CLASSIFICATION ON HIGH DIMENSIONAL METABOLIC DATA: PHENYLKETONURIA AS AN EXAMPLE

Christian Baumgartner, PhD[1], Daniela Baumgartner, MD[2], Christian Böhm, PhD[3]

[1]Institute for Information Systems, University for Health Informatics and Technology Tyrol, Innrain 98, A-6020 Innsbruck, Austria, [2]Department of Pediatrics, University of Innsbruck, Anichstrasse 35, A-6020 Innsbruck, Austria, [3]Institute for Computer Science, University of Munich, Oettingenstrasse 67, D-80538 Munich, Germany.

Abstract

Tandem mass spectrometry is a promising new screening technology which permits screening within one analytical run not only for phenylketonuria (PKU) but also for a wide range of other metabolic disorders in newborns.

We investigated two symbolic supervised machine learning techniques - logistic regression analysis (LRA) and decision trees (DT), where the knowledge is represented in an explicit way - to find classification rules for the presence of PKU. Our experiments were performed on pre-classified newborn screening data including a metabolite spectrum of 14 amino acids. LRA and DT classifiers showed high classification performance with a sensitivity of ≥ 97.7% and a specificity of ≥ 99.8%. In addition to the established diagnostic metabolites of phenylalanine and tyrosine, we also included alternative constellations of metabolites in our models showing comparable results in predictive power.

The presented machine learning techniques are appropriate to investigate metabolic patterns in newborn screening data for constructing classification models for PKU.

Key Words

Supervised machine learning, phenylketonuria, newborn screening, tandem mass spectrometry, metabolomics.

1. Introduction

Newborn screening involves laboratory testing of all newborn infants for certain genetic/metabolic or endocrine disorders of body chemistry. Inborn metabolism errors can hinder an infant's normal physical and mental development in a variety of ways. Recently, a new screening methodology based on the so-called tandem mass spectrometry (MS/MS) has been developed. This spectrometry technique can detect body fluids that are elevated or diminished in certain metabolic disorders. Thus, it is possible to screen for more than 20 inherited metabolic disorders with a single test. Considering the amount and complexity of data generated by MS/MS, it is increasingly difficult to derive medical interpretation by conventional means. Therefore, it might be indicated to apply machine learning techniques to discover and mine metabolic patterns in these data with respect to derive classification rules for the interpretation of high dimensional metabolic datasets.

As an example we investigated classification rules for testing every newborn for phenylketonuria (PKU, OMIM #261600), an amino acid disorder, where phenylalanine cannot be metabolized to tyrosine due to a blockade of the enzyme phenylalanine-hydroxylase. This results in excess levels of phenylalanine in body fluids with possible neurotoxic reactions [1]. We apply two well-established symbolic supervised machine learning techniques, logistic regression analysis (LRA) and decision trees (DT), to the problem of constructing a classification model in order to judge how well a newborn disorder can be predicted. In symbolic methods, the knowledge is represented in an explicit way, e.g. in a formula or in a tree-like structure, whereas non-symbolic methods (e.g. k-nearest neighbor classifier or artificial neural networks) keep the knowledge only implicitly in internal data structures which cannot directly be interpreted by a clinical expert.

To summarize, our task is to develop a symbolic classifier with highest classification performance from newborn screening data. With respect to the classifier's simple handling and understandability in the daily clinical practice we denounce classification rules of the proposed symbolic learning algorithms under the condition that they show high discriminatory performance with high predictive power.

2. Systems and Methods

Mass spectrometry

The mass spectrometer is a device that separates and quantifies ions based on their mass/charge (m/z) ratios. Characteristical patterns of fragments and relative peak intensities in the resulting spectrum allow qualitative as well as quantitative determination of chemical compounds. By coupling two mass spectrometers, usually separated by a reaction chamber or collision cell, the modern tandem mass spectrometry (MS/MS) allows simultaneous analysis of multi-compounds in a high-throughput process. MS/MS thus permits very rapid, sensitive and, with appropriate internal standards, accurate measurement of many different types of

metabolites with minimal sample preparation and with adequate throughput to handle the large number of samples that are processed in newborn screening programs.

