Optimal Data Analysis
2010, Vol. 1, Release 1 (September 17, 2010), 10-22

2155-0182/10/$3.00

Optimal Data Analysis: A General Statistical Analysis Paradigm

Paul R. Yarnold, Ph.D., and Robert C. Soltysik, M.S.
Optimal Data Analysis, LLC

Optimal discriminant analysis (ODA) is a new paradigm in the general statistical analysis of data, which explicitly maximizes the *accuracy* achieved by a model for every statistical analysis, in the context of exact distribution theory. This paper reviews optimal analogues of traditional statistical methods, as well as new special-purpose models for which no conventional alternatives exist.

Rarely does a technical report concerning an apparently focused and arcane classification methodology, such as optimal discriminant (data) analysis—ODA, stand a realistic chance of appealing to a diverse scientific community. Even more rarely, however, does one have the opportunity to report the emergence of a new paradigm in the statistical analysis of data.[1] ODA is a highly intuitive, powerful, and exact methodology for the general statistical analysis of data, and this paper reports on the emergence of this paradigm.

ODA is *the* methodology that explicitly maximizes the accuracy of any type of statistical model for the training sample—that is, for the data upon which statistical analysis is performed and upon which the statistical model is based. An increasing awareness of the intuitive appeal of maximizing accuracy (and minimizing errors), and commercial availability of dedicated software, are fueling increasingly widespread application of ODA.[1] Nevertheless, because ODA is relatively new, and therefore relatively few introductory and review resources covering the paradigm are yet widely available, this paper introduces many major concepts and methods of the ODA paradigm.

Initial Assumptions

An ODA model explicitly maximizes the number of correctly classified observations for a specific application. Observations are considered correctly classified when the model assigns them to the class of which they are, in reality, a member, and are misclassified otherwise. The number of misclassifications arising in a given analysis is referred to as the "*optimal value*." It is clear that derivation of a distribution theory for ODA requires investigation of distributions underlying optimal values. Using the *simplest possible data structure* to illustrate derivation of exact distribution theory, imagine a hypothetical application having the following three features.

First, assume one binary class variable. In ODA, a class variable is what one is trying to predict, discriminate, or classify. Examples of binary class variables might include gender (male, female), therapy (drug, placebo), or outcome (success, failure). Class variables may of

2155-0182/10/$3.00

course consist of more than two levels, but two levels is the simplest case.

Second, assume one random continuous attribute. In ODA, an attribute is a variable that will be employed in an effort to predict the class variable. The continuity assumption implies that every observation will achieve a unique score on the attribute (no ties). Nothing is assumed about the shape of the distribution underlying scores on the attribute, but only that the scores are random—for example, uniform or normal. Single-attribute ODA analyses are referred to as univariable ODA, or UniODA. Because the present case involves a continuous attribute, we are discussing a "continuous UniODA design".

Finally, assume three observations: two from one class, and one from the other class (three observations are required because with two the problem is trivial: the mean of two observations' scores on a continuous attribute is a perfect discriminant classifier for those two observations). Though it is arbitrary, refer to these as classes "1" and "0", respectively. Hereafter, the total number of observations is referred to as n, and the number of observations in class c as n_c.

Note that only the continuity assumption is capable of being violated by "real-world" data (we return to this point later). The first (binary class variable) and third (n in each class) assumptions can never be violated because they exactly define the structure of the design. That is, we are considering a UniODA design with a binary class variable, and with n_1=2 and n_0=1: any deviation from this structure, such as more than two class levels or different sample sizes, simply defines another specific UniODA design.

The UniODA Model

For clarity we give an example of a two-category continuous UniODA model. Imagine that a cardiologist wished to determine if heart rate variability (HRV)—the standard deviation of one's heart rate over a 24-hour period (the continuous attribute), can discriminate patients who die (class 0) versus live (class 1). For a given sample UniODA would provide at least one optimal model, consisting of a *cutpoint* and a *direction*, which when used together explicitly maximize forecasting accuracy: percent accurate classification, or PAC. For example, a UniODA model could be: "if HRV score is greater than (direction) 12.2 (cutpoint), assign that person to class 0; otherwise, assign that person to class 1."

A UniODA model is said to be optimal because the total number of misclassifications resulting from application of the model to the data is minimized, and the number of correct classifications is maximized. In the example, no alternative combination of HRV cutpoint and direction would yield fewer misclassifications than the model which UniODA identified.

