The Relative Performance of Corporate Bond Funds 1991-2003: A Perspective from Pattern Recognition

James Moore

Indiana University – Purdue University Fort Wayne

This study employs a rule induction expert system to search for patterns among characteristics of individual mutual bond funds that could predispose the fund to a specific measure of long-term performance relative to its peer index composite. Individual bond funds are categorized as either "Outperforming", "Matching", "Underperforming" a benchmark index, or as "No-Longer-Traded". The system's training period is the nine-year interval from year-end 1991 through year-end 2000. The rule induction results are compared with those of multiple discriminant analysis, with the induced rule structure generating a considerably better classification "hit-rate" than multiple discriminant analysis in this application. The induced pattern is then used to forecast individual bond fund performance beyond the training cases and beyond the original training period, through year-end 2003. The pattern recognition capability of rule induction detects a substantive role for short-term persistence, portfolio maturity, the percentage change in the fund's asset holdings, and portfolio yield. These factors are operational surrogates for recent historical performance, the activity level of actively managed funds, and asset portfolio characteristics. All such critical factors are readily accessible to the ordinary mutual fund investor.

Mutual funds are investment companies that pool the resources of numerous investors so as to invest in diversified holdings of stocks, bonds, real estate, precious metals, and/or other securities that meet the fund's investment objectives and criteria. Collectively, mutual funds have grown to become prominent components in the long-term investment strategies of institutional investors, company pension plans, 401(k) plans, as well as individual investors. Over the last two decades of the recent millennium, the collective assets of mutual funds have exploded nearly 50-fold to a record 7.5 trillion dollars, held in more than 8000 funds. One major contributor to this explosive growth has been the technology revolution that began in the early 1980's; this provided funds with a cost-effective way to sell shares directly to investors and to handle the volume of smaller-sized accounts. Certainly, not all mutual funds are created equal; otherwise there would be no need for so many distinct funds.

The composite performance of a particular fund, over a specific period of time, can be assessed via comparison to a benchmark index. While Patel et al., (1992) found that investors are typically prepared to respond to recent measures of relative performance, actually predicting such performance for individual funds is clearly an exercise of ever growing impact and importance, not to mention complexity. While many investors attempt to predict just the over-performers, there may be comparable merit in identifying, and consequently, avoiding the underperformers as well. What are the attributes and practices of corporate bond funds that outperform (underperform) the bond market over a long-term horizon? This is the issue under examination by this study. Moreover, can long-run successful (and unsuccessful) corporate bond mutual funds be accurately identified by recurring patterns among their characteristics?

This study will use both rule induction, as a form of machine learning, and multiple discriminant analysis to look for patterns among corporate bond mutual fund characteristics and operating practices that will classify the individual fund's performance, relative to that of a bond market benchmark, over a nine year period (12/1991 to 12/2000). The learned pattern from rule induction will then be used to forecast individual bond fund performance beyond the training cases and beyond the original training period, through year-end 2003.

Literature Review

Numerous performance measures for mutual funds have been developed and discussed in the literature. Early works by Lintner (1965), Treynor (1965), Sharpe (1966) and Jensen (1968) were grounded in the capital asset pricing model (CAPM) and employed some measure of risk-adjusted return. Treynor's and Sharpe's were composite return-to-risk ratios. Jensen's Alpha was the difference between ex-post and CAPM returns; a positive (negative) alpha was an indicator of above (below) average performance. Since mutual funds may specialize in holding only corporate bonds, or municipal bonds, or small cap stocks, or international stocks, etc., performance measures have been developed that use indices tied to the differential risk exposures of these pools (Elton, Gruber & Blake, 1996a, b; Grinblatt & Titman, 1993; Sharpe, 1988). Bers and Madura (2000) offer further evidence for risk-adjusted performance persistence of domestic closed-end mutual funds being attributable to the unique characteristics of such funds.

Market practitioners argue that selecting mutual funds for the near term on the basis of past performance simply does not work (Bogle, 1999). The academic literature remains unresolved on the issue of performance persistence. Hendricks et al., (1993) examined open-ended, no-load funds over the 1978-1988 period. They documented a persistence effect for both the relatively superior and inferior performing funds in the near term (one to eight quarters and peaking near the fourth quarter). Consequently, a strategy of selecting every quarter, the top performers based on the last four quarters can outperform the average mutual fund but does not necessarily outperform a broad market index. Through their subsample analysis, they also conclude that the issue of survivorship bias was not important for studying persistence in mutual fund performance. This finding conflicts with studies which conclude that survivorship bias can contaminate studies of performance persistence among mutual funds (Brown et al., 1992; Brown & Goetzmann, 1995). Carhart (1997) explains short-term persistence in equity mutual funds in terms of common factor and cost-based issues after controlling for survivorship bias. He finds that investment costs of expense ratios, transactions costs and load fees all have a direct and negative impact on fund performance.

While the benefits of diversification and liquidity are justifiably attributed to mutual funds due to the breadth and active secondary market of their holdings, the issue of actual value added due to professional management remains clouded at best. Actively managed mutual funds will seek to time the market as they alter the composition of the fund's portfolio. Doing so requires additional liquidity through cash reserves than would otherwise be the case. Clearly, such reserves impose an opportunity cost for earnings. The continued coexistence of both managed and indexed funds must be evaluated in the context of additional expenses vs. additional return. Virtually all of the difference in the return to participants between managed funds and index funds is attributable to the higher costs imposed by actively managed funds. It is inherently more expensive to make active investment decisions than to follow a preset portfolio match rule. These management costs include load fees, transaction costs associated with high asset turnover, expense ratios, and the opportunity cost of holding cash reserves. Nevertheless, the security selection ability, or inability, of the mutual fund managers is frequently examined as a potential influence on the performance of their respective fund. Umamaheswar (2001) examines the performance of the managers of mutual funds and finds that they do not appear to possess any special information about the future direction of the market, and this is reflected in their security selection performance. Carhart (1997) finds the one-year momentum effect to be of more importance in explaining persistence in mutual fund performance than in the fund's component holdings. His results do not support the presence of particular skills or information possessed by mutual fund portfolio managers which would result in superior security selection. Fortin et al., (1999) examines the relationship between mutual fund performance and fund manager tenure. Their sample of 800 equities and bonds spanning ten years reveals no relationship between manager tenure and fund performance. Consequently, they conclude that investors should consider the fundamental investment parameters such as consistency of return, expense ratio, asset turnover, cash reserve and fund size.