Metabolic data

Our experimental dataset was anonymously provided from the newborn screening program in Bavaria, Germany [2][3]. From the given database a total amount of 1599 clinically validated newborn datasets in two classes were selected. 1241 randomised controls and 307 cases designated as PKU patients were sampled within two weeks after birth. Table 1 summarizes all analysed metabolites from a single blood spot representing a spectrum of 14 metabolites in amino acid metabolism.

Table 1: Metabolites from MS/MS analysis of a single blood sample

Amino acids (symbol)	PKU	Controls
Alanine (Ala)	364.6 ± 118.7	508.9 ± 210.7
Arginine (Arg)	734.6 ± 496.2	90.9 ± 49.7
Argininosuccinate (Argsuc)	0.22 ± 0.80	0.01 ± 0.02
Citrulline (Cit)	45.1 ± 37.4	28.7 ± 39.9
Glutamate (Glu)	2595 ± 1790	235.9 ± 74.0
Glycine (Gly)	248.2 ± 179.7	624.2 ± 315.9
Methionine (Met)	29.2 ± 12.9	24.3 ± 7.4
Ornitine (Orn)	146.5 ± 86.2	85.2 ± 60.7
Phenylalanine (Phe)	721.6 ± 426.3	57.9 ± 17.9
Pyroglutamate (Pyrglt)	19.6 ± 13.7	51.8 ± 31.6
Serine (Ser)	710.5 ± 362.3	400.6 ± 358.2
Tyrosine (Tyr)	83.5 ± 36.0	97.2 ± 64.2
Valine (Val)	203.9 ± 64.9	170.6 ± 61.3
Leuzine+Isoleuzine (Xle)	144.8 ± 68.3	264.5 ± 107.7

Concentrations of amino acids (mean ± sd) for PKU and controls classes are denounced in µmol/L.

Supervised machine learning techniques

Usually, for a supervised classification problem, the training data sets are in the form of a set of tuples $\{(y_1, x_{1,j}),\ldots, (y_n, x_{n,j})\}$ where y_i is the class label and x_{ij} is the set of attributes for the instances. The task of the learning algorithm is to produce a classifier (model) to classify the instances into the correct class.

Both learning methods (LRA and DT) used in this study were obtained from the WEKA machine learning package (http://www.cs.waikato.ac.nz/~ml/weka). An established methodology to evaluate the robustness of the classifier is to perform a cross validation on the classifier. Ten fold cross validation has been proved to be statistically good enough in evaluating the performance of the classifier [4].

Logistic regression analysis (LRA), a learning technique, which is widely used in medical applications, construct a separating hyperplane between the two datasets which have to be distinguished by the classifiers. This hyperplane is described by a linear discriminant function $z = f(x_1,\ldots x_n) = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c$. Here, $x_1,\ldots x_n$ are the input variables (in our case the metabolite concentrations), and $b_1,\ldots b_n$ as well as the constant c are the coefficients which have to be learned by the method. Once the coefficients have been learned according to the training set, each new individual $(x_1,\ldots x_n)$ can be classified by substituting the variables in the discriminant function. Additionally, a logistic function is used to consider the distance from the hyperplane as a probability measure of class membership. *Logit(p)* is the log (to base *e*) of the likelihood ratio that the resulting class is 1. In symbols it is defined as: *logit(p)=log(odds)=log(p/(1-p))*. Whereas p can only range from 0 to 1, *logit(p)* ranges from negative infinity to positive infinity. There is a (relatively) simple exponential transformation for converting log-odds back to probability:

$$p = \frac{1}{1+e^{-z}} \qquad (1)$$

where p is the conditional probability of the form $P(z=1| x_1,\ldots,x_n)$ and z the discriminant function. The class membership to class "0" is indicated by $p < 0.5$, to class "1", which represents the presence for a disorder, by $p \geq 0.5$. LRA uses a maximum likelihood method, which maximises the probability of getting the observed results given the fitted coefficients [5].