Multiple optimal models which all yield the same maximum PAC may occur for a given data set. For example, two different HRV cut-points might result in the same overall number of misclassifications, yet one model may have greater sensitivity (ability to accurately classify members of class 1) and lower specificity (ability to accurately classify members of class 0) than the other model. In such cases, it is necessary to select one optimal model, preferably before conducting the analysis, using an appropriate decision heuristic.[1] Examples of such selection heuristics include the sensitivity or specificity heuristic (select the model having greatest sensitivity or specificity, respectively), or the balanced performance heuristic (select the model with the smallest difference between sensitivity and specificity).[1]

Exact Distribution Theory

We are now ready to derive the theoretical distribution of optimal values for a two-category continuous UniODA design with n_1=2 and n_0=1. First, it is necessary to determine the set of all possible outcomes that could occur if the attribute were continuous and random. In order to differentiate the two observations from class 1, they will be called "1A" and "1B."

2155-0182/10/$3.00

There are six possible outcomes: one is that the value of the attribute for observation 1A is greater than that for observation 1B, which in turn is greater than that for the observation from class 0. Symbolically, {1A > 1B > 0}. The five other possible outcomes are: {1A > 0 > 1B}; {1B > 1A > 0}; {1B > 0 > 1A}; {0 > 1A > 1B}; and {0 > 1B > 1A}. Because the attribute was random, each of these six possible outcomes is equally likely, with a probability of 1/6.

Next, it is necessary to determine the optimal value for each of the six possible outcomes. This, of course, means that UniODA must be performed for each of the six possible data configurations.[1] Two of the six possible outcomes (those in which the attribute of the class 0 observation lies between the attributes of the two class 1 observations) have an associated optimal value of 1 misclassification, because at least one observation will be misclassified regardless of where the cutpoint is placed). The other four possible outcomes (in which the two class 1 observations can be perfectly separated via a cutpoint from the class 0 observation) have an optimal value of 0 misclassifications. Cumulating optimal values over the set of possible outcomes gives the theoretical distribution of optimal values for this UniODA design: the probability of an optimal value of 0 is 4/6, and the probability of an optimal value of 1 is 2/6.

Enumerating in this manner the theoretical distribution of optimal values for balanced (equal number of class 0 and 1 observations), continuous, two-tailed (no *a priori* hypothesis was specified) UniODA designs required a CRAY-2 supercomputer—which only achieved results for $n \leq 30$ due to exponential increases in the number of combinations.[2] Examination of the resulting table of optimal values for *post hoc* UniODA revealed organization which motivated discovery[3] and proof[4] of a closed-form solution for one-tailed confirmatory UniODA.

Inexact Measures

What if data aren't continuous, and there are ties—violating the continuity assumption? Discontinuity in empirical data is thought to reflect imprecise measurement, and not as compromising of theoretical probabilities[5], but this begged the question of exactly how imprecise can measurement become before the theoretical probabilities become compromised? This line of thinking naturally led to the question of what would occur for a binary attribute—and it was then that we understood that the binary attribute problem was the optimal analogue to chi-square analysis, and the continuous attribute problem was the optimal analogue to *t*-test. Proceeding with binary enumeration we found the binary and continuous distributions differ. This finding motivated two important insights.

First, there is a theoretical dimension—which we call *precision*—which may be used to describe the metric underlying the attribute for any specific UniODA problem. The precision dimension is bounded at the extremes by binary data (least precise) versus continuous data (most precise). Just as specific distribution theory can be derived for the extreme poles of the precision dimension, so too can *exact* distribution theory be derived for every specific attribute measure metric: for example, if the attribute is measured using a 7-point Likert scale, then derive distribution theory by assuming a 7-point Likert scale was used. As it is possible to derive distribution theory that assumes that the specific measure metric actually used in a given application was in fact used, distribution theory for ODA can be based strictly on structural features of a problem, and such distribution theory *will never be violated* by data for a given application.

The second insight is that UniODA is an optimal alternative to common conventional statistical methods: Student's t-test is often used to analyze data involving a binary class variable and a continuous attribute, and chi-square is often used to analyze data involving a binary class variable and a binary attribute. UniODA

2155-0182/10/$3.00

may also be used, and exact distributions may be determined for, designs that lie anywhere on the precision dimension—anywhere between the binary and continuous polar extremes. This is not true for conventional statistical procedures.

ODA as an Alternative to Conventional Statistical Methodologies

Encouraged by early success, we began programmatic research to assess the domain of experimental designs and data configurations that may be addressed using UniODA. We next investigated multicategory problems involving class variables with more than two levels. For a continuous attribute, multicategory UniODA is analogous to oneway analysis of variance, and for a binary attribute it is analogous to log-linear analysis.[6,7]

UniODA, and other models within the ODA paradigm, clearly can be used to analyze different data configurations that are evaluated using a host of different conventional statistical methods. Why should ODA be used rather than a host of conventional methods?

