European Symposium on Computer Aided Process Engineering – 15
L. Puigjaner and A. Espuña (Editors)

Multivariate Decision Trees for the Interrogation of Bioprocess Data

Kathryn Kipling, Gary Montague, Elaine Martin, Julian Morris

School of Chemical Engineering and Advanced Material, University of Newcastle, Newcastle upon Tyne, NE1 7RU, United Kingdom

Abstract

This paper introduces a method for analysing correlated process data to extract useful information with multivariate decision trees. The techniques have been applied to a bioprocess data set and have been shown to provide insight into the causes of process variation. The differences between the univariate and multivariate methods are highlighted and the tree interpretations are discussed. It is observed that there is very little difference in the classification abilities of the tree types from a prediction of the outcome perspective but the information from the multivariate trees is more informative with regard to cause of deviation rather than simply identifying the observed effects.

Keywords: multivariate decision trees; correlated data; bioprocess data

1. Introduction

Effective data analysis is key to understanding the causes of variation within process operation. Extracting information from continuous process operation is challenging but batch processes raise further complications. Batch processes are highly interdependent systems and the relationships between variables is an important consideration in data analysis. However, when analysing bioprocess data these relationships are often not considered although they can contribute critical information to data interpretation. Ignoring the interdependencies can lead to unnecessarily complex and possibly incorrect conclusions being drawn from the data.

Multivariate statistics provides techniques for analysing dependent and independent variables that are correlated to each other to varying degrees. The techniques that are considered under the banner of multivariate statistics include: canonical correlation, multiway frequency analysis, analysis of covariance, discriminant analysis, principal component analysis (PCA) and partial least squares (PLS). The commonly used techniques are PCA (Francis and Wills, 1999) and PLS (Wold, 1985).

Decision trees are a commonly used technique for data analysis / mining and rule induction. In its simplest form, the concept allows the approximation of a discrete valued target function through splitting the data based on information measures and this is presented in a tree structure. It is widely used for inductive inference and the types of applications that have been investigated using decision trees include: the increasing chemical process yield in nuclear power plants by assessing the quality of the pellets (Langley and Simon, 1995); classification of celestial objects (Fayyad et al., 1995);

improving oil and gas separation by suggesting separation vessel dimensions (Guilfoyle, 1986) and making credit decisions (Michie, 1989).

There are many algorithms that can be used to perform a decision tree analysis and the choice of algorithm depends greatly upon the application and the type of data that is to be interrogated. It is not the aim of this paper to discuss the relative merits of one technique over another but comparison studies are available in the literature (Lim et al., 2000). The algorithms that have been applied or adapted in this work are ID3 (Quinlan, 1986) and CART (Breiman et al., 1984). These are commonly used techniques that are well understood and are adaptable for multivariate techniques. By using such methods the interpretation of results is simplified and the techniques used are explainable.

There are other methods available for classification of a data set and discriminant analysis is one such technique. Here a model is built from a set of training data to form a relationship between the input variables and the target variable. This can then be tested on an unseen data set to verify its ability to correctly classify the data. The main differences between this technique and a multivariate approach is that the models are formed through multivariate linear regression rather than using a method that orthogonalises the data to create independent variables. Orthogonalisation is a critical transformation for information extraction.

2. Multivariate Decision Trees

The concept of a multivariate decision tree has been discussed in the literature. Larson and Speckman (2002) considered the use of multivariate regression trees in ecology for the analysis of plant abundance data. They conclude that the resultant trees extract some of the important features that distinguish where species occur. These techniques and other literature (Brodley and Utgoff, 1992) consider the concept of a multivariate response and use methods that delineate between classes in the output. However, there is little consideration given to the idea that the predictor variables may be related and the correlation between variables may be important. This paper discusses these topics.

The first method considered was the removal of the correlation between the predictor variables to provide an independent set for analysis. Here the data was pre-processed to remove any missing data or anomalous values, then the data was analysed using the PCA technique. The resultant principal components were used as the inputs into the decision tree program. To assess the results the loadings plots are required: if a variable has a large absolute loading value then it is considered to be influential to the principal component and is a key decision making variable in the tree.

The second technique uses the concept of PLS to relate the input values to the outcome to force a significant link between the variables and more accurately assess which variables influence the outcome. Initially the technique is similar to the pre-processed decision tree; the data is analysed using the PLS approach and the resultant information used to feed into the decision tree program. At the end of an iteration the error matrix is considered. The information contained in the errors is unmodelled hence at the next iteration it is this data that is analysed. The error matrix becomes the new data set and the process is repeated until there are no errors or there are no samples to be classified.

The split methods that are used to interpret the outcome are the Gini index and entropy technique. There are many techniques that exist and these are discussed in Mingers,

(1989). Many of these methods are based in univariate statistics but since the attributes used in the decision tree algorithm are independent of each other then these are reliable methods to use. Other strategies do exist and have been used in multivariate response decision trees. These include linear machines (Duda and Hart, 1973), the mean structure of the data set and the covariance structure of the data set (Segal, 1992). Since these methods are considered for multivariate response trees they are not discussed here but they do form the basis of a possible extension to the work.

