Comput Manage Sci (2008) 5:317–335
DOI 10.1007/s10287-007-0056-x

ORIGINAL PAPER

Using economic and financial information for stock selection

I. Roko · M. Gilli

Published online: 19 April 2007

Abstract A major inconvenience of the traditional approach in portfolio choice, based upon historical information, is its inability to anticipate sudden changes of price tendencies. Introducing information about future behavior of the assets fundamentals may help to make more appropriate choices. However, the specification and parameterization of a model linking this exogenous information to the asset prices is not straightforward. Classification trees can be used to construct partitions of assets of forecasted similar behavior. We analyze the performance of this approach and apply it to different sectors of the S&P 500.

Keywords Portfolio optimization · Decision trees · Factor models

JEL Classification G12 · C35

1 Introduction

The classical approach in portfolio choice balances risk and return in order to determine optimal asset allocations. This approach relies on past information and is generally unable to capture variations in investment opportunities. To overcome this inconvenience one approach is to model the time-varying behavior of mean returns and variances and covariances (e.g., Bollerslev et al. 1988). Another way is to use models

I. Roko (✉)
Department of Econometrics, University of Geneva, Bd du Pont d'Arve 40, 1211 Geneva 4, Switzerland
e-mail: Ilir.Roko@metri.unige.ch

M. Gilli
Department of Econometrics, University of Geneva and Swiss Finance Institute, Geneva 4, Switzerland
e-mail: Manfred.Gilli@metri.unige.ch

where the returns are explained by either statistical, macroeconomic or fundamental factors. Statistical factor models work like a black box and do not allow for an interpretation of the relations in the model. What will be considered here are models where the factors are economic, fundamental and technical variables allowing for interpretation of the results.

Many of the relations between returns and the fundamental factors may exhibit nonlinearities and therefore linear models have to be discarded. Possible approaches are, among other, artificial neural networks and recursive partitioning techniques such as classification trees. We will apply the latter to identify the set of outperforming stocks of several sectors in the S&P 500 index.

Several papers have been published addressing similar problems. Albanis and Batchelor (2000) investigate several linear and nonlinear classification techniques including artificial neural networks and recursive partitioning methods to separate underperforming from outperforming assets. Kao and Shumaker (1999) use classification trees to explain relationships between macroeconomic variables and performance of timing strategies based on market size and style. Sorensen et al. (2000) also use classification trees to identify outperforming stocks to enter a portfolio. In a different context Velikova and Daniels (2004) use classification trees to model housing prices.

The paper is organized as follows: Section 2 describes the factor model and gives a condensed description for the methodology used to build the classification tree; Section 3 presents the application where portfolios composed by the set of outperforming stocks are compared to the market index, and Sect. 4 concludes.

2 Factor models and classification trees

In the proposed factor model the returns for period $t + 1$ on an asset i are explained by characteristics of this asset observed at period t

$$r_{i,t+1} = f(z_{1t}, z_{2t}, \ldots, z_{mt}) \tag{1}$$

where the variables $z_{1t}, \ldots, z_{mt}$ are the factors such as balance sheet average, liquidity, incomes, price-to-earning ratios or earnings growth. To model the nonlinear relations existing between the explanatory variables and the returns we use a classification tree where the returns are assigned to three classes of state: outperforming, neutral and underperforming. For the explanatory variables the classes will be defined by the quartiles.

The methodology for classification trees has been well established by Breiman et al. (1984). The following small example recalls the principles of the procedure for building a classification tree. We consider the observations of returns and two corresponding explanatory variables for nine successive time periods. Returns are assigned to the three classes of state `O`, `N` and `U` and the explanatory variables z_1 and z_2 to the ordered classes 1, 2, 3, 4 and 5.

t	1	2	3	4	5	6	7	8	9
r_{t+1}	O	N	O	U	U	N	O	N	U
z_{1t}	5	1	2	1	2	3	4	4	1
z_{2t}	1	1	2	3	4	5	2	4	5

In the next tables the data are reordered corresponding to the ascending order of the explanatory variables z_1 respectively z_2 and the vertical bars indicate the different possible partitions (splits) of the data.

t	2	4	9	3	5	6	7	8	1
r_{t+1}	N	U	U	O	U	N	O	N	O
z_{1t}	1	1	1	2	2	3	4	4	5

t	1	2	3	7	4	5	8	6	9
r_{t+1}	O	N	O	O	U	U	N	N	U
z_{2t}	1	1	2	2	3	4	4	5	5

All these partitions have to be explored in order to find the one which maximizes the homogeneity of the two sets created. For the sake of brevity the definition of the criterion to be maximized is not given here and the reader is referred to (Martinez and Martinez 2002, p. 346–347). For this example the optimal split is given by $z_1 < 5$. The set verifying $z_1 < 5$ is then partitioned according to $z_2 < 3$ and so on. Figure 1 shows the complete classification tree with the corresponding splitting rules. All terminal nodes contain observations from only one class. Such nodes are also called pure nodes.

Such a complete tree certainly overfits the data and will not generalize well to new observations. Breiman et al. (1984) suggests to find a nested sequence of subtrees by pruning branches. The best subtree of this sequence minimizes the misclassification estimated by ten-fold cross-validation (see Han and Kamber 2001). In the very simple case of our example such a subtree could be obtained by pruning the branches from node (2,3,7). Node (2,3,7) would then become an impure node and we would classify an observation at this node with the class O using the plurality rule. Again the reader is referred to (Martinez and Martinez 2002, p. 352–364) for a detailed description of these procedures.

We can now use this tree to predict future classes of state for the returns, given new observations of the explanatory variables. Thus if we observe $z_{1,10} = 4$ and $z_{2,10} = 2$ we expect the return to fall into class O. Recall that the branches leaving node (2, 3, 7) have been pruned.

An additional technique introduced by Breiman (1996) used to reduce the variance of the predictions is bootstrap aggregation (bagging). The technique consists in

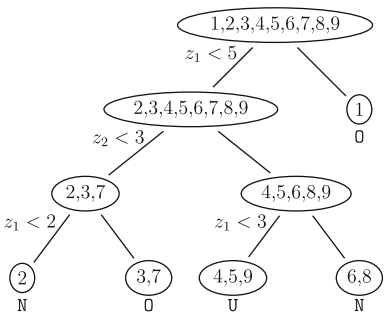

Fig. 1 Complete classification tree

building trees from a certain number of samples obtained by bootstrapping. The final classification model will be an aggregation of the bootstrapped trees. Details of this procedure can be found in (Sutton 2005, p. 318–323).