First, only ODA explicitly maximizes (weighted) classification accuracy and provides a forecasting model for every application. Not only do conventional methods fail to explicitly maximize PAC, but many, such as *t*-test or chi-square, also fail to provide a forecasting model.

Second, no matter what the nature of a particular data configuration might be—for example, the number of class levels, attribute metrics, or class sample-size imbalances, the classification performance of every ODA model is summarized using a normed measure of effect strength, called effect strength for sensitivity, or ESS.[1] On this index 0 represents the level of classification performance that is expected by chance, and 100 represents perfect, errorless classification. No such intuitive, universal index can be used to compare the effect strength of different conventional methods such as analysis of variance, logistic regression, and tau.

Third, conventional methods require assumptions regarding the nature of the data. Unlike ODA—for which distribution theory is exact for every design, conventional methods are inappropriate when the data violate their assumptions. Whereas the assumptions of ODA must conform to the data, data must conform to the assumptions of conventional methods.

Finally, with ODA a *single methodology* may be *optimally applied* to analyze a *host of problems*, while with the conventional approach a *host of methods* may be *suboptimally applied* to analyze *a single problem*. ODA is therefore simultaneously more unique *and* parsimonious than conventional methods.

To illustrate the flexibility and power of ODA as a general statistics paradigm, below we describe different common data configurations and conventional methods often used in their analysis, and the corresponding ODA model.

Binary Class Variable and BinaryAttribute

The most common conventional method for analyzing data of this type is chi-square analysis: the ODA analogue is two-category binary UniODA. Chi-square is an approximate statistic that should not be used when the expected value for a given cell (cells are formed by cross-tabulating the class variable with the attribute) is less than five.[8] In contrast, binary UniODA is an exact statistic with no such restriction: one- and two-tailed estimated *p* by UniODA and Fisher's exact test are isomorphic except in a hypothetical degenerate condition.[1]

It is easy to show that UniODA may be particularly useful in small sample designs. For example, imagine a problem with $n = 6$, three observations from class 0 all scoring 0 on the attribute, and three observations from class 1 all scoring 1. Chi-square can't be used to analyze this problem, as the expected value is less than five in all four cells. When analyzed using two-tailed binary UniODA, a single optimal model (if attribute < 0.5 then class = 0; else class = 1)

2155-0182/10/$3.00

emerged that achieved 100% PAC, $p<0.032$. No systematic review/comparison of chi-square versus binary UniODA has yet been reported.

Binary Class Variable and Multiple Binary Attributes

The most common linear methods for analyzing data of this type include log-linear or logistic regression analysis. Completely binary problems are easiest for ODA to solve, but can be problematic for conventional methods, with aspects including marginal imbalance, sparse cells, singularities, and structural zeros (some design cells don't exist), for example, rendering binary data difficult or impossible to analyze. The optimal linear analogue is binary MultiODA—a linear model which uses two or more attributes to explicitly maximize classification accuracy (discussed ahead).

For example, we reanalyzed data from a study designed to predict if 120 persons with AIDS would require home care or structured long-term care (the class variable) on the basis of three binary attributes which assessed the attitudes of patient and physician towards long-term care, and whether the patient had mental impairment.[9] The data were "ill-condiioned" and thus could not be analyzed by log-linear or logistic regression methods. MultiODA, however, found a two-attribute model that achieved 93.3% PAC in <1/20 CPU second on a 33MHz 386 microcomputer running a special-purpose ODA search algorithm (discussed ahead).

Binary Class Variable and Continuous Attribute

Among the most frequently reported of statistical tests, Student's *t*-test is a common conventional procedure for analyzing data of this type. The ODA analogue is two-category continuous UniODA.

It is easy to construct a hypothetical problem for which *t*-test fails to find a significant intergroup mean difference on the attribute, while UniODA detects nearly perfect intergroup discriminability. Imagine that ten class A observations each score a value of 0 on the attribute; nine class B observations all score 1, and a tenth class B observation scores -9. Because the mean difference on the attribute between groups is zero, *t*-test would conclude that the groups can't be discriminated whatsoever by the attribute. But, with UniODA, 95% of the observations are correctly classified—nearly perfect intergroup discriminability. Systematic research contrasting UniODA and *t*-test is not yet available.

