach to predict the classes based on the time series data. An induced classifier uses the class predictions for each metabolite to produce a final prediction (see also Figure 1). In our experiments we use kNN and NB as classification methods.

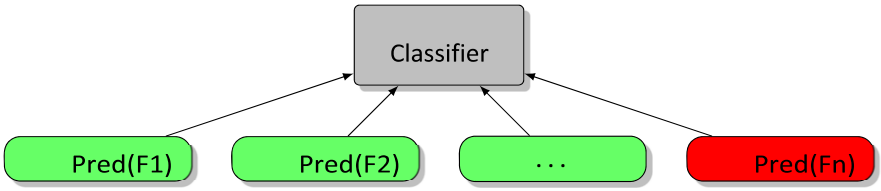

Figure 1. Stacking approach using the class prediction of each metabolite. The color represents the class prediction considering each metabolite (green = class 1, red = class 2).

3. Results

3.1. Statistical Evaluation and Clustering

Figure 2 visualizes metabolites and corresponding FDR adjusted p-values for comparing groups (i.e., average vs. competitive athletic activity) and time (i.e., varying workload). Table 1 depicts the p-values for change of time, group differences and interaction of both of these variables.

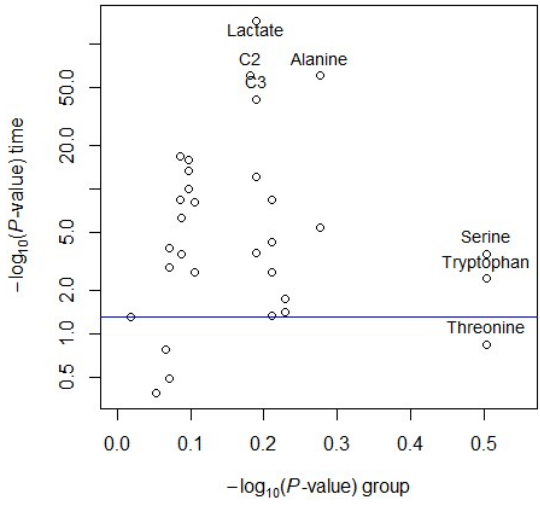

Figure 2. Scaled p-values comparing groups and time (workload). The y-axis is plotted in log scale. Features above the horizontal blue line significantly change over time ($p < 0.05$).

Table 1. FDR adjusted P-values for change of time, group differences (average vs. competitive athletic) and interaction of both variables (group:time) calculated using non-parametric tests for repeated measures data [7]. The majority of metabolites significantly change over time (i.e., P-value for time < 0.05).

	time	**group**	**group:time**
Lactate	< 0.01	0.65	0.88
C2	< 0.01	0.66	0.59
Alanine	< 0.01	0.53	0.87
C3	< 0.01	0.65	0.59
Arginine	< 0.01	0.82	0.88
Glycine	< 0.01	0.80	0.88
C4	< 0.01	0.80	0.78
Tyrosine	< 0.01	0.65	0.87
Glutamic Acid	< 0.01	0.80	0.88
Phenylalanine	< 0.01	0.62	0.88
C3 DC M	< 0.01	0.82	0.78
C5 OH	< 0.01	0.78	0.78
C5	< 0.01	0.82	0.88
Glucose	< 0.01	0.53	0.87
Methionine	< 0.01	0.62	0.87
C18 2	< 0.01	0.85	0.78
Ornithine	< 0.01	0.65	0.78
Serine	< 0.01	0.31	0.88
Histidine	< 0.01	0.82	0.88
C18	< 0.01	0.85	0.78
C16	< 0.01	0.78	0.64
xLeucine	< 0.01	0.62	0.88
Tryptophan	< 0.01	0.31	0.88
Lysine	0.02	0.59	0.86
Valine	0.04	0.59	0.88
Proline	0.05	0.62	0.88
C18 1	0.05	0.96	0.86
Threonine	0.14	0.31	0.78
C0	0.17	0.86	0.88
Aspartic Acid	0.33	0.85	0.87
Citrulline	0.41	0.88	0.88

3.2. Classification Performances

Figure 3 shows the classification performances comparing the ensemble-based methods using 10-fold cross validation. The number of top ranked features (n_{top}) for the feature selection approach (Mj. + FS) was set to $\frac{number\ of\ metabolites}{3}$. The black lines indicate the median accuracy of each method.