Mutual fund performance is typically judged in relative terms. Morey and Morey (1999) endogenously determine a custom-tailored benchmark portfolio against which each mutual fund's performance is compared. In contrast, this study will examine how well a particular fund has performed over a specific period of time via comparison to an exogenous benchmark index as it examines the attributes and practices of corporate bond funds that outperform (underperform) the bond market over a long-term horizon. Moreover, can long-run successful (and unsuccessful) corporate bond funds be accurately identified by recurring patterns among their characteristics?

Classification Methodologies

Classification situations arise in many business environments including credit scoring, default prediction, bond ratings, and performance assessment among others. The solution to such classification problems is a discriminant function that maps the component variable space into an outcome set. Since the pioneering work of Fisher (1936), numerous stochastic techniques have been used for classification purposes including canonical correlation, discriminant analysis, and logistic regression. Such approaches typically impose rather restrictive conditions on the underlying population distribution and data measurement scales, such as assumptions about normality and equal dispersion. Unfortunately, violations of these assumptions, due to bounded variables or categorical variables, are not uncommon, and thus, potentially compromise the application of the stochastic tool (Deakin, 1976; Eisenbeis, 1977). This study will examine and compare the two classification techniques of rule induction and multiple discriminant analysis over the expansionary period of year-end 1991 to year-end 2000.

Pattern Recognition Through Rule Induction

Recently emerging technologies for the purpose of classification through pattern recognition include forms of artificial intelligence, known as machine learning. These techniques are robust in that they do not presuppose any underlying probability distribution or dispersion equality. Machine learning is a data driven approach to extracting expertise from prior cases, under the presumption that future relationships will follow the patterns of past outcomes. Moreover, the underlying assumption is that patterns can be inferred from representative examples of prior behaviors. Machine learning systems use training examples to induce classification heuristics which map sets of input attributes into classification outcomes. These mappings can be accomplished with expert systems that employ rule induction or through artificial neural networks. Rule induction is an automated case-driven method of expert system knowledge acquisition. A major strength of rule induction relative to direct articulation is that experts often find it easier to provide representative cases from prior situations than to actually reconstruct their decision process. Often they can be more confident and hence more deterministic about the "what" than the "how". Messier and Hansen (1988) compared the quality of the results of a rule induction algorithm with those of MDA and various statistical tools. Using financial bankruptcy cases, they found that the induction model outperformed the competing models.

The rule induction routine employed in this study is an optimization procedure based upon Quinlan's (1983 and 1986) Iterative Dichotomizer, ver. 3 (i.e. ID3 induction algorithm). This data-driven induction method examines a set of prior cases and seeks to identify the relevant attributes and patterns among them which have led to the recorded findings; the induction algorithm generates the most parsimonious system of production rules which result in the known outcomes. Thus, ID3 seeks to minimize the number of attributes in the final decision rule, and consequently, find the most efficient path to the conclusion. Its mechanism for discovering a set of classification rules and organizing them into an efficient decision tree is based on a measure of the entropy of each attribute; where the higher the entropy of an attribute, the more uncertainty there is regarding its linkage to a particular outcome value.

As an information theoretic measure, entropy is computed for attribute A_k as:

$$H(C \mid A_k) = \sum_{j=1}^{M_k} p(a_{k,j}) \ * \ [- \sum_{i=1}^{N} p(c_i \mid a_{k,j}) * \log_2 p(c_i \mid a_{k,j})]$$

where:

$H(C \mid A_k)$ = entropy of the classification property of attribute A_k
$p(a_{k,j})$ = probability of value j for attribute k
$p(c_i \mid a_{k,j})$ = probability that the decision choice is i when attribute k has value j
M_k = number of values for attribute k ($j=1$ to M_k)
N = number of decision choices ($i=1$ to N)
K = number of attributes ($k=1$ to K)

Each attribute is examined for the internal split of its values which leads to the largest decrease in entropy. The root node of the decision tree is built around the attribute with the least uncertainty and hence the least measurable entropy. This process is repeated for each node of the tree, with each such node being associated with a specific attribute. Moreover, the algorithm identifies the remaining factor that has the least uncertainty about its association with an outcome value. It then builds a decision junction around such a factor which effectively is most discriminating amongst changes in the final outcome. Subsequent attributes are selected in order of increasing entropy; the iterative nature of these tests continues to form the hierarchy of a tree structure which ultimately has zero remaining entropy and thus correctly classifies all cases in the training set (Quinlan, 1983).

It is apparent from the construct for entropy above, that there is a dimension of Bayesian probability revision, (i.e. conditional updating based upon new information about a value for attribute 'k'), incorporated in the optimization algorithm of rule induction. In fact, rule induction shares a fundamental underpinning with Bayesian decision theory. Central to both rule induction and Bayesian theory is the reverse process of deduction, namely inductive reasoning. Bayes was interested in the inverse of deducing the consequences of specified hypotheses; specifically, he sought to address how one might make inferences from observed sample data about the populations that gave rise to these data i.e. drawing conclusions about hypotheses from observations of consequences; such reasoning from the specific (observations) to the general (hypotheses) is the process of induction. In parallel fashion, rule induction is a data driven approach to extracting patterns from prior cases by reasoning from the specific (examples) to the general (rules) and thus, closely represents the synthesis activities performed by the human brain.

While there are substantive commonalities between rule induction and Bayesian decision theory, there are real differences as well. Bayesian theory is predominantly a prescriptive approach, more than a descriptive one. That is, it presents the principles and methods for making the best decisions under specified conditions. While Bayesian decision theory does not purport to present a description of how actual decisions are made in the real world, rule induction goes beyond prescription and does attempt to offer a possible process that would be consistent with the final outcomes found in the real world cases. While both tools generate visual decision trees, Bayesian theory uses probabilities to compute an expected value for each branch, whereas rule induction uses an interconnected sequence of "If/Then" relationships to seek a deterministic outcome for each terminal branch.