3. Data Analysis

The data that was analysed using these techniques was taken from an antibiotic fermentation. The data comprised seed, final stage and biochemical data. Due to the problems of using time series batch data for analysis, point values were taken from the data and used as indicators of the process behaviour. These point values are maxima, minima, time values and rates of change in variable values. In total there were 36 predictor variables and 1 target variable. This target variable is an indicator of the quality of the batch and is classed as "good" or "poor".

For comparison purposes the data was analysed using the simple univariate tree. The data that was used for training comprised 66% of the total number of batches and was chosen based on the extremities of the variables values: some of the batches that were chosen exhibited the maximum and minimum values for a variable while other batches were more representative of the whole set. This training set was used for all three tree types. The remaining 34% of the available data was used for validation. The data was analysed using PCA and 9 principal components were retained describing 72% of the data variation. The tree was then built using the CART technique (Breiman et al., 1984) and was trained using the same 66% of the available batches as defined above.The iterative decision tree technique was used on the same 66% of the batches for training. The iterative method uses the ID3 algorithm (Quinlan, 1986). In this technique, the variables are all continuous and to split these values at appropriate points the method described by Fayyad and Irani (1992) is used.

4. Results

4.1 The univariate decision tree

Figure 1 shows the univariate decision tree pertaining to the training data. This tree indicates that a lower value of variable 27 suggest a better batch and a higher value of variable 15 produces a poorer batch. The information contained in this tree and Table 1 shows that it can accurately classify the unseen data set although the limited number of batches in this set means that some of the branches are not evaluated.

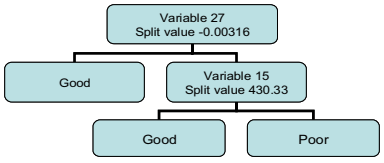

Figure 1- Univariate decision tree for bioprocess data

Table 1- Accuracy of the univariate decision tree

Leaf Number	Number of Samples	Number of Samples Correctly Classified	% Accuracy
1	12	11	92
2	0	N/A	N/A
3	5	2	40

4.2 Pre-processed decision tree

Figure 2 shows the decision tree produced using the PCA pre-processing technique. It can be that PC1 is the key discriminator of the data for the training set. The loadings plot for PC1 indicates that variables 6, 18, 26 and 27 contribute most to the variation.

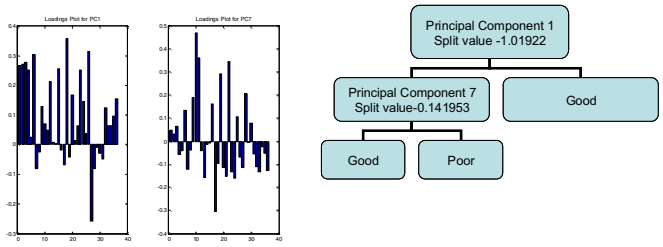

Figure 2- PCA pre-processed decision tree

Larger values of these variables suggest that there is a higher probability of following the right hand branch and ending in a good quality product. However, smaller values of these variables will force the decisions down the left hand branch and the discrimination is then made by considering PC7. Table 2 presents the results of the validation analysis showing that the root node was successful at separating the good from poor batches but again there are issues with the limited number of validation batches. The results in Table 2 are similar to those of the univariate tree but this multivariate approach provides greater insight into the process and which combinations of variables are influential.

Table 2- Accuracy of the pre-processed decision tree

Leaf Number	Number of Samples	Number of Samples Correctly Classified	% Accuracy
1	0	N/A	N/A
2	3	1	33
3	14	12	86

4.3 Iterative decision tree

The iterative approach was used to produce the tree shown in Figure 3. All of the decision nodes of the tree use the first latent variable of the iteration. This is not unexpected since the outcome variable, the quality, has been considered throughout the

process of building the model. Considering iteration 1, it can be seen from Figure 3 that variable 18, 26 and 27 contribute most to the model and for iteration 2 variables 10, 21, 22, 33 and 35 show the greatest values in the loadings and for iteration 3 variables 5, 11, 14, 20, 28, 32, and 33 are the most important in the model. Hence for decision 1, if the values of variables 18, 26 and 27 are large then the batch has a greater probability of having a low quality. Similarly for iteration 2, higher values of the key variables lead to poorer batches. Good batches are produced when the values of the key variables in iteration 3 are low. The tree was tested using an unseen data set.

