A Comparison of Nonlinear Methods for Predicting Earnings Surprises and Returns

Vasant Dhar and Dashin Chou

***Abstract*—We compare four nonlinear methods on their ability to learn models from data. The problem requires predicting whether a company will deliver an earnings surprise a specific number of days prior to announcement. This problem has been well studied in the literature using linear models. A basic question is whether machine learning-based nonlinear models such as tree induction algorithms, neural networks, naïve Bayesian learning, and genetic algorithms perform better in terms of predictive accuracy and in uncovering interesting relationships among problem variables. Equally importantly, if these alternative approaches perform better, why? And how do they stack up relative to each other? The answers to these questions are significant for predictive modeling in the financial arena, and in general for problem domains characterized by significant nonlinearities. In this paper, we compare the four above-mentioned nonlinear methods along a number of criteria. The genetic algorithm turns out to have some advantages in finding multiple "small disjunct" patterns that can be accurate and collectively capable of making predictions more often than its competitors. We use some of the nonlinearities we discovered about the problem domain to explain these results.**

***Index Terms*—Data mining, earnings surprise, genetic algorithms, machine learning, nonlinear systems.**

I. Introduction

THIS study compares the performance of four nonlinear methods in predicting earnings surprise. It has been observed that earnings expectations tend to be an important determinant of future equity prices. On average, there tends to be a "preannouncement drift" where prices of companies who end up delivering "positive earnings surprise" drift upwards, adjusting for the overall market, and the prices of companies who deliver a negative surprise are relatively weaker. There is then a "post announcement drift," where the prices of firms who had an actual positive surprise show above market returns and vice-versa for the negative surprise firms.

Our experience with building systematic investment models in finance leads us to believe that predictive modeling is challenging for a number of reasons. First, financial markets are inherently "noisy" with several types of nonlinearities. For example, the degree to which a company's reported earnings exceed its expected earnings has a nonlinear relationship with its stock price. Sometimes the effect on price kicks in only after an expectation threshold has been *significantly* exceeded, and the strength of the relationship increases rapidly thereafter. For example, prices are often unaffected when earnings expectations are exceeded marginally, but react very strongly if earnings exceed a high threshold. But this effect can also be tempered by other variables, such as the industry involved, whether the company is a "growth" or "value"[1] stock, whether it surprised positively or negatively last time, or more macroeconomic variables. There is also an observed "price creep" effect, upwards for companies that deliver positive surprises and negative for companies that deliver negative surprises. Again, this preannouncement effect is also tempered by other variables. Discovering an underlying structure that works generally in predicting pre-and postannouncement drift for such problems is therefore challenging. Simple models just do not work most of the time, so finding models that make reliable predictions even occasionally is useful.

The problem of predicting earnings surprise and postannouncement drift is also interesting because it has been well studied in the accounting and finance arenas. Understandably, prior research has modeled this problem using linear statistical approaches, where expected interaction effects are modeled using dummy variables. With this approach, the expected interaction effect is a part of the problem formulation, not a relationship that is discoverable through an intelligent search. One of the desirable aspects of nonparametric machine learning methods is that they are able to discover the interesting interaction effects automatically. If such methods turn out to be significantly better in terms of predictive accuracy and are easy to understand, they should be considered seriously as standard tools to be used in exploratory research in investment decision making, and for nonlinear prediction problems in general. As one experienced money manager put it, "patterns often emerge before the reasons for them are apparent."

The area of earnings surprise prediction is also a good testbed for nonlinear methods because a number of studies have shown that earnings expectations often affect markets and therefore there is the possibility of predictive power in this area. In other words, while difficult, prediction is not unequivocally an exercise in futility. Whether nonlinear methods perform better than linear methods is in itself an open question in this arena, and if they do, which methods perform better and why is of particular interest.

The four methods we compare to linear regression (and of course, random guessing) are the artificial neural network

Manuscript received September 13, 2000; revised February 22, 2001 and March 11, 2001.

The authors are with the Department of Information Systems, Stern School of Business, New York University, New York, NY 10003 USA.

Publisher Item Identifier S 1045-9227(01)05008-1.

[1]Growth and value are standard nomenclature in the investment community. A value approach places emphasis on the strength of a company's current financial statements whereas a growth approach places emphasis on a company's potential for long-term growth and ignores many aspects of its current financials. Internet stocks, for example, are currently evaluated from a growth perspective although investors are beginning to look more closely at company fundamentals more recently in that sector.

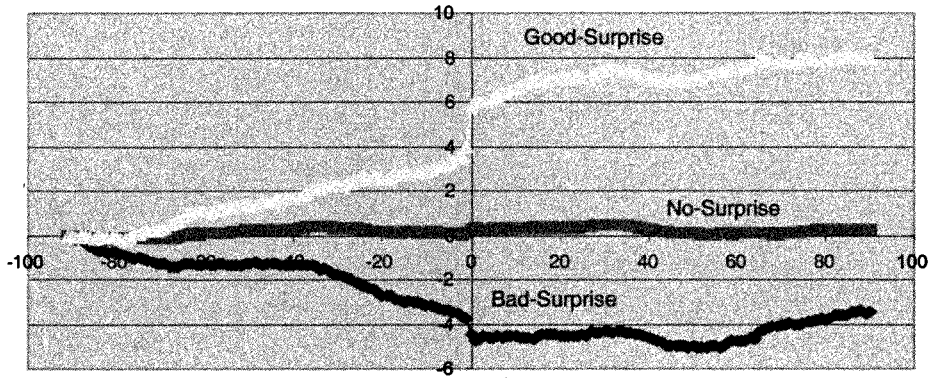


Fig. 1. CAR-SP500 (90–97).

(ANN), genetic algorithm (GA), classification and regression tree (CART), and Naïve Bayes (NB).

In the next section, we review prior research on event studies and establish definitions for earnings surprise. In Section III, we describe the variables and the selection process used in the study. In Section IV we provide a brief description of the genetic algorithm, in particular, its concept class representation. Results are presented in Section V. We conclude with an explanation of the observed results.

II. Prior Research

Fig. 1 shows the relationship between prices and expected and actual earnings. It shows the cumulative abnormal return[2] (CAR) of SP500 firms from 90 days before the earnings announcement to 90 days after the announcement. The chart is based on eight years of data from 1990 to 1997 of S&P500 firms, eliminating survivorship bias. A similar effect was noted by [19] between January 1989 and December 1993 on stocks in the Dow Jones index. Differences between the reported earnings and consensus forecast earnings serve as the basis for classifying stocks into three groups: good-news firms, no-news firms, and bad-news firms. It has been demonstrated that each group exhibits distinctive cumulative abnormal returns in the proximity of the earnings announcement day in the following ways:

1) *Preannouncement drift:* on average, 30 days (or more) before the earnings announcement day, the CARs of good-news firms gradually drift up to the announcement day, the CARs of the bad-news firms drift down, and the CARs of the no-news firms are around zero.
2) *Announcement day effect:* on average, good-news firms have big price jumps, bad-news firms have big price drops, and no-news firms have no big price movements on announcement day.
3) *Postannouncement drift:* on average, the CARs of the good-news firms drift up all the way to 70 days after the earnings announcements. The drift of the bad-news firms is less obvious. Again, the CARs of the no-news firms are almost zero.

[2]CAR is defined formally in Section II-C (8). It is the return of a stock adjusted by factors such as the return of the overall market.

It has been pointed out that the abnormal CAR following announcement is a violation of semistrong market efficiency.[3] A number of explanations have been proposed for this phenomenon, which has been persistent for a long period of time in financial markets (see [33]). This research focuses on preannouncement, where the objective is to classify a company as positive, neutral, or negative surprise a specific number of days before announcement. In particular, we address the following questions:

1) How accurately do various methods classify a case into one of the three categories prior to announcement? That is, which method is the most accurate in estimating the probability that a firm will have good news or bad news D days prior to the announcement, i.e.,
   P (good news at event day T | Information available until day T−D)
   P (bad news at event day T | Information available until day T−D)
2) How accurately do the various methods forecast the magnitude of returns between now and announcement day? The number of days before announcement was varied between ten and 40.

A. Event Studies Relating Change of Earnings and Change of Returns

The fact that prices change in anticipation of earnings has been apparent for some time. For evidence in the accounting literature, see [8], and [22]. They argue that events such as long-term sales contracts affect the market's expectation of future earnings and cash flows, and stock prices reflect these revised beliefs. It is believed that growth opportunities facing a firm are reflected in current stock prices, but are not generally reflected in current earnings. Therefore, returns are expected to lead earnings changes. In other words, price changes could have some predictive power with respect to future earnings. These price changes happen while analysts are revising earnings estimates and the company itself is releasing information to the market.

Abarbanell [1] shows a statistically significant positive association between the sign of cumulative returns and the sign of

[3]Semi-strong market efficiency means that all publicly available information is reflected in prices.

subsequent earnings revisions for value line data. He also designs a statistic to test whether analysts' forecasts fully reflect the information in prior price changes

$$freq[Error < 0 \,|\, Return > 0] - freq[Error < 0 \,|\, Return < 0] > 0. \quad (1)$$

The above expression says that analysts tend to revise estimates upwards when returns are positive and downward when they are negative.

