Optimal Data Analysis
Vol. 4 (May 20, 2015), 81-86

2155-0182/10/$3.00

Distance from a Theoretically Ideal Statistical Classification Model Defined as the Number of Additional Equivalent Effects Needed to Obtain Perfect Classification for the Sample

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

A method for computing the distance between an empirically-derived statistical classification model and a corresponding theoretically ideal classification model is described. Use of the distance index to identify and to compare globally optimal classification models, within and between descendent families, is illustrated with an example using ethnicity to parse the incidence of different types of cancer.

For applications involving a sample of classical data, a mathematical-programming method called optimal data analysis (ODA) is used to identify the non-parametric, exact (non)linear statistical classification model that *explicitly maximizes classification accuracy* for the sample considered as a whole, or when selectively weighting specific outcomes of interest.[1,2] In an empirical investigation, it is possible, of course, that statistical analysis may find, for example, that in the sample no reliable (non)linear model exists for predicting X on the basis of Y. And, it is also a possibility that statistical analysis may identify a reliable model, or may even identify *multiple* reliable models—which together are known as the *descendant family*[3]—for predicting X on the basis of Y.

The present paper discusses how to identify the "best" model in the descendant family. The best model in the descendant family is known as the "globally-optimal" model (or GO model) for the specific sample and application.

An *ideal* statistical classification model achieves perfect (errorless) classification of every observation in the sample, and accomplishes this using the smallest possible number of model endpoints.[3,4] Fewer model endpoints are desirable because an increasing number of endpoints reflects increasing model complexity—manifest in the increasing domain of unique *strata* that exist within the sample.[3,4] In contrast, minimizing the number of model endpoints (strata) maximizes model *parsimony*.[1-4]

Comparing the quality of an empirical model to a corresponding theoretically ideal model has been described conceptually in terms of assessing Euclidean distance of the empirical result from the upper right-hand corner of a unit square Cartesian space defined by two orthogonal axes, with accuracy (ESS) as abscissa, and parsimony—quantified as ESS divided by the number of strata in the model—as ordinate.[3]

This conceptual perspective is unproductive as a means of computing the distance between an empirical and a theoretically ideal model. This is because if an interactive transformation[1,3] is used to obtain a unit efficiency scale separately for problems of varying complexity (number of strata), then all models in the descendant family lie along the proper diagonal between chance (0,0) and the theoretically ideal model (1,1). The distance of the empirical model from the theoretically ideal model in this approach is a perfect function of ESS. The use of interactive transformations to standardize efficiency to unit scale separately by number of strata is necessary to obtain a unit square space for models differing in complexity, but this standardization ignores the role of model complexity.

Distance of an Empirical Classification Model from a Theoretically Ideal Classification Model

Distance from a theoretically ideal statistical classification model is heuristically defined here as the number of additional equivalent effects needed to obtain perfect classification for the sample. Imagine that a 3-strata model achieved overall ESS = 75, with efficiency of 75 / 3 = 25. If one additional attribute is identified—that produces an equivalent effect having efficiency of 25, then overall ESS = 100 and the ideal model for this sample is identified. Distance from a theoretically ideal statistical classification model is computed using the formula (here, Strata is the number of strata in the model): [100 / (ESS / Strata)] – Strata. Note that this heuristic evaluates both accuracy (ESS) and parsimony (strata) in computing the distance of an empirically-obtained classification model from the corresponding theoretically ideal model, for a given sample and application.

Empirical Example: Parsing Cancer Incidence by Ethnicity

Cancer types for which no parsing model was obtained, and for which only one parsing model was identified, were ignored for this exposition because no between-model comparisons were possible within such cancer types.[3]

Table 1 presents selected examples of application of the present algorithm, for selecting globally optimal models, for an application involving multi-model parsing of cancer incidence as a function of ethnicity—white and African-American (Appendix 1 presents results for all multi-model applications).[3] Separately by type of cancer, Table 1 first reports the number of *Strata* parsed by the model; second the number of observations in the smallest strata—this endpoint parameter is known[2] as the minimum denominator or *MinD*; third the overall model accuracy indexed as *ESS* with 0 = the level of accuracy that is expected by chance, and 100 = perfect accuracy; fourth model parsimony indexed as *Efficiency* = ESS / Strata; and finally, fifth, the difference between the empirical model and the theoretical ideal, indexed as the number of additional equivalent effects needed to obtain perfect classification in the sample: (100 / Efficiency) – Strata.