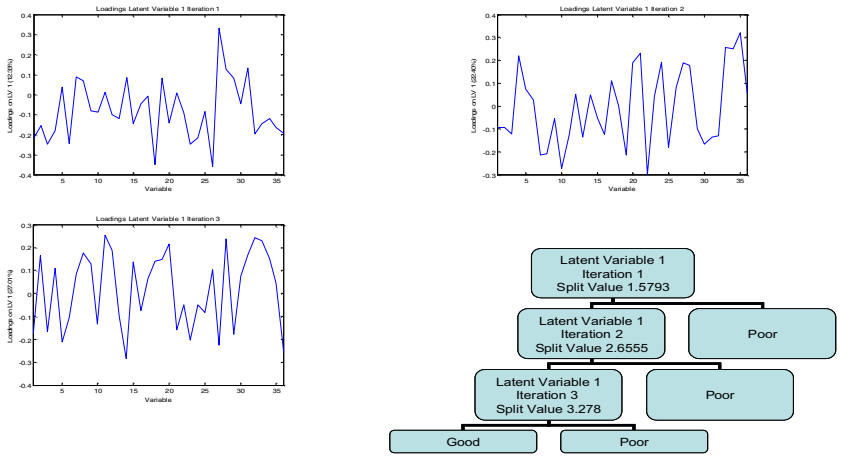

Figure 3- Iterative decision tree

Table 3- Accuracy of the iterative decision tree

Leaf Number	Number of Samples	Number of Samples Correctly Classified	% Accuracy
1	12	10	83
2	4	1	25
3	0	N/A	N/A
4	1	0	0

Table 3 shows the results of the validation of the iterative decision tree. These results are comparable with both the univariate and pre-processed decision trees but the process understanding that the multivariate approach provides is much more useful to the end user than the univariate tree. The concept of using the errors as input to the next iteration of the tree means that all of the available information can be used in the development of the tree improving the information that is available to the end user.

5. Conclusions

This paper has shown that where there are relationships between variables it is beneficial from a process understanding perspective to consider combinations of these variables to eliminate the correlation and assist in the decision making process. The technique suggested here is to pre-process the data using a multivariate technique such as principal components analysis and use the result of this analysis as the input into the

decision tree. The use of such a technique orthogonalises the data and as a result the data fed into the tree is independent of the other variables. The second method described first produces a model relating the input to the output and then using this model, where again the inputs are orthogonal to each other, determines a decision node. The residuals from the model are then used to build another model and another node until the residuals are too small to be considered significant. Three decision tree techniques have been compared on the same data sample and it has been shown that the multivariate techniques are comparable to the univariate method in classification ability but it is important to appreciate that decisions are rarely taken in isolation and that many variables are considered in parallel when interpreting data. The multivariate tree techniques give the user this ability and consider which variables are most influential on the outcome and why this is the case. The results of the analysis indicate that for the technique to be successful there need to be many samples for training and testing and although this is a common disadvantage of using decision tree methods for data mining, the results of the validation presented here are promising.

6. References

Breiman, L., J. Freidman, R. Olshen and C. Stone (1984). Classification and Regression Trees. California, Wadsworth International.

Brodley, C. E. and P. E. Utgoff (1992). Multivariate Decision Trees. Amherst, University of Massachusetts: COINS Technical Report 92-82

Duda, R. O. and P. E. Hart (1973). Pattern Classification and Scene Analysis. New York, Wiley-interscience.

Fayyad, U. M. and K. B. Irani (1992). "On the Handling of Continuous-Valued Attributes in Decision Tree Generation." Machine Learning 8(1): 87-102

Fayyad, U., P. Smyth, N. Weir and Djorgovski (1995). "Automated Analysis of Image Databases: results, progress and challenges." Journal of Intelligent Information Systems 4: 1-19

Francis, P. J. and B. J. Wills (1999). Introduction to Principal Components Analysis. in Quasars and Cosmology. eds: G.J.Ferland and J.A.Baldwin. San Fransico, Astronomical Society of the Pacific. **CS-162**.

Guilfoyle, C. (1986). Ten Minutes to Lay the Foundations. Proceedings of Expert Systems User, August, 16-19.

Langley, P. and H. Simon, A (1995). "Applications of Machine Learning and Rule Induction." Communications of ACM 38: 54-64

Larson, D. R. and P. L. Speckman (2002). Multivariate Regression Trees for Ananysis of Abundance Data. Columbia, University of Missouri: 21

Lim, T.-S., W.-Y. Loh and Y.-S. Shih (2000). "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classifcation Algorithms." Machine Learning 40: 203-229

Michie, D. (1989). Problems of Computer-Aided Concept Formation. in Applications of Expert Systems. eds: J. R. Quinlan. Wokingham, Addison-Wesley. 2.

Mingers, J. (1989). "An Emprical Comparison of Selection Measures for Decision-Tree Induction." Machine Learning 3: 319-342

Quinlan, J. R. (1986). "Induction of Decision Trees." Machine Learning 1(1): 81-106

Segal, M. R. (1992). "Tree-Structured Methods for Longitudinal Data." Journal of the American Statistical Association **87**(418): 407-418

Wold, H. (1985). Partial Least Squares. in Encyclopaedia of Statistical Sciences. eds: S. Kotz and N. L. Johnson. New York, Wiley. **6:** 581-591.