Elgers and Murray [34] evaluated the relative performance of financial analysts' forecasts (FAF) and price-based forecasts (PBF) of expected earnings and abnormal returns. Their first model tests the forecasts of expected earnings

$$EPS_{j,t} = \beta_{0,t} + \beta_{1,t} * FAF_{j,t} + \beta_{2,t} * PBF_{j,t} + \mu_{j,t} \quad (2)$$

where

$EPS_{j,t}$ percentage earnings change for firm j in year t;
$PBF_{j,t}$ price-based prediction of percentage earnings change for firm j in year t;
$FAF_{j,t}$ financial analysts' forecasts for firm j in year t.

They show that both forecast sources do in fact contain unique information that is relevant to explaining actual changes in earnings.

Their second model addresses the potential combination of financial analysts and price-based proxies for unexpected earnings

$$SAR_{j,t} = \beta_{0,t} + \beta_{1,t} * UE_{FAF,j,t} + \beta_{2,t} * UE_{PBF,j,t} + \mu_{j,t} \quad (3)$$

where $SAR_{j,t}$ is the size-adjusted abnormal return for firm j in year t, and UE is the unexpected earnings measured using FAF or PBF forecasts. They show that neither forecast source is superior in any size group. The two forecast sources are useful as complements; a linear composite of both measures accounts for a greater proportion of the cross-sectional variability in security returns than does either individual measure of unexpected earnings.

B. CAPM and Single Factor Pricing Models

Sharpe [69] and Lintner [56] show that if investors have homogeneous expectations and optimally hold mean-variance efficient portfolios then, in the absence of market friction, the market portfolio will be a mean-variance efficient portfolio. Here the mean-variance efficient portfolio is a portfolio with the highest expected return for a given level of variance. For this version of the capital asset pricing model (CAPM) we have for the expected return of asset i

$$E[R_i] = R_f + \beta_{im}(E[R_m] - R_f) \quad (4)$$

$$\beta_{im} = \text{Cov}[R_i, R_m]/\text{Var}[R_m] \quad (5)$$

where R_m is the return on the market portfolio, and R_f is the return on the risk-free asset. The CAPM implies that the expected return of an asset is linearly associated with β_{im}, the covariance of its return with the return of the market portfolio.

Basu [12] reports the price-earnings-ratio (P/E) effect, where firms with low P/E ratios have higher sample returns, and firms with high P/E ratios have lower sample returns than would be the case if the market portfolio was mean-variance efficient. Banz [10] documents the size effect: small size firms have higher sample mean returns than would be expected if the market portfolio was mean-variance efficient. Fama and French [36] find that firms with high book-to-market ratios have higher average returns than is predicted by the CAPM. Jegadeesh and Titman [51] find that a portfolio formed by buying stocks whose value has declined in the past and selling stocks whose value has risen in the past has a higher average return than the CAPM predicts.

The single factor model states that the return of a stock is linearly related to the market return.

$$R_{jt} = \alpha_j + \beta_j * R_{mt} + \varepsilon_{jt} \quad (6)$$

where the R_{mt} is the market return at time t and the ε_{jt} is the error term for time period t. These models are not intended to provide any predictive power of the earnings surprise events which can be explained by the following two reasons:

- Both models state the "current" relationship between a stock's return and the market return. Since this is not a leading or lagging relationship, we can not use this period's market return to forecast next period's stock return.
- Even if we use this period's market return to forecast next period's stock return, we cannot expect the forecast to be accurate since it ignores events specific to a stock such as earnings revisions or announcements.

The above limitations have motivated the development of multifactor models, such as the BARRA model [11], that incorporate additional risk factors such as firm size, industry sector, etc into the model.

C. Multifactor Pricing Models

Prior research indicates that the CAPM beta does not completely explain the changes of expected asset returns. Rather, certain groups of stocks are affected by common "risk" factors. For example, weather patterns would affect commodity stocks more strongly than automobile stocks. Ross [66] introduced arbitrage pricing theory (APT) as an alternative to the CAPM by allowing for multiple risk factors other than just "the market." Unlike the CAPM, the APT does not require the identification of the market portfolio, but that

$$R_i = a_i + b_i * f + \varepsilon_i \quad (7)$$

where

R_i return for asset i;
a_i intercept of the factor model;
b_i $(K * 1)$ vector of factor sensitivities for asset i;
f $(K * 1)$ vector of common factors;
ε_i error term.

The underlying theory of the multifactor pricing models does not specify the number of required factors. One empirical approach is to repeat the estimation and testing of the model for a variety number of factors and observe if the tests are sensitive to increasing the number of factors. For example, Lehmann and Modest [54] demonstrate empirical results for five, ten, and 15 factors. They show minimal sensitivity when the number of factors increases from five to ten to 15. Similarly, Connor and

Korajczyk [24] find little sensitivity to increasing the number of factors beyond five.

Multifactor models of security market returns can be divided into three types: macroeconomic, fundamental, and statistical factor models. Macroeconomic factor models use observable economic time series, such as inflation and interest rates, as measures of the pervasive shocks to security returns. Fundamental factor models use the returns of portfolios with observed attributes such as dividend yield, the book-to-market ratio, and industry membership. Statistical factor models derive their factors from a factor analysis of the covariance matrix of security returns.

D. Other Multifactor Models

Reference [15] proposed a two-factor model to test the inverse relation between drift and firm size:

$$CAR = b_0 + b_1 * SUE + b_2 * SUE * Size + error. \quad (8)$$

In (8), CAR is the postannouncement drift, SUE measures the standardized unexpected earnings, and $Size$ represents firm size. The coefficients b_1 and b_2 were statistically significant with b_1 positive and b_2 negative, confirming the previously documented inverse relation between size and drift.

Foster *et al.* [37] use the following two-factor regression to probe the relative importance of the earnings forecast error variable and the firm size variable in explaining the sign and magnitude of CAR:

$$CAR = \alpha + \beta_1 * FEP + \beta_2 * FSQ + error \quad (9)$$

where FEP is coding from 1 to 10 of the earnings forecast error, and FSQ is a coding of the firm size quintile.

They find that the more positive the unexpected earnings change, the more positive the post-announcement CAR. Furthermore, the smaller the firm size, the larger the post-announcement CAR.

Francis and Soffer [38] use a multifactor model to investigate market reaction to earnings forecast revisions, stock recommendations, revisions in stock recommendations, and earnings forecast revisions. They find that market reactions to analysts' reports are significantly related to both earnings forecast revisions and stock recommendations.

Abarbanell and Bernard [2] use the multifactor regression model to present evidence that analysts' forecasts under-react to recent earnings

$$FE = \alpha + \beta_1 * FE_{-1} + \beta_2 * FE_{-2} + \beta_3 * FE_{-3} + \beta_4 * FE_{-4} \quad (10)$$

where FE_{-n} represents the forecast error N quarters ago. They show that earnings forecast errors are positively autocorrelated over the first three lags, with declining magnitude.

E. Market Model CAR

The market model for the return of security j in event time t is in (6). Prior research, such as [37], [25], and [68], defines the abnormal return, AR_{jt}, to be the difference between the estimated normal returns and the actual returns and the cumulative abnormal return to be the sum of the abnormal returns

$$AR_{jt} = R_{jt} - (\alpha_j + \beta_j * R_{mt}) = \varepsilon_{jt}. \quad (11)$$

$$CAR_j(T_a, T_b) = AR_{j,Ta} + AR_{j,Ta+1} + \cdots AR_{j,Tb}. \quad (12)$$

We, however, define the cumulative abnormal return from day T_a to T_b to be

$$CAR_j(T_a, T_b) = (1 + AR_{j,Ta}) * (1 + AR_{j,Ta+1}) * \cdots * (1 + AR_{j,Tb}) - 1. \quad (13)$$

We do not choose the additive form of CAR since the CAR should be multiplicative, which can be significantly different from the sum of abnormal returns, especially as the holding period becomes longer.

F. Earnings Surprise

There are several different definitions of earnings surprises. Foster *et al.* [37] propose the forecast error measure:

$$Forecast\ Error_{it} = (Q_{it} - E(Q_{it}))/|Q_{it}| \quad (14)$$

where Q_{it} is the actual quarterly earnings of firm i in period t and $E(Q_{it})$ is its expected earnings. The problem with this measure is that it gives highly positive or highly negative values when Q_{it} is close to zero. Thus even when expected and actual values are only slightly different but close to zero, the measure is highly unstable. Accordingly, we modified their measure by adding the $E(Q_{it})$ term in the denominator, as follows:

$$FE_{it} = (Q_{it} - E(Q_{it}))/(|Q_{it}| + |E(Q_{it})|). \quad (15A)$$

This is a better "normalized" variable whose value varies between -1 and $+1$, Also, when Q_{it} is close to zero and $E(Q_{it})$ is not close to zero, the denominator will not be close to zero. A small disadvantage of this measure is that it takes the extreme values of $+1$ and -1 even when expected and actual earnings are of opposite sign regardless of magnitude. It, therefore, emphasizes the fact that a surprise is "positive" or "negative" regardless of the degree of the surprise.

Mendenhall [60] defines the earnings forecast errors (ERR) to be

$$ERR_{it} = (Q_{it} - E(Q_{it}))/P_{it} \quad (15B)$$

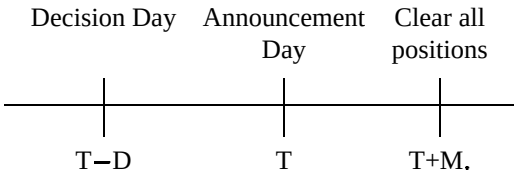
where P_{it} is the common stock price.