Quinlan's ID3 algorithm seeks to minimize the number of attributes in the final decision rule, and consequently, find the most efficient path to the conclusion while correctly classifying all cases in the training set. This optimization algorithm eliminates redundancies by screening out those factors that are not necessary for the minimal decision tree. By seeking to minimize entropy, this algorithm will minimize the number of discriminating factors needed to reach an outcome. Thus, it operates on the premise that the "ideal" rule is one with as few attributes / factors as possible that will successfully distinguish among the different possible outcomes. Bundy, Silver, and Plummer (1985) compared the major inductive algorithms and found that ID3 was able to learn disjunctive concepts that are more general than those that can be learned by most other algorithms. Braun and Chandler (1987) found that the ID3 algorithm performed better than other induction methods in the development of a production system for aggregate stock market behavior.

Multiple Discriminant Analysis

Another classification technique for situations when the outcome of interest is a truly categorical variable, is multiple discriminant analysis (MDA). Like rule induction, MDA seeks to construct a predictive model of group membership based on observed characteristics of existing cases. Each tool provides both a "profile development" and a "predictive" capability which can be compared across techniques. Linear discriminant analysis can be traced to the pioneering work of Fisher (1936) and has been employed extensively in the financial literature to investigate, among other finance-related applications, bond yields and ratings (Brister, Kennedy & Liu, 1994; Chan & Jegadeesh, 2004; Watson, et al., 1983), commercial paper ratings (Chandy & Duett, 1990), consumer credit (Clark & McDonald, 1992; Malhotra & Malhotra, 2003; corporate bankruptcy (Anandarajan et al., 2001; Moyer, 1977; Piesse & Wood, 1992; Pompe & Bilderbeek, 2005; stock repurchases (Medury, Bower & Srinivasan, 1992), and loan risk (Epley, Liano & Haney, 1996; Kumar & Haynes, 2003; Long, 1976).

Unlike rule induction, MDA requires that all of the predictor variables and the criterion variable must be numeric; consequently, all the non-numeric variables of this model were recoded as integer variables for the use of multiple discriminant analysis. The stochastic MDA is more restrictive than the deterministic rule induction in that it assumes the discriminating variables within each classification group are distributed as multivariate normal, and the variance-covariance matrices across all groups are equal. When these are satisfied, Fischer's linear discriminant function is optimal in the sense of minimizing the rate of misclassification.

The presence of several categorical driver variables in our model clearly violates the MDA assumption of multivariate normality. These restrictive conditions and assumptions have long been investigated in the literature for sensitivity to their relaxation. While Eisenbeis (1977) argued strongly for adherence to the underpinnings of multiple discriminant analysis, others have shown that MDA is a relatively robust technique with respect to selected violations of its underlying assumptions. Following Gilbert (1968) and Lachenbruch (1975), who first demonstrated that MDA is not acutely sensitive to departures from the normality assumption, this study will retain the categorical predictor variables in an effort to provide a common-base comparison to the performance of rule induction, which is designed to work with categorical discriminating variables.

The Bond Fund Model

The risk adjusted rate of return of a fund is the relevant performance indicator for this study. Net Asset Value (NAV) is the per share price of the fund. Reinvested distributions of dividends and capital gains will supplant the change in NAV to give a total (composite) yield that can be compared to the benchmark index. The relative performance of each fund will capture the total advance of an investment in that fund as reflected in the percentage change in an initial investment over a specified period of time. Thus, the composite return on a corporate bond mutual fund includes dividend and capital gain distributions from the securities comprising the fund's portfolio as well as the share price adjustments of the fund itself; all components must be adjusted for share splits.

This study segments the twelve-year period from 12/31/91 to 12/31/03 into three intervals. The pattern recognition model will be built from the nine-year period (year-end 1991 to year-end 2000) based on the performance of each corporate bond mutual fund relative to that of an index fund for corporate bonds. This will enable the classification labeling to be done on the basis of comparison with a broad market benchmark with a similar risk / reward profile. The Lipper Corporate A-Rated Debt Index was selected as the relevant corporate bond benchmark over the horizon of this study. From year-end 1991 through year-end 2000, this index advanced by 79.07 percent. Thus, the individual fund performance metrics, as a measure of composite advance, are compared to this baseline. Once built, the model will be tested on cases in two subsequent intervals; year-end 2000 to year-end 2002, which captured most of the economic contraction, and calendar year 2003, which reflected a struggling attempt at economic recovery. Cases in each interval will be independently subjected to the pattern detected during the earlier training period to assess the model's capabilities.

Mutual funds are discontinued for both positive and negative reasons. We will not impute a positive or negative connotation on such discontinued funds; rather, in order to control for survivorship bias, funds that existed at the beginning of the period under review but did not exist at the end of the period, will be retained in this analysis and labeled as no longer traded. Thus, each mutual fund will be assigned a categorical label of either:

i) **Outperforming** (> +3%) its benchmark index,
ii) **Matching** (+/- 3%) the performance of its benchmark index,
iii) **Underperforming** (< -3%) its benchmark index, or
iv) **No Longer Traded (NLT)**

These four values become the possible outcomes of the pattern recognition process. Following Indro et al., (1999) and Bers and Madura (2000), the current study seeks to forecast performance in terms of fund-specific characteristics as drivers; yet in this model the measure of performance will be a relative measure of the individual bond fund vis-a-vis an appropriate market benchmark. This study will extend the holding period from the one year of Indro et al. to a nine year period (year-end 1991 to year-end 2000) that encompasses elements of both bull and bear markets through this time. The conceptual model relates the above categorical measure of relative performance to descriptive attributes and operating practices of the mutual fund.

The theory of efficient markets contends that all available information relevant to an asset will be reflected in its market value. Since a mutual fund is simply a collection of individual component assets, the current value of the fund should also reflect the available information about the fund and its operating practices. Thus, we will examine the descriptive information about the individual fund, available at the beginning of the period, for clues as to the pattern of its future performance.