To evaluate the accuracy of the aggregated classification tree one can measure the percentage of observations that are correctly classified providing an estimation of the probability of correctly classified cases. This enables the use of different statistics, like the hit-ratio or R^2-type measures as well as Pearson chi-square test or versions of likelihood ratio chi-square tests (see Arentze and Timmermans 2003; Ritschard and Zighed 2003). Another possibility is to extend the Hosmer and Lemeshow's goodness-of-fit test (Hosmer and Lemeshow 2000) to ordinal categorical data like in the binary choice models. Our objective is to find the model with the best forecasts and therefore the accuracy will be defined by the performance of the portfolio formed by the predicted outperformers.

3 Application

As mentioned earlier the application mainly yields the identification of assets which are likely to outperform, for the period ahead, the index of the sectors composing the S&P 500. These assets are then chosen to form an equally-weighted portfolio. The selection process and the rebalancing of the portfolio are repeated monthly and its performance is compared to the corresponding index. This is done by applying the methodology described in the previous section. An additional benefit of using classification trees is that the splitting rules, defining the optimal classification tree, reveal valuable information about the driving forces in the market.

3.1 Data set

We consider monthly observations of returns and financial and economic factors from different sectors of the S&P 500 for the period from January 1999 to July 2006 (observations prior to 1999 are incomplete). The composition of the index is as of July 15, 2006. The definition of the sectors is based on MSCIs Global Industry Classification Standards (GICS). The data have been provided by Factset Research Systems Inc.

Our objective is to identify assets with higher returns relative to the other assets. Therefore we consider for a given month the empirical distribution of the returns of the set of all assets in the sector. The lower quantile Q_ℓ corresponding to the $\ell = 0.40$ percentile and the higher quantile Q_u corresponding to the $\ell = 0.60$ percentile provide the boundaries that define the response variable y ($r_{t+1,i} < Q_\ell$ is classified underperformer, i.e. `U`, $r_{t+1,i} > Q_u$ as outperformer `O` and else as neutral `N`). The explanatory variables are discretized according to the quartiles of their empirical distribution.

Sectors are composed by 10 to 87 assets and therefore the number of observations available for the estimation of the tree is insufficient. To overcome this bottleneck Sorensen et al. (2000) suggest pooling the observations for successive months and consider the data as a cross-section. The pooling is justified by the hypothesis that all assets of a sector are driven by the same mechanisms. In order to clarify this procedure

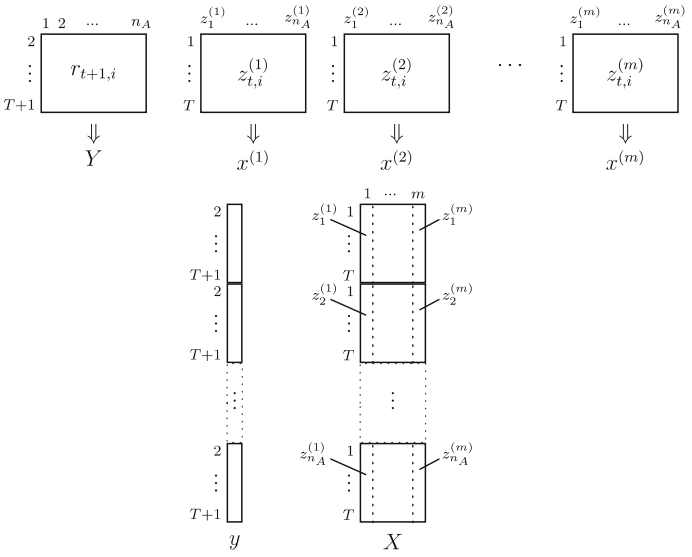

Fig. 2 Data and pooled model

Fig. 2 first represents the data available over the T periods and then shows how the data are organized in the pooled model.

The columns of matrix $[r_{t+1,i}]$ collect the returns at time $t+1$ for the n_A assets of the sector. The columns of matrix $[z_{t,i}^{(k)}]$, $k = 1, \ldots, m$ represent the k-th factor for each asset. The values of each row of matrices $[r_{t+1,i}]$ and $[z_{t,i}^{(k)}]$ are then discretized following the procedure described above. The discretized variables are denoted Y respectively $x^{(k)}$, $k = 1, \ldots, m$ and are reorganized as follows,

$$\underbrace{\text{vec}(Y)}_{y} \simeq f\left(\underbrace{[\text{vec}(x^{(1)})\ \text{vec}(x^{(2)}) \cdots \text{vec}(x^{(m)})]}_{X}\right).$$

Finally the variables y and X are the input for the construction of the classification trees.

3.2 The set of factors

There is some evidence that excess returns can be realized by exploiting links between movements in individual share prices and key accounting ratios. For instance, assets with a low price-to-book ratio or assets with low price-to-earnings tend to outperform the index (see Fama and French 1992 respectively Basu 1977). The literature extensively discusses growth or value strategies defined by value or growth type factors. Some authors (Fama and French 1992; Capual et al. 1993) find that in various periods value stocks tend to exhibit higher returns than growth stocks. Nevertheless, it seems that shifting from one strategy to the other even leads to improved performance

(Kao and Shumaker 1999). Our approach combines the two strategies by introducing variables of growth or value type. The model will then move gradually over time between the different strategies according to the importance of the variables in the splitting rules.

Another strategy largely used in practice relies on technical analysis. This approach uses momentum in price or price patterns to forecast future performance. De Bondt and Thaler (1985, 1987) report that long term past losers outperform past winners over the subsequent three to five years. Jegadeesh (1990) and Lehmann (1990) find short term reversals on returns. George and Hwang (2004) observe that 52-week high price explains the profits from momentum investing.

Various explanations are given to justify the efficiency of these strategies, for instance under- or overreaction of the market to new information (e.g. Jegadeesh and Titman 2001, Chan et al. 1996). The common denominator of these explanations is lack of market efficiency.

We use trend-following indicators like percentage price oscillator (PPO) or relative strength index (RSI). More elaborate technical trading strategies, based on kernel regression can be considered (see Lo et al. 2000).

Hereafter the complete list of the 49 factors used in our model:

- Value and growth factors[1]:
  - Price-to-Earnings (`PE`, `PE1M`, `PE3M`, `PFE`, `PFE1M`, `PFE3M`), value, one and three month changes, trailing and forward.
  - Earnings Momentum (`E1M`, `E3M`, `FE1M`, `FE3M`), one and three month change ratios on trailing and forward earnings.
  - Price-to-Earnings Growth rate (`PEG`, `PEG1M`, `PEG3M`, `PFEG`, `PFEG1M`, `PFEG3M`), value, one and three month changes, trailing and forward. Computed as, (price / $\text{earnings}_{\text{trailing or forward}}$) / annual earnings growth.
  - Price-to-Earnings Growth rate / Dividends (`PEGD`, `PEGD1M`, `PEGD3M`, `PFEGD`, `PFEGD1M`, `PFEGD3M`), value, one and three month changes, trailing and forward.
  - Return on Equities (`ROE`, `ROE1M`, `ROE3M`, `FROE`, `FROE1M`, `FROE3M`), value, one and three month changes, trailing and forward.
  - Price-to-Cash flow (`PCF`, `PFCF`), trailing and forward.
  - Price-to-Sales (`PS`, `PFS`), trailing and forward.
  - Price-to-Book (`PB`, `PFB`), trailing and forward.
  - Debt ratio (`DebtR`, `FDebtR`), computed as, total $\text{debts}_{\text{trailing or forward}}$ / total assets.
  - Net Profit Margins (`NPM`), computed as, net profits after taxes / $\text{sales}_{\text{trailing}}$.
  - Mean Long Term Earnings Growth rate (`MLTEG1M`), one month changes.
  - Dividend Yield (`DIVY1M`, `FDIVY1M`), one month changes on trailing and forward yields.