Most findings were similar to the pattern of results that emerged for small intestine cancer. As seen in Table 1, the efficiency of 27.6 / 4 = 6.90 for the more complex 4-strata model corresponds to a distance of [(100 / 6.90) – 4] = 14.49 – 4 = 10.49 additional equivalent effects needed to obtain perfect classification for the sample. And, the efficiency of 22.7 / 3 = 7.57 for the less complex 3-strata model corresponds to a distance of [(100 / 7.57) – 3] = 10.21 addi-

Copyright 2015 by Optimal Data Analysis, LLC
2155-0182/10/$3.00

tional equivalent effects needed to obtain perfect classification for the sample. Because the distance of the less complex 3-strata model from a theoretically ideal classification model is less than the corresponding distance of the more complex 4-strata model, the less complex 3-strata model is selected as the best model in the descendant family—the globally optimal or GO model for this analysis, for this sample.

Table 1: Identifying the Globally Optimal Model for Parsing the Incidence of Selected Types of Cancer by Ethnicity for This Sample[3]

Small Intestine

Strata	MinD	ESS	Efficiency	Distance
4	59	27.6	6.90	10.49
3	155	22.7	7.57	10.21
Tongue				
5	29	35.5	7.11	9.06
4	122	29.3	7.32	9.66
3	135	20.7	6.91	11.47
Prostate				
4	35	32.9	8.22	8.17
3	36	23.7	7.89	9.67
2	71	19.1	9.54	8.48
Esophagus				
5	39	29.6	5.92	11.89
3	184	24.0	8.00	9.50
2	221	16.8	8.39	9.92
Melanoma of the Skin				
5	25	75.3	15.06	1.62
4	63	71.4	17.85	1.62
2	234	67.1	33.55	0.98

Note: The minimum distance of the empirical model from the theoretically ideal model is highlighted using red font within each descendant family.

In contrast, for tongue cancer even though the most complex 5-strata model wasn't the most efficient model in the descendant family, it nevertheless was the closest model to the theoretically ideal model for this sample. Similarly, for prostate cancer even though the most complex 4-strata model wasn't the most efficient model in the descendant family, it was the closest model to the theoretically ideal model for this sample. And, for esophagus cancer, the 3-strata model of intermediate complexity did not have highest efficiency, but it was the closest model to the theoretically ideal model for this sample.

The strongest result obtained occurred for melanoma of the skin: the single threshold-based 2-strata model (white, African American) achieved ESS of 67.1, corresponding to an efficiency of 33.55, and a distance of less than one additional equivalent effect needed to obtain perfect classification for the sample.

References

[1]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*, Washington, DC, APA Books.

[2]Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis, 1*, 144-160. URL: http://odajournal.com/2013/09/19/62/

[3]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis*, *3*, 55-77. URL: http://odajournal.com/2014/08/18/globally-optimal-statistical-classification-models-i-binary-class-variable-one-ordered-attribute/

[4]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, II: Unrestricted class variable, two or more attributes. *Optimal Data Analysis*, *3*, 78-84. URL: http://odajournal.com/2014/08/25/globally-optimal-statistical-models-ii-unrestricted-class-variable-two-or-more-attributes/

Appendix

Results for All Multi-Model Applications

Cancer Incidence Parsed by SEX

All Sites Combined

Strata	MinD	ESS	Efficiency	Distance
6	2	33.2	5.53	12.1
5	63	33.2	6.64	10.1
3	80	31.9	10.63	6.4

Bones and Joints

4	58	20.4	5.10	19.6
2	251	14.8	7.40	11.5

Non-Hodgkin Lymphoma

4	73	19.1	4.77	17.0
2	299	12.8	6.42	13.6

Leukemia

4	41	24.3	6.08	12.4
2	84	15.1	7.56	11.2

Cancer Incidence Parsed by ETHNICITY

Tongue

Strata	MinD	ESS	Efficiency	Distance
5	29	35.5	7.11	9.06
4	122	29.3	7.32	9.66
3	135	20.7	6.91	11.47