We have experimented with both the FE and ERR measures. It turns out that they are highly correlated, i.e., greater than 0.9. In this research, we therefore use FE as the dependent variable for earnings surprise.

III. Research Design

The time line below shows the relevant decisions, events and actions. Day T is the earnings announcement day. Day T−D is the day on which we forecast earnings surprise and the future risk-adjusted return. If we wished to trade, we would initiate a

position on day T−D and liquidate on day T+M. For the trading strategy, the "holding period" is therefore M+D



In this study, we varied D between ten and 40 days. For an investment manager, the choice of D is critical. Predicting too early will results in a high error rate while waiting too close to announcement day is easier but foregoes potential returns.

A. Data Sources

The data used in this research were from commonly used industry sources, namely, CRSP, I/B/E/S, DAIS (now part of Bridge), and HOLT Analytics. Each of these sources provides different market indicators including balance sheet data, analyst expectations and revisions, and prices and volumes. The various indicators of interest generally have a low degree of correlation with FE. Our objective is to discover those models that combine the indicators in a way that maximizes predictive accuracy and coverage, i.e., you want to make predictions when you are most likely to be accurate, and at the same time you want to make predictions as often as you can.

The data cover 12 years, from 1986 to 1997, of stocks in the I/B/E/S universe, which consisted of roughly 3600 stocks in 1986 and 5700 stocks in 1997. The stocks covered most commonly among the sources tend to be the ones followed most by analysts. This includes the S&P500, covered by all, plus roughly 200 additional stocks that have sufficient market capitalization and liquidity to be of interest to institutional investors. The following data sources were used.

- **CRSP:** provides information on individual U.S securities traded on the New York Stock Exchange, the American Stock Exchange and the Nasdaq Market. For each security there are five kinds of data: identifying data (i.e., name, cusip, ticker, SIC code), price history, trading volume, outstanding shares, and a returns history.
- **I/B/E/S:** the I/B/E/S historical database provides quarterly earnings information, such as quarterly earnings and earnings report dates, and monthly data, such as number of earnings estimates, earnings forecast revisions, and the consensus earnings forecast. I/B/E/S also provides detailed earnings estimate history (identifying each analyst) which contains over 12 years of forecast changes for U.S. companies. It encompasses earnings estimates from more than 200 brokerage houses and 2000 individual analysts.
- **DAIS:** DAIS provides metrics based on the moving average of analyst earnings revisions, and adjusts these based on the return behavior of the stock [26], [27]. These metrics are based on multivariate regression models that take analyst expectations and recent price action as inputs.
- **HOLT:** HOLT is one of the more comprehensive sources of fundamentals-based information. HOLT's basic premise is that the market ultimately sets prices based on cash flow, not on traditional accounting measures of corporate performance such as price/earnings ratios since such measures are highly amenable to manipulation. HOLT argues that over the past three decades, historical cost accounting conventions and large fluctuations in the level of inflation have greatly reduced both the comparability and usefulness of earnings based ratios. Accordingly, HOLT's measure of performance is a cash flow return on investment called CFROI [58]. CFROI is defined as the inflation adjusted gross cash flow of the business vs. the inflation adjusted gross investment of the business (cash in vs. cash out).

The complete dataset consisted of 36 204 records, which was partitioned into two subsets, a training and a test set. The training set consisted of 24 140 records and was used for variable selection, and for model building. The test set consisted of 12 164 records. The models generated from the training set were tested on randomly selected samples from the test set, thereby giving us a performance distribution for each model on data that were not used to build the models.

B. Variable Definitions

The potential number of input variables that could be included in this research was very large. For example, [3] describes hundreds of technical variables that can be constructed from a time series. If several different interval lengths are used for each, we can end up with thousands of independent variables. In order to keep the problem tractable, we focused on about a dozen technical, fundamental, and "expectation-based" variables commonly followed in the investment community, and used a heuristic for variable selection from within this set. The basic selection mechanism we used was to include variables that demonstrated some degree of consistent correlation with the dependent variable, even though the correlation might be weak. In this section, we describe the chosen variables.

The variables can be classified roughly as *expectation-based*, *fundamentals-based*, and *technicals-based*. The first set consists of aggregated statistics about expected prospects expressed by a group of analysts. The second are derived from accounting statements, which are meant to express the current financial health and risks of a company. The third set consists of moving averages based on historical prices and volumes, which express the "relative strength" of a stock or sector.

1) Dependent Variables: Two dependent variables were used, namely, forecast error and the future D-day risk-adjusted return, where D was varied between 10 and 40.

a) FE: Forecast Error: was previously defined in (15A)

$$FE_{it} = (Q_{it} - E(Q_{it}))/(|Q_{it}| + |E(Q_{it})|) \qquad (15)$$

where Q_{it} is the quarterly earnings of the ith firm in period t and $E(Q_{it})$ is the expected earnings of the ith firm on the earnings announcement day. This measure gives values between -1 and $+1$, where the extreme values occur when Q_{it} and $E(Q_{it})$ are of opposite sign.

b) F20: Future 20-day risk adjusted return

$$F20_{it} = Cumret_i(t+1, t+20)/Stdev_i(t+1, t+20) \quad (16)$$

where $Cumret_i(t+1, t+20)$ is the cumulative return of the ith firm from day $t+1$ to $t+20$ and $Stdev_i(t+1, t+20)$ is the standard deviation of return of the ith firm from day $t+1$ to $t+20$. This variable measures each model's forecast ability on the profit and loss of a one month holding period.

2) Expectation-based Independent Variables:

c) ER: Estimate Revision Index: The DAIS group's global estimate revision score is a predictive multivariate earnings momentum model which uses three month changes in consensus earnings estimates, the net change of raised versus lowered estimates, and the changes in the high and low estimates to forecast unanticipated earnings [27]. The research in this area suggests that changes in consensus earnings are weakly correlated with companies' actual performance. Three time series values were used, the value of the ER index in the current month, one month ago, and two months ago.

d) F1 and F2: Forecast Risk Factor Indexes: The DAIS group also produces a "Forecast Risk Factor Model" index based on consensus estimate revisions and earnings surprises over the last three months. Two forecasts, F1 and F2 are produced where F1 is the value for the current fiscal year and F2 is the F value for the next fiscal year. We use three values for F1 and F2 based on the current month, one month ago, and two months ago.

e) Analyst Forecast differences: I/B/E/S provides the number of analysts that forecast a specific company's next quarterly earnings. I/B/E/S also provides the number of analysts that recently made upward earnings estimate revisions and the number of analysts that recently made downward earnings estimate revisions. We constructed this independent variable, Analyst Forecast differences, to be the number of upward revisions minus the number of downward revisions. We included three periods of this variable: current month, one month ago, and two months ago.

3) Fundamentals-based Independent Variables:

f) CI and CF: HOLT value associates computes the cash flow return on investment, CI. The CI number, which is adjusted for inflation and for material accounting distortions, calculates company performance in terms of a cash-flow-return-on-investment. It provides an indication of the degree of latitude that a company's management has in deploying its "free cash" which is the cash left over after expenses and necessary expenses and investments. The details of CI are beyond the scope of this paper, and the reader is referred to [58] for a full discussion of this variable. HOLT also computes the momentum of CI, based on the CI history that is referred to as CF.

CI is a "fundamental" variable in a sense that it is based on the accounting data. Rather than relying on measures of performance like EPS and ROE, company income statements and balance sheets are converted into a cash flow based measure of performance. We chose three values of CI and CF, based on the current month, one month ago, and two months ago.

g) Discount Rate Differential: HOLT also computes a discount rate differential, defined as the difference between the market discount rate and the company-specific discount rate. Conceptually, this measures the additional return a company would be required to pay on debt relative to the average return required for that market. For example, companies with higher leverage and weak balance sheets would have higher discount rate differentials. The reader is referred to [58] for a discussion of this variable. We used three values, from the current month, one month ago, and two months ago.

h) Size: Each stock's market capitalization was computed and percentiled. Size has been shown to be moderately correlated with firm performance during several periods in the last two decades. Several hypotheses have been proposed for this effect, the most recent one in the practitioner community being that investment fund managers tend to favor larger stocks because they tend to be more liquid, which has a significant impact on the costs of accumulation and divestment. In effect, liquidity itself has economic value.

4) Technicals-based Independent Variables:

i) Industry trend: is defined as the risk-adjusted return of an industry group. To compare the industry trends among industry groups, we grouped stocks according to the S&P industry classification and ranked these risk-adjusted returns every month. We selected ten, 20, 30, and 60 days industry trends as the input variables. Industry trend has been shown to exhibit serial correlation. An explanation for this correlation is that business cycles tend to last several years, and sectors therefore tend to do well or poorly for several quarters in a row.

C. Variable Pruning Heuristics

In order to make the problem more tractable, we attempted to reduce the search space by eliminating variables that by themselves, or in interaction with others, are unlikely to have a major impact on the dependent variable. We used the training data to remove independent variables that are highly correlated with others and/or have low correlation with the dependent variable. Furthermore, independent variables that did not have persistent correlation with the dependent variable in this dataset were also removed.

