Optimal Data Analysis
Vol. 8 (April 6, 2019), 68-71

2155-0182/10/$3.00

Novometric Stepwise CTA Analysis Discriminating Three Class Categories Using Two Ordered Attributes

Paul R. Yarnold, Ph.D. and Ariel Linden, Dr.P.H.
Optimal Data Analysis, LLC Linden Consulting Group, LLC

The adaptability of novometric analysis is illustrated for an example involving three class categories and two ordered attributes.

ODA tests exploratory as well as confirmatory hypotheses for data configurations traditionally evaluated vis-à-vis a smorgasbord of historical statistical methods. Classic methods require assessing whether sample data violate specific distributional assumptions which are required to obtain valid findings.[1,2] In contrast, without any distributional assumptions ODA finds the model that is consistent with the hypothesis and maximizes predictive accuracy for sample data.[3-6]

This paper illustrates the malleability of ODA using a simulated analogue of a *novel data configuration* discussed online.[7] Data for this exposition (Table 1) are illustrated in Figure 1.

A *novel "stepwise" analysis* is used to identify the model which best discriminates all three class categories. The first (initial) step in stepwise analysis uses ODA to evaluate every possible comparison of three class categories first treating variable x as the model attribute (i.e., "independent variable"), and then treating variable y as the attribute (nodes of maximum-accuracy classification trees consist of a single attribute). The comparison having greatest effect strength in cross-generalizability analysis is used as the root node of the stepwise model.

Table 1: Simulated Data

Class 1		Class 2		Class 3	
x	y	x	y	x	y
3	9	2	11	6	8
4	8	4	11	7	7
6	7	5	12	8	13
7	11	5	7	9	15
7	9	6	12	9	7
7	8	6	11	11	15
7	5	8	12	11	10
8	8	8	11	11	5
8	7	9	14	12	13
9	11	9	13	12	8
9	8	9	12	14	13
11	12	9	13	15	8
8	15	9	9	16	14
12	6	11	8	16	11
5	9	11	7	16	6

Table 2 gives findings of the initial set of exploratory hypotheses evaluated for attribute x and then for attribute y.

2155-0182/10/$3.00

Figure 1: Simulated Three-Class Data Configuration in Two-Dimensional Space
(**Blue**=Class 1; **Red**=Class 2; **Black**=Class 3)

