Andriyashin, Anton

Working Paper

Stock picking via nonsymmetrically pruned binary decision trees

SFB 649 Discussion Paper, No. 2008,035

Provided in Cooperation with:
Collaborative Research Center 649: Economic Risk, Humboldt University Berlin

Suggested Citation: Andriyashin, Anton (2008) : Stock picking via nonsymmetrically pruned binary decision trees, SFB 649 Discussion Paper, No. 2008,035, Humboldt University of Berlin, Collaborative Research Center 649 - Economic Risk, Berlin

This Version is available at:
https://hdl.handle.net/10419/25277

SFB 649 Discussion Paper 2008-035

Stock Picking via Nonsymmetrically Pruned Binary Decision Trees

Anton Andriyashin*

* Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Stock Picking via Nonsymmetrically Pruned Binary Decision Trees

Anton V. Andriyashin

CASE – Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin,
Spandauer Straße 1, 10178 Berlin, Germany

Abstract

Stock picking is the field of financial analysis that is of particular interest for many professional investors and researchers. In this study stock picking is implemented via binary classification trees. Optimal tree size is believed to be the crucial factor in forecasting performance of the trees. While there exists a standard method of tree pruning, which is based on the cost-complexity tradeoff and used in the majority of studies employing binary decision trees, this paper introduces a novel methodology of nonsymmetric tree pruning called Best Node Strategy (BNS). An important property of BNS is proven that provides an easy way to implement the search of the optimal tree size in practice. BNS is compared with the traditional pruning approach by composing two recursive portfolios out of XETRA DAX stocks. Performance forecasts for each of the stocks are provided by constructed decision trees. It is shown that BNS clearly outperforms the traditional approach according to the backtesting results and the Diebold-Mariano test for statistical significance of the performance difference between two forecasting methods.

JEL classification: C14, C15, C44, C63, G12

Keywords: decision tree, stock picking, pruning, earnings forecasting, data mining

This paper was presented at the *20th Annual Australasian Finance and Banking Conference* in Sydney, Australia in December 2007.

Acknowledgements: The support from DekaBank and the Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko", Humboldt-Universität zu Berlin is gratefully acknowledged.

1 Introduction

Professional capital management involves numerous forms of asset allocation and employment of various financial instruments. Trying to obtain better risk-return characteristics, available funds are frequently invested into different stocks constituting a diversified portfolio. The components of such a portfolio are to be regularly revised, and at this point the individual stock performance is what counts.

There is a lot of research evidence supporting the fact that stock returns can effectively be forecasted – consider, for instance, the studies of Fama and French (1988b) or Keim and Stambaugh (1986). Moreover, as in Fama and French (1988a) and Balvers et al. (1990), it is concluded that predictability is not necessary inconsistent with the concept of market efficiency. Fama (1991) examines the links between expected returns and macro-variables and acknowledges the existence of connection between expected returns and shocks to tastes or technology (changes of business conditions). Chen (1991) continues the work in this direction and concludes the consistency of the link between excess return macro-variables and growth rates of output with intertemporal asset-pricing models.

This applied paper, partly motivated by the valuable collaboration with one top financial services company, focuses on the ability of effective forecasting of future stock price movements based on available market data using the so called *binary decision trees*. Decision trees are a classification method of nonparametric statistics that was introduced in 1980s by a group of American scientists and is thoroughly described in Breiman et al. (1987).

Many studies like Ferson and Harvey (1991) or Campbell and Hamao (1992) employ standard statistical and econometric methods to examine predictability of excess stock returns. However, the special properties of decision trees create notorious distinction among the pool of other available classification techniques. Unlike parametric methods, which are quite sensitive to issues of misspecification, one of the advantages of decision trees (or Classification and Regression Trees – CART – as

they are called alternatively) is the ability to handle specification issues much smoother. Moreover, the nature of the method provides substantial benefits for classification result interpretation, see Breiman et al. (1987) for more details. Steadman et al. (2000) emphasizes the practical importance and flexibility of decision trees in the way that this method poses contingent – and thus, possibly different – questions to classify an object into a given set of classes, while the traditional parametric regression approach employs the common set of questions for each classified object, and final classification score is produced by weighting every answer. Moreover, parametric regression relies on a particular error distribution assumption (e.g. Gaussian model), and decision trees become particularly useful when the data do not meet this assumption (Feldman et al., 1972).