1) Removing Low Correlation Variables: The Rank Order Correlation, which does not assume normality of the data, was used to measure correlation among the variables. Fig. 2 shows the correlation coefficients between independent variables and one dependent variable, FE on the training data. These charts show that CI and discount rate differential have low rank-order-correlation coefficient with FE, suggesting that they could be removed from the attribute set without sacrificing predictive power. In retrospect this makes sense given the relatively short holding period (i.e., 20 days) and the fact that these fundamentals based variables change only periodically, when a company discloses financial statements. All variables with correlation below 0.05 were removed, narrowing the set to 24 variables.

2) Removing Mutually Correlated Variables: The following variables were observed to be significantly correlated:

Var1	Var2	Correlation-Coefficient
Feqn1	ERRqn1	0.967 380
Feqn2	ERRqn2	0.967 150
F1n0	F2n0	0.614 062
F1n1	F2n1	0.610 702
F1n2	F2n2	0.586 567.

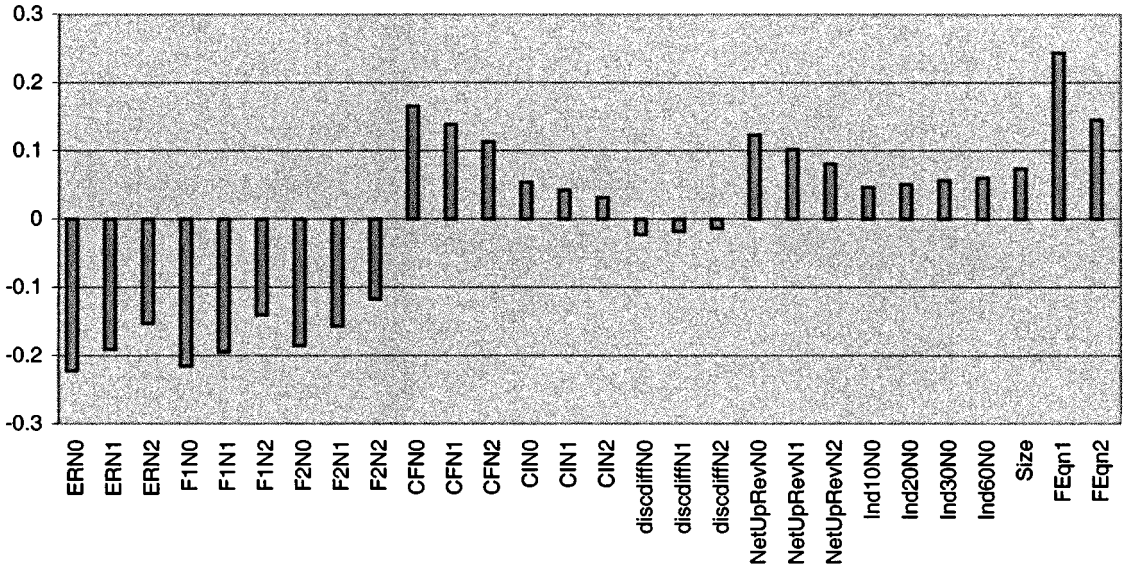


Fig. 2. Correlation between forecast error and independent variables.

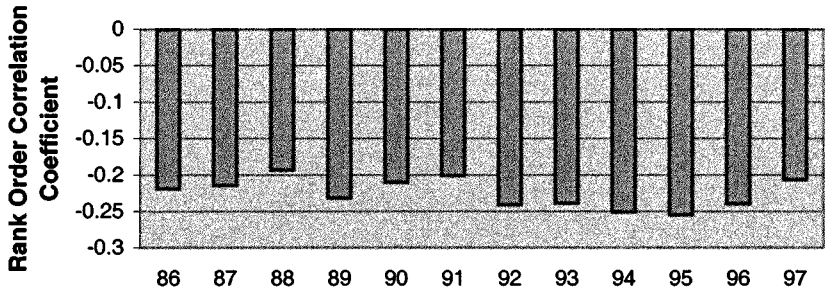


Fig. 3. Yearly correlation between forecast error and the most recent earnings revision.

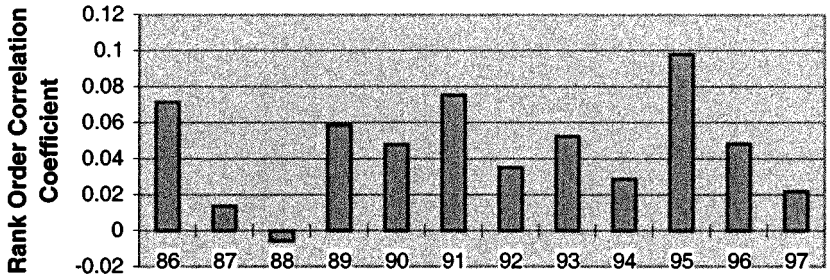


Fig. 4. Correlation between forecast error and ten-day industry strength.

Accordingly, ERRqn1, ERRqn2, F2n0, F2n1, and F2n2, were removed from consideration, leaving us with 19 variables.

3) Removing Variables with Low Persistent Correlation: We would like to keep variables that have relatively higher and persistent correlation with the dependent variable. Fig. 3 shows the yearly correlation on in-sample data between FE and one of the independent variables, namely, the most recent earnings revision score, ERn0. Their yearly correlation coefficients varied between -0.195 and -0.257, reflecting a fairly high and stable yearly correlation. A similar analysis was done for the other candidate variables.

Fig. 4 shows an opposite case, of relatively low correlation and low persistence in correlation, in this case between FE and the ten-day industry strength.

More formally, persistency was measured as the ratio of the average yearly correlation over a period to the standard deviation during this period

$$Persistency\ Measure\colon Average\ (Yearly\ Correlation)/ Std\ (Yearly\ Correlation).$$

Fig. 5 shows this persistency measure of the candidate variables. We dropped variables with absolute values below 2. The three lowest values are for ten-day industry trend, CI, and discount rate differential. The last two were already removed earlier because of low correlation with FE. In this step, we therefore removed the ten-day industry trend, leaving a total of 18 variables.

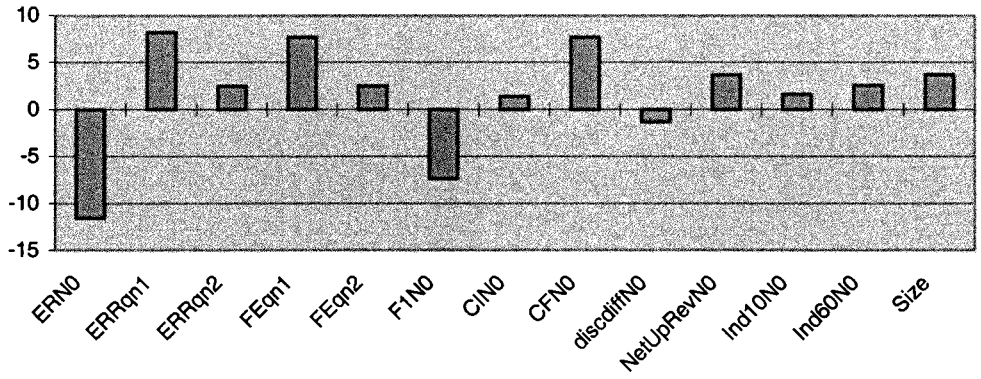


Fig. 5. Correlation persistency between forecast error and independent variables.

In summary, we started with 30 variables that are followed commonly in the industry and narrowed down this set to 18 using the filters described above.

IV. Nonlinear Methods

Linear regression is used as the benchmark since it is a standard econometric technique used in virtually every previous event study. We compare linear regression with an ANN, a genetic learning algorithm, a tree induction algorithm called CART, and a "Naïve Bayesian" learning algorithm.

One of the basic differences in the above methods is the functional form of the learned model. For regression and neural networks, the data are used to estimate the parameters of a prespecified functional form (linear for linear regression and piece-wise nonlinear for neural networks). With tree induction and genetic algorithms, the output consists of explicit "rules" whose left-hand side corresponds to Boolean expressions defined over the independent variables and the dependent variable is a prediction such as "positive surprise" or "3% expected return over the next 20 days."

The tools used for performing regression, neural modeling, Bayesian learning, and tree induction are widely available, and more importantly, the representation of the learned model is standard (i.e., weights for linear regression and neural nets, trees for tree induction, and conditional distributions for Naïve Bayes). For linear regression and neural networks, the software used was from Kennedy *et al.* [73]. For the neural network, the standard backpropagation algorithm was used with one hidden layer and double the number of hidden nodes as inputs (the default), and three output nodes for earnings surprise prediction and one for returns. For tree induction, we used the Salford Systems version of CART, which is described by [17]. The Naïve Bayes algorithm was implemented in Perl.