[1] All forward variables used in this study are 12-month forward consensus estimates provided by *FactSet JCF Estimates database* propriety of *Factset Research Systems Inc.*

- Payout Ratio (PR1M, PR3M, FPR1M, FPR3M), one month and three month changes. Computed as, $\text{dividends}_{\text{trailing or forward}}$ / $\text{earnings}_{\text{trailing or forward}}$.

- Technical or momentum factors:
  - Close Location Value (CLV), computed as,

$$\frac{(C-L)-(H-C)}{(H-L)}$$

    where C is the current price and H and L are the highest, respectively the lowest monthly closing prices for the preceding twelve months.
  - Percentage Price Oscillator (PPO), computed as,

$$\frac{EMA_{\text{short}}(P)-EMA_{\text{long}}(P)}{EMA_{\text{long}}(P)}$$

    where $EMA_{\text{short}}(P)$ and $EMA_{\text{long}}(P)$ are three respectively twelve month exponential moving averages of prices.
  - Relative Strength Index (RSI), defined as,

$$\texttt{RSI}_t = 100 - \frac{100}{1+\texttt{RS}_t} \quad \text{where} \quad \texttt{RS}_t = \frac{\sum_{i=1}^{12} \mathbb{I}_{\{t-i\,|\,r_{t-i}>0\}}\, r_{t-i}}{\sum_{i=1}^{12} \mathbb{I}_{\{t-i\,|\,r_{t-i}<0\}}\, |r_{t-i}|}$$

    and where the indicator function $\mathbb{I}$ defines the *up*, respectively *down* closing months and r_{t-i} is the return.
  - Price Momentum (PMom1M, PMom3M), one and three month changes in prices.

3.3 Tree construction

The process how the market appreciates our predictors is evolutive. For instance, the assets will not outperform always for the same reasons. In order to capture this dynamics we consider a relatively short sample period which is successively moved forward. We tested different lengths of the sample period—called sample window—and found best results for the 12 month length. To clarify the way the backtesting has been carried out Fig. 3 illustrates the successive estimation and prediction steps over time.

The first sample window contains predictors observed from January 15, 2000 to December 15, 2000 and observations on returns from February 15, 2000 to January 15, 2001. At January 15, 2001 this sample window is used to construct the bagged decision tree. Forecasts of the outperforming assets are then obtained using the observations of predictors as of January 15, 2001.

The bagging procedure, in a nutshell, is the following. We bootstrap B samples from our original sample window and construct the corresponding B trees. Then the predictors observed at January 15, 2001 are fitted into each tree and one possible criterion for the final classification of each observation is the class to which the observation has been attributed with the highest frequency. Another possible criterion is to

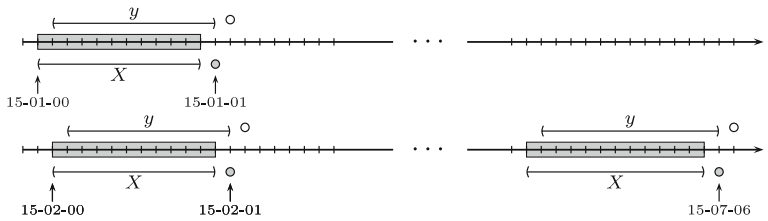

Fig. 3 Displacement of the sample window over the testing period

aggregate (bagg) the probability estimates given by the misclassification rates of each terminal node. Hastie et al. (2001) reports that this criterion tends to perform better for small values of B. We tested both approaches for different values of B and adopted the latter for $B = 50$.

This methodology has been applied to the ten sectors composing the S&P 500. We will first discuss in greater detail the results for the Health-Care sector and then describe more generally the outcomes for the remaining sectors.

Figure 4 reproduces one of the 50 bootstrapped trees obtained for the Health-Care sector using the second sample window. Notice that the splitting rules define the first class of outperforming assets as those for which the price-to-forward sales (PFS) lies in the first four quantiles and the price-to-forward cash flow (PFCF) in the first quantile.

The hierarchy in the tree of the factors in the splitting rules reflects their capability in distinguishing the performance classes of the response variable. The evolution in time of this hierarchy is important to gain insight into the dynamic of the market's driving forces. Recall that the bagging technique constructs for each learning sample $B = 50$ trees, therefore the representation of the hierarchy has been limited to the

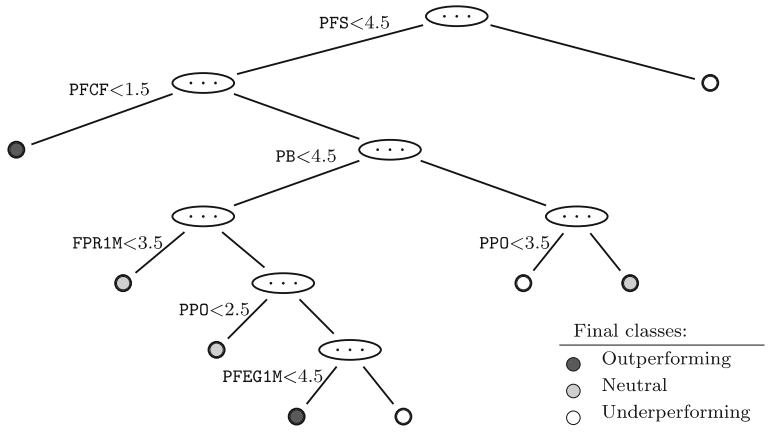

Fig. 4 Particular tree for the Health-Care sector (learning sample 15-02-00 to 15-01-01)