In the recent years several financial services companies (e.g. *JPMorgan* and *Salomon Smith Barney*) showed their interest in applying decision trees for stock picking by issuing a number of press releases for professional investors (Brennan et al., 1999; Seshadri, 2003; Sorensen et al., 1999). The reports provided valuable feedback on the method performance potential when decision trees are applied to the US stock market. This study extends the geography of the method application and focuses on German XETRA DAX stock market.

Decision tree financial applications are not limited solely by the stock selection challenge. In Schroders (2006) the selection of underperforming and outperforming Pan-European banks was achieved with the help of decision trees, and asset allocation to shares, bonds or cash was also derived with the help of CART in Harman et al. (2000).

The majority of studies employing CART uses the industry-standard approach of tree building described in Breiman et al. (1987). However, prior simulations and the architecture of the method have suggested (Kim and Loh, 2001) that due to the specific nature of financial markets, it might be reasonable to change the classical approach and introduce potentially a more effective technique of tree building.

Tree pruning is considered to be the most important step (Breiman et al., 1987) in obtaining a

proper decision tree, which potentially can have various sizes. Overfitting or underfitting directly affect, and affect negatively, the forecasting power of such a decision rule. In Schroders (2006) it is mentioned that the traditional tree pruning approach (Breiman et al., 1987) used by the authors in the past is now substituted with a set of three rules based on different decision tree characteristics. Although these algorithms are not revealed explicitly, this statement creates additional motivation to search for a more effective decision tree pruning technique for financial applications.

The main contribution of this paper is the presentation of the novel methodology of *nonsymmetric decision tree pruning* called *Best Node Strategy* (BNS). While the traditional cost-complexity approach operates only with node triplets when pruning, BNS allows for a more flexible tree optimization and focuses on individual node characteristics rather than an integral measure of quality of a given subtree. The efficiency of the new method is examined on XETRA DAX stock market via backtesting of the stock picking algorithm employing available XETRA DAX company data for the period of 2002–2004. One important theoretical property of BNS is proven, and backtesting results are compared with the similar trading strategy that relies on canonical version of the tree pruning described in Breiman et al. (1987). According to the Diebold-Mariano test, the economic performance difference between the two forecasting methods proved to be significant at the 0.1% confidence level in favor of the novel methodology.

The paper is organized as follows. Section 2 provides a short introduction on decision trees. Section 3 describes the traditional way of optimizing the tree size. Afterwards, BNS is introduced in Section 4 as an alternative to some limitations of the traditional cost-complexity approach. The second part of the work focuses on backtesting: in Section 5 a brief overview of available data and calibration algorithm is provided. Section 6 describes the performance part of the study including the formal statistical testing of the significance of the forecasting performance difference of the two methods via the Diebold-Mariano test. Finally, Section 7 concludes the study. Proofs of some important properties of the proposed tree pruning method are available in the Appendix.

2 Decision Tree Basics

Classification trees are a nonparametric method of data classification. One of its peculiarities is the special form of produced decision rules – binary decision trees. These trees are constituted by *nodes*, and each node carries a "yes-no" question. When new data are to be classified, they are processed by sequential posing of tree questions: left branches stand for positive answers and right branches – for negative ones. Every node of a tree in the bottom has a class tag, in this way classified data are assigned to one of the predefined groups. This type of nodes is called *terminal*.

Figure 1 introduces a simple two-dimensional data structure. Its observations are of one of five predefined classes, which are marked with different colors. Each split clearly separates one homogenous data cluster that constitutes a terminal node with a respective class tag.