Following Fortin et al., (1999), the operational model for bond funds excludes a specific discriminator for the influence of management and instead focuses on such fundamentals as: i) fund size, ii) asset portfolio characteristics, iii) fund expenses, and iv) historical performance. The operational predictor variables are those fund attributes/descriptors believed to have an impact on the multi-year performance outcome; these include the following:

I) FUND SIZE as represented by:
Assets: holdings in millions of dollars
PctChg: percentage change in assets from prior year
These serve as proxies of critical mass and the activity level of actively managed funds.

II) PORTFOLIO CHARACTERISTICS as represented by:
Yield: the income earned in the prior year expressed as a percentage of the fund's year-end net asset value per share,
Mature: the average maturity (years) of the securities in the portfolio weighted according to the market value of those securities.

III) FEES AND EXPENSES as represented by:
LoadPct: percentage sales charge including deferred charges and redemption fees,
ExpRat: expense ratio expressed as a percentage
The existence of load fees and a higher expense ratio are each posited to degrade the fund performance on behalf of the shareholder.

IV) HISTORICAL PERFORMANCE as represented by:
1Retrn: total return for most recent year.
3Retrn: avg annual total return w/ reinvestment of dividends for the most recent 3 years,
5Retrn: avg annual total return w/ reinvestment of dividends for the most recent 5 years,
BWRate: five year historical risk adjusted return relative to S&P 500 as established by Business Week with possible values:
Superior, Very Good, Above-Avg, Average, Below Avg, Poor, Very Poor

The trend of historical performance relative to all other funds is captured in three potential discriminators. The three year period prior to the investment decision is broken into three equal periods of 12 months, as represented by:

1Trend: relative performance in the most recent 12 month period,
2Trend: relative performance in the second most recent 12 month period,
3Trend: relative performance in the third most recent 12 month period,
Possible values for each of the above variables are:
Top Quartile, Second Quartile, Third Quartile, Bottom Quartile

These variables (capturing absolute and relative performance) test the influence of the persistence effect over an immediate history period of variable length from one to five years.

The current study will not seek to resolve the issue of the effect of survivorship bias, but rather will control for its impact by including both cases that did, and others that did not, survive the multi-year period under review. In sum, mutual fund performance is projected to be directly impacted by the absence of load fees, a low expense ratio, portfolio characteristics, and potentially the fund's own historical performance. The above 13 predictor variables include nine continuous numeric and four categorical variables, which, collectively, become the candidate variables for the ID3 rule induction.

The Case-Based Data

Annually each February, *Business Week* publishes its Mutual Fund Scoreboard with vital statistics on over 500 distinct bond funds including corporate, government, municipal, international, and convertible funds. These fund specific data include:

i) returns for the immediate prior one, three, and five year periods when available,
ii) sales charges and expense ratios, and
iii) portfolio data such as total assets, percentage change in assets, portfolio yield and maturity.

This data is precisely the candidate discriminators discussed in the previous section. Clearly, the set of candidate predictors employed here is not exhaustive of those with potential influence on fund performance; in fact, it is a relatively commonplace set of possible drivers, without the use of weighted average portfolio metrics (other than Mature). But this is consistent with the stated underlying intent of the study, which is simply that the model should be built with data that is readily accessible to the ordinary mutual fund investor. Since the actual set of candidate discriminator variables was driven by, and limited to, the set of attributes reported in the annual Mutual Fund Scoreboard issue of *Business Week*, the operational goal is to explore whether data available in the popular press is sufficient to induce a meaningful decision structure.

A fifty percent random sample of the reported corporate bond fund cases was drawn for this study. Following some minimal culling due to incomplete data, we were left with 102 complete cases. These were split randomly into two equally sized subsets by each classification technique, one for development/training of the systems, and the other for contemporary validation.

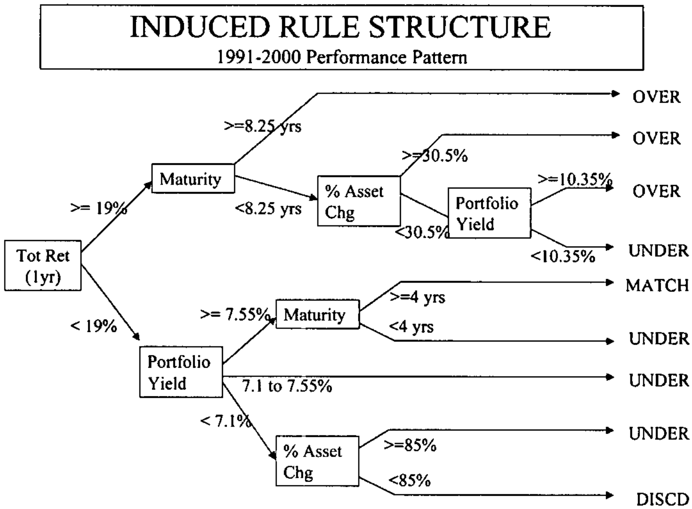
Results

Induced Rule Structure

The output of the ES development stage is the configuration of the rule structure that is compatible with the expertise of the contributing domain experts or, in the case of rule-induction, with the training cases. Such a rule structure seeks to describe the process by which the outcome under examination may be achieved. The induced rule system of Figure 1 below was developed using a training set of 51 cases, all of which are found to be consistent with this resulting rule structure.

The induced results show that a classification pattern was indeed present in the data for the nine-year period under review. The pattern required, at most, four of the thirteen candidate factors in order to correctly classify the cases of the training set. The primary discriminating variable was determined by ID3 to be the historical total return for the most recent year (1Retrn). If this recent return was below 19 percent, then the best performance that could be expected of the firm is to match the composite market performance, with most paths resulting in underperformance. If the one year total return met or exceeded the 19 percent threshold, then the maturity of the fund's holdings (Mature), the fund's percentage change in assets from the prior year (PctChg), and the prior year's ending ratio of earned income to net asset value (Yield) each played a role in determining relative performance. Moreover, when the total return exceeded 19 percent, the typical outcome was overperformance of the fund relative to the benchmark unless the portfolio maturity was short (<8.25 years) and the percentage change in assets was relatively low (<30.5%) and the portfolio yield was relatively low (<10.35%). Thus, it took a unique combination of three adverse events to reverse a favorable prognosis.

