BestClass: A Robust Nonparametric Classification Package
A. Pedro Duarte Silva, Terry College of Business, The University of Georgia, Athens, GA
Antonie Starn, Terry College of Business, The University of Georgia, Athens, GA

ABSTRACT

In this paper, we introduce BestClass, a set of mainframe-based SAS® macros designed for solving two-group classification problems using a class of recently developed nonparametric classification methods. The criteria used to estimate the classification function are based on either minimizing a function of the absolute deviations from the surface which separates the groups, or directly minimizing a function of the number of misclassified observations in the training sample. The solution techniques used by BestClass to estimate the classification rule utilize the mathematical programming routines of the SAS/OR® software (SAS Institute Inc. 1989a).

Recently, a number of research studies have reported that under certain data conditions this class of classification methods can provide more accurate classification results than existing methods, such as Fisher's linear discriminant function and logistic regression. However, these robust classification methods have not yet been implemented in the major statistical packages, and are hence beyond the reach of those statistical analysts who are unfamiliar with mathematical programming techniques. Thus, we believe that BestClass contributes significantly to the field of nonparametric classification analysis, in that it provides the statistical community with convenient access to this recently developed class of methods. BestClass is available from the authors.

INTRODUCTION

The problem of assigning (classifying) entities (observations) to exactly one of several well-defined mutually exclusive groups or classes, based on their characteristics on a set of relevant attributes, is important in almost any field of applied science. Many different approaches have been proposed for solving the classification problem. Denote the vector of attributes by \( \mathbf{x} \), and indicate membership in group \( j \) by \( G_j \). The most widely used classification approach is to first estimate the probability density functions or the probability functions \( p(\mathbf{x}|G_j) \) of \( \mathbf{x} \), conditioned on membership in group \( G_j \), and then derive the classification rule which minimizes either the probability of misclassification or the expected misclassification cost.

Another approach is to directly estimate the posterior probabilities \( p(G_j|x) \) of group membership given the observed attribute vector, and use a classification rule that weighs these probabilities by the appropriate misclassification costs. A third approach is to pre-specify a particular form of classification function, and determine the parameter values or coefficients of this functional form that optimize some measure of discrimination or classification accuracy in the training sample.

The origins of the latter approach can be traced back to Fisher's seminal paper (Fisher 1936), in which he introduces the classification rule known as Fisher's Linear Discriminant Function (LDF), derived as the linear function of the attributes that maximizes the ratio between among-group squared distances and within-group variances. The population LDF has been shown to be the optimal classification rule if the attributes in each group follow a multivariate normal distribution and the variance-covariances are equal across groups (Wald 1944; Welch 1939). If the attribute distributions are multivariate normal, but the variance-covariances are heterogeneous across groups, the population quadratic discriminant rule (QDF) (Smith 1947) is optimal. The sample-based LDF has been found to provide relatively robust classification results for problems with attributes that do not follow multivariate normal distributions, but do have similar variance-covariance matrices across groups (Lachenbruch, Sneeringer and Revo 1973). Likewise, the sample-based QDF is robust with respect to deviations from normality in the case of unequal variance-covariances across groups, as long as the sample size is reasonably large in comparison with the number of attributes (say, at least 50 observations per group, with 2 or 3 attributes).

However, research has shown that the LDF and QDF may not yield accurate classification results if the characteristics of the attribute distributions significantly deviate from
the assumptions for which the LDF and QDF are known to be optimal, for instance for highly skewed and/or heavy-tailed distributions (Fatti, Hawkins and Raath 1982; Lachenbruch, Sheeringer and Revo 1973).

As non-normal data conditions occur frequently in practice, it is of interest to explore distribution-free (nonparametric) classification methods that optimize the classification in the training sample with respect to some "desirable" criterion. The appropriate optimization criterion depends strongly on the nature of the particular data set to be analyzed. For instance, criteria that — like the Fisher criterion — use distances based on the $L_2$ norm tend to weigh the largest distances more heavily. Hence, extreme observations in the training sample, i.e., observations that are located much further from the mean or median than the majority of observations, can heavily influence the coefficient estimates of the classification function. Examples of situations where extreme observations occur include the case where the attribute distributions are contaminated by outliers, data sets with highly skewed distributions, and distributions with heavy tails.

The training sample is the sample of observations with known group membership that are used to estimate the classification rule. The training sample is also known as the development sample. It is well-known that the classification performance in the training sample of most classification rules is positively biased. Therefore, once a classification rule has been estimated, its classification accuracy is often evaluated using holdout or validation samples (with known group membership), which consist of observations that have not been used to estimate the classification rule. The estimated classification rule is then used to predict the group membership of observations for which the true group membership is unknown.