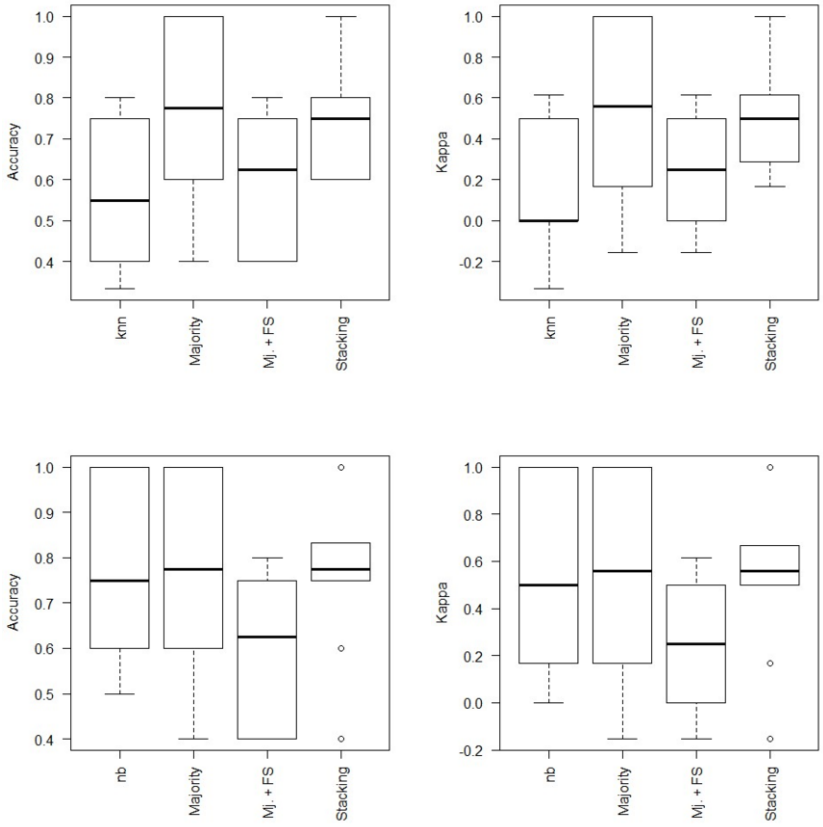

Figure 3. Boxplots of accuracy (left) and kappa (right) values for predicting average vs. competitive athletic using kNN (first row) and NB (second row) classifier.

4. Discussion and Conclusion

In this work, we analyzed metabolic changes by considering change over time (i.e., varying Watt levels) and group differences. In summary, a total of 25 metabolites changed significantly over time ($p < 0.05$). Similar to our previous work [5], the smallest p-values were observed for lactate, alanine, acetylcarnitine (C2) and related short-chain acylcarnitines (C3, C5). Interestingly, no significant changes were observed comparing average vs. competitive athletic groups. The smallest p-values were observed for serine,

tryptophan, and threonine. However, considering time charts, we identified a clear trend by observing systematically higher metabolite levels for all time points for these metabolites. The missing significance levels may be a result of the relatively high standard deviation due to the small sample size and heterogeneities within the groups (e.g., different individual maximum Watt levels and individual anaerobic thresholds). Considering these three metabolites biochemically, the carbon skeletons of serine, threonine and tryptophan are used to form pyruvate that is used as fuel in the mitochondria by conversion to acetyl CoA (TCA cycle), converted to lactate or utilized to produce glucose in the liver [10].

Even though the statistical approach revealed no significant metabolites when comparing the classes, we obtain accuracy values of 75%. The highest mean accuracy of 76.83% was obtained by using our stacking approach using NB as classifier. The standard deviations of the resulting performance values were also comparably low. Interestingly, the proposed feature selection step did not improve the performance. Our assumption is that the proposed feature ranking method is very prone to noise.

The degree of 9 used for polynomial fitting was based on our previous work, however this value can be further optimized to increase accuracy values.

In summary, we introduced a new ensemble-based classification method for time series metabolite data. For each metabolite, a class prediction is produced using polynomial fitting. The predictions are summarized by using an induced classifier to obtain a final classification of a new unlabeled sample. Note that this approach can be also applied to proteomic or genomic datasets.

5. Acknowledgements

Michael Netzer was supported by the Tiroler Wissenschaftsfond. The authors thank Prof. Dr. Elske Ammenwerth for her comments improving the paper.

References

[1] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the loop?, *Brain Informatics* 3(2) (2016), 119–131.

[2] D. Ravı, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo and G.-Z. Yang, Deep learning for health informatics, *IEEE journal of biomedical and health informatics* 21(1) (2017), 4–21.

[3] T.H. McCoy, A.M. Pellegrini and R.H. Perlis, Assessment of Time-Series Machine Learning Methods for Forecasting Hospital Discharge Volume, *JAMA network open* 1(7) (2018), 184087–184087.

[4] Y. Saeys, I. Inza and P. Larranaga,˜ A review of feature selection techniques in bioinformatics., *Bioinformatics* 23(19) (2007), 2507–2517.

[5] M. Breit, M. Netzer, K.M. Weinberger and C. Baumgartner, Modeling and Classification of Kinetic Patterns of Dynamic Metabolic Biomarkers in Physical Activity., *PLoSComputBiol* 11(8) (2015), 1004454. doi:10.1371/journal.pcbi.1004454. http://dx.doi.org/10.1371/journal.pcbi.1004454.

[6] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown and D. Botstein, Imputing missing data for gene expression arrays, Stanford University Statistics Department Technical report, 1999.

[7] K. Noguchi, Y.R. Gel, E. Brunner and F. Konietschke, nparLD: An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments, *Journal of Statistical Software* 50(12) (2012), 1–23. http://www.jstatsoft.org/v50/i12/.

[8] F. Konietschke, A.C. Bathke, L.A. Hothorn and E. Brunner, Testing and estimation of purely nonparametric effects in repeated measures designs, *Computational Statistics & Data Analysis* 54(8) (2010), 1895–1905.

[9] J. Wolfson, S. Bandyopadhyay, M. Elidrisi, G. Vazquez-Benitez, D.M. Vock, D. Musgrove, G. Adomavicius, P.E. Johnson and P.J. O'Connor, A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data, *Statistics in medicine* 34(21) (2015), 2941–2957.

[10] D.M. Medeiros, R.E. Wildman et al., *Advanced human nutrition*, Jones & Bartlett Publishers, 2013.