Threshold extraction in metabolite concentration data

André Flöter,
Institut für Informatik,
Universität Potsdam,
14439 Potsdam, Germany

Jacques Nicolas,
IRISA, Campus de Beaulieu,
35042 Rennes, France

Torsten Schaub,
Institut für Informatik,
Universität Potsdam,
14439 Potsdam, Germany

Joachim Selbig,
Max-Planck-Institut für
molekulare Pflanzenphysiologie,
14424 Potsdam, Germany

August 31, 2003

Abstract

The further development of analytical techniques based on gas chromatography and mass spectrometry now facilitate the generation of larger sets of metabolite concentration data. These are a prime source for the study of metabolic behaviour under different environmental conditions. In order to study the impact of environmental stimuli on organisms it is helpful to know about discrete states the concentrations adopt. A straightforward method to recognize such states is the identification of modes in the individual distributions of concentration variables. However, this approach does not fit well noisy, sparse or ambiguous data. General techniques for finding discretisation thresholds in continuous data also prove to be practically insufficient to detect states due to the weak conditional dependencies in the data. We address this problem by identifying significant thresholds in single variables through a global survey considering all variables. The technique is based upon a comparison of sets of decision trees that explain the potential states of variables. This way, we were able to find significant thresholds in metabolic data which could not be detected with conventional methods.

1 Introduction

In recent years it has become possible to effectively obtain various types of biological data at the molecular level. These give rise to new "post-genomic" studies. Metabolite concentration data is a yet little studied form of expression data [11]. It can be observed using high-throughput techniques that generate large data sets [5]. The main goal of such studies is to be able to reconstruct the dynamics of interaction between the metabolites. This paper proposes a contribution towards this goal, trying to detect significant thresholds for some concentration variables based on a global analysis of the complete set.

The basic assumption is that, as for any dynamical system, one can observe a finite set of "stable" states between which the system evolves. A state is considered to be a reasonably stable condition of any measurable variable, observed directly at the level of concentrations, in a (sub-)set of samples. A simple example of a distribution with two stable states is given in Figure 1. One observes an increased level of $NADPH_2$ in the leaves of a plant during daytime and a decreased level of it during nighttime. Thus, the plant can be considered as having two distinct states; we could label them as "night state" and "day state". There are two modes in the distribution of Figure 1 indicating each of the two states. Here, it is known that $NADPH_2$ increases with the amount of light

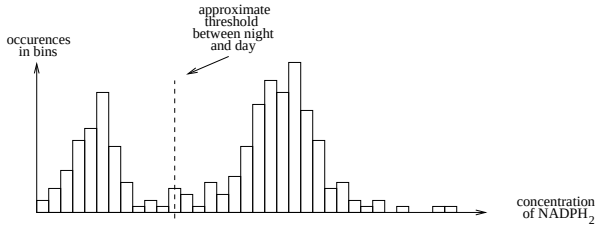

Figure 1: The bimodal distribution of $NADPH_2$.

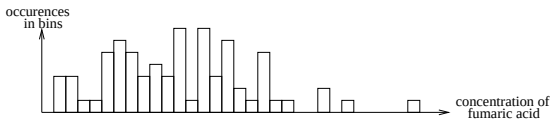

Figure 2: The distribution of fumaric acid does not provide any clear modes.

the leaf is exposed to. It is usually not that easy to relate the states of an organism to a variable.

Often the distributions of variables appear to be uniform, Gaussian or just random as in Figure 2. Thus, several distinct states (or modes) cannot be read off or found with conventional statistical methods (e.g. [12]). Nonetheless, there can still be several states which are just hidden in the sum of several modes or in the noise of the data. After all, despite substantial advances in analytical techniques, biological data has considerable variances.

We address this problem by developing a tool for identifying some of these hidden states in variables. Since functional dependencies (including states) cannot be derived reliably from single variables with few data points we use a global approach to increase robustness. It considers for any given target variable a set of thresholds and compares them in *quality* and *stability* through sets of decision trees. With this approach, it is possible to find robust and explainable states in variables. Once the states are identified, a direct examination can lead to further understanding of the organism's dynamics.

