Business phase classification and prediction:
How to compare interpretability of classification methods?

Claus Weihs and Ursula Sondhauss[1],
Department of Statistics, University of Dortmund, Germany,

When comparing methods for classification, often the rating relies on their prediction accuracy alone. One reason for this is that this is the aspect that can be most easily measured. Yet, often one wants to learn more about the problem than only how to predict. The interpretation of the relation of predictors and classes is often of high interest, but an unique accepted general formalization of "interpretability" relevant for many classification problems and measurable at least for a wide range of different classification methods does not exist, and - as we believe - is not really what is needed. Instead of trying to measure "interpretability" as such, standardizing and formalizing typical ways to interpret classification rules and finding performance criteria for this kind of outcomes leads to ratings of classification methods w.r.t. interpretability that can be tailored for the specific problem at hand and the subjective preferences of addressees of results. In this short paper, three results of this kind stemming from a comparative study of various classification methods applied to the classification of German business cycle phases based on 13 economic variables are exemplarily discussed .

Keywords: classification methods, performance measures, business cycle analysis

1. Data and Methods

The data set consists of 13 ‘stylized facts’ (cp. Lucas (1983)) for the (West-) German business cycle and 157 quarterly observations from 1955/4 to 1994/4 (price index base is 1991). The stylized facts (and their abbreviations) are real-GNP-gr (Y), real-private-consumption-gr (C), government-deficit (GD), wage-and-salary-earners-gr (L), net-exports (X), money-supply-M1-gr (M1), real-investment-in-equipment-gr (IE), real-investment-in-construction-gr (IC), unit-labor-cost-gr (LC), GNP-price-deflator-gr (PY), consumer-price-index-gr (PC), nominal short term interest rate (RS), and real long term interest rate (RL). The abbreviation 'gr' stands for growth rates relative to last year’s corresponding quarter.

For the investigation of the data with respect to business cycle phases we use the same 4-phase scheme as Heilemann and Münch (1996) where phases are called ‘upswing’ (up), ‘upper turning points’ (utp), ‘downswing’ (down), and ‘lower turning points’ (ltp).

The compared classification methods include classical standard procedures like Linear Discriminant Analysis without (LDA) and with variable selection (LDA-VS), analogously Quadratic Discriminant Analysis QDA and QDA-VS, as well as Classification and Regression Trees (CART). Compared modern standard procedures are Multi-Layer Perceptrons (NN) and the linear Support Vector Method (SVM). Two recent developments (Weihs, Röhl, and Theis (1999)) based on projection pursuit algorithms constructed to guarantee optimal error rates in linear projections are included also. Both methods, called Minimal Error Classifier 1 and 2, are using either LDA or QDA for classification in the projected space (MEC1-L, MEC1-Q, MEC2-L, MEC2-Q). MEC1 is assuming normality only in the projected space, whereas MEC2 assumes global normality of the observations in each group. Another new method is a Discrete Dynamic Bayesian Network (DDBN) with a certain ‘rake’-structure, tailored for classification in dynamic domains (Sondhauss and Weihs, 1999).

[1] This work has been supported by the Collaborative Research Center "Reduction of Complexity in Multivariate Data Structures" (SFB 475) of the German Research Foundation (DFG).

2. Important Variables

One obvious way to try to decide on the most important variables for prediction with a certain classification method is to perform a variables selection based on cross validated error rates. For LDA and QDA best subsets consisting of 1,2,...,12 variables were constructed. Considering the best variables combinations and practically equivalent ones for different subset sizes all variables are included, although more or less frequently . Thus, one still has to decide how to measure 'importance'. A first idea was realized by Weihs, Röhl, and Theis (1999) by using the number of the appearances of the variables within the best eight models of sizes one to four, both for LDA and QDA.

As a comparable measure for the importance of variables in CART we ranked variables by their distance to the root-node and, within decision nodes on the same level, by their corresponding number of observations. Highest rank of 12 is given to the variable closest to the root with the highest number of observations, rank 0 to variables that do not appear in the tree. These rankings were summed up over trees gained from a leave-1-(business-) cycle-out analysis. As an alternative ranking on the same level we could have also used the splitting index.

All other methods were performed without variables selection. Thus, comparable rank-based measures of importance could not be used. Alternatives might be measures based on sensitivity analysis of the output of the rule learnt in a leave-1-predictor-out analysis.

Exemplifying outputs

Histograms are shown in figures 1 and 2 to display the frequency of appearance of stylized facts within the eight best models with 1 to 4 variables of LDA-VS and QDA-VS. The overall best four variables were wage-and-salary-earners-gr, unit-labor-cost-gr, real-investment-in-equipment-gr, and the GNP-price-deflator-gr.