Figure 1



The actual 19 percent threshold for the most recent one-year return, as determined by the pattern recognition technology to create the greatest discrimination among final outcome values, must be interpreted in the context of its relevant time frame. Given the data set used here, the total return for the most recent one-year period would refer to that for calendar year 1991. During 1991, the Lipper Corporate A-Rated Debt Index, used as the base reference throughout this study, advanced by 17.47 percent. In that context, the 19 percent established by the pattern recognition algorithm is not extraordinary, but simply in excess of that benchmark. Thus, having an opportunity for a favorable long-term performance prognosis is predicated on beating the appropriate market index in the most recent year. This prominent role at the root node of the decision tree for the historical return over the most recent year supports the case for short-term persistence as argued by Carhart (1997). The secondary roles for portfolio yield and maturity and tertiary role for the percentage change in assets are still indicative of the importance of these attributes as perceived by bond fund investors, particularly in light of the fact that the pattern recognition algorithm did not find a pivotal role for nine of the 13 candidate descriptors. Of notable interest is the observation that at least one candidate variable from each of the conceptual categories of "Fund Size", "Portfolio Characteristics", and "Historical Performance" did emerge as a critical performance discriminator; yet a role for the category of "Fees and Expenses" was not evident, as neither the load fees nor expense ratio was found to be an effective discriminator among performance outcomes over this period.

It is important to stress that for each numeric discriminator, the numeric thresholds for all branches of the induced tree are determined by the ID3 induction algorithm, not by the investigator. Since the rule induction procedure seeks the most efficient path to the final outcome, it did not include all candidate variables in order to explain the fund's performance relative to the relevant index composite. Once the above four variables were built into the rule structure, the remaining candidate variables were not necessary in explaining the relative performance of the funds over this period.

Rule-Induction Validation Test

The rule structure's ability to classify bond funds beyond the training set was tested in multiple ways via the accompanying "consultative" module of the expert system software. In one test, contemporary cases, not used in the development of the rule, were individually subjected to the rule and the system's performance was monitored; a parallel test was conducted with the MDA model for comparison. Later, a second test will use the total set of cases, but extend the induced rule structure, developed for the 1991-2000 expansion period, to each of the 2001-2002 economic "contraction" period and the calendar year 2003 "recovery" period.

In the contemporary environment, the classification performance of rule induction, for the previously unseen cases over the original development period, is summarized in Table 1 below.

Table 1: *Induced Rule Structure Applied to Hold-Out Sample Over 12/31/91 - 12/31/00 Period (entries are number of occurrences)*

Predicted as: Actual:	**Outperforming**	**Matching**	**Underperforming**	**No Longer Traded**	**Correct Pctg**
Outperforming (n=19)	16	2	1	0	16/19 = 84.2%
Matching (n=6)	1	3	2	0	3/6 = 50.0%
Underperforming (n=14)	0	2	11	1	11/14 = 78.6%
No Longer Traded (n=12)	0	2	6	4	4/12 = 33.3%
Total (n=51)	17	9	20	5	34/51 = 66.7%

While the ability to accurately identify those bond funds that will become "No Longer Traded" is certainly not a strength of this model, the diverse reasons for the cessation of trading for a particular fund may make this particular aspect of classification accuracy problematic. In contrast, the expert system's induced rule structure is quite successful in identifying those funds in the two key groups of "Outperforming" (84.2% accuracy) and "Underperforming" (78.6% accuracy). Collectively, the system's rule structure successfully predicted the outcome in 34 of the 51 firms in the contemporaneous hold-out sample for a 66.7% "hit rate" and, thus, an 83.3% (= (51 + 34) / 102) overall success rate for the induced model.

MDA Results

The non-numeric variables of BWRATE, each of the three historical trend variables, and the criterion performance variable were all recoded as integer variables to satisfy the numeric requirement of MDA. The four-group MDA model of this application will generate a set of three discriminant functions; these functions are based on linear combinations of the predictor variables that generate the best discrimination among the groups of the criterion variable. Wilk's lambda and canonical correlation are indicators of the degree of separation achieved by an individual discriminant function. Wilk's lambda can be converted into a χ^2 value for a test of how well the function distinguishes among groups. The χ^2 significance level for the package of all three discriminant functions was .09; for just the second and third functions, the significance level was .57, and for the third function alone, it was .76. Thus, while substantive differences exist among groups, only the first function contributes substantial discriminating power to the model.

Unlike the parsimonious theme driving the rule induction algorithm, which limited the resulting discriminating variables to just those necessary to reach an outcome for the training set of cases, MDA finds a role, albeit frequently inconsequential, for each candidate discriminating variable. The interpretation of the MDA functions describes how fund performance is related, positively or negatively, to the set of attributes.

The outperforming funds have been assigned a {positive/negative} centroid value in the first function; consequently, we view the MDA interpretative statistics in light of the group centroid values. Standard MDA output includes:

i) the standardized discriminant coefficients (Stan), whose relative absolute values reflect the contribution of each predictor variable to the separation of groups while taking into consideration the simultaneous influence of all other independent variables, and

ii) the structure correlations (Corr), which are analogous to factor scores or "loading factors", and relate the discriminant scores of a function with the predictor elements.

When sample sizes are small or the potential for collinearity among discriminating variables is high, the use of structure correlations, which are simple bivariate correlations and thus are not affected by relationships with other variables, is more stable and defensible (Cooley & Lohnes, 1971; Klecka, 1980). Thus, our MDA interpretation will focus on structure correlations rather than standardized coefficients; these are reported in Table 2, where like-signs among structure correlations and centroids indicate a direct relationship and opposite-signs reflect an inverse relationship.

The group centroids, reported in Table 2, identify the degree of separation among the groups achieved by each function. The first MDA function, which distinguishes Underperforming funds (centroid = -2.290) from those outcomes with positive centroids, is dominated by the variable for the performance trend extending 3 years back. The inverse relationship with the Underperforming group's centroid indicates that those funds with a strong 3-year historical track record are less likely to underperform in the subsequent period. This conclusion lends support to a longer version of the "persistence" argument. Variables with lesser roles in the first MDA function include the historical risk-adjusted return relative to S&P 500 (BWRate) and the weighted average maturity of the portfolio (Mature). Their inverse relationships with the Underperforming group's centroid suggest that funds with longer average maturities or higher risk-adjusted relative returns are also less likely to underperform.