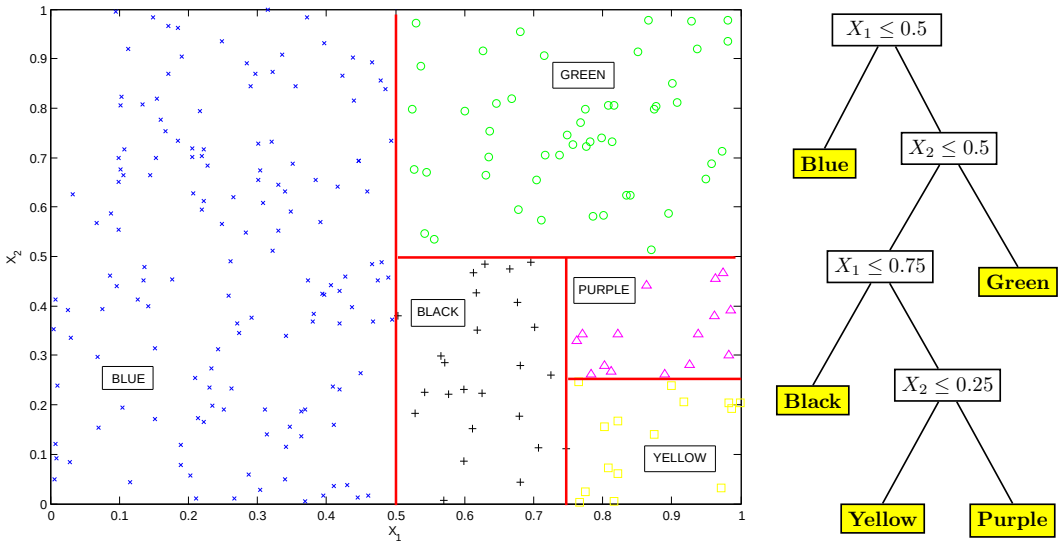

Figure 1: Application of CART to an artificial two-dimensional data set. The *root node* at the top contains a filter $X_1 \leq 0.5$. There are five terminal nodes in this tree and five classes: *blue*, *green*, *black*, *yellow* and *purple*. Left branches stand for positive answers, rights ones – for negative answers

Decision trees can be created from the available data, e.g. data from the past. If a certain link

between some objects is assumed, then the first step to build a tree is to create a *learning sample*. In the framework of stock picking, future stock price fluctuations are assumed to be driven by present changes of fundamental or technical company indicators like Earnings Per Share. Then factors like Earnings Per Share (Cash Flow, Return on Equity, Sales etc.) are grouped into explanatory variable set $X \in \mathcal{R}^P$ (where P is the overall number of explanatory factors) while the target characteristic – the next period stock price yield – is characterized by the class vector Y. The natural range of values $y \in Y$ in this particular case is $\{long, short, neutral\}$ standing for undervalued, overvalued and fairly priced stocks respectively.

The application of decision trees to a data set with observations of unknown class implies three major steps to be conducted:

- construction of the so called *maximum tree* T_{MAX}
- choice of the right tree size (tree pruning) T^*
- classification of new data using the constructed tree T^*

A maximum tree is the one containing observations of the same class at each of the terminal nodes. The *root node* – the one at the top of any tree – resembles the whole learning sample. After that it is being split recursively in a way that more homogenous clusters of observations are separated into tree nodes. This can be achieved as follows.

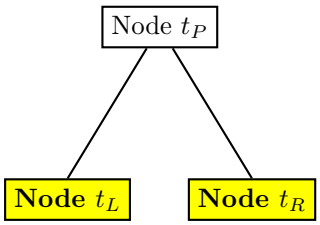

Figure 2: The triplet of nodes: t_P – the parent node, t_L – the left child node and t_R – the right child node

At each node a univariate filter of the form $X_p \leq x$, $p = \overline{1, P}$ (x is some constant) is posed where particular p and x are selected as a result of an optimization procedure to be described later in this section. Let t_P be the parent node and t_L, t_R – the left and right child nodes of the parent node t_P respectively so that a fraction p_L of observations from the node t_P follows to the left child node, and a fraction $p_R = (1 - p_L)$ – to the right one. If n_P is the number of observations in t_P and n_L, n_R – in t_L and t_R respectively, then

$$p_L = \frac{n_L}{n_P}, \quad p_R = \frac{n_R}{n_P} \tag{1}$$

Let the class labels represented by the variable Y be denoted as j. Then the conditional probability of an observation to belong to node t given that its class is j is computed as follows:

$$p(j|\, t) = \frac{n_t(j)}{n_t} \tag{2}$$

i.e. proportion of observations of class j in the node t. It is straightforward that $\sum\limits_{j=1}^{J} p(j|\, t) = 1$ where J is the number of classes in the learning sample. In the described stock picking setup where $y \in \{long, short, neutral\}$, J is equal to three.