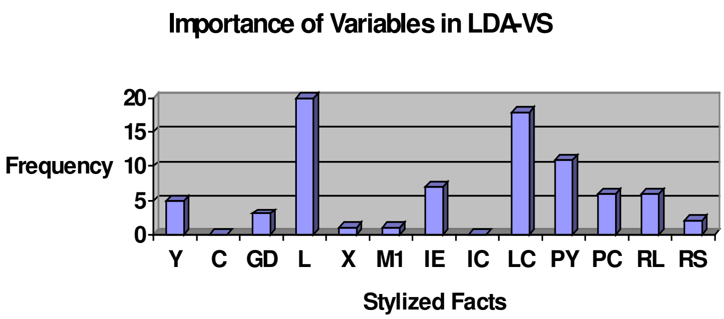

Figure 1: Appearances of variables within the best eight models of LDA

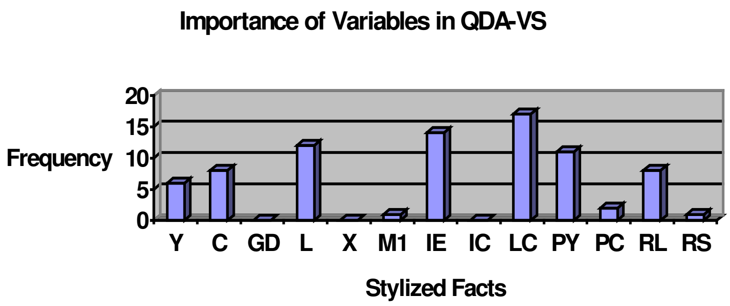

Figure 2: Appearances of variables within the best eight models of QDA

In figure 3 we show a histogram of the sum of ranks of the variables in CART. This is motivated by the need for a comparable graphic but can also be justified by interpreting ranks as weighted frequencies. With all this arbitrariness in measuring importance, the outcome is astonishingly robust: Most important were the same variables as in LDA-VS and QDA-VS, only the fourth variable was real-GNP-gr instead of its price deflator.

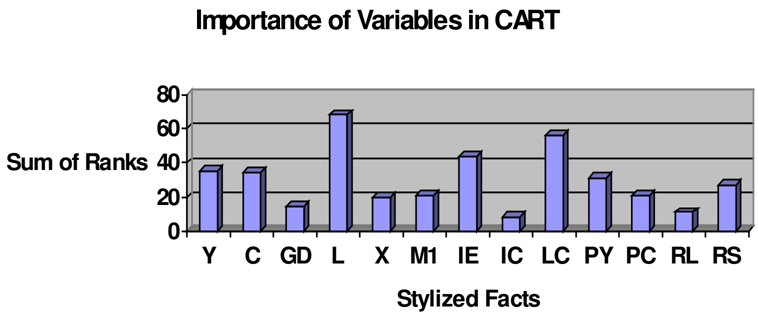

Figure 3: Sum of ranks of variables in leave-1-cycle-out analysis of CART

Learning from Important Variables

> *Astonishing enough it is not GNP which appears to be most important for business phase characterization in contrast to standard business cycle descriptions. Instead, two variables representing availability and cost of labor can be used to determine the business cycle phases.*

Rating according to Important Variables

If nothing is known about the relation of predictors and classes, no obvious performance measure for the goodness of a certain group of important variables exists. Domain knowledge though, might lead to a preference ranking of variables.

> *By means of economic arguments one might come to the conclusion that the result of CART should be preferred to the outcome of the discriminant analyses. Indeed, the usage of GNP instead of its price deflator should be preferred since otherwise the demand side of the economy is only represented by investment in equipment and prices are not expected to be such important to appear in the four most important stylized facts (Tichy, 1994).*

3. Partitions

Any ‘crisp’ classification rule partitions the observation space into sections corresponding to the different classes. These partitions characterize the connection of the class variable and the predictor variables. The understandability of the partitions of a classification rule thus is a relevant performance aspect for a rating of methods.

Exemplifying outputs

When only two variables are used for prediction, the rule can be visualized in diagrams as in figures 4 and 5 for LDA and QDA-VS. Note that for LDA the projection on the first two discriminant coordinates is displayed, whereas for QDA-VS the partition of the selected optimal 2D space in original variables is shown. The regions corresponding to the business phases are indicated by the abbreviations noted above. Including the data of the predictor variables of a cycle (here the cycle that was not used for generating the rule) helps to understand the progression of the business cycle with respect to these variables, and visualizes difficult regions where prediction errors occurred. Note that misclassified observations are marked by red arrows pointing at them.