Naïve Bayes makes the assumption that given a class, the value of each variable is *independent* of the value of other attributes (this is the naïve assumption). So, if our objective is to classify a datum into a class, where the datum is described in terms of attributes $A_1, A_2, \cdots, A_k$, and we wish to predict its class value c, we want

$$\Pr(C = c | A_1 = a_1 \wedge \cdots \wedge A_k = a_k)$$

to be maximal. By Bayes rule

$$\Pr(C = c | A_1 = a_1 \wedge \cdots \wedge A_k = a_k) = \frac{\Pr(A_1 = a_1 \wedge \cdots \wedge A_k = a_k | C = c) \cdot P(C = c)}{\Pr(A_1 = a_1 \wedge \cdots \wedge A_k = a_k).}$$

Since the denominator is the same for each class, we can ignore it and rewrite the conditional as

$$\Pr(A_1 = a_1 \wedge \cdots \wedge A_k = a_k | C = c) = \Pr(A_1 = a_1 | A_2 = a_2 \wedge \cdots \wedge A_k = a_k, C = c) \cdot \Pr(A_2 = a_2 \wedge \cdots \wedge A_k = a_k | C = c).$$

Recursively, the second factor above can be written as

$$\Pr(A_2 = a_2 | A_3 = a_3 \wedge \cdots \wedge A_k = a_k | C = c) \cdot \Pr(A_3 = a_3 \wedge \cdots \wedge A_k = a_k | C = c)$$

and so on. With the naïve assumption, we assume that

$$\Pr(A_1 = a_1 | A_2 = a_2 \wedge \cdots \wedge A_k = a_k, C = c) = \Pr(A_1 = a_1 | C = c).$$

Thus

$$\Pr(A_1 = a_1 \wedge \cdots \wedge A_k = a_k | C = c) = \Pr(A_1 = a_1 | C = c) \cdot \Pr(A_2 = a_2 | C = c) \cdots \Pr(A_k = a_k | C = c)$$

and each product above is estimated from the data simply as

$$\Pr(A_j = a_j | C = c) = \frac{count(A_j = a_j \wedge C = c)}{count(C = c)}$$

As can be seen, the relevant conditional probabilities are computed easily from the learning dataset.

In contrast, the genetic algorithm uses rules as its representation, where the left-hand side is a conjunction of Boolean expressions and the right-hand side is the predicted class. For example, a discovered rule is of the form:

> "if the ER index is greater than the Pth *percentile*, and
> the *N day industry trend* exceeds *S standard deviations*, then
> the company will deliver *an earnings surprise of type T*."

Given this type of rule template, a discovery algorithm would attempt to fill in the blanks denoted by the italicized phrases. For example, it might find the best results occur when P is 90, when N is equal to 60, S is 1.5, T is positive, and so on.

As we can see, the search space of possible rules is extremely large even for this trivial example with only a few variables. It should also be apparent that the representation used by the discovery algorithm can deal easily with *inequality* conditions such as "at least 2%," "between 2 and 5%," "less than 2 or greater than 10," and so on. In other words, it does not require us to discretize the data into buckets *a priori* like the Naïve Bayes algorithm which is suitable for dealing with equality conditions only.

Table I shows the representation of patterns used by the genetic learning algorithm. Each pattern, a *chromosome*, represents a specific rule. The genetic algorithm works with multiple hypothesized rules, which make up its *population*. A population at any point in the search is a snapshot of the solutions the algorithm is considering. The algorithm iterates, modifying its population with each *generation*. Each pattern (chromosome) is an expression defined over the problem variables and constants using the relational operators *equal, greater-than,* and *less-than,* and the Boolean operators *and, or,* and *not*. At the implementation level, chromosomes are queries issued to a database. Chromosomes in turn are composed of constraints defined over individual problem variables. These are represented as sets of *genes*. At the lowest level, a gene represents the smallest element of a constraint, namely, a variable, constant, or an operator.

The above representation is equivalent to that of tree induction algorithms such as CART in that a chromosome (rule) is equivalent to a path from a root to a terminal node in a decision tree. In addition, however, a constraint on a single variable can be a disjunct, such as "$MA_{30} < 10$ *OR* $MA_{30} > 90$." It should be noted that our system *does not* represent knowledge in the manner commonly associated with *genetic classifier systems* [44], [48] where individual chromosomes may represent subcomponents or interim results that make up some larger chain of reasoning represented by groups of chromosomes. We view genetic search simply as an alternative algorithm for searching the rule space—one with particular, attractive properties that we discuss more fully in the results section.

For determining fitness, chromosomes can be evaluated based on criteria such as entropy, support, confidence, or some combination of such metrics. By controlling the numbers of constrained variables in chromosomes, genetic search can be biased to look for patterns of a desired level of specificity. This is a parameter that we can manipulate to control (to some extent) the degree of variable interactions, or nonlinearity, we want the algorithm to be capable of discovering in the data.

While the recursive partitioning and GA classifiers are equivalent in terms of how they impose splits on the data, one of the major differences is in how they search. The most important difference is that the GA is *multivariate* in nature in that it can try several splits at the same time instead of relying on a greedy heuristic that results in an irrevocable split. This is how the chromosome is able to evaluate an entire path in a decision tree from root to leaf node, in effect imposing splits on variables in parallel instead of sequentially. By manipulating hundreds of these at the same time and exchanging pieces of them, the GA is able to conduct a more thorough search. This additional power in search occurs at the expense of speed. For more details and discussion of the genetic algorithm, see [30].

TABLE I
THE CONCEPT CLASS REPRESENTATION

Concept	*Representation*
Variable|Constant|Operator Example: *30-day-moving average of price (MA_{30})*	Gene
Univariate predicate (Single "conjunct") Example: *$MA_{30} > 10$* Example: *$MA_{30} < 10$ OR $MA_{30} > 90$*	Set of Genes
Multivariate predicate (Conjunctive pattern) Example: *$MA_{30} > 10$ AND $MA_{10} < 5$*	Chromosome
Multiple Patterns	Population

V. RESULTS

We report the results from two sets of experiments that were run on the test set consisting of 12 164 records. In the first set of experiments we randomly selected roughly half the companies for predicting surprise 20 days before the earnings announcement, using forecast error as the dependent variable. We repeated these experiments 50 times using random subsets of companies in order to generate each method's performance distribution.

In the second set of experiments we varied the forecast day to be ten, 20, 30, and 40 days before the earnings announcement and chose two dependent variables: forecast error, and the risk-adjusted return from the forecast day to the earnings announcement day. The main objective of these experiments was to compare each method's average performance over problems of varying difficulty.

Dependent variables were preprocessed into 3 bins. A dependent variable was assigned to value 1, 2, or 3 if its Z score, or normalized value, was less than -0.5, between -0.5 and 0.5, or greater 0.5, respectively. We categorized the dependent variables because Naïve Bayes does not handle continuous values.

A. Results on 20-day FE

The testing data set for these experiments had 12 164 records. The dependent variable, FE, was categorized into three bins. Bin 1, representing negative earnings surprise, had 1624 records, or 13.35% of the data. Bin 3, positive earnings surprise, had 1546 records, or 12.71% of the data. Bin 2 was no surprise and contained 8994 records, or 73.94% of the data. In other words, the unconditional distribution (or priors) was roughly 1:6:1. We focused only on bin 1 and bin 3 forecasts since they represent earnings surprises. All results below are reported on the test data.

Fig. 6(a) and (b) shows the means and standard deviations of the predictive models from each of the methods on the testing data for negative surprises. The bars in Fig. 6(a) add up to the total percentage of negative forecasts made by each method. For example, the third set of columns in Fig. 6(a), for CART, add up to roughly 6%, which is the sum of CART's forecasts of negative earnings surprises. Of these, the actual distribution was 2.58% in bin 1 (correct), 2.02% in bin 2, and 1.38% in bin 3 (Table II). If we assume that the serious errors are those where negative forecasts are actually positive and vice-versa (ignoring neutral actuals), we would like to see this ratio of the first to the third bar be as high as possible.

There are three commonly used criteria to evaluate models: accuracy, cover ratio (also called support), and variance. High accuracy means that the percentage of correct forecasts is high. A high cover ratio means that the model applies to a larger proportion of the database. A low variance shows that the forecast results are similar across experiments. From the above figures, we can see that linear regression gives a very low support for negative earnings surprise. The neural network and CART have slightly higher support, but less than half of that produced by Naïve Bayes and the GA. Compared to NB, the GA has higher support and comparable accuracy for negative surprises.

Fig. 6(c) and (d) display the means and standard deviations of the models on testing data when the forecast is a positive surprise. The linear regression method fails to forecast bin 3 records because it predicts the majority class in this case. ANN and CART again have low support, about half that of NB. The GA has a much higher support than the other methods and much higher accuracy as well, although it has a high misclassification error on bin 2.

It is instructive to check whether the forecasts of these methods are any better than pure guessing. The chi-square test is suitable for testing this hypothesis. Let us first consider CART's forecasts as an example to test this hypothesis.

Tables II and III show the mean forecasts of CART on FE. For example, CART forecasts 728 records to be bin 1, negative earnings surprise. Of these, 314 records are actually bin 1, 246 records are bin 2, and 168 records are bin 3.

Table IV displays the FE distribution in the testing data. There are 13.35% data in bin 1, 73.94% data in bin 2, and 12.71% data in bin 3.

Table V shows the expected results from pure guessing. If we randomly pick 728 records from the testing data, we expect to get 97 records of bin 1538 records of bin 2, and 93 records of bin 3. Similarly, if we randomly select 206 records from the testing data, we expect to get 28 records of bin 1, 152 records of bin 2, and 26 records of bin 3.