A functional that determines the question at each tree node – split s^* – is the maximum value of the one-level decrement of an *impurity function* $i(t)$, which can be defined for an arbitrary node t. Impurity is a measure of class heterogeneity for a given cluster of data (Breiman et al., 1987). One of its important properties is that $0 \leq i(\cdot) \leq 1$. Therefore, one can identify the optimal split s^* for a given node t and $i(\cdot)$ as follows:

$$\begin{aligned} s^* = \underset{s}{\operatorname{argmax}}\, \Delta i(s,t) = \underset{s}{\operatorname{argmax}}\, \{-p_L i(t_L) - p_R i(t_R)\} = \\ = \underset{s}{\operatorname{argmin}}\, \{p_L i(t_L) + p_R i(t_R)\} \end{aligned} \tag{3}$$

where t_L and t_R are implicit functions of s.

While different definitions of $i(t)$ lead to different questions in a tree, i.e. different optimal values in (3), in Breiman et al. (1987) it is argued that the tree shape is relatively robust to the choice of an impurity function. In financial applications of CART, the *Gini index* is frequently used as the measure of class heterogeneity (Kolyshkina and Brookes, 2002) so that $i(t) = 1 - \sum_{j=1}^{J} p^2(j|t)$.

Employing the Gini index as $i(\cdot)$, the optimal choice of tree questions is equivalent to

$$s^* = \underset{s}{\operatorname{argmax}} \left\{ p_L \sum_{j=1}^{J} p^2(j|t_L) + p_R \sum_{j=1}^{J} p^2(j|t_R) \right\} \tag{4}$$

In this way the maximum tree T_{MAX} can be built where each terminal node contains observations of only one class.

3 Cost-Complexity Tradeoff as a Traditional Way of Finding Optimal Tree Size

Although it is possible to grow a maximum tree for a given learning sample using (4) sequentially, its direct application for classification is far not always desirable because of the frequent overfitting – the training error reaches zero, but the validation error is usually much greater than its minimum level, which is feasible with a smaller tree. Note, however, that for some rare examples like on Figure 1 this is not the case – T_{MAX} is the best choice there.

One way to achieve the reasonable value of the validation error could be the employment of some kind of an *early stopping rule*. Since the growth of a tree is controlled by the decrement of an impurity function, the following criterion could be introduced to stop expanding the tree size:

$$\Delta i(t_P, s^*) < \bar{\beta} \tag{5}$$

for some $0 < \bar{\beta} < 1$.

However, $i(\cdot)$ is usually a non-monotone function of the tree size (Breiman et al., 1987), therefore a signal to stop could be premature.

Breiman et al. (1987) introduced a method that is based on the idea of optimizing the trade-off between the tree complexity and its size. Let $e(t) = 1 - \max_j p(j \mid t)$, $\tilde{T}$ be the set of terminal nodes and $\left|\tilde{T}\right|$ – the number of terminal nodes. Then $E(t) = e(t)p(t)$ and $E(T) = \sum_{t \in \tilde{T}} E(t)$ are the so called *internal misclassification errors* of a node t and tree T. For a given tree T the cost-complexity function $E_\alpha(T)$ to minimize takes the following form:

$$E_\alpha(T) = E(T) + \alpha \left|\tilde{T}\right| \tag{6}$$

where $\alpha \geq 0$ is a complexity parameter and $\alpha \left|\tilde{T}\right|$ is a cost component: the more complex is the tree (the higher is the number of terminal nodes) – the lower is $E(T)$, but at the same time the higher is the penalty $\alpha \left|\tilde{T}\right|$, and vice versa.

Although α can have infinite number of values, the authors of the method prove that the number of subtrees of T_{MAX} resulting in minimization of $E_\alpha(T)$ is finite. The traditional method employs cross-validation for a drastically reduced set of optimal subtrees (compared to the number of all possible subtrees of a given tree T_{MAX}) to select the optimal one with the balanced complexity (training error) and validation error.

It is claimed that the minimum value of all $E_\alpha(T)$ is not always desirable since results are frequently unstable. An empirical "one standard error" rule is employed instead, see Breiman et al. (1987) for more details.

4 Best Node Strategy – An Alternative Way of Tree Pruning

By its architecture, the cost-complexity approach ultimately operates with triplets of nodes $\{t_P, t_L, t_R\}$, which are parts of optimized subtrees of T_{MAX}. The decision whether to employ the selected triplet or not is based on the *joint performance of two child nodes in the triplet*, refer to the definition of the "weak link" in Breiman et al. (1987) for more details.