Learning from partitions

Interpretations of partitions in latent variable spaces, resulting, e.g., from LDA, depend heavily on the latent variables. Therefore, no real understanding is gained if no textual interpretation is available. With very different latent variables, a comparison of the partitions of different methods is equally useless. Preferable for the interpretation of partitions are projections of rules at observed coordinates. From the partition resulting from QDA-VS (figure 5) one may learn the following.

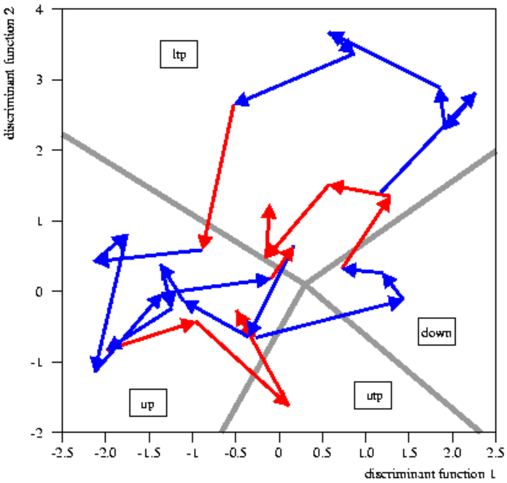

Figure 4: LDA partition in 2D

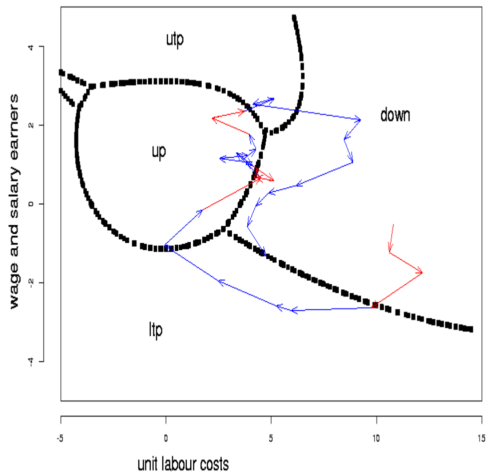

Figure 5: Partition from QDA-VS

> *Small absolute values of both wage-and-salary-earners-gr and unit-labor-cost-gr lead to an upswing, more extreme ones lead to a different phase. Upper turning points can only be reached if the growth rate of employment exceeds 3% and at the same time the growth rate of unit labor cost does not rise above 7%. A change of employment lower than –2% leads to a lower turning point, nearly no matter how unit labor cost behaves. Downswing can approximately be characterized by growth rates of employment bigger than –2% and simultaneous growth rates of unit labor costs of more than 5%.*

Rating according to Partitions

> *The partitions in figures 4 and 5 indicate that LDA has lesser problems with classifying upswing and QDA-VS with downswing. Moreover, on the one hand with QDA-VS the chance seems to be bigger to change from upswing erroneously directly to downswing without having touched the upper turning points because of the long common border of upswing and downswing. On the other hand with QDA-VS the upper and lower turning points do not have any border in common. Overall, however, the distinction of upswing and downswing is much more important, and thus the LDA partition might be preferred to the QDA-VS partition.*

4. Standardized Partition

For a standardized comparison of rules from very different classification methods we propose a new method using a diagram that is well known in experimental design, and that was used e.g. by Anderson (1958) to display regions of risk for Bayes classification procedures. Essential to our idea is the fact that almost all classification methods - all the above in any case - finally decide for a certain class using an argmax rule (e.g. the Bayes rule) based on transformations of the observations individual for each class. Interpreting these transformations as coordinates and standardizing in [0,1]-cubes of dimension g:= 'number of classes' means to use barycentric coordinates for the representation of membership measures of the observations to classes. A diagram showing the allocation of test set observations in a corresponding equilateral simplex leads to a comparable representation of very different rules.

An obvious measure of performance with respect to separation of groups is the average distance of test set observations to their corresponding true corner. This is the same as the root of half of the quadratic score on the test set. Note that this measure of inaccuracy overcomes the potential weakness of the error rate not taking into account how far from the thresholds the estimated memberships functions lay (cp. Hand, 1997, 100-101).

Exemplifying outputs

In figures 6-11 we show the simplexes gained by DDBN applied to the different training- and test sets of the leave-1-cycle-out analysis. Note that regions corresponding to classification into each of the four business phases are indicated by separating planes inside the simplex. Moreover, note that filled markers present misclassified observations where colors correspond to their true class.