In the next section we discuss related work on finding thresholds in continuous data without considering biological issues. The subsequent sections introduce our proposal and discuss why this new technique seems more suitable for present metabolite concentration data than previous approaches. First results on real data are given and we conclude with a preliminary analysis of its significance in a metabolic context.

2 Related work

The problem of finding significant thresholds in continuous data is mostly equivalent to the problem of discretisation. This has been vastly researched in the past. And though discretisation is often considered a pre-processing for further examination, it is also accepted as a standalone analysis [8].

An older but comprehensive synopsis of existing discretisation techniques has been given by Dougherty et al. [3]. To our knowledge, they were the first to introduce a systematic categorisation of techniques. The three proposed dimensions were *global* vs. *local*, *supervised* vs. *unsupervised*, and *static* vs. *dynamic*. An additional category has been introduced by Kwedlo & Kretowski [9]: *univariate* vs. *multivariate*. A recent overview of discretisation techniques with the goal of constructing better Bayes classifyers can be found in Yang et al. [15].

Strengths and weaknesses of new techniques belonging to particular categories have been discussed for many discretisation problems [7, 6, 14, 9, 1]. Generally, supervised methods are said to deliver more useful results than unsupervised techniques [3]. Supervised techniques make use of a class label attributed to every sample in the data set. However, they strictly require the presence of such a preclassified variable, which is usually not given with metabolite concentration data. Our work tries to keep the advantages of supervised discretisation in such an unsupervised context by conducting an exhaustive search through possible class labelings.

Ho and Scott [7] argue about advantages and disadvantages of *global* vs. *local* discretisation. Global discretisation performs the discretisation of all continuous values in one step, while local discretisation processes only subsets of the data at a time. They state that local discretisation can lead to more accurate results at the cost of higher computation time. But they also note that local discretisation might deliver ambiguous results which are harder to interpret. There is no hard evidence of whether the one or the other category is better fitted to discretize metabolite concentration data. We prefer a global approach, because

interpretability might still be of interest in our analysis.

The main difficulty of the discretisation of metabolic data stems from the conjunction of a high amount of noise with a relatively low number of available samples. It is thus of utmost importance to make the most out of the available information and dependency structure of the data. Kohavi et al. [8] and Bay [1] argue that *dynamic* and *multivariate* discretisation is best fitted to satisfy this need. Dynamic methods consider interdependencies between variables in the feature space; multivariate techniques do the same but for all variables simultaneously.

Starting from this background, we now introduce a new discretisation technique which is, in terms of prior work, *global*, *unsupervised*, *dynamic*, and *multivariate*, but tries also to make biologically plausible discretisation choices.

3 Growing and comparing Decision Forests

3.1 Decision Trees

Decision trees can be interpreted as functions that allow for classifying data objects into discrete target classes [2]. They classify objects on the basis of a set of selected attributes. Each internal node represents a test of the value of an attribute, branches correspond to different possible values for these attributes, and leaves specify the object's target class.

Trees on specific classification problems can be built automatically with induction algorithms. For this purpose they need a set of preclassified data objects (often referred to as *training data*). A hierarchically ordered set of tests is then learned which allows for classifying new observations.

3.2 Modeling states of an organism

In order to identify possible states of an organism we try to detect significantly stable conditions of concentration variables. Such conditions can be modelled by decision trees in the following way:

If we knew about two states comprised in a given variable, we could dichotomise this variable into the classes "state 1" and "state 2". Largely, this dichotomisation can be performed by finding the concentration threshold dividing the two states. Literature refers to such a threshold as a *cut point* [4]. With the obtained two classes a decision tree can then be induced as a model for explaining these states (e.g. with C4.5 [10]).

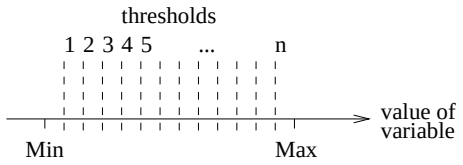

Figure 3: Determining thresholds 1...n by dividing the variable's domain into uniform intervals.