To test the hypothesis, we used Table V as the expected results of pure guessing and Table III as the observed results of CART's forecasts, ignoring the middle column. The chi-square value of CART's forecasts on negative earnings surprises is 704.42 and on positive earnings surprises is 173.54. From the chi-square distribution table, the probability that the chi-square value is greater than 9.2 due to chance alone is 1% and the probability that the chi-square value is greater than 19 due to chance alone is equal to 0.01% at 2 degrees of freedom. Clearly, the values of

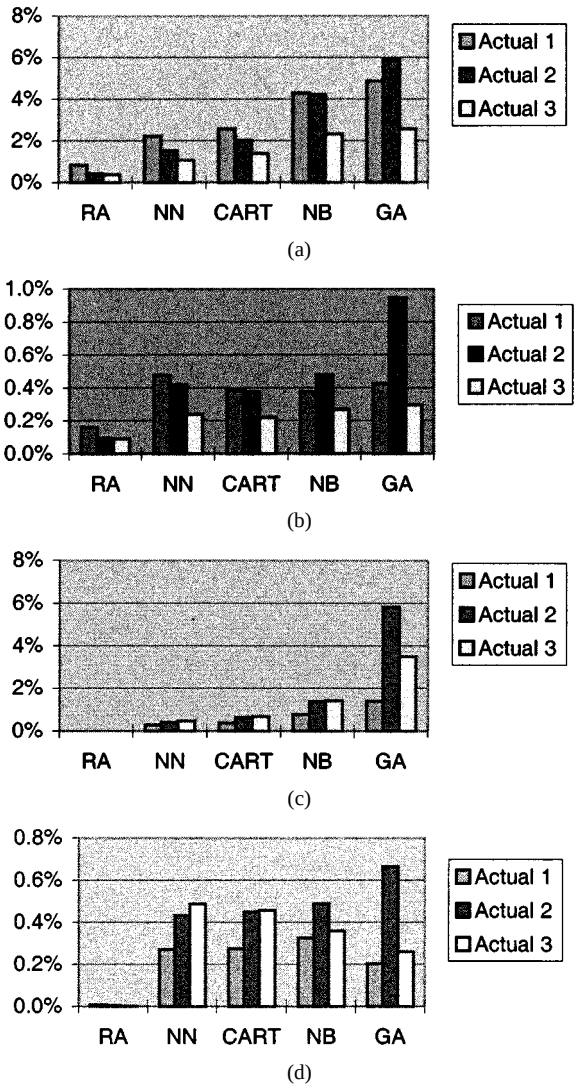


Fig. 6. (a) Forecast mean, negative surprise. (b) Forecast standard deviation, negative surprise. (c) Forecast mean, positive surprise. (d) Forecast standard deviation, positive surprise.

TABLE II
CART'S FE FORECAST IN %

		Actual			
		1	2	3	total
Forecast	1	2.58%	2.02%	1.38%	5.98%
	3	0.38%	0.63%	0.68%	1.69%

704.42 and 173.54 are strongly suggestive that CART performs significantly better than random guessing.

The other Chi-square values are as follows: RAs are 289 for negative and 0 for positive; NNs are 652 for negative and 127 for positive; NBs are 979 for negative and 342 for positive; GAs are 900 for negative and 477 for positive. In summary, all are significant, except for linear regression on positive surprises, where it makes no predictions.

While the nonlinear methods all provide significant results, there is substantial difference between how often the various

TABLE III
CART'S FE FORECASTS IN # OF RECORDS

		Actual			
		1	2	3	total
Forecast	1	314	246	168	728
	3	46	77	83	206

TABLE IV
ACTUAL FE DISTRIBUTION

	1	2	3	total
%	13.35%	73.94%	12.71%	100.00%
# records	1624	8994	1546	12164

TABLE V
PURE GUESSING RESULT DISTRIBUTION

	1	2	3	total
Guess 728 records	97	538	93	728
Guess 206 records	28	152	26	206

TABLE VI
RA'S FE FORECASTS IN # OF RECORDS

		Actual			
		1	2	3	total
Forecast	1	101	51	46	198
	3	0	0	0	0

TABLE VII
NB'S FE FORECASTS IN # OF RECORDS

		Actual			
		1	2	3	total
Forecast	1	522	512	282	1316
	3	96	168	172	436

TABLE VIII
NN'S FE FORECASTS IN # OF RECORDS

		Actual			
		1	2	3	total
Forecast	1	270	186	129	585
	3	34	50	57	141

TABLE IX
GA'S FE FORECASTS IN # OF RECORDS

		Actual			
		1	2	3	total
Forecast	1	592	718	313	1623
	3	170	707	426	1303

methods make forecasts. In this respect, the GA does a lot better than other methods. Our interpretation is that the search space consists of several disjoint areas that provide a desirable distribution of the dependent variable and the genetic search manages to find more of these areas than the other methods. It therefore manages to find more combinations of variables where it is worth making a positive or negative surprise forecast.

B. Results on Accuracy and Support with 10 to 40-Day Forecasts

In this set of experiments we varied the forecast day to be ten, 20, 30, and 40 days before the earnings announcement. To compare the methods, we used the following evaluation functions that take into account both predictive accuracy and coverage of the discovered patterns:

$$\begin{aligned}\text{Score\_bin1} = (&\#\text{forecast\_bin1} - \#\text{forecast\_bin3} \\ &- \#\text{forecast\_bin2} * \text{bin2\_penalty})/ \\ &\#\text{actual\_bin1} \\ \text{Score\_bin3} = (&\#\text{forecast\_bin3} - \#\text{forecast\_bin1} \\ &- \#\text{forecast\_bin2} * \text{bin2\_penalty})/ \\ &\#\text{actual\_bin3}\end{aligned}$$

where #forecast_bin1, #forecast_bin2, and #forecast_bin3 are the actual number of bin 1, bin 2, and bin 3 records in a forecast. The #actual_bin1 and #actual_bin3 are the total number of bin1 and bin 3 records in the data. Let us take CART's forecast on bin 1 of FE 20 days before the announcement as an example. From Tables III and IV, we have the following data:

$$\#\text{forecast\_bin1} = 314, \#\text{forecast\_bin2} = 246,$$
$$\#\text{forecast\_bin3} = 168, \#\text{actual\_bin1} = 1624.$$

If we set the bin2_penalty to be 0.2, Score_bin1 is $(314 - 168 - 246 * 0.2)/1624 = 0.0596$.

This evaluation function takes into account both the support and misclassification across all categories. The more accurate and the higher the support, the higher the score. Also, a zero bin2_penalty indicates that the misclassification error associated with bin 2, the neutral bin, can be ignored, whereas a value of 1 means that the misclassification error associated with bin 2 is as serious as the misclassification error associated with the other bins.

We used 0, 0.1, and 0.2 as three bin2_penalty values. In Fig. 7, eval 1, eval 2, and eval 3 represent the evaluation scores at bin2_penalty of 0, 0.1, and 0.2, respectively.

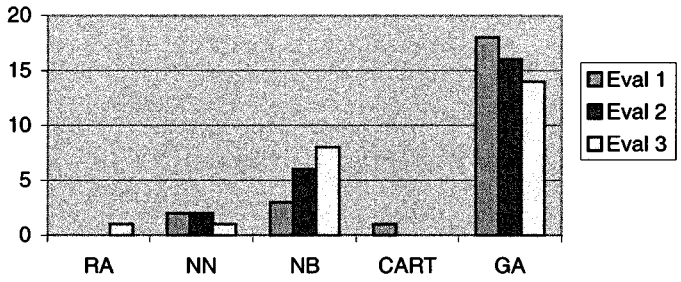


Fig. 7 Number of timesmethod has highest score.

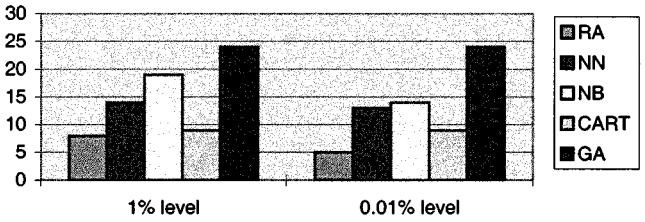


Fig. 8. Number of times method produces statistically significant results.

Fig. 7 shows the number of times that each method has the highest score. It is clear that the GA outperforms other methods on this measure although it is interesting that NB does relatively better when the penalty associated with the neutral bin is higher.

Another way to compare the methods is to count the number of times that each method gives statistically significant forecasts, namely, those where the null hypothesis is violated at the 1% or 0.01% levels using the Chi-square test (again, ignoring bin 2). Fig. 8 shows the number of times that each method gives significant forecasts for two significance levels. Again, the GA significantly outperforms other methods and Naïve Bayes slightly outperforms the neural net, which in turn outperforms CART. Again, the linear model is the worst performer. These results are consistent with the ones in Fig. 7 and earlier in this section.

VI. Discussion

We have offered a preliminary explanation for the GAs ability to find multiple sweet spots in a search space. We now use two nonlinearities in the earnings surprise prediction problem to offer a fuller and more concrete explanation for the observed results. We discovered these nonlinear relationships by examining the outputs produced by the genetic algorithm. Similar relationships have been observed in the literature. We consider some of their interactions and discuss why a method that produces "multiple small disjuncts" can be effective in revealing interesting nonlinear relationships.

We discovered an interesting relationship between earnings forecast error and company size. Specifically, smaller companies tend to have larger *absolute* earnings surprises than bigger firms. One reason for this could be that larger companies have more analysts covering them, so there is "more information" available about them. Hence earnings forecasts for them should be more reliable than those for smaller companies. This result is in accordance with prior research, where for example, [6] argues that the amount of private predisclosure information production and dissemination is an increasing function of the firm size. Furthermore, [7] shows that the amount of "unexpected" information conveyed to the market by actual earnings reports is inversely related to firm size, other things being equal.