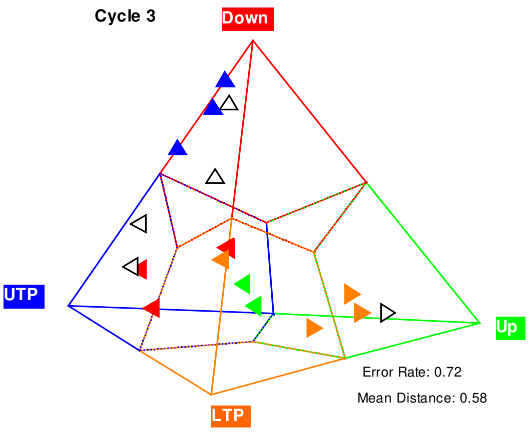

Figure 6: Standardized partition for DDBN

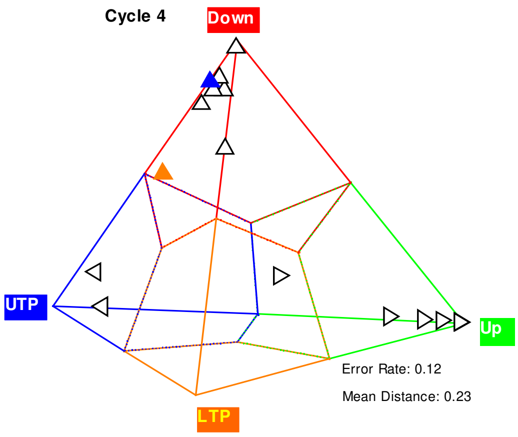

Figure 7: Standardized partition for DDBN

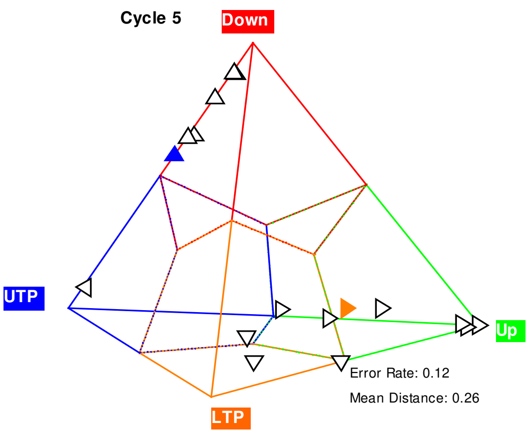

Figure 8: Standardized partition for DDBN

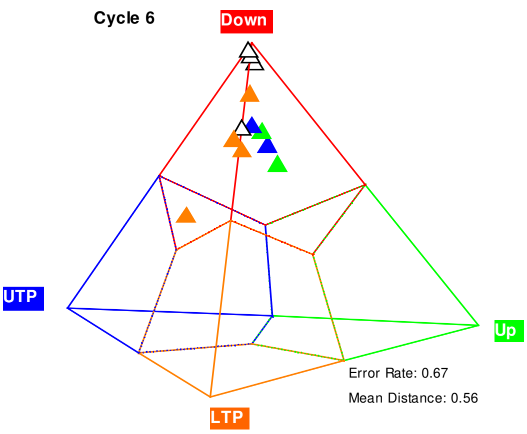

Figure 9: Standardized partition for DDBN

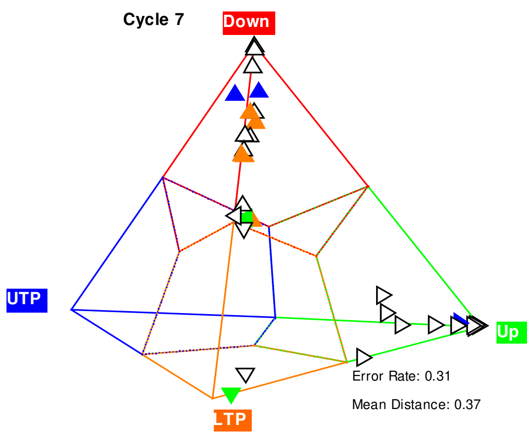

Figure 10: Standardized partition for DDBN

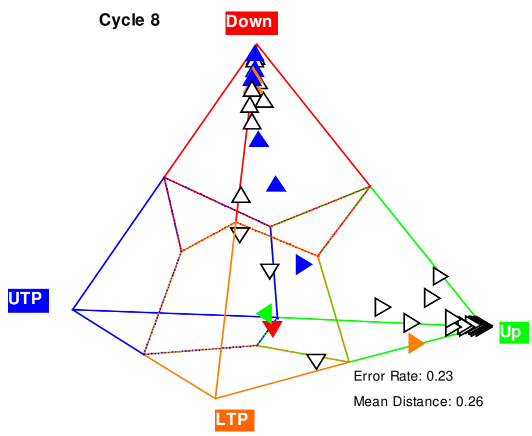

Figure 11: Standardized partition for DDBN

In figures 12 and 13 we compare the simplexes of DDBN and MEC2-Q on the 7th cycle.