However, there are many cases when only one of the child nodes contains homogenous data while the second one is filled with points belonging to various classes. Performing validation of the tree containing both child nodes, which is done traditionally, frequently results in a mediocre performance of the triplet as a whole.

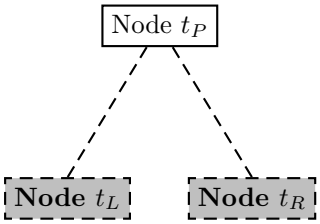

Figure 3: Traditional CART pruning operates only with both child nodes simultaneously – both child nodes are pruned here

Hence, there are serious reasons to concern that at least for selected types of classification tasks, for instance, in stock picking, the traditional cost-complexity balance approach does not provide the best feasible results.

Best Node Strategy (BNS) analyzes the individual node performance and provides an opportunity to prune only one child node if necessary, at the same time pruning both child nodes simultaneously is also an option.

While the traditional approach relies solely on the cross-validation performance of a given subtree, the "quality" of individual nodes is ignored. The tree "quality" is estimated via an *integral*

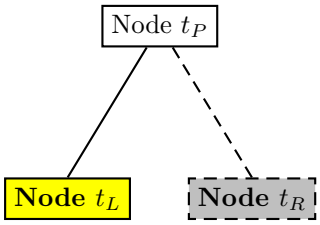

Figure 4: Situation that is infeasible for the traditional cost-complexity approach – only one child node is pruned here

characteristic (6), therefore individual nodes have only a minor impact on the overall result.

BNS reverses this approach and assumes that good performance of the tree is driven by good *individual* performance of nodes. This method of tree pruning takes into account *only* individual node characteristics and does not perform cross-validation to obtain an integral measure of the tree performance.

Let us consider a slightly modified example from Figure 1 and introduce some overlapping of the elements in a three-class problem depicted on Figure 5. Classes *Black* and *Green* are not linearly separable anymore, and that creates a challenge for the canonical cost-complexity approach, which is not able to keep only one of the child nodes in the decision tree.

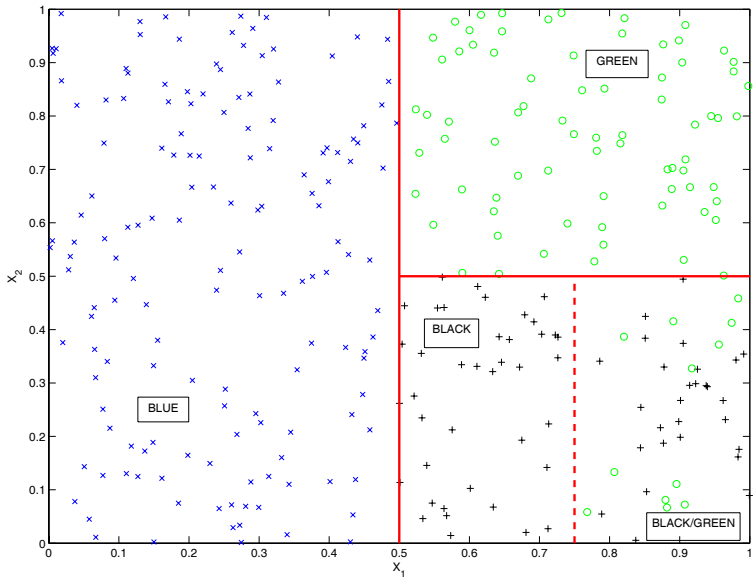

Figure 5: Modified example with three-class data and a cluster of linearly non-separable data. Solid lines refer to recursive partitioning suggested by the canonical cost-complexity approach, the dashed line indicates another partitioning that is missing and might be useful to separate a lot of points belonging to class *Black*.