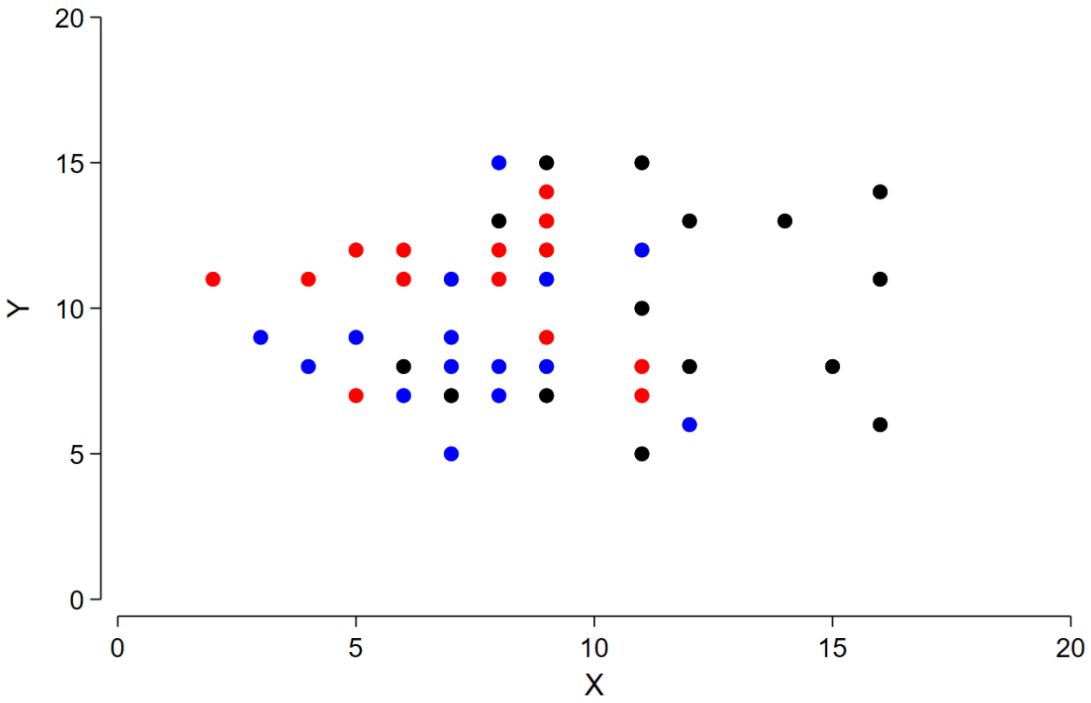


Table 2: Initial Set of Analyses

Attribute	Hypothesis	ESS	$p<$
x	$1 \neq 2 \neq 3$	36.67	0.022[a]
y	$1 \neq 2 \neq 3$	33.33	0.060
x	$1 \neq (2 = 3)$	36.67	0.060[b]
y	$1 \neq (2 = 3)$	36.67	0.067[b]
x	$(1 = 2) \neq 3$	53.33	0.0018
y	$(1 = 2) \neq 3$	26.67	0.281
x	$(1 = 3) \neq 2$	26.67	0.279
y	$(1 = 3) \neq 2$	36.67	0.066

[a]Leave-one-out (LOO) single case jackknife analysis is conducted when training $p<0.05$. Unless otherwise noted, LOO is stable. For this analysis, LOO ESS=20.00, $p<0.185$.
[b]The discriminant threshold for the model for x was 8.5 (stable in LOO), and 9.5 for y (not stable in LOO).

Table 2 indicates the model contrasting values on x between class 3 *vs*. the combined 1 and 2 classes yielded relatively strong ESS (no model having $p<0.05$ emerged for attribute y). The model was: IF x≤10 then predict class=3; OTHERWISE predict class=1 or class=2. Classification performance obtained in training and LOO analysis is given in Table 3.

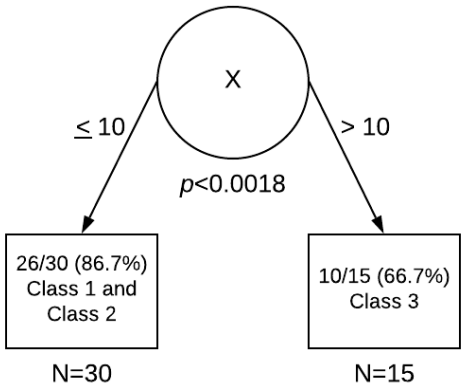
Table 3: Classification Performance of Initial Analysis for Attribute x: $(1 = 2) \neq 3$

		Predicted Category		
		1,2	3	*Sensitivity*
Actual	1,2	26	4	86.67
Category	3	5	10	66.67

The model identified in the first step is seen in Figure 2. Five misclassified observations for the right-hand endpoint renders inadequate statistical power to support further development of this branch—which is terminal.

2155-0182/10/$3.00

Figure 2: Initial Model: First Step



The second step of the analysis entails discriminating combined classes (1 and 2) on the left-hand side of the model (Figure 2). Table 4 summarizes findings of exploratory analysis for x, and then for y.

Table 4: Second Set of Analyses

Attribute	Hypothesis	ESS	*p*<
x	1 ≠ 2	20.00	0.721
y	1 ≠ 2	46.67	0.046

As seen, the model contrasting classes 1 and 2 on attribute y was statistically significant with moderate ESS (the model for attribute x had $p<0.721$). The model was: IF y≤10 then predict class=1; OTHERWISE predict class=2. Classification performance of this model in both *training and LOO* analysis is given in Table 5.