For the two-group classification problem, some authors have proposed classification functions that optimize robust accuracy criteria in the training sample (Mangasarian 1965; Smith 1968; Liittschwager and Wang 1978; Freed and Glover 1981; Baigier and Hill 1982; Glover, Keene and Duea 1988; Stam and Joachimsthaler 1989; Duarte Silva and Stam 1994a). The classification performance of these methods is mixed, but nevertheless somewhat promising for non-normal data conditions (Joachimsthaler and Stam 1990). Although many different criteria have been proposed, the most important ones can be divided into two major categories, A) criteria based on some function of misclassification cost and the absolute values of the deviations from the surface which separates the two groups, i.e., distances based on the $L_1$ norm, or B) criteria based on the number of misclassified cases, or total misclassification cost in the training sample.

For two-group problems, criteria of type A can be optimized by formulating and solving a linear program, while criteria of type B can be solved using mixed-integer programming methods. Type A criteria use the sample information more fully than criteria of type B. However, for a wide class of classification functions, including all polynomial functions, optimizing a criterion of type B yields a function that asymptotically minimizes the probability of misclassification or the expected misclassification cost among all functions of the same form, as the size of the training sample tends to infinity (Glick 1976). For an overview of the most important concepts, methods and classification results associated with the type A and B criteria, we refer you to Joachimsthaler and Stam (1990) and Erenguc and Koehler (1990).

Due to limitations of current computer technology (hardware and software), procedures which seek to solve type B criteria to optimality (such as mixed-integer programming procedures) can handle only small and medium-size problems (with up to 150—200 observations and 4—6 attributes). The reason is that type B problems are NP-hard. Several researchers have recently developed special-purpose heuristic computer codes that provide close approximations of the optimal type B solution for larger size data sets. We will not concern ourselves with these special-purpose codes, but limit ourselves to the mixed-integer programming subroutines available within the SAS/OR System, which solve to optimality. However, we stress once more that these SAS/OR-based techniques cannot be used to solve large problems for type B criteria.

GENERAL OVERVIEW OF BESTCLASS

In this paper, we describe BestClass, a software package that implements the most widely used two-group classification methods with type A and B criteria in the SAS System (SAS Institute Inc. 1989a, 1989b, 1989c, 1990). BestClass is designed for analyzing two-group
classification problems only. BestClass is almost entirely written in the SAS macro language (SAS Institute Inc. 1990), and uses the SAS/OR system (1989a) to solve the relevant mathematical programming (MP) models. The few files not written in the SAS macro language are system-dependent, and provide the interface with several host operating systems. At the time of this writing, interfaces are available for the VMS-TSO and CMS operating systems.

BestClass provides the following major features:

- BestClass can be used in two operating modes, interactive and batch.
- BestClass facilitates an analysis based on three types of classification function: linear, quadratic, and quadratic without cross-products.
- BestClass offers a choice between several nonparametric accuracy criteria. The nature of these criteria and classification rules is described in Duarte Silva and Starn (1994b).
- BestClass allows for assigning individual costs to each observation.
- BestClass can accept input and direct output through either text files or the SAS data set format.
- BestClass allows for the retrieval of previously fitted classification functions, which can subsequently be used to classify new data sets.
- BestClass offers an easy interface with other SAS programs, and can easily be embedded in larger SAS programs.

BestClass is implemented as a series of SAS macros. The %control macro controls the flow of the program under interactive mode, and %bestc controls the flow under batch mode. BestClass requires at least one input data set, which is either the “training sample” data set, which contains a sample of observations that will be used to fit the classification function, or a “current sample” data set, consisting of observations that are to be classified according to some already existing classification function, and (3) the parameter data set, which contains the parameter values needed for a BestClass analysis.

INTERACTIVE MODE

In interactive mode, BestClass is activated by calling an executable file (i.e., a CLIST under TSO, or a REXX programming file under CMS), which first allocates all necessary source files and then invokes the SAS system and the macro %control. BestClass prompts you for the name and location of the input data set (either in text file or SAS data set format), and for the number of attributes per observation. Once you have provided this information, the main menu of BestClass will be displayed. This menu offers the following options:

D: Define data sets.

This option allows you to define a SAS data set of one of the following types: (1) the training sample data set, which contains a sample of observations that will be used to fit the classification function, (2) the current sample data set, consisting of observations that are to be classified according to some already existing classification function, and (3) the parameter data set, which contains the parameter values needed for a BestClass analysis.

E: Create new data sets from External files.

Using this option, you can convert text files to SAS data set format. In order to make the converted data set available to BestClass, you should always follow this option by invoking the Define Data Sets option.

F: Choose classification function Form.

This option enables you to choose between three types of classification function: linear, quadratic, and quadratic without cross-products. A linear functional form is recommended for problems where differences across groups in location are more pronounced than differences in dispersion. A quadratic form is appropriate in the case of problems for which differences in dispersion are more pronounced than differences in location. Cross-products should be included in the quadratic form if the attributes are correlated, while a quadratic form without cross-products is preferred if the attributes are uncorrelated.

I: View the Individual classification results.

This option presents the individual classification results for the most recently classified data set. This data set can either be the training sample or the current sample. The output displayed by this option reports 1) the
observation number, 2) the absolute distance from the surface that separates the groups, 3) the original group membership (if known), and 4) the predicted group membership, for each observation.