Binary Class Variable and Multiple Continuous Attributes

Common linear methods for analyzing data in this configuration are linear discriminant analysis, logistic regression analysis, and one-way multivariate analysis of variance.[6,7] The linear ODA analogue is continuous MultiODA, but UniODA has been used with great success to maximize accuracy achieved by suboptimal models.[10,11]

Monte Carlo research is often used to contrast continuous MultiODA versus conventional statistical methods.[12,13] A difficulty with such simulation research is that the experimental data are generated using idealized routines that meet criteria—such as normally distributed data and coincident covariance, which are important for conventional statistical methodologies but which are no substitute for "real-world" data collected by naturalistic empirical observation. Our strategy has been to analyze a variety of different applications using MultiODA, and then compare the performance against suboptimal methods such as Fisher's discriminant or logistic regression analysis, using training and validity data. Early results are encouraging, but more research is needed to compare "in the field" classification performance of MultiODA versus conventional procedures.[9,14,15]

2155-0182/10/$3.00

Binary or Multicategory Class Variable and Continuous and Binary Attributes

Multinomial logistic regression analysis is a commonly employed conventional analysis for problems of this type. The linear optimal analogue is MultiODA, with weights used to reduce problem size by eliminating redundant data profiles (discussed ahead). Little research using either approach is available, and to our knowledge no prior research comparing these approaches has yet been published (until now).

Analyzing credit screening data for a British bank, our objective was to develop a model to predict credit worthiness (the class variable) for a sample of 325 credit applicants. Attributes were two binary variables and a third 4-point ordinal attribute. A nonparametric classification methodology that performed sample stratification based on a recursive chi-square procedure identified four interaction terms used as attributes in follow-up analysis. With these data logistic regression analysis and MultiODA both achieved 90.5% PAC in training analysis, but the latter model used one less term (and thus was more efficient and parsimonious) than the former model. Comparing the two models using jackknife validity analysis revealed that PAC for the MultiODA model was stable, but regressed to 83.1% for the obviously over-determined logistic regression model.

Multicategory Class Variable and Polychotomous Attribute

Common conventional methodologies for analyzing these designs include chi-square, log-linear, or multinomial logistic regression analysis. The optimal analogue is multicategory UniODA. As was true for designs that involved one binary class variable and multiple binary attributes, issues such as structural zeros, sparse cells, imbalanced marginal distributions, small samples, and multicollinearity may spell disaster for conventional designs. As discussed earlier, these are *not* problems for ODA.

It is easy to construct an example for which conventional analyses are inappropriate, but for which multicategory UniODA is ideal. For example, imagine a problem with a three-category (A, B, C) class variable, with each category having three observations. Further imagine all three class A observations scored a value of 1 on the attribute; all three class B observations scored a 2, and all three class C observations scored a 3. Although the small sample renders conventional methods inappropriate, a multicategory UniODA achieved 100% PAC, two-tailed $p<0.01$.

Multicategory Class Variable and Continuous Attribute

The most common conventional analysis used for such designs is oneway analysis of variance, and the optimal analogue is multicategory UniODA. As for *t*-test, distribution theory for analysis of variance is highly sensitive to assumption violations.[5] Such data can present insurmountable problems for multinomial logistic regression, because of small samples, sparse cells, and marginal imbalance, particularly when polychotomous attributes are thrown in the mix: for example the analysis will fail if a degenerate attribute—which has fewer response categories than the class variable has levels—is included in the analysis.

As an example of a three-category UniODA, imagine the following hypothetical data set, problematic for conventional methods due to the small sample, the presence of outliers, heterogeneity, the presence of zero variance for one group, and non-normality (in Table 1, *X* is the attribute).

2155-0182/10/$3.00

TABLE 1: Hypothetical data set for three-category UniODA

Class	*X*	Class	*X*	Class	*X*
A	29	B	35	C	5
A	30	B	35	C	42
A	31	B	35	C	43
A	50	B	35	C	50

In this example the mean *X* of classes A, B, and C is exactly equal, so F=0. However, the UniODA model (if $X < 33$ then class = A; if $X > 38.5$ then class = C; else class = B) correctly classified 10 of the 12 data points: overall and mean PAC over all three groups is 83.3%, two-tailed $p<0.05$.

Ordered Class Variable and Continuous and/or Binary Attributes

Among the many types of nonparametric methods in use, Kendall's tau is arguably the least problematic procedure conventionally used to evaluate associations among ordinal (ranked) data.[16] Tau is a *computed* index for evaluating the relationship between *two* ordered variables: collect data, compute tau, and "it is what it is." Ahead we show that multicategory MultiODA can be used to determine criterion weights for two *or more* attributes to generate a summary score which explicitly *maximizes* tau.