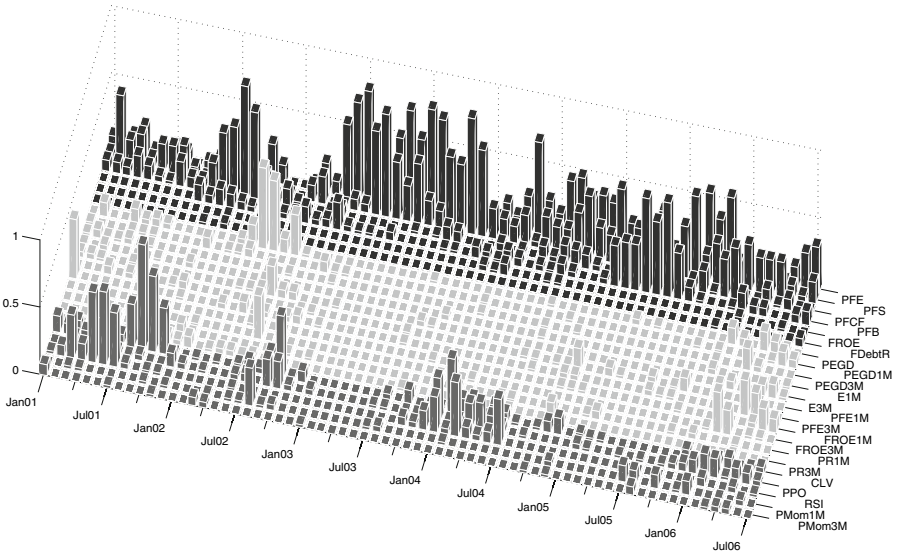

Fig. 5 Frequency of factors appearing in the root node over time (Health-Care sector)

most important ones, i.e. the factors appearing in the root node[2]. Figure 5 reproduces the frequency with which the factors appear in the root node in the successive time periods. Value factors seem to be the most important ones for the Health-Care sector. The mostly used are price-to-forward earnings (`PFE`), price-to-forward sales (`PFS`), price-to-forward cash flow (`PFCF`) and price-to-forward book (`PFB`). Notice how these variables appear and disappear in time and how they are occasionally replaced by growth or technical analysis type factors. Also notice, the periods where technical analysis factors are important follow periods of up or down trends.

This same methodology has also been applied to the other sectors composing the S&P 500 and Table 1 reproduces the factors appearing at each period with the highest frequency in the root node. Hence, this table reveals the most relevant factors for the different sectors and periods.

Recall that the model highlights the characteristics of the outperformers of the past twelve months and assumes that these driving forces remain the same for the month that follows. Changes in the way the market reacts to the factors will be perceived gradually by the model. Typically new factors will enter in the lower part of the decision tree and move toward the root if they remain dominant. Therefore the table only tells when a factor has become predominant in the preceding 12 months. In order to anticipate this information we need to analyze the complete bagged trees.

What follows are a few comments for the sectors where the model performed best. In the Energy sector the most used factors are of value or technical analysis type. The first become important mostly in periods without significant trends. The period from

[2] Another possibility would consist in constructing an index of relative importance of the factors accounting for the position in the tree and the frequency they appear in the 50 bootstrapped trees.

Table 1 Factors appearing in the root node over time, by sector

	Jan01	Feb01	Mar01	Apr01	May01	Jun01	Jul01	Aug01	Sep01	Oct01	Nov01	Dec01
Energy	PFS	PFS	PFS	PFS	PE	PR3M	PMom3M	PEG	PMom3M	PMom1M	PE1M	PE
Materials	PS	PS	PS	PS	PB	FE1M	PFE1M	RSI	FE1M	PS	PCF	PMom1M
Industrials	PS	PS	PCF	PS	PS	PB	RSI	PS	PCF	PCF	PS	PS
Cons.-discret.	PFCF	PFCF	PFCF	PFE	PFCF	PFCF	PFB	PFB	PFB	PFB	PFB	PFB
Cons.-staples	PFS	PFE	CLV	PFS	PFB	PFS	PFS	PFB	PFE	PFB	PFS	PFS
Health-Care	PFE3M	PFS	RSI	PFS	RSI	RSI	RSI	CLV	CLV	CLV	CLV	PFCF
Financials	PFS	PFS	PFS	PFS	PFS	PFS	PFB	PFB	PFS	PFS	PFS	PFS
Info.-tech.	PCF	PCF	PCF	PCF	RSI	PB	PCF	PS	PB	PCF	PE	PS
Telecom.-serv.	PS	PS	PPO	RSI	PFE	CLV	PPO	PB	PPO	PPO	DebtR	PFE
Utilities	PE	PS	PS	PE	PS	PE3M	PE3M	RSI	RSI	RSI	RSI	PE1M

	Jan02	Feb02	Mar02	Apr02	May02	Jun02	Jul02	Aug02	Sep02	Oct02	Nov02	Dec02
Energy	PE	PFB	PE	FROE	FROE	PFS	PFS	PPO	PPO	PPO	PE	PE
Materials	PB	PB	PB	PS	PB	PB	PS	PB	PB	PS	PPO	PS
Industrials	PS	PCF	RSI	RSI	PCF	PPO	PPO	PS	PS	PS	PS	PS
Cons.-discret.	PFB	PFB	PFB	PFB	PFB	PFB	PFB	PFB	PFB	PFB	PFB	PFB
Cons.-staples	PFS	PFE	PFE	RSI	PFB	PFE	RSI	RSI	PFS	PFS	PFS	RSI
Health-Care	PFCF	PFCF	PFCF	PFE	PEGD	PEGD	FROE3M	PEGD	PFE	CLV	PFE	PFE
Financials	PFB	PFS	PFB	PFB	PFB	PFB	PFB	PFS	PFB	PFS	PFS	PFB
Info.-tech.	PFE3M	PS	PS	PCF	PS	PS	PS	PE1M	FROE3M	FROE3M	FROE3M	PB
Telecom.-serv.	PPO	PPO	PPO	PFE1M	PFE1M	CLV	CLV	PMom1M	PFE1M	PMom1M	PFE1M	PR3M
Utilities	PS	PS	PE	PS	PE	PE	PS	PE	PFCF	PFCF	PFCF	PE

Table 1 continued

	Jan03	Feb03	Mar03	Apr03	May03	Jun03	Jul03	Aug03	Sep03	Oct03	Nov03	Dec03
Energy	PPO	PPO	PPO	PFCF	PFS	PFCF	PFCF	CLV	PFS	PMom3M	PMom3M	PMom3M
Materials	PS	PS	PB	PB	PS	PS	PS	PS	PPO	PPO	PPO	PS
Industrials	PB	PS	PS	PS	PS	PS	RSI	PS	RSI	RSI	PS	PE
Cons.-discret.	PFB	PFB	PFB	PFB	PMom3M	PMom3M	PFB	PFB	PFB	PPO	PPO	RSI
Cons.-staples	PFS	PFS	RSI	PFE3M	PFS	PFE	PFE	PFS	RSI	RSI	RSI	RSI
Health-Care	PFE	PFS	PFS	PFE	PFE	PFE	PFE	PFE	PFS	PFS	PFS	PFS
Financials	PFB	PFB	PFS	PFS	PFB	PFB	PFB	PFB	PFS	PFS	PFS	PFS
Info.-tech.	PB	PB	PB	PS	PFE3M	PFE3M	PB	PB	PB	PMom1M	PMom3M	RSI
Telecom.-serv.	PCF	PCF	E3M	PFE	PS	PCF	E3M	PMom1M	RSI	PMom1M	RSI	PPO
Utilities	PE	PFCF	PFCF	PFCF	PFCF	PFCF	PFCF	PFCF	PFCF	PFCF	RSI	RSI