L: List external file.
   This option allows you to view text files, without leaving BestClass.

M: Choose Model.
   This option allows you to choose between a number of different mathematical programming-based models, each implementing a different accuracy criterion. These models are detailed in Duarte Silva and Stam (1994b). For a discussion of these models and their properties, we refer you to the review articles by Joachimsthaler and Stam (1990) and Erenguc and Koehler (1990).

N: Choose Normalization option.
   The fitted classification functions are unique up to a scaling constant. A unique function can be achieved by normalization. In fact, for several models a normalization of the data is necessary to ensure that a meaningful (non-trivial) classification rule will result from the analysis. A number of different normalization schemes have been proposed in the literature (see, e.g., Freed and Glover 1986, Mahmood and Lawrence 1987, Koehler 1989, Glover 1990, and Ragsdale and Stam 1991). The Choose Normalization option allows you to choose between several of the most important normalization schemes, as described in detail in Duarte Silva and Stam (1994b).

O: Switch to Operating system shell.
   This option enables you to switch to the operating system shell without leaving BestClass. The way to switch back to BestClass is operating system-dependent.

P: Change BestClass Parameters.
   A list and description of BestClass parameters is provided in Duarte Silva and Stam (1994b).

R: Run model.
   This option invokes the PROC Ip of SAS/OR to solve the mathematical programming formulation selected using the Choose Model option.

S: Save output.
   This option saves the results of the BestClass analysis. The following information can be saved, in any combination: 1) the coefficients of the classification function, 2) summary statistics of the classification results for the training sample and/or the current sample, 3) classification results for the individual training sample and/or current sample observations.

   Any of the above output can be directed to new text files or SAS data sets. Alternatively, if preferred the output can be appended to existing text files or SAS data sets. These options are controlled by a set of interactive menus.

AC: Apply Classification function to current data set.

MS: Manage/list SAS data sets and libraries.
   This option enables you to access basic SAS library management utilities, such as utility programs to create, allocate, delete, merge and list SAS libraries and data sets, without leaving BestClass.

RF: Retrieve classification Function from SAS data set.
   This option retrieves a classification function previously saved as a SAS data set, and makes it available to classify the current data set.

X: EXit BestClass.
   This option terminates BestClass, and exits you to the operating system.

**BATCH MODE**

In batch mode, you can activate BestClass by calling the macro %bestc from a SAS program. When invoking %bestc, you can optionally specify parameters that, among others, control the origin and form of the input, the amount and destination of the output, the accuracy measure to be used, and the form of the classification function.

Except for those options that are relevant in interactive mode only (such as the listing of external files, switching to the operating system, and the interactive management of SAS libraries), you can also access all options available in the interactive mode of BestClass in batch mode, by selecting an appropriate
combination of parameter values.

External calls of BestClass to or from a non-SAS System environment are best accommodated through text files. However, SAS data sets are more efficient if BestClass communicates with other programs within the SAS System environment. A complete list of %bestc options, their use and default values is given in Duarte Silva and Stam (1994b).

CONCLUSIONS

In this paper, we introduce BestClass, a set of SAS-based macros which facilitate the use of a recently developed class of nonparametric classification methods. BestClass requires the mainframe version of the SAS System statistical software, including the SAS/OR System. BestClass does not require any additional software and, as it is menu-driven, the user does not need to be familiar with mathematical programming techniques. Hence, BestClass offers the statistical analyst a convenient tool for analyzing classification problems with certain data conditions for which parametric methods do not classify well.

To date, BestClass has been implemented in the VMS-TSO and CMS operating systems. Useful extensions would be to extend the platform to other operating systems and to include other relevant nonparametric classification methods. Recently, several special-purpose and/or LINDO®-based (Schrage 1991) microcomputer software packages for classification methods with type A and B criteria have been developed (Banks and Abad 1991; Lam and Choo 1991; Soltysik and Yarnold 1993; Stam and Ungar 1994). An extension of BestClass to the SAS System for PCs would be useful as well. We are hopeful that eventually the class of nonparametric classification methods that can be analyzed using BestClass will be integrated in the SAS System and other major statistical software products.

REFERENCES


Koehler, G. J. (1989), “Characterizations of
Unacceptable Solutions in LP Discriminant Analysis,” *Decision Sciences*, 20, 239–257.


**ACKNOWLEDGMENTS**

We would like to thank Francis J. Kelley of The University of Georgia Computing and Network Services for his time and effort in solving many problems during the development of BestClass.

The first author was supported in part by the Junta Nacional de Investigação Científica e Tecnológica, Programa Ciência, Lisbon, Portugal, and the Universidade Católica Portuguesa, Centro Regional do Porto, Porto, Portugal.

The second author was supported in part by a Terry and Selig research grant from the Terry College of Business, The University of Georgia, and gratefully acknowledges the support provided by the International Institute for Applied Systems Analysis in Laxenburg, Austria.

SAS, SAS/OR and SAS/STAT are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are registered trademarks of their respective companies.