Fig. 9 shows this effect in terms of the distribution of FE across different firm size quintiles. The five bars correspond to the sizes of the surprises, with FEQ1 being the largest negative and FEQ5 being the largest positive surprise.

Notice how the distribution in SizeQ1 (smallest companies) is U-shaped, meaning that the smallest companies have the largest negative and positive surprises, with somewhat more negative (FEQ1) than positive (FEQ5) surprises. This gradually shifts into a flat distribution for the middle size quintile into an inverted U-shape for the largest quintile, meaning that the largest companies have the fewest extreme surprises, and in this case, more positive than negative ones.

Suppose we used the following linear form to model the above relationship:

$$FE = A_1 + B_1 * SIZE. \tag{20}$$

It would not work. The more appropriate form would be

$$|FE| = A_1 - B_1 * SIZE \tag{20A}$$

which states that the *absolute* value of FE and firm size have linear relationship.

Although the nonlinear relationship between FE and firm sizes can be modeled using several "piecewise linear" functions, prior knowledge of this relationship is needed to set up the model correctly. We would prefer a method that *finds* the relationship for us instead of requiring us to know it. The GA and CART do this for us, while the other learning methods require more inspection and analysis of the outputs for comparable insight into the problem.

Consider a second example, namely, the forecast error this quarter, FE, as a function of last quarter's forecast error, FEqn1. Fig. 10 shows that on average the most negative surprise companies last quarter (FEqn1Q1) surprise negatively again this quarter as represented by the leftmost bar, FEQ1. This effect is also known in the literature. However, the interesting thing is that this relationship is nonlinear since many of the largest negative surprises from last quarter become the largest positive surprises this quarter. A similar U-shaped distribution is also observed for the symmetric case evident in the rightmost set of bars of Fig. 10: some of the largest positive surprises from last quarter give the largest negative surprises this quarter.

Equation (21), produced by linear regression on the data states that the forecast error, FE, increases when FEqn1 and/or Size increase

$$FE = 2.1824 + 0.23453 * FEqn1 + 0.041691 * Size. \tag{21}$$

This equation shows a directionally reasonable relationship in the sense that on average FE increases when FEqn1 and Size increase. However, (21) does not capture the fact that the variance of FE is large when Size = Q1 and the variance of FE is small when Size = Q5. Nor does it recognize the fact that the variance of FE is large when FQqn1 is in the extreme quintiles

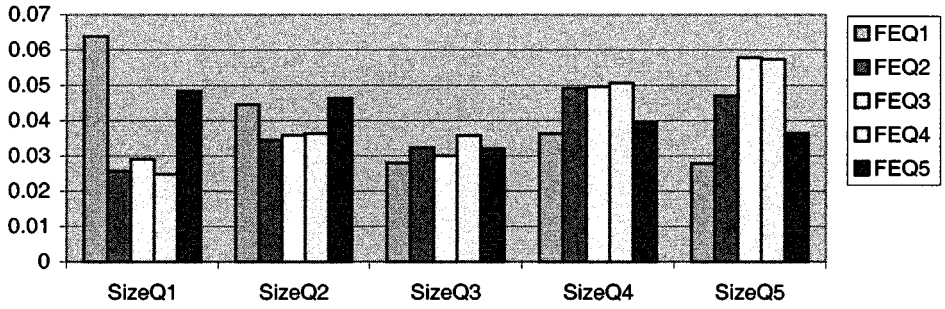


Fig. 9. Forecast error versus firm size (1986–1997).

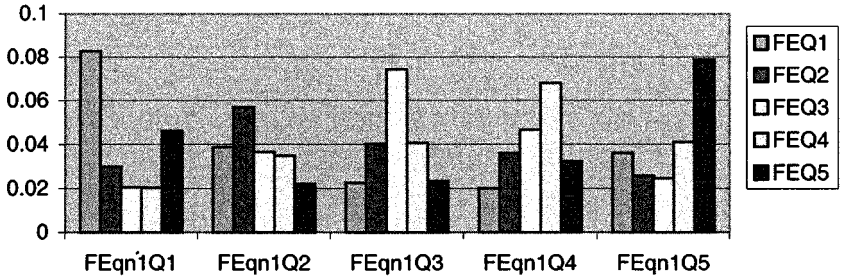


Fig. 10. Current forecast error versus last quarter's forecast error (1986–1997).

as discussed earlier. It is not surprising that it does poorly at predicting FE in general.

The GA does well because it is able to find the "small disjuncts" corresponding to positive and negative surprises. Indeed, the relationships above became apparent in examining the rules produced by the GA. These rules looked something like:

> IF FEqn1 is highly negative and Size is small → FE is negative,
>
> IF FEqn1 is highly positive and Size is small → FE is positive,
>
> If FEqn1 is highly negative and CFROI is negative → FE is negative

The first rule says that highly negative surprises are followed by a negative surprise this quarter *for small companies*. The second rule says that a previous extreme positive surprise *for a small* (probably growth) *company* is followed by a positive surprise. The third rule says that a negative surprise last quarter for a company is likely to be followed by another negative surprise *for companies with weak cash-flow based return on investment*.

The GA has a major advantage over competitors such as CART that also produce explicit rules. CART works by imposing successive splits on the data, resulting in a tree where each path from root to leaf is a rule. But there are two main drawbacks of imposing successive splits in the data. First, the algorithm is not able to recover from a suboptimal split earlier in the tree. Equally importantly, since each split results in smaller subsets of the data, the variance of the dependent variable within each cluster gets larger. In contrast, the GA examines an entire path from leaf to terminal node as a rule. In essence, it searches for splits in parallel and equally importantly, combines parts of entire paths with others to search for better overall solutions.

The search in the GA is targeted to find explicit rules that provide a high fitness, meaning low/high means and low variance on the dependent variable. The GA focuses naturally on the more interesting cases such as highly positive or negative forecast errors, searching for the conditional distributions that are different from those of the overall data. In this way, it can quite easily find the above rules. The advantage of the GA over Naïve Bayes is that it shows the rules explicitly instead of requiring us to construct a theory by examining and interpreting conditional probability distributions. Interestingly enough, once we knew about the rules, we could verify them roughly by examining the distributions produced by Naïve Bayes.

Finally, it is also worth remarking that methods such as the GA are particularly useful when the objective is to make predictions only *some* of the time and be agnostic otherwise. The method is effective at finding several "small disjuncts" where each disjunct (rule) only covers a small range of conditions. In our representation language, these small disjuncts are hypercubes. If the algorithm is able to find enough such hypercubes, it can cover a reasonable amount of the search space. Clearly, in our problem, it was able to cover more regions of the search space than the other techniques at comparable levels of accuracy.

REFERENCES

[1] J. S. Abarbanell, "Do analysts' earnings forecasts incorporate information in prior stock price changes?," *J. Accounting and Economics*, vol. 14, pp. 147–165, 1991.

[2] J. S. Abarbanell and V. L. Bernard, "Test of analysts' overreaction/under-reaction to earnings information as an explanation of anomalous stock price behavior," *J. Finance*, vol. 47, no. 3, pp. 1181–1207, 1992.

[3] S. B. Achelis, *Technical Analysis From A to Z*. Chicago, IL: Irwin, 1995.

[4] A. Alford and P. Berger, The Association Between Analysts' Underreaction to Earnings and Postearnings-Announcement Drift, 1997.

[5] Asquith, Paul and David W. Mullins, Jr., "Equity issues and offering dilution," *J. Financial Economics*, vol. 15, pp. 61–89, 1986.

[6] A. Atiase and R. Kwame, "Predisclosure Informational Asymmetries, Firm Capitalization, Financial Reports, and Security Behavior," Ph.D. dissertation, Univ. California, Berkeley, CA, 1980.

[7] ——, "Predisclosure information, firm capitalization, and security price behavior around earnings announcements," *J. Accounting Res.*, vol. 23, no. 1, 1985.

[8] R. Ball and E. Bartov, "An empirical evaluation of accounting numbers," *J. Accounting Res.*, vol. 6, pp. 159–178, 1968.

[9] ——, "How Naïve is the stock market's use of earnings information?," *J. Accounting Economics*, vol. 21, no. 3, 1996.

[10] R. Banz, "The relation between return and market value of common stocks," *J. Financial Economics*, vol. 9, pp. 3–18, 1981.

[11] *Risk Model Handbook*, 1998.

[12] S. Basu, "The investment performance of common stocks in relation to their price to earnings ratios: A test of the efficient market hypothesis," *J. Finance*, vol. 32, pp. 663–682, 1997.

[13] V. Benard and J. Thomas, "Post-earnings announcement drift: Delayed price response or risk premium?," *J. Accounting Res.*, vol. 27, pp. 1–36, 1989.

[14] ——, "Evidence that stock prices do not fully reflect the implications of current earnings for future earnings," *J. Accounting Economics*, vol. 13, pp. 305–340, 1990.

[15] R. Bhushan, "An informational efficiency perspective on the post-earnings announcement drift," *J. Accounting and Economics*, vol. 18, pp. 45–65, 1994.

[16] M. E. Bradbury, "Voluntary semiannual earnings disclosures, earnings volatility, unexpected earnings, and firm size," *J. Accounting Res.*, vol. 30, no. 1, Spring 1992.

[17] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth, 1984.

[18] H. M. Cartwright and G. F. Mott, "Looking around: Using clues from the data space to guide genetic algorithm searches," in *Proc. 4th Int. Conf. Genetic Algorithms*, 1991.