Decision trees (DT) are rooted, usually binary trees, with simple classifiers placed at each internal node and a class label at each leaf. For most DT algorithms, these simple classifiers associated with the internal nodes are comparisons between an input variable and a fix value. Decision trees are generally trained by means of a top down growth procedure, which starts from the root node and greedily chooses a split of the data that maximizes some cost function, usually a measure of the class purity of the two subgroups defined by the split. After choosing a split, the subgroups are mapped to the two child nodes. This procedure is then recursively applied to the children, and the tree grows until some stopping criterion is met. If the resulting tree is too complex (and, therefore, often overfitted) some of the branches can be pruned. For instance, in figure 1 variable m_1 is chosen as the first split variable, and a value of 0.6 is chosen for the split, because we obtain an optimal class purity (100% positives) on the left side and a fairly good class purity (only three positives in a high number of negatives) on the right side. No other split with a higher class purity is possible in this figure.

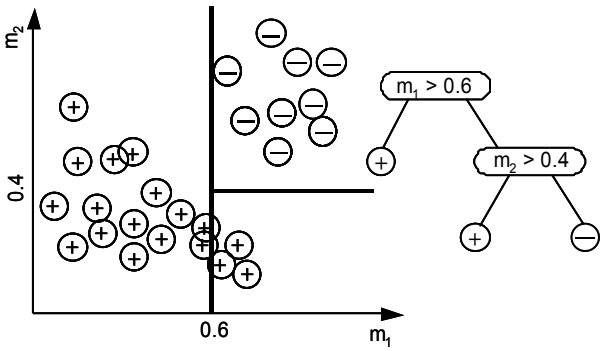

Figure 1: Example for splitting strategy of a DT classifier. Variable m_1 is chosen as first split variable.

The algorithm most often used to generate decision trees is ID3 or its successors C4.5 and C5.0 respectively [6][7].

This algorithm selects the next node to place in the tree by computing the information gain for all candidate features and then choosing that feature that gains the most information about the output category of the current example. Information gain is thus a measure of how well the given feature separates the remaining training data by minimizing the entropy [8]. In this work we used C4.5 for tree construction without pruning (cf. figure 3).

Evaluation of classifier's performance

We evaluated the discriminatory power of the investigated methods constructing a so-called classification or contingency table for our binary class problem. The evaluation measure most frequently used in classification is accuracy (Acc) which describes the proportion of correctly classified instances: Acc = (TP+TN)/(TP+FP+TN+FN). Measures which more precisely consider the influence of the class size are sensitivity (S_n) or recall, specificity (S_p), positive predictive value (PPV) or precision and negative (NPV) predictive value. S_n = TP/(TP+FN) measures the fraction of actual positive instances that are correctly classified; while S_p = TN/(TN+FP) measures the fraction of actual negative examples that are correctly classified. The PPV (or the reliability of positive predictions) is computed by PPV=TP/(TP+FP), the NPV is defined as NPV=TN/(TN+FN).

3. Results

Both presented supervised learners, LRA and DT, first ran on newborn screening data including the entire feature spectrum of 14 amino acids. From these results, we observed that the LRA classifier showed a high classification performance with a S_n of 97.7%, a S_p of 99.6% and an Acc of 99.2% (table 2). Applying feature selection methods, irrelevant and redundant features (metabolites) can be removed in order to retain or improve the predictive power of our classifiers. We performed a feature selection technique, which is based on information gain [8], an approach also implemented in the C4.5 learner. The ranking results are illustrated in figure 2.

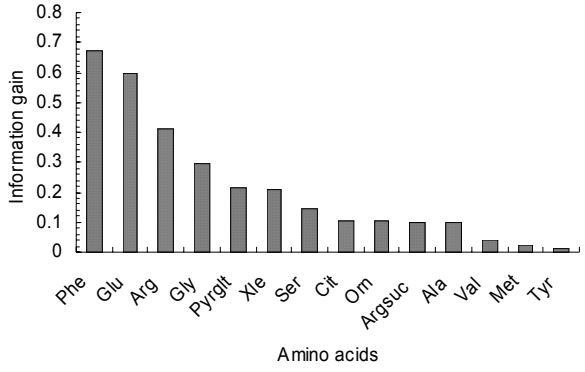

Figure 2: Ranked features after running the information gain algorithm on the metabolite spectrum of 14 amino acids.