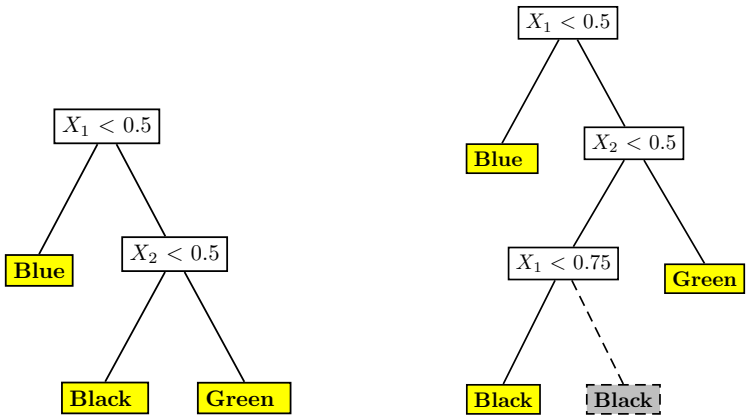

Figure 6: Two trees produced by the cost-complexity approach (left) and the novel Best Node Strategy (right). The grey dashed node on the right tree indicates the noisy part of the data in the learning sample and suggests this cluster of the decision rule to be excluded when classifying new data.

The key idea of BNS is to analyze *node reliability* in terms of their *purity* and *size* and allow nonsymmetric pruning when necessary. If a terminal node, which is potentially the crucial element of the final classification of new data, contains mixture of observations belonging to various classes, its presence in the decision rule is desirable only if one class clearly dominates others, because otherwise the reliability of the classification decision may be compromised. At the same time, such node should not contain only a minor number of observations from the learning sample; put differently, it should be representative.

Let us consider the tree produced by the cost-complexity approach (the left one on Figure 6) and its terminal node with the class tag *Black*. It partially corresponds to the mixed area on Figure 5 where observations of both *Black* and *Green* classes are present. It contains 68 points of class *Black* and only 12 points of class *Green*, and the risk of making the wrong classification decision, all other things being equal, is $\frac{12}{12+68} = 15\%$. The aim of BNS is to reduce or, when possible, to avoid fully this risk – BNS results in a slightly different tree (the right one on Figure 6) where another perfectly homogenous cluster of points is separated, which corresponds to the auxiliary condition $X_1 < 0.75$. Now when the data to classify appear to be in the unreliable node of the BNS tree (put differently, when these data meet the conditions $X_1 \geq 0.75$ and $X_2 \leq 0.5$), the decision rule considers this area of the learning sample unreliable (the risk of misclassification, all other things being equal, is already $\frac{12}{40} = 30\%$) and suggests no reliable classification can be performed. In the realm of stock picking that would mean the recommendation to hold the neutral position. At the same time the tree produced via the cost-complexity approach is not able to differentiate between two terminal nodes with the same class *Black* (as on the tree produced by BNS) that implies taking extra risk of misclassification. This happens because only symmetric pruning is available to the cost-complexity approach.

The more balanced control – individual node versus node triplet – comes at cost of introducing two degrees of freedom. Let $\bar{n}$ be the minimum required number of observations of the *dominating class* j^* in a node t, $j^* = \underset{i}{\operatorname{argmax}}\, p(i|\, t)$, and $\bar{p}$ – the minimum required proportion of the dominating

class j^* observations. Assuming that there are J classes, if the following conditions hold:

$$\begin{cases} n_t(j^*) \geq \bar{n} \\ \frac{n_t(j^*)}{n_t} \geq \bar{p} \geq \frac{1}{J} \end{cases} \tag{7}$$

then the node t is called *reliable* and marked as such via the *boolean function* $v(t) = 1$, which takes the zero value if (7) does not hold. If n_{LS} is the size of the learning sample, the feasibility constraint for (7) is straightforward:

$$\begin{cases} \frac{\bar{n}}{\bar{p}} \leq n_{LS} \\ \bar{p} \leq 1 \end{cases} \tag{8}$$

Let $T(n)$ be the tree where each terminal node contains *at most* n observations unless all observations in the node belong to the same class. To obtain a classification decision for a given new observation $x \in \mathcal{R}^P$ using BNS, it suffices to build the tree $T(\frac{\bar{n}}{\bar{p}})$, locate the terminal node $t[x]$ of the observation x and check if this node is reliable. If it is, then the optimal decision is the one produced by that node. If not, then assuming that the Gini index is employed as the impurity function, all parent nodes of the given node $t[x]$ are also unreliable, therefore tree pruning results in an empty tree and *it is advised by BNS procedure to perform no classification based on the provided tree as chances for misclassification are considered to be rather high.* This can be called the *inverse propagation property* of BNS – if a child node is unreliable, its parent is unreliable as well.