Receiver Operator Curve (Signal Detection) Analysis

Bayesian classification methods are commonly used to evaluate the discriminating power of attributes.[17] Such procedures typically aim to maximize the sensitivity, specificity, or some combination of sensitivity and specificity achieved using an attribute. Since ODA models may be derived which explictly maximize sensitivity, specificity, or any weighted composite of sensitivity and specificity, either for individual attributes or for sets of attributes, we call this application "optimal signal detection analysis."

In summary, it is a common practice to employ multiple different statistical methods, each requiring data to satisfy different essential assumptions, to analyze a given sample of data in numerous "different" (actually related) ways. We recommend using a single statistical method to analyze data with one objective function in mind: maximizing accuracy. The utility of this approach will undoubtedly receive increased attention as researchers learn more about the unrivaled generalizability and power of ODA across different data configurations.

Fast MultiODA Solutions

Early research was highly productive, and new applications for UniODA models were discovered routinely as new data structures were considered.[1] As data configurations became increasingly complex, so did ODA models, and researchers began formulating and investigating optimal linear models for designs with a binary class variable and two or more ordinal and/or binary attributes: an optimal analogue of logistic regression or Fisher's discriminant analysis. These multivariable ODA models are called "MultiODA," for short.

Although UniODA problems can easily be solved for enormous samples, MultiODA problems may be computationally intractable for tiny samples, even on the fastest computers. Several procedures affording reductions of an order of magnitude or more in solution time for

2155-0182/10/$3.00

MultiODA problems were recently developed, and analysis is feasible for enormous samples in favorable circumstances. Review of MultiODA here will be brief: so much work has focused on MultiODA models that a review is warranted. Below we review two fast new methods to solve MultiODA problems: MIP45 is a mixed integer formulation, and WARMACK a special-purpose search algorithm. These methods are extended for nonlinear and multicategory MultiODA.

MIP45

The first approach to computing a MultiODA solution that we shall discuss is a mixed integer linear programming formulation called MIP45, in which the discriminant function is normalized so the sum of the absolute values of the coefficients adds to one.[18] This enables one to determine, for each constraint, a lower bound for the value of the problem parameter, *M*. This is in distinction to previous formulations of this problem, where *M* is defined as "a very large number." Since the value of *M* can be kept low for each constraint, the branch-and-bound procedure can fathom branches more quickly than other formulations. Also, fewer branches need to be stored in memory, and computation time is reduced.

We compared computational resources needed to solve a problem in classification of medical residency applicants using MIP45 and a recent formulation that did not limit *M*. The problem had 3 attributes and 49 observations. Running the SAS/OR optimization package on an IBM 3090/600 mainframe computer, MIP45 solved the problem in 48 CPU seconds, versus 268 CPU seconds using the other formulation: MIP45 analyzed 2,896 branches, versus 14,549 branches using the other formulation.

MIP45 can be extended to obtain MultiODA solutions which maximize the weighted number of satisfied inequalities. As for UniODA, this is useful in two different contexts.

First, the weights may represent the return obtained in the correct classification of an observation. For example, consider the problem of predicting whether the price of a stock will rise or fall over a given time horizon, given a series of market indicators and price measurements. If the prediction is for a rise in the stock price, the stock will be purchased. Conversely, if a fall in the price is predicted, the stock will be sold short. The weighted MultiODA solution of this problem would maximize the trading return over the set of observations.

The other context in which the weighted criterion is useful occurs when the number of observations in each class differs. In this case, the weighted MultiODA solution balances the number in each class by maximizing the mean PAC over the two classes.

A useful extension of MIP45 involves fixing the sign of the discriminant coefficients (e.g., in a confirmatory design). In fact, bounds or any linear constraints on the coefficients may be imposed. Yet another type of constraint which can be modeled is any Boolean function of actual or predicted class membership among the observations. One example of this would be forcing certain observations to be classified correctly in the MultiODA solution (if this is feasible). Another example would be forcing observation A to be assigned to a certain class only if observation B is similarly classified.

Finally, a method for reducing the problem size can be applied when multiple observations share identical values for all attributes. In this case, these observations may be aggregated into a single observation, with a weight applied to the objective function. This procedure is especially useful with binary attributes: we solved binary MultiODA problems having five attributes and one *million* observations in less than ten CPU seconds on an IBM 3090/600.

WARMACK

A second approach to obtaining fast solutions to MultiODA problems involves our adaptation of a fast search algorithm initially developed by Warmack and Gonzalez (hence

2155-0182/10/$3.00

the origin of the name we use to refer to the method).[19] With this method we obtained a reduction of an order of magnitude or more in computation time versus the MIP45 approach.