For instance, the samples used to provide the distribution of $NADPH_2$ in Figure 1 can be classified into "night state" and "day state" according to their $NADPH_2$ level. In fact, we discretize this variable into 0 ("night state") and 1 ("day state") according to a chosen threshold. A decision tree grown on this target variable can classify new samples as belonging either to class 0 or class 1 without considering the concentration level of $NADPH_2$. This classification is based only upon the remaining variables of the training set.

The last issue is to find an appropriate threshold for the discretisation of the target variable. As mentioned in Section 1, most distributions do not allow for a clear distinction between two modes (respectively states). Thus, we have to find another way to pick an appropriate threshold out of the many possibilities.

3.3 Growing Decision Forests

We propose to grow sets of decision trees for each considered discretisation threshold and compare them. Sets of decision trees are also referred to as decision forests. To get candidate thresholds the domain of the target variable is uniformly divided as indicated in Figure 3. The size of the intervals is chosen so that on average a "sufficient" number of samples is occurring in each of them (value is set to 5 for our experiments). The end of each interval marks one candidate discretisation threshold. This procedure is known as *uniform binning*.

For each possible threshold, a decision forest is grown

with an embedded decision tree induction algorithm. We used C4.5, one of the most established algorithm for this task [10]. Initially, the set of available variables contains all measured variables minus the target variable. Then, the following procedure is used:

- While variables are present in the data set do
  1. Grow a decision tree with C4.5 on the discretized target variable and add it to the forest.
  2. Remove the variable occurring at the top of the tree from the set of available variables.
- Sort the trees of the forest according to their predictive accuracy and keep the k best trees in the forest ($k = 3$ in our experiments).

This way, we obtain a forest of varying trees with highest predictive accuracy for each target discretisation threshold.

Here, we gain the possibility of using a supervised learning approach in an unsupervised process by systematically using *all* candidate thresholds and constructing models for them.

3.4 Finding a threshold

At this point a particular decision forest has been produced for each of the considered discretisation thresholds. Each forest is evaluated in turn through comparison with the forests of the two neighbouring thresholds. More precisely, an evaluation function grants a score of 1 for each tree in the neighbouring forests which is *similar* to one in the evaluated forest. We use a "syntactical" similarity criterion. Two trees are similar if the attributes used in the nodes of the first two levels of both trees are the same. That way, high scores are given to forests with similar neighbours.

With this "smoothening" process thresholds are found that promote environments of *stable* models of the data. If the scores are plotted into a curve we can identify regions of stable forests (see Figure 4). Stable forests indicate robust models for the explanation of the target variable. We can assume that robust models indicate a biologically feasible choice of the target classes and thus the discretisation threshold.

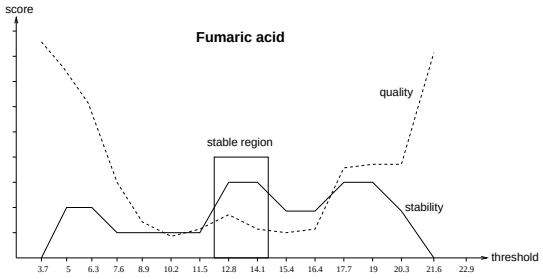

Figure 4: Peaks or elevated plains in the score functions indicate regions of stable models.

Another way to compare the forests is by their *predictive quality*. To measure this quality we propose the following function:

Definition 1 *Let T be a decision tree of depth n and let D be a set of objects with known classifications. For $1 \leq i \leq n$, let C_i be the set of objects from D, being correctly classified by T at depth i. Then, define the quality of T by means of the following function:*

$$quality(T) := \sum_{i=1}^{n} |C_i| \cdot \frac{1}{2^i}$$

This function delivers high values for trees classifying the training samples with little error and few decisions. It can also be understood as a measure of the *effectiveness* of a decision tree in solving the classification problem. For comparing forests we use the arithmetic average of qualities of the trees in the forests and compare them. We use this measure only to be able to compare forests and find those with an increased quality.

As a matter of principle, this function produces peaks for discretisation thresholds close to the boundaries of the target variable's domain. This is due to the very asymmetric distribution of samples in the target classes when discretising is done with a marginal threshold. We call these peaks *sparse data peaks*, because one of the two target classes contains very few samples. These peaks will not be considered for the determination of high quality forests. Instead, we look for local peaks of the function. These indicate a significant gain of quality against neighbouring thresholds.