	Jan04	Feb04	Mar04	Apr04	May04	Jun04	Jul04	Aug04	Sep04	Oct04	Nov04	Dec04
Energy	PMom3M	PE	PE	PE	PMom3M	PFS	PFS	PE	RSI	RSI	RSI	RSI
Materials	PFE3M	PMom3M	PMom3M	PMom3M	PMom3M	PS	PS	PS	RSI	PS	PS	PS
Industrials	PE	RSI	RSI	PS	PS	PS	PPO	PPO	PS	PS	PS	PE
Cons.-discret.	PPO	PFE	PFE	PFE	PFE	PFE	PFB	PFB	PFB	PFB	PFB	PFB
Cons.-staples	RSI	RSI	PFE	PFE	CLV	PFE	PFS	PFS	PFE	PFS	PFS	PFS
Health-Care	PPO	CLV	RSI	PFE	PFE	PFE	RSI	PFE	PFE	PFS	PFE	PFE
Financials	PFB	PFS	PFB	PFB	PFB	RSI	RSI	PFS	PFS	PFS	PFB	PFB
Info.-tech.	RSI	PMom1M	PFE1M	PB	PB	PB	PS	PB	PS	PS	PB	PS
Telecom.-serv.	PCF	PMom1M	PFE1M	PCF	PFE3M	RSI	PB	PPO	CLV	CLV	PS	PS
Utilities	PE	PFCF	PFCF	PE	PFCF	PFCF	PFCF	PFCF	PFCF	PFCF	PFCF	PFCF

Table 1 continued

	Jan05	Feb05	Mar05	Apr05	May05	Jun05	Jul05	Aug05	Sep05	Oct05	Nov05	Dec05
Energy	RSI	RSI	RSI	RSI	RSI	RSI	RSI	PMom1M	RSI	PFCF	PFCF	PPO
Materials	PS	PS	PS	CLV	RSI	PS	PS	PS	PS	PS	PEG3M	RSI
Industrials	PS	PS	PE	PS	PS	RSI	PE	RSI	PS	RSI	PS	RSI
Cons.-discret.	PFE	PFE3M	PFE	PFE	PFE	PFE	PFE	PFE	PFE	PFB	PFE	PFE
Cons.-staples	PFS	PFS	PFS	PFS	PFB	PPO	PPO	RSI	PFE	PPO	PFS	PFS
Health-Care	PFS	PFB	PFCF	PFB	PFB	PFB	PFS	PFS	PFE	PFE	PFE	PFCF
Financials	PFS	PFS	PFS	PFS	PFS	PFB	PFS	PFS	PFS	PFS	PFS	PFS
Info.-tech.	PS	PS	PS	PS	PB	PS	PS	PS	PS	PS	PS	PS
Telecom.-serv.	CLV	PFE3M	CLV	E3M	PFE3M	PCF	PPO	PFE3M	PEGD3M	PPO	PPO	PPO
Utilities	PS	RSI	PFCF	RSI	PE	PE	PS	PE	PS	PE	PE	PS

	Jan06	Feb06	Mar06	Apr06	May06	Jun06
Energy	PMom3M	PMom3M	PMom3M	PMom3M	PMom3M	PFB
Materials	PS	PB	PB	PS	RSI	PS
Industrials	RSI	RSI	RSI	PE	RSI	PS
Cons.-discret.	PFE	PFE	PFE	PFE	PFE	PFE
Cons.-staples	PFS	PFS	PEGD1M	PFS	PFS	PFS
Health-Care	PFS	FROE3M	PFS	PFE3M	PFE3M	PFE
Financials	PFS	PFS	PFS	PFS	PFS	PFS
Info.-tech.	PS	PS	PS	PE	RSI	RSI
Telecom.-serv.	PPO	PFE	PPO	PFE	RSI	PFE
Utilities	PFCF	PS	PE	PE	PS	PS

2003 on is characterized by a significant growth of the sector and the trend following indicators become predominant. In the Materials sector the most used factors are price-to-sales and price-to-book. Notice that these factors account for realized numbers and not for forecasted ones. Another characteristic of this sector is the absence of factors based on earnings. Sales also dominate the industrial and consumer staples sectors. As it appears already in Fig. 5 concerning the Health-Care sector the factors based on forward earnings are the most significant of the sector. In the Utility sector price-to-forward cash flow is important between the last quarter of 2002–2004 and is then replaced by price-to-earnings and price-to-sales. Notice that we do not find many growth factors in the table. However, they are present throughout the periods and sectors, but they appear in lower levels of the trees.

3.4 Portfolio construction and transaction costs

Once the set of outperforming assets has been identified for the first period we construct an equally-weighted portfolio and invest an initial wealth of one. In the successive periods the portfolio is rebalanced in order to correspond to the updated set of outperformers. In the literature it has been often argued that good theoretical results are in reality annihilated by transaction costs or spreads. To better adapt our backtests to reality we applied a cost of 10 bp per transaction.

Given the transaction costs the quantities of assets to be sold and bought cannot be determined directly if we want to respect the weights in the portfolio. The procedure for computing these quantities of assets is detailed hereafter. We denote x_{it} the quantity of asset i in the portfolio at time t. The set of indices of assets appearing in the portfolio at time t is $J_t = \{i \mid x_{it} \neq 0\}$ and the nominal value of the portfolio is,

$$v_t = \sum_{i \in J_t} x_{it}\ p_{it} \tag{2}$$

where p_{it} is the price of asset i at time t. The value of the portfolio at time t just before rebalancing is $v_t^- = \sum_{i=1}^{n_A} x_{i,t-1}\ p_{it}$. In the absence of transaction costs, we have $v_t^- = v_t$ and if transaction costs occur they are deducted from the portfolio value,

$$v_t = v_t^- - C_t \tag{3}$$

where the transaction costs C_t are

$$C_t = \sum_{i \in J_{t-1} \cup J_t} \left(v\ |x_{it} - x_{i,t-1}|\ p_{it}\right). \tag{4}$$

As indicated earlier v has been fixed to 10 bp. Replacing (2) and (4) in (3) we get,

$$\sum_{i \in J_t} x_{it}\ p_{it} + \sum_{i \in J_{t-1} \cup J_t} \left(v\ |x_{it} - x_{i,t-1}|\ p_{it}\right) = v_t^- \tag{5}$$