We conducted Monte Carlo research to investigate the computer resources required by this algorithm as a function of *n*, the number of attributes, and the relative discriminability of the data. Problems having 2 attributes and 700 observations can be solved in less than one CPU minute on an IBM 3090/600. This is also true for problems with 3 attributes and 200 observations, or 4 attributes and 100 observations. Our findings show that the number of attributes exerts greater influence on computation time than *n* or relative discriminability of the data.

Extension of MultiODA to Nonlinear and Multicategory Problems

MultiODA may be extended to a large class of nonlinear separating surfaces. This is accomplished by defining attributes which are polynomial functions of the original attributes. Any nonlinear function which is linear in the parameters of the original attributes may be modeled in this manner.

It is also possible to solve multicategory problems involving more than two class levels using either MIP45 or WARMACK. There are two ways to accomplish this. If there are *k* class categories, the first method is to determine the ODA solution obtained with *k*-1 separating surfaces in parallel with each other. From a computational standpoint, this is equivalent to adding an extra attribute for each additional class.

The second method involves the determination of *k* different discriminant functions: an observation is assigned to the class for which the maximum value is obtained over these functions. If there are *p* original attributes, this is equivalent to a MultiODA problem with *p* times *k* attributes.

In conclusion, MIP45 and WARMACK make feasible the solution of much larger MultiODA problems than have been possible to solve previously, particularly for binary problems. Optimal analogues to conventional statistical methods are now available to researchers. However, ODA is far more than simply an optimal analogue to conventional statistics.

Special-Purpose ODA Models

The flexibility of the ODA methodology lends itself to special-purpose classification applications for which there are no alternative conventional statistical procedures. Indeed, the number of different ODA models that may be created is limitless, due to the inherently infinite number of possible unique classification applications. Nevertheless, below we describe some specialized ODA models that should be of great utility across a variety of applications.

Minimizing the Number of Terms in a MultiODA Solution

When performing an analysis, it is desirable to obtain a solution with as few terms as possible, in light of the principle of parsimony. This can be achieved in the context of the MIP45 formulation: an upper bound is set on the number of misclassifications, and the number of attributes used in the solution is minimized. This results in a more parsimonious model, with a corresponding increase in statistical power.

Optimal Attribute Subsets

A related problem is the determination of an optimal subset of attributes with exactly *k* members. This also is an extension of MIP45. This procedure is useful when the ratio of number of attributes to number of observations is too high to yield a meaningful model, or when redundant (multicollinear) attributes are present.

For example, we used this procedure to discriminate 15 Type A from 15 Type B (class variable) undergraduates using a subset of 20 items (attributes) from the Bem Sex-Role Inventory. With *k* specified at 2 attributes, MultiODA

2155-0182/10/$3.00

identified a single solution that achieved 93.3% PAC; with *k* specified at 3 attributes MultiODA identified a single solution with 100% PAC. These problems required 91.9 and 73.9 CPU seconds to solve on an IBM 3090/600 computer running SAS/OR. When the attributes selected by MultiODA were evaluated using logistic regression analysis, 90% PAC was achieved for both the 2- and 3-attribute models. The best 2-attribute model identified using stepwise logistic regression achieved 90% PAC, and the best 3-attribute model achieved 93.3% PAC.

Integer-Valued Coefficients

UniODA may be used to solve MultiODA problems in which the model weights for the attributes (the discriminant coefficients) are constrained to take on a small set of values. For example, in a problem having *p* attributes, the discriminant coefficients restricted to the values 0, 1, or -1, and the threshold coefficient unconstrained, all optimal solutions may be found by solving $3^p/2$ UniODAs. In general, for *k* possible coefficient values and *p* attributes, $k^p/2$ UniODAs are solved. If *k* and *p* are relatively small, then few computational problems arise due to the fast speed of UniODA. An additional benefit of this analysis is that optimal attribute subsets of every size are evaluated. We solved a problem with 3 coefficient values, 8 attributes, and 900 observations in 716 CPU seconds on a 33Mhz 386 microcomputer.[15]

Optimal Selection of Observation Subsets with Unknown Class Membership

In some problems, observations are available for which class membership is unknown. Typically, exactly *k* of these observations are to be acted upon in some manner. The initial phase of the MultiODA approach to this problem involves partitioning observations into two sets: the decision set, consisting of observations with unknown class membership, and the evaluation set, consisting of observations with known class membership.