The second MDA function achieves maximal group separation between the set of Outperforming funds and those groups with positive centroids for this function, particularly those "No-Longer-Traded". Unfortunately, as a collective, this function did not prove to be significant. A similar conclusion is noted about the significance of the third MDA function. Consequently, neither the second nor the third function are of much interpretative interest.

It is interesting to note the difference in "driver variables" recognized by the two techniques employed. While MDA acknowledged the 3-year historical trend variable, and to a lesser extent, the historical risk-adjusted relative return, and the portfolio maturity, rule induction recognized a pattern involving four fund attributes, with only maturity having any commonality with the MDA results. The theme of the variables selected by the rule induction algorithm is more "current" than is the "historical" theme of the MDA drivers.

Table 2: *Multiple Discriminant Analysis Results: 12/31/91 – 12/31/00*

	Function #1	Function #2	Function #3
Group Centroids:			
Outperforming	.682	- .814	- .190
Matching	.906	.632	.951
Underperforming	- 2.290	.119	.008
No Longer Traded	1.007	1.680	- .972
	Corr	Corr	Corr
Assets	.077	- .402	.007
PctChg	- .074	.049	.449
Yield	.211	- .277	.123
Mature	.329	.335	- .088
LoadPct	.282	- .156	- .015
ExpRat	.184	.299	- .180
1Retrn	.279	- .391	- .048
3Retrn	- .168	.285	.189
5Retrn	- .194	.206	.380
BWRate	.436	.284	.184
1Trend	- .104	.308	- .288
2Trend	- .167	.346	.150
3Trend	.558	- .194	.136
Canonical Correlation	.821	.662	.506
Eigenvalue	2.064	.779	.344
Relative % of Eignevalue	64.8%	24.5%	10.8%
^{2} Significance	.09 for Functions 1- 3	.57 for Functions 2 & 3	.76 for just Function 3

MDA Validation Test

Like rule induction, multiple discriminant analysis extends the profile development capability to include a predictive capacity. Individual cases are classified into the pre-defined performance groupings on the basis of their discriminant score. As was the case with rule induction, reclassification of the same cases used in the identification phase (a.k.a. "training") is vulnerable to a favorable bias, so for each technique, the database of cases was split into distinct subsets for training and hold-out validation. In the MDA prediction phase, the sample priors were used rather than the naïve assumption of equal group priors, in order to minimize the rate of misclassifications should the estimated model be applied to observations from the underlying population of corporate bond funds. The MDA classification results on the hold-out sample are reported in Table 3 below.

Table 3: *MDA Classification Results Applied to Hold-Out Sample*
12/31/91 - 12/31/00 Period
(entries are number of occurrences)

Predicted as: Actual:	**Outperforming**	**Matching**	**Underperforming**	**No Longer Traded**	**Correct Pctg**
Outperforming (n=20)	11	5	0	4	11/20 = 55.0%
Matching (n=6)	2	1	3	0	1/6 = 16.7%
Underperforming (n=15)	2	6	5	2	5/15 = 33.3%
No Longer Traded (n=10)	0	3	4	3	3/10 = 30.0%
Total (n=51)	15	15	12	9	20/51 = 39.2%

The correct classification rate for Multiple Discriminant Analysis on its hold-out sample was 39.2 percent, with the majority of the misclassified cases (19 of 31) involving the "Matching" group in some capacity. When combining the cases used in the training and hold-out subsamples, the overall rate of correct classifications for MDA was 59.1 percent. Morrison (1969) argued that the overall predictive accuracy of classification models should be assessed relative to the percent correctly classified by chance. If the object of the study is merely to obtain the greatest overall percentage of correctly classified cases, then the "maximum chance criterion" would simply be the largest of the group sample proportions. Should the classification technique fail to achieve this benchmark, it should be dropped in favour of simply classifying all cases as belonging to that largest group. But, if the analysis is being conducted in order to classify members of each group correctly, then a "proportional chance criterion" is relevant. Following the development of such "proportional chance models" by Mosteller and Bush (1954) and Morrison (1969), we establish the chance benchmark to be 28.3 percent correct classifications for this four group model. By this standard, the linear MDA model results in correct predictions for the 102 cases at a rate of slightly more than twice that of the chance yardstick. The rule induction model, with an overall accuracy rate of 83.3 percent correct classification is nearly three-fold that expected by chance.

One issue that compromises the direct comparison of these two classification techniques is that, while they use the same original data of 102 cases, their mechanisms for selecting the subsets of "training" cases and "holdout" cases differ such that one can not insure that each technique is using the same subset for each purpose. Consequently, a performance comparison of the two techniques may be most valid at the aggregate level covering all 102 cases. In this regard, the overall success rate for rule induction of 83.3% convincingly exceeds both the 59.1 percent overall rate of MDA and the reference point of 76 percent overall achieved by Yoon et. al, (1994) with their expert system / neural network hybrid model for stock selection.

Temporal Extension Tests

Given the demonstrated strength of the rule-induction results relative to the MDA results for the contemporaneous period of 12/31/91 through 12/31/00, projections beyond that period will be limited to the rule-induction model. For the inter-temporal validation tests, one would posit that a model developed with historical data from an expansionary period would likely perform better during another expansionary period than it would during an economic downturn. The years immediately following the 1991-2000 development period provided the opportunity to examine this premise. The interval of year-end 2000 to year—end 2002 is treated as a period of economic "downturn", while calendar year 2003 is seen as a struggling attempt at "rebound". The rule structure's classification accuracy was examined independently in each post-development period. This study began the development period (1991-2000) with 102 mutual bond funds. During the nine year development period, 12 firms ceased trading in an organized market, leaving 90 firms at year-end 2000 for testing over the recessionary period of year-end 2000 to year-end 2002. During that 24-month period, eight additional funds ceased trading, leaving 82 of the original 102 at year-end 2002. Over the calendar year 2003, all 82 continued to be traded. Thus, our inter-temporal validation test involves the model's ability to predict the performance of 90 firms over the 24-month "recession" and 82 firms over the 12-month "recovery".