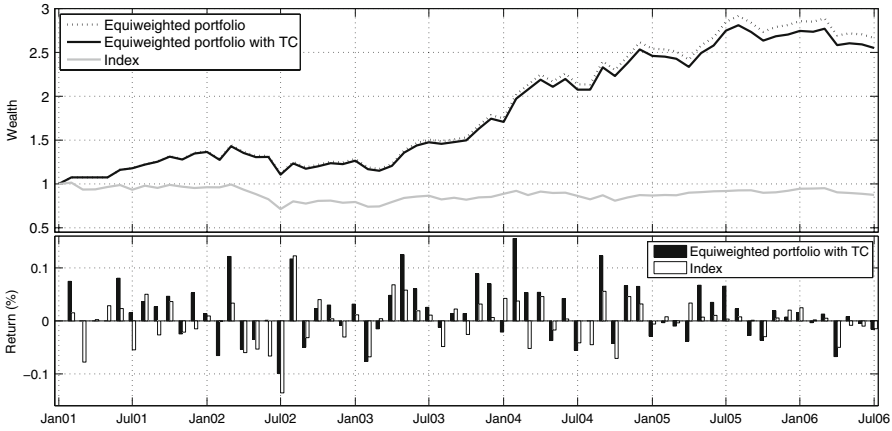

Fig. 6 Performance and returns of the monthly rebalanced portfolio (Health-Care sector)

and considering the portfolio weight constraints,

$$\frac{x_{i_j,t}\ p_{i_j,t}}{v_t} = \frac{1}{\#\{J_t\}} \qquad j = 1, 2, \ldots, \#\{J_t\} - 1 \tag{6}$$

we have a nonlinear system $F(x_{it}) = 0$, $i \in J_t$, given by (5) and (6). The quantities of assets in the rebalanced portfolio are then given by the solution of this nonlinear system.

As previously, we first discuss the results concerning the Health-Care sector and second summarize the findings of the other sectors. The lower panel of Fig. 6 confronts the monthly returns of our equally-weighted portfolio, including transaction costs, to the returns of the index of the Health-Care sector. Notice that in 40 out of 66 periods the model outperforms the index returns. The corresponding evolution of wealth is represented in the upper panel. The transaction costs reduce the overall performance by 11.4 %.

From now on all results include transaction costs. Relative to the index, the overall performance of this portfolio is 168 %, representing an annualized return of 30.5 %. However the performance varies for the different years as shown in Fig. 7. In this figure we also report the cardinality of the portfolios for the successive periods. Notice that in several instances the portfolio is fully invested in cash. An important fact is that the relative performance of the portfolios seem not to be influenced by the market tendency, i.e., whether the market is bullish or bearish.

A weakness of models based on fundamentals is their tendency to be overexposed to subsectors where the fundamentals, e.g., earnings, are higher than those of the other subsectors. A typical example in the Health-Care sector is Biotechnology. In our selection of outperformers this seems to be avoided. Figure 8 represents the subsectorial decomposition of the portfolios in time for the Health-Care sector. Generally, the portfolios are not concentrated on a particular subsector.

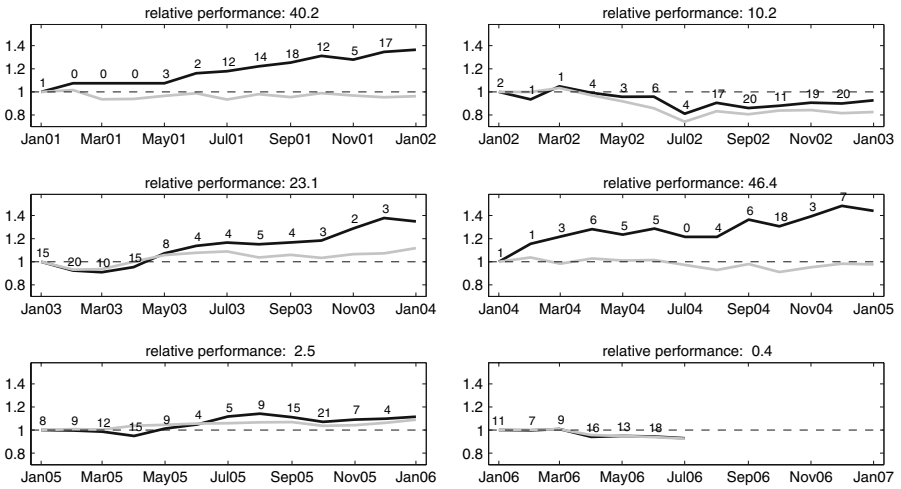

Fig. 7 Yearly performance of the monthly rebalanced portfolio including transaction costs. The index is drawn in *gray lines* and the numbers indicate the varying cardinality of the portfolio (Health-Care sector)

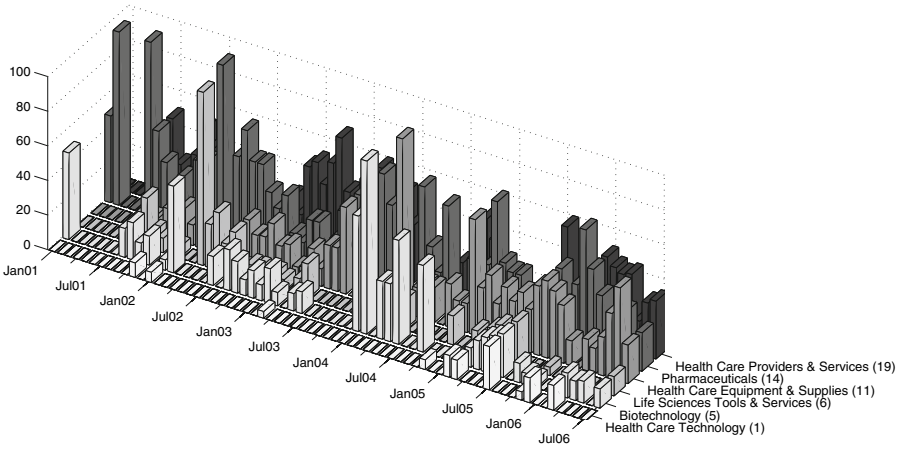

Fig. 8 Evolution of the subsectorial decomposition of the portfolio (Health-Care sector). Cardinalities of subsectors are in parenthesis

We now consider all the different sectors composing the S&P 500. Figure 9 presents the overall performance of the sector portfolios together with the corresponding indexes. The best performing sectors in terms of portfolio performance relative to the index are Energy, Materials, Industrials and Health-Care. Indeed 60–70% of their monthly returns are higher than the index returns. The overall return of the Energy sector is 282% including 10 bp of transaction costs. Relative to the index its performance is 196%, representing a relative annualized return of 32.6%. The absolute returns of Materials and Industrials sector portfolios are 254% respectively 186%, net of transaction costs and their returns relative to the index 172% respectively 163%.