To illustrate this, consider the problem of selecting *k* job applicants from a pool of applicants. The attributes may reflect measures of previous employment experience and skills required to perform the job task. The evaluation set is comprised of previously hired individuals who have been measured on these attributes. Each individual in the evaluation set is weighted by a performance index, in this case a measure of job performance. The decision set is comprised of the pool of job applicants, *k* of whom are to be selected for employment, and all of whom have been measured on the attributes. MultiODA identifies a solution which maximizes the weighted number of inequalities in the evaluation set, such that exactly *k* inequalities in the decision set are satisfied.

Or, consider the problem of selecting prisoners to be released under a court mandate which requires that exactly *k* must be released, due to overcrowding. Here the decision set is the current population of prisoners, and the evaluation set are those prisoners who previously were released. The performance index, which is to be minimized, is a measure of mayhem produced by the previously released prisoners.

Other interesting applications of this method lie in the areas of market research, investment selection, and pattern recognition.

Ordered Class Variables

Another fruitful area of investigation relates to the use of MultiODA in analysis of data which have been sorted into ordered (ranked) categories. MultiODA is used to maximize the goodness of fit between the actual and predicted category assignments. Kendall's tau is a similarity index widely used for comparison of two ranked sequences, and is proportional to the number of satisfied inequalities between paired observations. Thus, MultiODA finds a linear discriminant function which *maximizes* the value of Kendall's tau. It is worthwhile to

2155-0182/10/$3.00

note that this situation differs from the multi-category case in that the latter corresponds to the analysis of *unordered* categories.

Optimal Nonparametric Linear Multiple Regression

A distribution-free approach to multiple linear regression is available using the Kendall's tau procedure. Initially observations are ranked according to their values on the dependent measure. MultiODA is then used to find the optimal predicted rank sequence. As a final step, an inequality-constrained multiple linear regression problem is solved for each optimal rank sequence. This quadratic program uses sum-of-squared-error as the objective function, and the inequalities corresponding to the paired observations as constraints. The linear model produced by this procedure is the model with the highest R^2 for which the value of Kendall's tau is the maximum achievable overall. If multiple optimal sequences exist, the solution with the highest R^2 is selected. We have solved such a problem with 3 independent variables and 22 observations in 49 CPU seconds on a 50 MHz 486 microcomputer.

Optimal Templates

Another interesting application of Multi-ODA lies in the design of optimal templates. To illustrate this, imagine an individual is given a list of questions and set of possible responses for each question, one of which is to be selected as the individual's answer to that question. Each question is answered by filling in a circle (e.g., on an "IBM answer form") corresponding to a selected answer. The class membership of each individual is known. The objective of this MultiODA procedure is to produce a template, that is, a series of holes on an opaque sheet, so that overlaying the template on the answer sheet and counting the number of filled-in holes produces a discriminant score for the individual. This score is compared to the cutpoint obtained by MultiODA in order to assign class membership to individuals. This assignment minimizes the number of classification errors.

This problem was formulated as a pure integer program. As an example, consider the application of creating a template for personnel selection purposes. A 38-item questionnaire, with each item answered as true or false, was completed by 107 employees of a corporation, 70 of whom were known desirable workers, and 37 of whom were known undesirable workers. MultiODA identified a template which resulted in 74.8% PAC, requiring 26 CPU minutes on an IBM 3090/600 running SAS/OR.

MultiODA with Boolean Attributes

The ODA approach of minimum error may also be applied to classification problems with purely logical attributes. In this case, the decision rule involved in the assignment of an observation to a class is a Boolean function of logical attributes which have been measured for that observation. We wish to find a Boolean function with at most *k* terms which minimizes the number of misclassifications. Alternatively, we may look for a function with at most *k* misclassifications which minimizes the number of logical terms. These problems can be formulated as integer programs, or solved in crude brute force manner via exhaustive enumeration.

Consider the following application as an example of this procedure. A pair of emergency physicians independently diagnosed 51 patients with hip trauma for bony abnormality. Each physician rated each patient as abnormal or normal based a measure of sound conduction, and also based on physical inspection. Presence of bony abnormality (the class variable) was independently determined radiographically. A Boolean MultiODA identified a single optimal solution that achieved 96% overall PAC. The optimal decision rule was: if either physician rates either attribute as abnormal, then classify the observation as abnormal; else classify the observation as normal.

2155-0182/10/$3.00

Classification Tree Analysis

Hierarchically optimal classification tree analysis, or CTA, is an algorithm which chains UniODA analyses together so as to stratify the sample in a manner that explicitly maximizes ESS.[20] As for MultiODA, discussion of CTA lies outside the domain of this manuscript: sufficient work using CTA has accumulated so that a comprehensive review is warranted.

