BIOINFORMATICS

Vol. 20 no. 10 2004, pages 1491–1494
doi:10.1093/bioinformatics/bth107

Threshold extraction in metabolite concentration data

A. Flöter[1,*], J. Nicolas[3], T. Schaub[1] and J. Selbig[2]

[1]University of Potsdam, Institute for Computer Science, August-Bebel-Str. 89/Hs. 4, 14482 Potsdam, Germany, [2]Max-Planck-Institute of Molecular Plant Physiology, 14424 Potsdam, Germany and [3]Institut de Recherche en Informatique et Systèmes Aléatoires, Campus Universitaire de Beaulieu, 35042 Rennes cedex, France

Received on October 12, 2003; accepted on February 3, 2004

ABSTRACT

Motivation: Continued development of analytical techniques based on gas chromatography and mass spectrometry now facilitates the generation of larger sets of metabolite concentration data. An important step towards the understanding of metabolite dynamics is the recognition of stable states where metabolite concentrations exhibit a simple behaviour. Such states can be characterized through the identification of significant thresholds in the concentrations. But general techniques for finding discretization thresholds in continuous data prove to be practically insufficient for detecting states due to the weak conditional dependences in concentration data.

Results: We introduce a method of recognizing states in the framework of decision tree induction. It is based upon a global analysis of decision forests where stability and quality are evaluated. It leads to the detection of thresholds that are both comprehensible and robust. Applied to metabolite concentration data, this method has led to the discovery of hidden states in the corresponding variables. Some of these reflect known properties of the biological experiments, and others point to putative new states.

Availability: An implementation of this approach can be obtained from the authors upon request.

Contact: floeter@cs.uni-potsdam.de

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 are a yet little studied form of expression data (Rößner *et al*., 2001). They can be observed using high-throughput techniques that generate large data sets (Fiehn *et al*., 2000). The long-term goal of such studies is to be able to reconstruct the dynamics of interaction between 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.

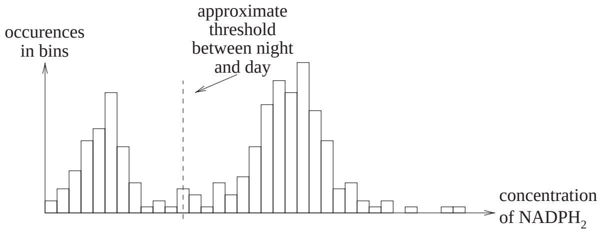

Fig. 1. The bimodal distribution of $NADPH_2$.

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 the daytime and a decreased level at night time. 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 the two states. Here, it is known that $NADPH_2$ increases with the amount of light 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. Silverman (1981)]. Nonetheless, there can still be several states that 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 have considerable variances.

We address this problem by developing a tool for identifying some of these hidden states in variables. Since functional dependences (including states) cannot be derived reliably

[*]To whom correspondence should be addressed.

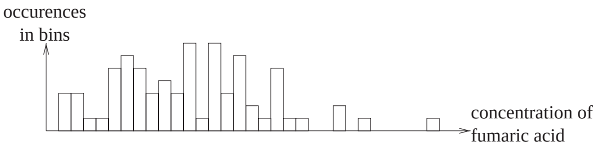

Fig. 2. The distribution of fumaric acid does not provide any clear mode.

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 a 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.

SYSTEM AND METHODS

Data discretization

The problem of finding significant thresholds in continuous data is largely equivalent to the problem of discretization. This field has been intensely investigated in the past. Valuable synopses of the techniques have been given by Dougherty *et al*. (1995) and Yang and Webb (2002). The important axes for classifying discretization techniques are global versus local, supervised versus unsupervised and univariate versus multivariate (Kwedlo and Kretowski, 1999). Metabolic concentration data usually do not provide previously known target classes, nor do they facilitate univariate discretization because of unclear distributions. Thus, our problem demands a multivariate and unsupervised approach. To our knowledge, no such discretization technique is currently available that we could apply to our data.