As was done with the system's original development, the Lipper Corporate A-Rated Debt Index was used as the relevant corporate bond benchmark over the post-development periods of this study. From year-end 2000 through year-end 2002, this index advanced by 17.2 percent. Whether based on either the proportionate growth in the Lipper index or the relative length of the time horizon, the proportionate window of advances, comparable to the +/- 3% used to define "Matching" in the nine-year developmental model, became 16.55% to 17.85% for this period's "Match" range. Thus, the individual fund performance metrics, as a measure of composite advance, are compared to this baseline range. The classification performance of the original system extended to the "recession" of 2001 and 2002 is summarized in Table 4 below.

Table 4: *Induced Rule Structure Applied to "Recession" Period of 12/31/2000 to 12/31/2002 (entries are number of occurrences)*

Predicted as: Actual:	**Outperforming**	**Matching**	**Underperforming**	**No Longer Traded**	**Correct Pctg**
Outperforming (n=6)	3	2	1	0	3/6 = 50%
Matching (n=0)	0	0	0	0	n.a.
Underperforming (n=76)	41	9	19	7	19/76 = 25%
No Longer Traded (n=8)	1	3	3	1	1/8 = 12.5%
Total (n=90)	45	14	23	8	23/90 = 25.6%

As is evident, the rule structure developed by pattern recognition over 12/31/91 to 12/31/00 did not perform well over the subsequent economic downturn, with only a 25.6 percent overall success rate. While funds in none of the four groups were classified well, the most disturbing misclassification was the preponderance of funds classified as "Outperforming" that were, in fact, "Underperforming" during this period of economic contraction. This evidently reflects the expansion orientation captured by the system during the original development time period.

The next aspect of this temporal extension was to consider the economically stronger environment of 12/31/2002 to 12/31/2003. For this "recovery" period , the Lipper index advanced 4.92 percent. Again, whether based on either the proportionate growth in the Lipper index or the relative length of the time horizon, the proportionate window of advances, comparable to the +/- 3% used to define "Matching" in the original nine-year developmental model, became 4.73% to 5.11% for this period's "Match" range. Thus, the individual fund performance metrics for calendar year 2003, as a measure of composite advance, are compared to this baseline range of performance. The classification performance of the original system extended to the "recovery" of 2003 is summarized in Table 5 below.

In this period of struggling recovery, the expert system's rule structure performed much better with an overall accuracy rate of 74.4 percent. The strength of the model is clearly its accuracy in identifying 39 of 41 that outperformed the index range. The weakness is the parallel identification of true underperformers. Fortunately, the occurrence of polar misclassification errors (outperforming funds predicted as "Underperformers" (n=1) and underperforming funds predicted as "Outperformers" (n=3)) is a modest 4.9 percent (=3/82). Most of the errors during this period are underperforming funds being cast as either "Matching" or "NLT".

Table 5: *Induced Rule Structure Applied to "Recovery" Period of Calendar Year 2003 (entries are number of occurrences)*

Predicted as: Actual:	**Outperforming**	**Matching**	**Underperforming**	**No Longer Traded**	**Correct Pctg**
Outperforming (n=41)	39	0	1	1	39/41 = 95%
Matching (n=1)	1	0	0	0	0/1 = 0%
Underperforming (n=40)	3	8	22	7	22/40 = 55%
No Longer Traded (n=0)	0	0	0	0	n.a.
Total (n=82)	43	8	23	8	61/82 = 74.4%

While both the expert system's rule structure and the MDA results were tested on hold-out cases in a contemporaneous period, the rule-induction results were further validated by extension to time periods beyond that of the original development. The forms of validation confirm the merit of the induced rule structure for periods of economic strength. The further generalization of that same structure to periods of economic contraction is not defensible; indicating the need for a distinct model for such periods.

Summary and Directions for Future Research

This study has sought to employ a form of machine learning (a rule induction expert system) to search for patterns among readily available descriptors of mutual bond funds that would predispose the fund to a specific measure of long-term performance relative to its peer index composite. The classification performance of rule-induction and multiple discriminant analysis were compared over the 12/31/91 to 12/31/00 period. Rule-induction outclassified MDA for each of the four groups of bond funds (Table 1 vs. Table 3). The rule-induction model was further tested beyond the original developmental period and found to perform poorly in a period of economic contraction but well in another period of economic expansion. This suggests that an underlying structural change may exist between periods of differing economic conditions, and distinct rules are warranted for such dissimilar economic environments.

In addition to classification, the technology of rule induction is able to identify the key factors contributing to the relative performance of domestic mutual bond funds. The pattern recognition capability of rule induction has detected a substantive role for short-term persistence, portfolio maturity, the percentage change in the fund's asset holdings, and portfolio yield. These factors are serving as operational surrogates for historical performance; the activity level of actively managed funds, and asset portfolio characteristics. All such critical factors are readily accessible to the ordinary mutual fund investor.

References

Anandarajan, M., Lee, P. & Anandarajan, A. (2001). Bankruptcy prediction of financially stressed firms: An examination of the predictive accuracy of artificial neural networks. *International Journal of Intelligent Systems in Accounting, Finance, and Management*, 10, 69-81.

Bers, M. & Madura, J. (2000). The performance persistence of closed-end funds. *The Financial Review*; 35, 33-52.

Bogle, J. (1999). *Common sense on mutual funds: New imperatives for the intelligent investor*, John Wiley & Sons Pub.

Braun, H. & Chandler, J.S. (1987). Predicting stock market behavior through rule induction: An application of the learning-from-example approach", *Decision Sciences*, 18, 415-429.

Brister, B.M., Kennedy, R.E. & Pu, L. (1994). The regulation effect of credit ratings on bond interest yield: The case of junk bonds. *Journal of Business Finance and Accounting*, 21, 511-531.

Brown, S. J. & Goetzmann, W. (1995). Performance persistence. *Journal of Finance*, 50, 679-698.

Brown, S. J., Goetzmann, W., Ibbotson, R. & Ross, S. (1992). Survivorship bias in performance studies. *Review of Financial Studies*, 5, 553-580.