[19] J. Campbell, Y. Lo, W. Andrew, and A. C. MacKinlay, *The Econometrics of Financial Markets*. Princeton, NJ: Princeton Univ. Press, 1997.

[20] A. E. Chambers and S. H. Penman, *Journal of Accounting Research*, vol. 22, no. 1, Spring 1984.

[21] V. V. Chari, R. Jagannathan, and A. R. Ofer, "Seasonalities in security returns," *J. Financial Economics*, vol. 21, pp. 101–121, 1988.

[22] D. Collins and S. P. Kothari, "An analysis of the intertemporal and cross-sectional determinants of the earnings response coefficients," *J. Accounting Economics*, vol. 11, pp. 143–181, 1989.

[23] G. Connor, "The three types of factor models: A comparison of their explanatory power," *Financial Anal. J.*, May/June 1995.

[24] G. Connor and R. Korajczyk, "Risk and return in an equilibrium APT: Application of a new test methodology," *J. Financial Economics*, vol. 21, pp. 255–290, 1988.

[25] J. Conrad and G. Kaul, "Long-term market overreaction or biases in computed returns?," *J. Finance*, vol. XLVIII, no. 1, 1993.

[26] DAIS, "Forecast risk factor model: Overview and backtests," in *Quantitative Investment Analytics*: The DAIS Group, Aug. 1997a.

[27] ——, "Global estimate revision model: Overview and backtests," in *Quantitative Investment Analytics*: The DAIS Group, June 1997b.

[28] L. A. Daley, J. S. Hughes, and J. Rayburn, "The impact of earnings announcements on the permanent price effects of block trades," *J. Accounting Res.*, vol. 33, no. 2, Autumn 1995.

[29] V. J. Defeo, "An empirical investigation of the speed of the market reaction to earnings announcements," *J. Accounting Res.*, vol. 24, no. 2, Autumn 1986.

[30] V. Dhar, D. Chou, and F. Provost, "Discovering interesting patterns in investment decision making with GLOWER: A genetic learner overlaid with entropy reduction," *J. Data Mining Knowledge Discovery*, Oct. 2000.

[31] V. Dhar and R. Stein, "Intelligent decision support methods," in *The Science of Knowledge Work*. Upper Saddle River, NJ: Prentice-Hall, 1997.

[32] P. Domingos and M. Pazzani, "Beyond independence: Conditions for the optimality of the simple Bayesian classifier," in *Proc. 13th Int. Conf. Machine Learning*. San Mateo, CA, 1996, pp. 105–112.

[33] A. Dontoh, J. Ronen, and B. Sarath, Postannouncement Drift in Rational Expectations Models, Ross Institute Working Paper Series ACC-97–4, Stern School of Business, 1997.

[34] P. Elgers and D. Murray, "The relative and complementary performance of analyst and security-price-based measures of expected earnings," *J. Accounting Economics*, vol. 15, pp. 303–316, 1992.

[35] C. Elkan, *Boosting and Naïve Bayesian Learning*: KDD CUP Entry, 1997.

[36] E. Fama and K. French, "Common risk factors in the returns on stocks and bonds," *J. Financial Economics*, vol. 33, pp. 3–56, 1993.

[37] G. Foster, C. Olsen, and T. Shevlin, "Earnings releases, anomalies, and the behavior of security returns," *Accounting Rev.*, vol. 59, no. 4, 1984.

[38] J. Francis and L. Soffer, "The relative informativeness of analysts' stock recommendations and earnings forecast revisions," *J. Accounting Research*, vol. 353, no. 2, Autumn 1997a.

[39] R. N. Freeman and S. Y. Tse, "A nonlinear model of security price responses to unexpected earnings," *J. Accounting Res.*, vol. 30, no. 2, 1992.

[40] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proc. 2nd Europ. Conf. Comput. Learning Theory*, 1995, pp. 23–37.

[41] J. H. Friedman, *Local Learning Based on Recursive Covering*: Dept. Statist., Stanford Univ., 1996.

[42] N. Friedman and M. Goldszmidt, "Building classifiers using Bayesian networks," in *Proc. Nat. Conf. Artificial Intell.*, 1996, pp. 1277–1284.

[43] E. I. George, H. Chipman, and R. E. McCulloch, "Bayesian CART," in *Proc. Comput. Sci. Statist. 28th Symp. Interface*, Sydney, Australia, 1996.

[44] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

[45] R. L. Hagerman, M. E. Zmijewski, and S. Pravin, "The association between the magnitude of quarterly earnings forecast errors and risk-adjusted stock returns," *J. Accounting Res.*, vol. 22, no. 2, Autumn 1984.

[46] J. Hekanaho, "Background knowledge in GA-based concept learning," in *Proc. 13th Int. Conf. Machine Learning*, 1996.

[47] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: Univ. Michigan Press, 1975.

[48] ——, "Adaptation in natural and artificial systems," in *An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge, MA: MIT Press, 1992.

[49] ——, *Hidden Order: How Adaptation Builds Complexity*. Reading, MA: Perseus, 1995.

[50] HOLT, "The CFROI life cycle," *J. Investing*, vol. 5, no. 1, Summer 1996.

[51] N. Jegadeesh and S. Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *J. Finance*, vol. 48, pp. 65–91, 1993.

[52] S. A. Kauffman, "Whispers from Carnot: The origins of order and principles of adaptation in complex nonequilibrium systems," in *Complexity: Metaphors, Models, and Reality*, G. A. Cowan *et al.*, Eds. Reading, MA: Addison-Wesley, 1994.

[53] S. P. Kothari and R. G. Sloan, "Information in prices about future earnings," *J. Accounting and Economics*, vol. 15, pp. 143–171, 1992.

[54] B. Lehmann and D. Modest, "The empirical foundations of the arbitrage pricing theory," *J. Financial Economics*, vol. 21, pp. 213–254, 1988.

[55] S. C. Linn and J. J. McConnell, "An empirical investigation of the impact of 'Antitakeover' amendments on common stock prices," *J. Financial Economics*, vol. 11, pp. 361–399, 1983.

[56] J. Lintner, "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets," *Rev. Economics Statist.*, vol. 47, pp. 13–37, 1965.

[57] B. Madden, "The CFROI life cycle," *J. Investing*, vol. 5, no. 1, Summer 1996.

[58] B. Madden and S. Eddins, "Different approaches to measuring the spread of return on capital in relation to the cost of capital," *Valuation Issues*, July/Aug. 1996.

[59] S. W. Mahfoud, "A comparison of parallel and sequential niching methods," in *Proc. 6th Int. Conf. Genetic Algorithms*, 1995.

[60] R. Mendenhall, "Evidence of possible underweighting of earnings-related information," *J. Accounting Res.*, vol. 29, pp. 170–180, 1991.

[61] N. Packard, "A genetic learning algorithm," Univ. Illinois Urbana Champaign, Tech. Rep., 1989.

[62] J. Quinlan, *Machine Learning and ID3*. San Mateo, CA: Morgan Kauffman, 1992.

[63] ——, "Boosting, bagging, and C4.5," in *Proc. Nat. Conf. Artificial Intell.*, 1996, pp. 725–730.

[64] D. B. Parker, "Learning logic," Center for Computational Research in Economics and Management Science, MIT, Cambridge, Tech. Rep. TR-47, 1985.

[65] F. Rossi, C. Petrie, and V. Dhar, On the equivalence of constraint satisfaction problems (a comprehensive and more formal version of proceedings publication #8 below), , 1999.

[66] S. Ross, "The arbitrage theory of capital asset pricing," *J. Economic Theory*, vol. 13, pp. 341–360, 1976.

[67] D. Rumelhart and J. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, I & II*. Cambridge, MA: MIT Press, 1986.
[68] R. Schweitzer, "How do stock returns react to special events?," *Business Rev.*, July/Aug. 1989.
[69] W. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk," *J. Finance*, vol. 19, pp. 425–442, 1964.
[70] D. Shores, "The association between interim information and security returns surrounding earnings announcements," *J. Accounting Res.*, vol. 28, no. 1, 1990.
[71] UCI Repository of machine learning databases, , University of California, Department of Information and Computer Science, Irvine, CA, 1995.
[72] P. Werbos, "Beyond regression: New tools for prediction and analysis in behavior sciences," PhD. dissertation, Harvard Univ., Cambridge, MA, 1974.
[73] R. L. Kennedy, Y. Lee, B. Van Roy, C. Reed, and R. Lippman, *Solving Data Mining Problems Through Pattern Recognition*. Upper Saddle River, NJ: Prentice-Hall, 1998.

Vasant Dhar has held a number of academic and industry appointments, including the Stern School of Business, New York University, New York, and Morgan Stanley and Company. He has pioneered the use of artificial-intelligence-based machine learning methods in the financial services industry. He has developed a number of published and proprietary algorithms that are suitable for revealing patterns in large amounts of time series data. He has written many articles on the use of artificial intelligence methods in business. He has also authored *Seven Methods for Transforming Corporate Data Into Business Intelligence* (Upper Saddle River, NJ: Prentice-Hall, 1997).

Dashin Chou received the Ph.D. degree from the Stern School of Business, New York University, New York, in 1999.

He was with Morgan Stanley and Company in information technology, where he implemented and tested a number of machine learning algorithms on financial problems. He is currently a Senior Scientist at Data Mining Systems, New York.