Decision trees

Decision trees are representations of classifiers (Breiman *et al*., 1984). On the basis of a set of selected attributes, they classify objects. Each internal node represents a test on 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 that allows classifying new observations.

Modelling 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 contained in a given variable, we could dichotomize this variable into the classes 'state 1' and 'state 2'. Largely, this dichotomization can be performed by finding the concentration threshold dividing the two states [in the literature referred to as a cut point (Fayyad and Irani, 1993)]. With the obtained two classes, a decision tree can then be induced as a model for explaining these states [e.g. with C4.5 (Quinlan, 1993)].

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 the $NADPH_2$ 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 to either 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 discretization of the target variable. As mentioned in the previous section, most distributions do not allow 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.

IDENTIFYING STABLE STATES

Growing decision forests

For the identification of thresholds indicating stable states, we propose to grow sets of decision trees for each discretization threshold considered 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 divided into several intervals. The end of each interval marks one candidate discretization threshold. There are several known strategies for choosing the number and sizes of intervals, e.g. uniform binning, equal frequency binning and exhaustive binning. We used uniform binning with an average number of five experiments per interval for our application.

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 algorithms for this task (Quinlan, 1993). 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 dataset, 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 discretization 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.

Finding a threshold

At this point a particular decision forest has been produced for each of the considered discretization 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 that is similar to one in the evaluated forest. We use a ‘syntactical’ similarity measure. Two trees are similar if the attributes used in the nodes of the first level 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 (Fig. 3). 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 discretization threshold.

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 binary decision tree of depth n and let D be a set of classified objects. 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:*

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

This function delivers high values for trees classifying the training samples with little error and few decisions. For comparing forests, we use the arithmetic mean of qualities of the trees in the forests and compare them.

As a matter of principle, this function produces peaks for discretization 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 discretizing 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 are not considered for the determination of high quality forests.

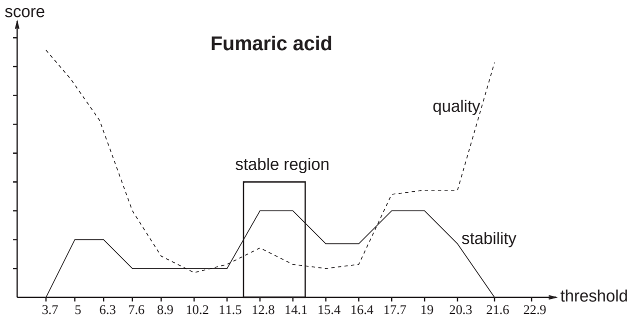

Fig. 3. Peaks or elevated plains in the score functions indicate regions of stable models.

With the two measures stability and quality, it is possible to find discretization 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.

RESULTS

We applied our technique to a set of metabolite concentration data of potato plants with 73 samples and 117 metabolites. Of the samples, 37 were treated to develop only low concentrations of phosphate. The other 36 were left unaffected. Thereby, two distinct states (‘presence’ and ‘absence’ of phosphate) are reflected in the data. Subsequently, these inherent states were searched for in the other variables.

About 5% of metabolite distributions exhibit two modes (similar to Fig. 1). We saw clear peaks in stability and quality between these modes. For them it is also possible to find discretization thresholds through conventional methods.

The rest of the metabolites do not exhibit clear modes and do not allow the determination of thresholds through conventional methods. For about 10% of these, it was possible to detect a threshold through our method (e.g. fumaric acid in Figs 2 and 3 or diaminovaleolactam in Fig. 4).

Roughly 30% of these additional thresholds split the samples into the known subsets of treated and unaffected experiments. The other 70% disclosed unexpected states (such as fumaric acid and diaminovaleolactam did) obviously not directly connected to the experimental treatment.

More than 85% of the metabolites, however, do not develop peaks in our score functions, showing that our method remains selective.