The discovered metabolic patterns (figure 2) showed that besides Phe, also Glu and Arg have high dominance in PKU data. However, these results correspond just partly with the established diagnostic criteria, according to which Phe and Tyr are the established diagnostic metabolites indicating the presence of PKU [9]. It is remarkable that Tyr, which is reduced in the metabolism of PKU, ranks at the final position (see figure 2). Table 2 summarizes the classification performance of LRA classifiers and DT learner running on the entire and reduced feature dimensionality. All combinations of top ranked metabolites yielded a S_n of $\geq$ 97.7% accompanied by various S_p values, but all above 99.7%. The Acc was $\geq$ 99.4%. The LRA learner including the primary diagnostic markers Phe and Tyr was given by a S_p of actually 100%. The highest S_n (98%) revealed the LRA learner with the metabolite combination of Phe and Arg.

Table 2: Classification performance of LRA and DT classifier

Model	S_n (%)	S_p (%)	Acc (%)
LRA (Phe, Arg)	98.0	99.9	99.7
LRA (Phe, Tyr)	97.7	100	100
LRA (Phe)	97.7	99.9	99.7
LRA (Phe, Glu)	97.7	99.9	99.7
LRA (all 14 acids)	97.7	99.6	98.4
DT (Phe, Glu, Xle, Gly)	97.7	99.8	99.3

Performance investigated on the entire and selected metabolite spectrum.

The DT learner, which selects the next node to place in the tree by computing the information gain of all candidate features, selected four amino acids for its resulting decision tree, i.e. Phe as root node, Glu, Xle and Gly as child nodes. It showed the same S_n as the LRA learner, but with a slightly lower S_p of 99.8%.

For clinical routine, feasible models have to ensure easy interpretation without losing predictive power. Equations 2 - 4 represent the best LRA models including Phe alone or combining Phe with Arg or Tyr as model parameters:

$$P(PKU = 1) = (1 + e^{-0.056 \cdot Phe + 8.9269})^{-1} \quad (2)$$

Odds ratios (OR): Phe = 1.0573
Overall model fit: χ^2= 1445; df = 1; p<0.001

$$P(PKU = 1) = (1 + e^{-0.0383 \cdot Phe - 0.0068 \cdot Arg + 8.524})^{-1} \quad (3)$$

OR: Phe = 1.0391, Arg = 1.0069
Overall model fit: χ^2= 1464; df = 2; p<0.001

$$P(PKU = 1) = (1 + e^{-0.0662 \cdot Phe - 0.0099 \cdot Tyr + 8.723})^{-1} \quad (4)$$

OR: Phe = 1.0684, Tyr = 0.9902
Overall model fit: χ^2= 1459; df = 2; p<0.001

The classification rule of the DT learner for predicting PKU is represented in figure 3. The tree is characterized by its size of 9 and its number of leaves of 5.

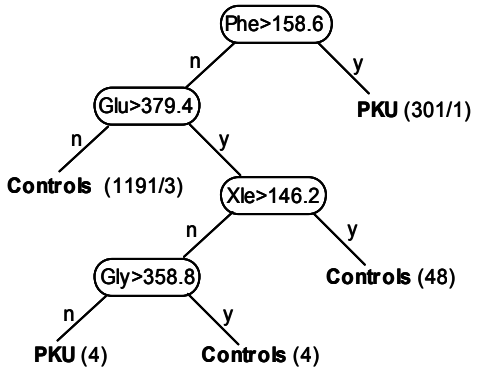

Figure 3: Classification rule of the DT learner for PKU. Values in brackets indicate the number of correctly and wrongly classified cases.

4. Discussion and Conclusion

We investigated two symbolic supervised machine learning algorithms for their suitability to construct classification models for high dimensional metabolic data. LRA models (equations 2 – 4) are easy to handle by physicians, who only have to substitute absolute amino acid concentrations in the equations. Similarly, the tree structure of DT models with their decision rules at any node is simple to interpret. Non-symbolic techniques such as the k-nearest neighbor classifier, artificial neural networks or support vector machines keep the knowledge only implicitly in internal data structures without any ability for the clinical expert to understand and interpret the learned knowledge. Our models, which were constructed on 307 PKU cases and a reduced number of randomly sampled controls, are characterized by a high S_n of ≥ 97.7%, that means that a small fraction of 2.3% newborns with PKU is incorrectly classified. The models ensure an overall fraction of invalid tests not exceeding 0.2%. However, in order to determine the real specificity, a much larger dataset would be required.