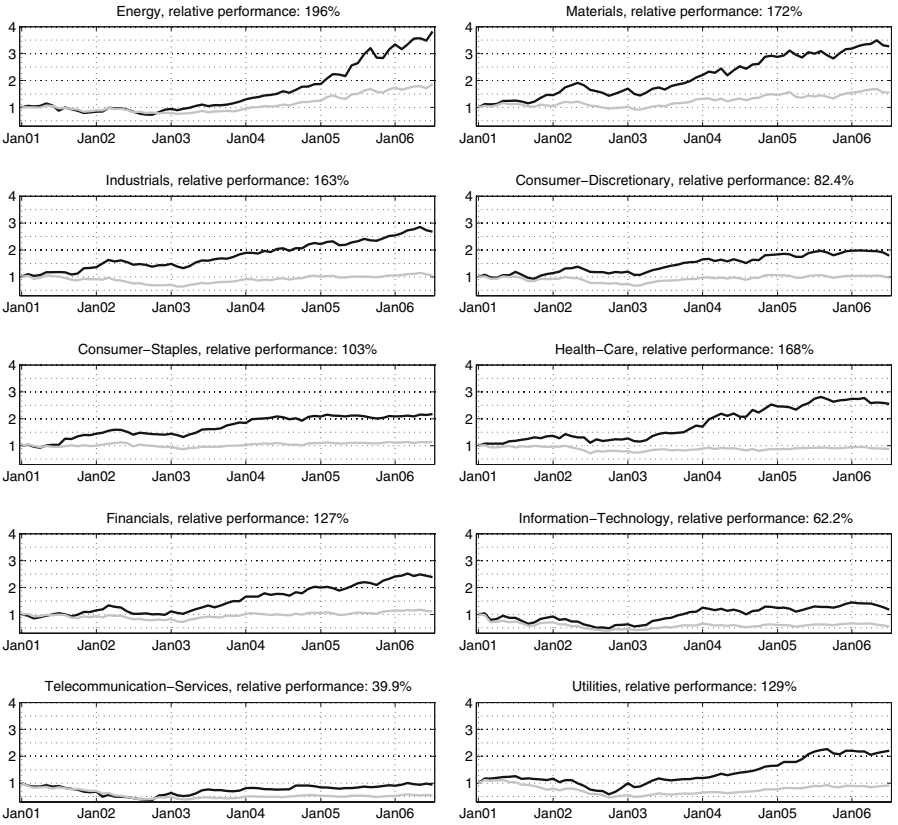

Fig. 9 Wealth evolution for all sectors, with transaction costs

In the Financial and Utilities sectors the portfolios also significantly outperform their relative indexes. Their annualized returns are 23.1% respectively 23.5%.

In order to reveal more information about the evolution of the portfolios in time, Table 2 presents the year by year performances for the different sectors. The table also contains Sharpe ratios based on a risk free rate of 5%. Notice that in our case years are defined from January 15th to January 15th, and therefore, annualized returns refer to this schedule. Highest Sharpe ratios are observed for the Energy, Materials and Industrial sector portfolios. The Sharpe ratios for Consumer-Discretionary, Information-Technology and Telecommunication-Services are the lowest over all sectors. Lack of homogeneity for the first sector and the particularity of the Information-Technology and Telecommunication-Services with respect to the other sectors may be an explanation for their weaker performance. Treating them apart from the point of view of the parameterization of the model might enhance their performance.

3.5 A real world test

The backtesting results presented in the previous section seem to assess the efficiency of our approach. However in reality it is not possible to replicate this policy exactly.

Table 2 Yearly performance of portfolios of outperformers (PO) by sectors and annualized Sharpe ratio

		2001	2002	2003	2004	2005	2006	Sharpe ratio
Energy	PO	−12.1	12.3	37.9	43.9	78.7	11.6	1.15
	Index	−11.0	−11.6	21.2	30.3	39.6	4.4	0.75
Materials	PO	45.6	16.6	29.7	30.6	10.7	1.5	1.21
	Index	3.9	−1.8	28.9	10.8	6.0	−0.7	0.56
Industrials	PO	35.9	9.3	27.5	17.3	14.5	5.2	1.19
	Index	−8.4	−23.2	30.1	10.3	4.1	−0.1	0.13
Cons.-discret.	PO	14.3	4.6	38.0	11.3	7.6	−8.7	0.66
	Index	−7.3	−20.2	32.0	8.8	−2.6	−6.1	0.04
Cons.-staples	PO	44.3	−0.1	27.8	13.1	0.1	3.9	0.99
	Index	0.2	−4.5	6.5	7.7	1.4	2.3	0.27
Health-Care	PO	36.4	−7.3	34.9	44.1	11.6	−6.8	0.98
	Index	−3.7	−17.5	11.8	−2.3	9.0	−7.2	−0.10
Financials	PO	15.2	−3.3	49.1	20.4	20.5	0.9	0.94
	Index	−6.3	−12.1	23.2	3.3	9.2	−2.7	0.20
Info.-tech.	PO	−7.7	−31.7	97.9	−0.9	16.9	−17.3	0.25
	Index	−29.1	−34.3	40.1	−7.6	11.2	−17.1	−0.30
Telecom.-serv.	PO	−29.8	−7.1	28.5	3.2	7.4	5.0	0.11
	Index	−28.8	−30.0	1.6	4.5	−1.4	4.9	−0.45
Utilities	PO	16.9	−14.6	19.5	38.6	33.5	0.8	0.69
	Index	−20.4	−30.2	17.3	18.3	16.8	1.6	0.01

Some of the reasons are that in the backtesting exercise the portfolio is computed and rebalanced using the closing prices, revenues from dividends are not taken into account, and spreads cannot be predicted. Even if we augment the transaction cost to cover the spreads the backtesting will not correspond to reality.

In order to provide some evidence that the model is also likely to perform well in the real world, we invested an initial wealth of 100,000 USD at January 16th, 2006.[3] The transaction costs paid are 0.005 per share with a maximum per order of 20 bp.

The first difference with respect to the backtesting exercise is that the database with the closing prices of the 15th of the month is available at the end of the day. We then execute the procedure for estimating the outperformers and rebalance the portfolio during the 16th day. Evidently, the prices for selling the old positions and buying the new ones do not generally correspond to the closing prices used in the selection of the outperformers. Another constraint is the integer nature of the number of traded assets, implying a loss of equal weight in the positions and thus the holding of some cash. Table 3 presents the performance of the real, backtested and index portfolios for the real test period.

The backtesting results seem to be in agreement with the results of the real world test even if the testing period is relatively short for drawing strong conclusions. Except for the Consumer discretionary and Information-Technology sectors all the other real and backtesting sector portfolios present the same tendencies in the year-to-date results. Generally the differences of the monthly returns are due to the trend and the volatility

[3] We thank Capital Strategy for providing funds and help for this experience and Factset Research Systems for the prompt supply of the data at the rebalancing days.