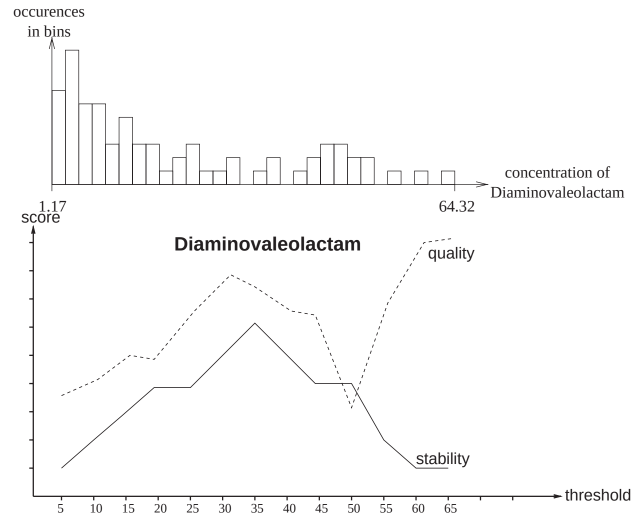

Fig. 4. Diaminovaleolactam develops clear peaks for thresholds between concentration levels 30 and 35.

DISCUSSION

In this paper, the problem of finding 'stable states' in metabolic data has been reduced to detecting discretization thresholds. However, these issues are not exactly equivalent because discretization simply aims at automatically mapping continuous 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. Propositional rules derived from the decision trees can deliver such information.

Numerous works have proposed methods for discretizing continuous data in the past [see Dougherty *et al.* (1995), Kwedlo and Kretowski (1999) and Yang and Webb (2002)]. These approaches have delivered feasible results for various types of data. The metabolic data we use demand specific capacities from the techniques as the data sets are rather small and contain a considerable amount of noise. Our approach considers 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 quantitative measures of the quality of a proposed threshold (quality and stability). These facilitate the recognition of the relevance of a proposed biological state. Such extra information is valuable, especially when experimental costs prevent an exhaustive examination of all hypotheses.

CONCLUSION

The application of our method to metabolite concentration data has led to the discovery of several unexpected thresholds. As an open issue remains a robust handling of the sparse data peaks when undertaking the peak analysis. Presently, we are developing a normalization strategy for coping with this problem. Further directions of our research are a direct interpretation of the decision trees and integration of that information into our method.

ACKNOWLEDGEMENTS

We would like to thank Georg Leggewie and Joachim Kopka for providing us with the data. This work was supported by a grant of the ministry of science, research and culture of the state of Brandenburg.

REFERENCES

Breiman,L., Friedman,J.H., Olshen,R.A. and Stone,C.J. (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

Dougherty,J., Kohavi,R. and Sahami,M. (1995) Supervised and unsupervised discretization of continuous features. In Prieditis, A. and Russell,S. (eds), *Machine Learning: Proceedings of the Twelfth International Conference*. Morgan Kaufmann Publishers, San Francisco, USA, pp. 194–202.

Fayyad,U.M. and Irani,K.B. (1993) Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the IJCAI'93*, Vol. 2. Morgan Kaufmann Publishers, Chambéry, France, pp. 1022–1027.

Fiehn,O., Kopka,J., Dörrmann,P., Altmann,T., Trethewey,R. and Wilmitzer,L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, **18**, 1157–1161.

Kwedlo,W. and Kretowski,M. (1999) An evolutionary algorithm using multivariate discretization for decision rule induction. *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*. Springer-Verlag, London, UK, pp. 392–397.

Quinlan,J. (1993) *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, USA.

Rößner,U., Lüdemann,A., Brust,D., Fiehn,O., Linke,T., Willmitzer,L. and Fernie,A. (2001) Metabolite profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, **13**, 11–29.

Silverman,B. (1981) Using kernel density estimates to investigate multimodality. *J. R. Stat. Soc.*, **43**, 97–99.

Yang,Y. and Webb,G. (2002) A comparative study of discretisation methods for naive-bayes classifiers. *Proceedings of Pacific Rim Knowledge Acquisition Workshop*. Tokyo, Japan, pp. 159–173.