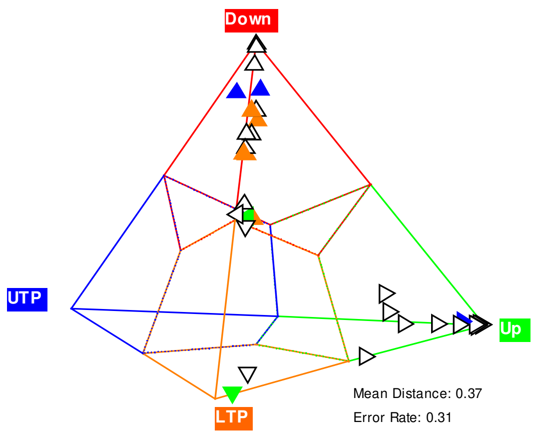

Figure 12: Standardized partition for DDBN

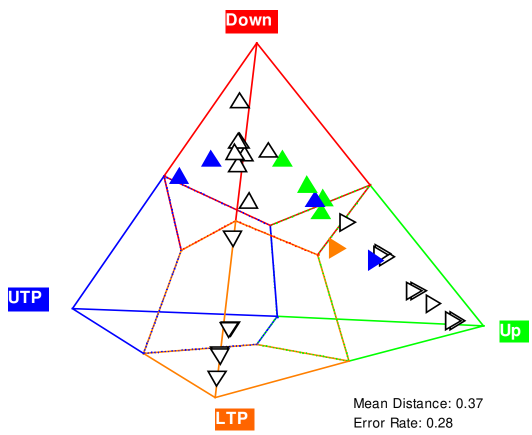

Figure 13: Standardized partition for MEC2-Q

Learning about Problem

On the one hand, standardized partitions can be interpreted best if the influence of the original variables can be identified. On the other hand, membership functions intrinsically have a textual interpretation. Additionally, because of space standardization we can learn about separability of groups by comparing different classifiers. In the present case, the standardized partitions of the leave-1-cycle-out analysis can be used to interpret deviations from the stability assumption of the data generating mechanism as special features of certain cycles:

> *Concentrating on those two cycles with poorest predictions one may identify two kinds of behavior. In cycle 3 no striking error structure is obvious, whereas in cycle 6 all observations are identified to be part of downswing. Thus, in cycle 3 there is no systematic deviation from the truth, whereas in the case of cycle 6 no cycle but only one phase is identified.*

Rating according to Standardized Partitions

Methods that lead to interpretable membership functions might be preferable to others. Another performance criterion is the above mentioned Euclidean distance to the true corner.

> *The position of the observations in figures 12 and 13 indicate that the separation in general and in particular the separation between upswing and downswing is stricter in DDBN than in MEC2-Q. Due to a higher misclassification rate of DDBN though, the performance w.r.t. Euclidean distance to the true corner of both classifiers is equal. Both classifiers have difficulties to identify upper turning points, MEC2-Q is better for lower turning points*

5. Conclusions

This paper gives examples for information other than error rates which might build a basis to compare the interpretability of different classification methods. It is demonstrated by means of some examples that such information refines the assertion upon the usability of a classification method in practice.

References

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, Canada.

Hand, D.J. (1997). *Construction and Assessment of Classification Rules*. John Wiley & Sons, Chichester

Heilemann, U. and Münch, H. J. (1996). West German Business Cycles 1963-1994: A Multivariate Discriminant Analysis, *CIRET--Conference in Singapore, CIRET--Studien 50*.

Lucas, R. E. (1983). Understanding business cycles. In *Studies in Business-Cycle Theory*. Cambridge, Mass.

Sondhauss, U. and Weihs, C. (1999). *Dynamic Bayesian Networks for Classification of Business Cycles*. Technical Report 26/1999, SFB 475, Department of Statistics, University of Dortmund.

Tichy, G. (1994). Konjunktur – Stilisierte Fakten, Theorie, Prognose. Springer Lehrbuch, Berlin

Weihs, C., Röhl, M. C., and Theis, W. (1999). *Multivariate Classification of Business Phases.* Technical Report 26/1999, SFB 475, Department of Statistics, University of Dortmund.