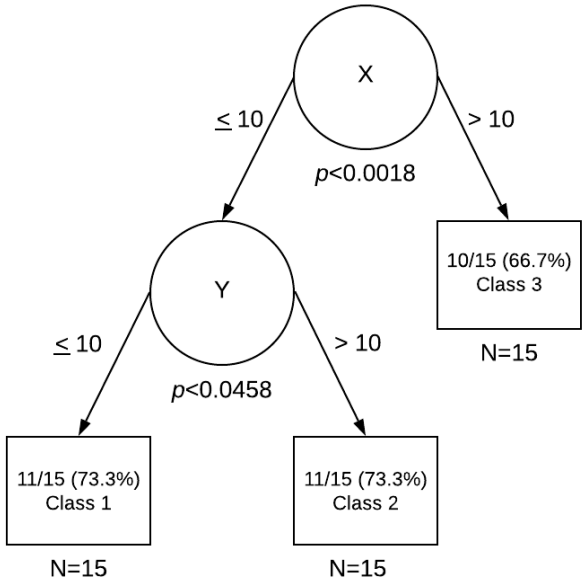
Table 5: Classification Performance of Second Analysis for Attribute y: 1 ≠ 2

		Predicted Category		
		1	2	*Sensitivity*
Actual	1	11	4	73.33
Category	2	4	11	73.33

Further improvement in discrimination of category 1 *vs*. 2 was infeasible because of inadequate statistical power afforded by the few remaining misclassified observations.[8]

Completing the composite model entails replacing the left-hand endpoint in Figure 2 with a model node representing variable y (Table 4). The final model discriminating all three of the class categories is presented in Figure 3.

Figure 3: Final Model: Second Step



Training and LOO classification performance of the final model is calculated for three class categories and sensitivities of 66.67% for Class 3 (Table 3) and 73.33% for both Class 1 and Class 2 (Table 5). Mean sensitivity across classes is 71.11%, so ESS=[(71.11-33.33)/(100-33.33)]x100%=56.67%, indicating a relatively strong effect which is stable in LOO analysis.

Discussion

It is natural to wonder how results obtained by novometric stepwise classification tree analysis (CTA) compare with results obtained by legacy statistical methods. We provide the data we used (Table 1) for this purpose.

2155-0182/10/$3.00

MANOVA is a popular parametric method which entails regressing x and y (dependent variables) on the 3-level class variable (treated as a categorical variable).[1] However, MANOVA does not assess the inter-dependencies between x and y in relation to (i.e., concurrently assessing) each's association with the 3-level class variable—thus omitting a crucial component of the analytic design. And, validity of its findings requires assumptions that may be violated in most samples, for example that data are multivariate normally distributed (for which there is no test). Other widely-used linear parametric methods include discriminant function analysis and multinomial logistic regression. Linear methods are not intrinsically (i.e., by design) designed to identify nonlinear effects such as was found by novometric stepwise CTA (Figure 3). Perhaps multidimensional scaling is better-suited to successfully contrast the three groups with respect to x and y.

This paper considered discriminating between levels of a categorical class variable having three levels. Novometric stepwise CTA can be extended to applications involving more categorical levels, and to applications involving an ordered ("continuous") class variable.[9] This paper dealt with the class variable abstractly without any assumptions of what the classes represent. In practice, this approach can be applied to problems of causal inference, most notably by using the x and y variables to ensure comparability between classes–which, in turn, can then be used in an outcomes analysis.[10,11]

References

[1]Grimm LG, Yarnold PR (Eds.) (1995). *Reading and understanding multivariate statistics*. Washington, D.C.: APA Books.

[2]Grimm LG, Yarnold PR (Eds.) (2000). *Reading and understanding more multivariate statistics*. Washington, D.C.: APA Books.

[3]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: Guidebook with software for Windows*. Washington, D.C.: APA Books, 2005.

[4]Yarnold PR, Soltysik RC (2010). Optimal data analysis: A general statistical analysis paradigm. *Optimal Data Analysis*, *1*, 10-22.

[5]Yarnold PR, Soltysik RC (2016). *Maximizing Predictive Accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[6]Yarnold PR (2017). What is optimal data analysis? *Optimal Data Analysis*, *6*, 26-42.

[7]https://www.researchgate.net/post/How_to_statistically_test_significance_between_scatter_groups

[8]Yarnold PR, Bryant FB (2015). Obtaining a hierarchically optimal CTA model via UniODA software. *Optimal Data Analysis*, *4*, 36-53.

[9]Yarnold PR, Linden A (2016). Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis. *Optimal Data Analysis*, *5*, 65-73.

[10]Linden A, Yarnold PR (2016). Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, 22, 868-874.

[11]Linden A, Yarnold PR (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*, *22*, 875-885.

Author Notes

Artificial data were analyzed. No conflict of interest was reported.