Table 3 Monthly performance of real, backtested and index portfolio, by sector, as of July 15, 2006

		15Feb06	15Mar06	15Apr06	15May06	15Jun06	15Jul06	Year-to-date
	Real PO	−6.7	7.1	6.0	0.0	−0.8	7.8	13.2
Energy	PO	−5.3	5.6	6.6	0.3	−2.5	7.1	11.6
	Index	−4.5	3.5	3.8	0.3	−4.1	5.8	4.4
	Real PO	2.6	2.2	1.3	2.1	−4.0	−3.0	1.0
Materials	PO	3.1	1.7	0.6	3.9	−5.3	−2.4	1.5
	Index	2.1	3.0	2.3	1.1	−7.5	−1.5	−0.7
	Real PO	3.1	3.6	1.1	3.5	−3.7	−3.2	3.7
Industrials	PO	2.9	4.3	1.5	3.2	−4.4	−2.2	5.2
	Index	1.6	3.5	0.9	3.0	−4.3	−4.5	−0.1
	Real PO	0.7	−1.0	0.4	0.1	−1.8	−4.1	−5.7
Cons.-discret.	PO	0.5	−0.3	−0.7	−0.5	−2.4	−5.5	−8.7
	Index	0.4	0.3	−0.7	1.6	−2.7	−5.0	−6.1
	Real PO	1.0	1.0	−1.1	3.0	−0.3	0.7	4.4
Cons.-staples	PO	0.4	0.9	−1.4	3.4	−0.6	1.2	3.9
	Index	0.3	1.6	−2.9	2.8	−0.7	1.4	2.3
	Real PO	−0.7	1.2	−4.9	0.2	−1.2	−1.8	−7.1
Health-Care	PO	−0.3	1.3	−6.7	0.8	−0.5	−1.2	−6.8
	Index	0.2	0.5	−5.0	−0.8	−1.0	−1.1	−7.2
	Real PO	1.0	3.6	−2.1	1.9	−2.0	−1.2	1.1
Financials	PO	1.4	3.1	−3.7	2.6	−2.1	−0.2	0.9
	Index	−0.8	2.0	−1.4	2.7	−3.8	−1.3	−2.7
	Real PO	−2.3	−0.1	−0.2	−3.8	−5.0	−5.6	−15.9
Info.-tech.	PO	−2.1	−0.5	−0.2	−4.2	−5.4	−6.1	−17.3
	Index	−3.1	1.1	0.0	−5.7	−4.5	−6.0	−17.1
	Real PO	5.0	6.5	−4.1	−1.9	5.3	−4.0	6.4
Telecom.-serv.	PO	4.7	6.5	−4.3	−2.2	4.6	−3.8	5.0
	Index	7.7	3.2	−3.5	−2.0	2.5	−2.6	4.9
	Real PO	−1.2	0.4	−5.3	2.8	2.1	2.1	0.7
Utilities	PO	−1.3	0.1	−5.6	3.1	2.2	2.6	0.8
	Index	−1.8	1.3	−5.8	3.6	2.7	1.8	1.6

of the market during the rebalancing day. Despite the fact that we pay relatively low costs for the transactions and that the spreads are not very important due to the small amount invested, the transaction costs applied in the backtesting seem to be justified.

4 Conclusions

We proposed a methodology for the selection of assets achieving future returns above average using classification trees improved by bootstrap aggregation (bagging). The model relies on factors with growth or value characteristics and technical analysis information. The approach has been applied to forecast outperforming assets for the different sectors of the S&P 500 index from January 2001 to July 2006. We conducted out-of-sample backtests by constructing equally-weighted portfolios composed by the outperforming assets and compared their performance relative to the index. The performance of these portfolios is significantly superior to the indexes even if more elaborate strategies than equal weighting can be implemented to improve the relative

Sharpe ratios. Finally, a test with a real investment has been performed in 2006, which seems to confirm the backtesting results.

References

Albanis G, Batchelor R (2000) Five classification algorithms to predict high performance stocks. In: Dunis C (ed) Advances in quantitative asset management. Kluwer Academic Publishers, Boston, pp 295–318

Arentze T, Timmermans H (2003) Measuring the goodness-of-fit of decision-tree models of discrete and continuous activity-travel choice: methods and empirical illustration. J Geograph Syst 5:185–206

Basu S (1977) Investment performance of common stocks in relation to their price-earnings ratios: a test of the efficient market hypothesis. J Finance 32(3):663–682

Bollerslev T, Engle R, Wooldridge J (1988) A capital asset pricing model with time-varying covariances. J Polit Econ 96(1):116–131

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth, Belmont, California

Capual C, Rowley I, Sharpe WF (1993) International value and growth stock returns. Financ Anal J 49(1): 27–36

Chan LKC, Jegadeesh N, Lakonishok J (1996) Momentum strategies. J Finance 51(5):1681–1713

De Bondt W, Thaler R (1985) Does the stocks market overreact? J Finance 40(3):1681–1713

De Bondt W, Thaler R (1987) Further evidence on investor overreaction and stock market seasonality. J Finance 42(3):557–581

Fama E, French K (1992) The cross-section of expected stock returns. J Finance 47(2):427–465

George TJ, Hwang C-Y (2004) The 52-week high and momentum investing. J Finance 59(5)

Han J, Kamber M (2001) Data mining: concepts and techniques. Morgan Kaufmann, San Francisco

Hastie T, Tibshirani R, Friedman J (2001) Elements of statistical learning: data mining, inference, and prediction. Springer, New York

Hosmer DWJ, Lemeshow S (2000) Applied logistic regression (Wiley series in probability and statistics - applied probability and statistics section), 2nd edn. Wiley–Interscience, New york

Jegadeesh N (1990) Evidence of predictable behavior of securities returns. J Finance 45(3):881–898

Jegadeesh N, Titman S (2001) Profitability of momentum strategies: an evaluation of alternative explanations. J Finance 56(2):699–720

Kao D, Shumaker R (1999) Equity style timing. Financ Anal J 55:37–48

Lehmann BN (1990) Fads martingales and market efficiency. Q J Econ 105(1):1–28

Lo AW, Mamaysky H, Wang J (2000) Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation. J Finance 55(4):1705–1770

Martinez WL, Martinez AR (2002) Computational statistics handbook with MATLAB. Chapman and Hall/CRC, London

Ritschard G, Zighed DA (2003) Goodness-of-fit measures for induction trees. In: Foundations of intelligent systems, 14th international symposium, ISMIS 2003, Maebashi City, Japan, October 28–31, 2003, Proceedings, pp 57–64, Springer, Heidelberg

Sorensen E, Miller K, Ooi C (2000) The decision tree approach to stock selection. J Portfolio Manage 42–52

Sutton CD (2005) Classification and regression trees, bagging and boosting. In: Rao C, Wegman E, Solka J (eds) Handbook of statistics: data mining and data visualization, vol 24. Elsevier, North Holland

Velikova M, Daniels H (2004) Decision trees for monotone price models. Comput Manage Sci 1(3):231–244