Summary

Research described herein, indeed the sum total of all of the world's knowledge in this field to date, merely scratches the surface of what ODA entails, what ODA offers. Although we can only imagine what we must be missing, it is clear to see that ODA is a powerful new paradigm in the statistical analysis of data. It is intuitively appealing, in the mathematical modeling of any process, that the model should make as few mistakes as possible. This is the essence of the ODA approach. Its fruitfulness, particularly in its application to the analysis of problems previously unanalyzable, is an indication of its value as a general-purpose problem-solving tool. Because ODA is inherently distribution- and metric-free, it avoids the necessity of making distributional assumptions required by conventional parametric methods. In ODA, powerful modeling capabilities of mathematical programming are joined with the inferential capabilities of statistics. Furthermore, one may combine different ODA methods so that every problem can be formulated in terms of its own unique characteristics. It thus seems appropriate to postulate that, in the area of optimal statistics, the best surely is yet to come.

References

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

[2]Yarnold PR, Soltysik RC (1991). Theoretical distributions of optima for univariate discrimination of random data. *Decision Sciences, 22*, 739-752.

[3]Soltysik RC, Yarnold PR (1994). Univariable optimal discriminant analysis: one-tailed hypotheses. *Educational and Psychological Measurement, 54*, 646-653.

[4]Carmony L, Yarnold PR, Naeymi-Rad F (1998). One-tailed Type I error rates for balanced two-category UniODA with a random ordered attribute. *Annals of Operations Research, 74*, 223-238.

[5]Bradley JV (1968). *Distribution-free statistical tests*. Englewood Cliffs NJ, Prentice-Hall.

[6]Grimm LG, Yarnold PR. (Eds.) (1995). *Reading and understanding multivariate statistics*. Washington DC, APA Books.

[7]Grimm LG, Yarnold PR. (Eds.) (2000). *Reading and understanding more multivariate statistics*. Washington DC, APA Books.

[8]Yarnold JK (1970). The minimum expectation of chi-square goodness-of-fit tests and the accuracy of approximations for the null distribution. *Journal of the American Statistical Association, 65*, 864-886.

[9]Yarnold PR, Soltysik RC, McCormick WC, Burns R, Lin EHB, Bush T, Martin GJ (1995). Application of multivariable optimal discriminant analysis in general internal medicine. *Journal of General Internal Medicine, 10*, 601-606.

[10]Yarnold PR, Soltysik RC (1991). Refining two-group multivariable classification models using univariate optimal discriminant analysis. *Decision Sciences, 22*, 1158-1164.

[11]Yarnold PR, Hart LA, Soltysik RC (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant

2155-0182/10/$3.00

analyses. *Educational and Psychological Measurement*, *54*, 73-85.

[12]Rubin PA (1990). Heuristic solution procedures for a mixed-integer programming discriminant model. *Mangerial and Decision Economics*, *11*, 255-266.

[13]Stam A, Joachimsthaler EA (1990). A comparison of a robust mixed-integer approach to existing methods for establishing classification rules for the discriminant problem. *European Journal of Operational Research*, *46*, 113-122.

[14]Yarnold PR, Soltysik RC, Martin GJ (1994). Heart rate variability and susceptibility for sudden cardiac death: An example of multivariable optimal discriminant analysis. *Statistics in Medicine*, *13*, 1015-1021.

[15]Yarnold PR, Soltysik RC, Lefevre F, Martin GJ (1998). Predicting in-hospital mortality of patients receiving cardiopulmonary resuscitation: Unit-weighted MultiODA for binary data. *Statistics in Medicine*, *17*, 2405-2414.

[16]Reynolds HT (1977). *The analysis of cross-classifications*. New York NY: Free Press.

[17]Kraemer HC (1992). *Evaluating medical tests: objective and quantitative guidelines*. Newbury Park CA, Sage.

[18]Soltysik RC, Yarnold PR (2010). Two-group MultiODA: a mixed-integer programming solution with bounded *M*. *Optimal Data Analysis*, *1*, 30-37.

[19]Soltysik RC, Yarnold PR (1994). The Warmack-Gonzalez algorithm for linear two-category multivariable optimal discriminant analysis. *Computers and Operations Research*, *21*, 735-745.

[20]Yarnold PR, Soltysik RC, Bennett CL (1997). Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine*, *16*, 1451-1463.

Author Notes

Address correspondence to the authors at: Optimal Data Analysis, 1220 Rosecrans Street, Suite 330, San Diego, CA 92106. Send Email to: Journal@OptimalDataAnalysis.com.