Bundy, A., Silver, B. & Plummer, D. (1985). An analytical comparison of some rule-learning programs. *Artificial Intelligence*, 27, 137-181.

Carhart, M. (1997). On Persistence in mutual fund performance. *The Journal of Finance*, 52, 57-82.

Chan, K. & Jegedeesh, N. (2004). Market-based evaluation for models to predict bond ratings. *Review of Pacific Basin Financial Markets and Policies*, 7, 153.

Chandy, P.R. & Duett, E.H. (1990). Commercial paper rating models. *Quarterly Journal of Business and Economics*, .29, (4), 79-101.

Clarke, D.G. & McDonald, J.B. (1992). Generalized bankruptcy models applied to predicting consumer credit. *Journal of Economics and Business*, 44, 47-62.

Cooley, W.W. & Lohnes, P.R. (1971). *Multivariate data analysis*, New York, Wiley Publishers.

Deakin, E. (1976). distributions of financial accounting ratios: Some empirical evidence. *Accounting Review*, 90-96.

Eisenbeis, R. (1977). Pitfalls in the application of discriminant analysis in business, finance and economics. *Journal of Finance*, 32, 875-900.

Elton, E., Gruber, M. & Blake, C. (1996a). The persistence of risk-adjusted mutual fund performance. *Journal of Business*, 69, 133-157.

Elton, E., Gruber, M. & Blake, C. (1996b). Survivorship bias and mutual fund performance. *Review of Financial Studies*, 9, 1097-1120.

Epley, D.R., Liano, K. & Haney, R. (1996). Borrower risk signaling using loan-to-value ratios. *The Journal of Real Estate Research*, 11, 71-86.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems, *Ann. Eugenics*, 7, 179-188.

Fortin, R., Michelson, S. & Jordan-Wagner, J. (1999). Does mutual fund manager tenure matter?" *Journal of Financial Planning*, 12, (7), 72-79.

Gilbert, E.S. (1968). On discrimination using qualitative Variables. *Journal of the American Statistical Association*, 63, 1399-1412.

Grinblatt, M. & Titman, S. (1993). Mutual Fund Performance: An Analysis of Quarterly Portfolio Holdings. *Journal of Business*, 66, 47-68.

Hendricks, D. Patel, J. & Zeckhauser, R. (1993). Hot hands in mutual funds: Short-run persistence of relative performance, 1974-1988. *Journal of Finance*, 48, 93-130.

Indro, D.C., Jiang, C.X., Patuwo, B.E. & Zhang, G.P. (1999). Predicting mutual fund performance using artificial neural networks. *Omega*, 27, 373-380.

Jensen, M.C. (1968). The performance of mutual funds in the period 1945-1964. *Journal of Finance*, 23, 389-416.

Klecka, W.R. (1980). Discriminant analysis, Beverly Hills and London, *Sage Publications.*

Kumar, K. & Haynes, J. (2003). Forecasting credit ratings using an ANN and statistical techniques. *International Journal of Business Studies*, 11, 91-108.

Lintner, J. (1965) The Valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47, 13-17.

Lachenbrach, P.A. (1975). Discriminant analysis, *Riverside New Jersey*, Hafner Press.

Long, M.S. (1976). Credit screening system selection. *Journal of Financial and Quantitative Analysis*, 11, 313-328.

Malhotra, R. & Malhotra, D. (2003). Evaluating consumer loans using neural networks. *Omega*, 31, 83.

Medury, P.V., Bowyer, L.E. & Srinivasan, V. (1992). Stock repurchases: A multivariate analysis of repurchasing firms. *Quarterly Journal of Business and Economics*, 31, 21-44.

Messier, W.F. & Hansen, J.V. (1988). Inducing rules for expert system development: An example using default and bankruptcy data", *Management Science*, 34, 1403-1415.

Morey, M. & Morey, R. (1999). Mutual fund performance appraisals: A multi-horizon perspective with endogenous benchmarking", *Omega*, 27, 241-258.

Morrison, D.G. (1969). On the interpretation of discriminant analysis. *Journal of Marketing Research*, 6, 156-163.

Mosteller, F., & Bush, R. (1954). Selective quantitative techniques, in Handbook for Social Psychology, G. Lindzey (ed.), Redding MA, Addison-Wesley Publishers.

Moyer, R.C. (1977). Forecasting financial failure: A re-examination. *Financial Management*, 6, 11-17.

Patel, J., Zeckhauser, R. & Hendricks, D. (1992). Investment flows And performance: evidence from mutual funds, cross-border investmenstsinvestments, and new issues, in R. Sato, R. Levich, & R. Ramachadran, eds.: *Japan and International Financial Markets: Analytical and Empirical Perspectives*, Cambridge University Press.

Piesse, J. & Wood, D. (1992). Issues in Assessing MDA Models of Corporate Failure: A Research Note. *The British Accounting Review*, 24, 33.

Pompe, P. & Bilderbeek, J. (2005). Bankruptcy prediction: The influence of the year prior to failure selected for model building and the effects in a period of economic decline. *Intelligent Systems in Accounting, Finance, and Management*, 13, 95.

Quinlan, J.R. (1983). Learning efficient classification procedures and their application to chess and games, in R.S. Michalski, J.G. Carbonell & T.M. Mitchell (ed), *Machine Learning: An Artificial Intelligence Approach*. Tioga Publishing Co. Palo Alto, CA.

Quinlan, J. (1986). Induction ofdDecisionof Decision trees. *Machine Learning*. 1, 81-106.

Sharpe, W. (1966). Mutual fund performance. *Journal of Business*, 39, 119-138.

Sharpe, W. (1988). Determining a fund's effective asset mix. *Investment Management* Review, 5-15.

Treynor, J.L. (1965). How to rate management of investment funds. *Harvard Business Review*, 63-75.

Umamaheswar, S.P. (2001). Mutual fund performance during up and down market conditions. *Review of Business*, 22, 62-65.

Watson, C.J. (1983). Multivariate normality and a bond rating decision model. *Decision Sciences*, 14, 513-526.

Yoon, Y., Guimaraes, T. & Swales, G. (1994) Integrating artificial neural networks with rule-based expert systems. *Decision Support Systems*, 11, 497-507.