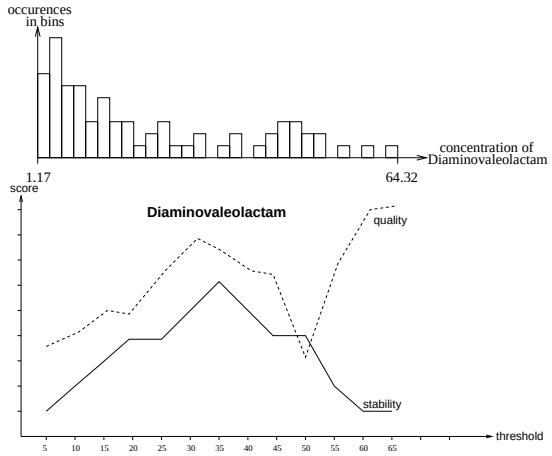

Figure 5: Diaminovaleolactam develops a clear peak approximately between the little visible modes.

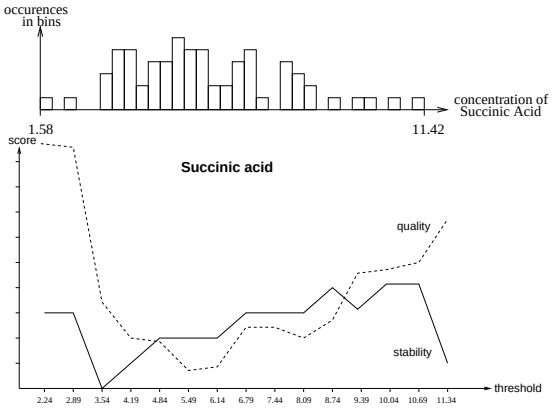

Figure 6: Succinic acid does not provide any clear proposition for a threshold.

With the two measures *stability* and *quality* it is possible to find one or more biologically motivated discretisation thresholds for any given variable based on peak analysis. If the measures lack remarkable peaks in their values it is assumed that there are no inherent stable states in the examined variable.

4 Results

We applied our technique on a set of metabolite concentration data of potato plants with 73 samples and 117 metabolites. 37 of the samples were treated to develop only low concentrations of Phosphoric acid. The other 36 were left unaffected. Thereby, we knew about two distinct states ("presence" and "absence" of Phosphoric acid) comprised in the data. Subsequently, these inherent states were tried to be found in the other variables.

On all clearly bimodally distributed metabolite concentrations (similar to that of Figure 1) we saw peaks in *stability* and *quality* for thresholds located directly between the modes. For those it would also be possible to find discretisation thresholds through conventional methods.

But some metabolites only exhibit barely visible modes and do not allow for a clear determination of thresholds. For those, it is possible to verify the significance of an assumed threshold with our method. In Figure 5 one observes the distribution of Diaminovaleolactam with only little visible modes. Though, the peaks in the score functions point out the very high significance of a threshold between those uncertain modes.

Furthermore, for a few other variables without apparent modes in the distributions we also saw peaks for certain thresholds. In Figure 2 we present the distribution of Fumaric acid which is supposedly random. But in Figure 4 it becomes obvious that there is a hidden threshold within the concentration levels of it. This is the clearest example of an unexpected threshold in our data.

More than 85% of the metabolites, however, do not develop peaks in our score functions. For instance, Succinic acid (as shown in Figure 6) has a Gaussian-like distribution. The score functions for it show an unstable curve. We assume that there is no significant threshold hidden in such metabolites.

5 Discussion

The problem we have addressed in this paper is finding "stable states" in metabolic data. If such states are characterised by reasonably stable conditions of variables, then finding them is closely related to detecting discretisation thresholds. However, it is not exactly equivalent, because discretisation simply aims at automatically mapping con-

tinuous numbers into discrete classes. Finding states in biological data, on the other hand, is a less distinct process benefiting from additional information about the qualities of a proposed threshold.