Nasopharynx

3	88	29.0	9.65	7.36
2	113	26.3	13.15	5.60

Tonsil

5	48	27.0	5.39	13.55
3	106	17.8	5.92	13.89
2	209	14.8	7.40	11.51

Oropharynx

4	41	32.9	8.22	8.17
3	158	26.3	8.77	8.40
2	291	21.4	10.70	7.35

Other Oral Cavity and Pharynx

6	35	29.9	4.99	14.04
4	37	27.0	6.74	10.84
3	118	26.6	8.88	8.26
2	271	17.4	8.72	9.47

Esophagus

5	39	29.6	5.92	11.89
3	184	24.0	8.00	9.50
2	221	16.8	8.39	9.92

Stomach

3	110	19.1	6.36	12.72
2	274	11.2	5.59	15.89

Small Intestine

4	59	27.6	6.91	10.47
3	155	22.7	7.57	10.21

Appendix

6	25	42.1	7.02	8.25
4	42	39.8	9.95	6.05
2	191	33.9	16.95	3.90

Hepatic Fracture

3	78	15.1	5.04	16.84
2	194	13.2	6.58	13.20

Splenic Flexure

3	193	22.0	7.35	10.61
2	195	11.5	5.76	15.36

Liver and Intra-Hepatic Bile Duct

4	2	19.4	4.85	16.62
3	54	18.8	6.25	13.00

Liver

7	2	28.0	3.99	18.06
4	55	25.3	6.33	11.80

Other Biliary

3	3	17.4	5.81	14.21
2	216	16.4	8.22	10.17

Other Digestive Organs

3	119	24.0	8.00	9.50
2	264	22.4	11.20	6.93

Larynx

5	29	31.6	6.32	10.82
3	170	24.0	8.00	9.50

Trachea, Mediastinum, Other

6	48	56.2	9.38	4.67
5	72	53.0	10.60	4.43
3	110	50.0	16.67	3.00
2	180	46.7	23.35	2.28

Skin excluding Basal and Squamous

5	14	66.1	13.22	2.58
4	51	63.8	15.95	2.25
2	229	61.5	30.75	1.25

Melanoma of the Skin

5	25	75.3	15.1	1.62
4	63	71.4	17.8	1.62
2	234	67.1	33.6	0.98

Cervix Uteri

3	84	40.8	13.6	4.35
2	90	39.5	19.7	3.08

Male Genital System

4	29	36.8	9.21	6.86
3	78	29.0	9.65	7.36
2	99	23.0	11.50	6.70

Prostate

4	35	32.9	8.22	8.17
3	36	23.7	7.89	9.67
2	71	19.1	9.54	8.48

Ureter

4	80	32.2	8.06	8.41
2	288	28.3	14.15	5.07

Other Urinary Organs

3	151	24.0	8.00	9.50
2	250	14.5	7.24	11.81

Eye and Orbit

7	27	71.1	10.16	2.84
6	47	68.1	11.35	2.81
5	51	62.8	12.56	2.96
2	202	62.5	31.25	1.20

Cranial Nerves, Other Nervous Systems

3	48	43.1	14.37	3.96
2	119	34.5	17.25	3.80

Other Endocrine including Thymus

4	31	33.6	8.39	7.92
3	79	32.6	10.87	6.20

Hodgkin Lymphoma

4	30	31.6	7.90	8.66
2	280	26.3	13.15	5.60

Hodgkin-Nodal

4	33	30.3	7.56	9.23
2	261	25.3	12.65	5.91

Acute Monocytic Leukemia

4	69	60.9	15.22	2.57
2	218	57.2	28.60	1.50

Chronic Myeloid Leukemia				
5	31	17.4	3.49	23.65
3	52	15.1	5.04	16.84
Other Acute Leukemia				
3	2	43.8	14.60	3.85
2	159	43.1	21.55	2.64
Aleukemic, Leukemic and NOS				
3	102	33.6	11.20	5.93
2	185	30.6	15.30	1.27

Note: The minimum distance of the empirical model from the theoretically ideal model is highlighted using red font within each descendant family.

Author Notes

This study involved secondary data analysis of published de-identified data and was exempt from Institutional Review Board review. The author declared no conflict of interest.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646
USA