In detail, the LRA classifier, which is a linear-in-parameter method using linear separating hyperplanes (e.g. support vector machines handle also non-linear separation problems), demonstrates superior classification performance . In this context it is to note that we also plan experiments on well established non-symbolic techniques as already discussed to investigate their discriminatory characteristics. In all model equations, Phe shows an OR > 1 which corresponds well with the elevated concentrations of Phe while e.g. Tyr demonstrates slightly decreased levels (OR < 1; cf. table 1). The model, implementing Arg (OR > 1) instead of Tyr, yields a comparable high classification accuracy. But the models in all show little differences in classification performance. However, our observations of strongly elevated Arg and Glu levels need to be discussed with clinical experts as factors like the date of sampling the newborn, the time of nutrition or a wrong handling to take a blood sample can additionally lead to abnormal changes of metabolite's concentrations. Comparing our investigated classification rules with the clinically established diagnostic flags (for PKU: *Phe ≥ 150µmol/L*, Wisconsin newborn screening program [10]), we investigated similar results. Thereby, model equation 2, which includes only Phe, indicates a comparable threshold of 159.5µmol/L, the decision rule at the root node of our DT model already classifies 301 PKU cases at a threshold of 158.6µmol/L.

Our results show that the use of symbolic machine learning techniques is appropriate to construct classifiers on high dimensional metabolic data. We have shown that the presented techniques enable us to investigate not only single pathway blockade disorders, moreover, these paradigms have a great potential to examine diseases based on more complex metabolic pathways.

5. Acknowledgement

We thank Dr. A.A. Roscher from Dr. von Hauner Children's Hospital, University of Munich, Germany for providing anonymous newborn screening data. Financial support for this study was provided by the Austrian Industrial Research Promotion Fund FFF (Grant No. HITT-10).

References

[1] D.H. Chace, D.S. Millington, N. Terada, S.G. Kahler, C.R. Roe, L.F. Hofman, Rapid diagnosis of phenylketonuria by quantitative analysis for phenylalanine and tyrosine in neonatal blood spots by tandem mass spectrometry. *Clin Chem, 39*, 1993, 66 –71.

[2] B. Liebl, U. Nennstiel-Ratzel, R. von Kries, R. Fingerhut, B. Olgemoller, A. Zapf, A.A. Roscher, Very high compliance in an expanded MS-MS-based newborn screening program despite written parental consent. *Prev Med, 34*, 2002, 127-131.

[3] B. Liebl, U. Nennstiel-Ratzel, R. von Kries, R. Fingerhut, B. Olgemoller, A. Zapf, A.A. Roscher, Expanded newborn screening in Bavaria: tracking to achieve requested repeat testing. *Prev Med, 34,* 2002, 132-137.

[4] I.H. Witten, E. Frank, *Data Mining: Practical machine learning tools and techniques with java implementations* Morgan Kaufmann, 2000.

[5] D.W. Hosmer, S. Lemeshow, *Applied logistic regression* 2nd edition, Wiley, New York, 2000.

[6] R.J. Quinlan, Induction of decision trees, *Machine Learning, 1*, 1986, 81-106.

[7] R.J. Quinlan, *C4.5: Program for Machine Learning* Morgan Kaufmann, San Mateo, CA, 1993.

[8] T.M. Mitchell, *Machine Learning* McGraw-Hill Boston, MA, 1997.

[9] Tandem mass spectrometry in newborn screening. American College of Medical Genetics/American Society of Human Genetics Test and Technology Transfer Committee Working Group. *Genet Med, 2*, 2000, 267-269.

[10] Health professionals guide to newborn screening. Wisconsin state laboratory of hygiene: www.slh.wisc.edu/newborn/guide/phenylketonuria.php.