5.1 New features of our approach

There has been plenty of work on discretising continuous data in the past [3, 7, 6, 14, 9, 1, 15]. These approaches have delivered feasible results for various types of data. The metabolic data we use demands new capacities from the techniques as the data sets are rather small and contain a considerable amount of noise. Our approach is among the very few techniques [9] to consider potential combinatorial relationships between *all* other variables at the same time, thus exploiting as much of the inherent information as possible.

Furthermore, our method provides two indicators for the quality of a proposed threshold (*quality* and *stability*). These let a scientist additionally recognise the significance of a proposed biological state. Such extra information is valuable, especially when experimental costs prevent an exhaustive examination of all hypotheses.

Finally, the decision tree approach yields the possibility to even interpret the classifyers used for the evaluation of states. Presently, we are just starting this interpretation work.

We believe our technique to be a suitable way to find significant states in metabolite concentration data. Indeed, the application of our method on metabolic data has led to the discovery of several unobvious thresholds.

5.2 Usefulness in metabolic data analysis

Biological systems like plants can adopt distinct states according to different environmental conditions. Each such state can lead to different activity in the metabolism of that organism [13]. It is therefore feasible to search for states in metabolite concentration data.

Most applications of GC/MS involve the analysis of samples taken from discriminative environmental conditions. The data we used comprised an unaffected condition and a condition of lowered Phosphoric acid. The objective of observing samples under different environmental conditions is always to see how these conditions affect the organism. A good way to notice an effect is by detecting a change of state in the system.

Because of the noise in concentration data, it is mostly hard to identify clear states within the concentrations. Our method now provides a possibility to track back such changes of states in concentration data.

5.3 Conclusion

Our technique potentially finds more discretisation thresholds in metabolic data than conventional discretisation methods, and they are obviously explainable through the data. Each proper threshold is an indicator for the presence of a state in the organism. Thus, the technique can be used to validate obvious or find hidden states. It is thereby possible to see better if and how an organism reacts to stimuli, or if an organism changes states subject to a hidden cause.

References

[1] Stephen Bay. Multivariate discretisation of continuous variables for set mining. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 315–319, 2000.

[2] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and regression trees*. Wadsworth International Group, Belmont CA, 1984.

[3] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In Armand Prieditis and Stuart Russell, editors, *Machine Learning: Proceedings of the Twelfth International Conference*, pages 194–202, San Francisco, USA, 1995. Morgan Kaufmann Publishers.

[4] Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the IJCAI'93*, volume 2, pages 1022–1027, Chambéry, France, 1993. Morgan Kaufmann Publishers.

[5] Oliver Fiehn, Joachim Kopka, Peter Dörrmann, Thomas Altmann, Richard Trethewey, and Lothar Wilmitzer. Metabolite profiling for plant functional genomics. *Nature Biotechnology*, 18:1157–1161, Nov 2000.

[6] João Gama, Luis Torgo, and Carlos Soares. Dynamic discretization of continuous attributes. In *Proceedings of the Sixth Ibero-American Conference on AI*, pages 160–169, 1998.

[7] K. Ho and P. Scott. Zeta: A global method for discretization of continuous variables. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 191–194, 1997.

[8] Ron Kohavi and Mehran Sahami. Error-based and entropy-based discretisation of continuous features. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 114–119, 1996.

[9] Wojciech Kwedlo and Marek Kretowski. An evolutionary algorithm using multivariate discretization for decision rule induction. In *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*, pages 392–397, 1999.

[10] J.R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo USA, 1993.

[11] Ute Rößner, Alexander Lüdemann, Doreen Brust, Oliver Fiehn, Thomas Linke, Lothar Willmitzer, and A.R. Fernie. Metabolite profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, 13:11–29, 2001.

[12] B.W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of Royal Statistical Society*, 43(Series B):97–99, 1981.

[13] Lubert Stryer. *Biochemistry*. W.H. Freeman and Company, New York, 4th edition, 1995.

[14] Ke Wang and Bing Liu. Concurrent discretization of multiple attributes. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, pages 250–259, 1998.

[15] Ying Yang and Geoffrey Webb. A comparative study of discretisation methods for naive-bayes classifiers. In *Proceedings of Pacific Rim Knowledge Acquisition Workshop*, pages 159–173, Tokyo, Japan, 2002.