These two statements about the optimal tree size $\frac{\bar{n}}{\bar{p}}$ and inverse propagation property need to be proven, of course. First of all, let us consider an arbitrary maximum tree $T_{MAX} = T(1)$ with exactly one observation at each terminal node. For a given observation $x \in \mathcal{R}^P$ to classify, the terminal node $t[x]$ of $T(1)$ is unreliable unless $\bar{n} = 1$ or, if $\bar{n} > 1$, the condition $n_{t[x]} > \bar{n}$ is violated since $t[x] \in \tilde{T}(1)$ and $n_{t[x]} = 1$ where $\tilde{T}(1)$ is the set of terminal nodes of a tree $T(1)$. It may happen though that $p(j^*|t[x]) = 1$ and $n_{t[x]} > 1$. Then still if $n_{t[x]} < \frac{\bar{n}}{\bar{p}}$, the node is unreliable,

because either the condition $n_{t[x]} \geq \bar{n}$ is violated from (7) or, if not, with the minimum required probability of the dominating class being equal to $\bar{p}$ and the number of observations of that class being equal to $\bar{n}$, the *minimum* feasible terminal node size not to violate (7) is $\frac{\bar{n}}{\bar{p}}$.

Therefore, if BNS-reliable nodes exist in T_{MAX}, *all of them* can be reached by building the tree $T(\frac{\bar{n}}{\bar{p}})$. The proof of Theorem 1 (in the Appendix) is in fact the proof of the inverse propagation property of BNS. Hence if $t[x] \in \tilde{T}(\frac{\bar{n}}{\bar{p}})$ is unreliable, then all the parent nodes of $t[x]$ are unreliable, too.

At the moment, the inverse propagation for an arbitrary impurity function is proven only for the case when dominating classes of the child and parent nodes coincide, refer to Lemma 2 (in the Appendix) for details. However, as it was mentioned before, since the choice of the impurity function does not change the configuration of the maximum tree a lot (refer to Section 2), this creates little or no limitations in practical applicability of the method depending whether there is a rigid constraint to employ a particular form of an impurity function. Even if there is one and an impurity function different from the Gini index must be used, that would only mean that all terminal nodes and their parents, if necessary, should be checked for condition (7) assuming that (8) holds.

To conclude, the traditional cost-complexity approach builds a maximum tree at the first step. Then a sequence of subtrees is found by minimization of the cost-complexity function. The last step is to find the optimal tree by employing cross-validation. There, a set of cost-complexity estimates for different subtrees is found and a rule of thumb is applied to select the optimal tree. On the other hand, BNS requires to build the tree $T(\frac{\bar{n}}{\bar{p}})$ using the Gini index. After that an observation to classify is to be processed by the tree in the following way – the decision rule produced by $T(\frac{\bar{n}}{\bar{p}})$ is valid only if the respective terminal node is reliable as indicated in (7) and (8). Otherwise, it is suggested to avoid conducting classification using the available tree as chances for misclassification are considered to be rather high. In the stock picking setup that would be equivalent to taking a neutral position.

Indicator	Company Name
ADS	ADIDAS-SALOMON AG
ALT	ALTANA IND-AKTIEN U. ANL AG
ALV	ALLIANZ AG
BAS	BASF AG
BAY	BAYER AG
BMW	BAYERISCHE MOTOREN WERKE AG
DCX	DAIMLERCHRYSLER AG
EOA	E.ON AG
LHA	DEUTSCHE LUFTHANSA AG
LIN	LINDE AG
MAN	MAN AG
SAP	SAP AG
SCH	SCHERING AG
SIE	SIEMENS AG
TUI	TUI AG

Table 1: List of available companies from XETRA DAX and their codes

5 Available Data and Calibration

To examine the adequacy of the nonsymmetric pruning via BNS, a stock picking algorithm, which operates with XETRA DAX stocks, was backtested. A similar algorithm but with the cost-complexity approach for tree pruning was backtested as the primary benchmark.

The available XETRA DAX market data for the analysis consist of the samples for 15 companies and for the time period of February 19, 2001 – May 31, 2004. 13 different fundamental and technical variables were at the disposal describing each of these companies, refer to Table 1 and Table 2 for more details. Time scale of the data is one week.

Every stock was analyzed independently meaning that each stock return was forecasted by using the individual company data.

The first model degree of freedom is the threshold value $\bar{R} \geq 0$ that defines the class y of each stock for a given next period return in the learning sample. Depending on the next period stock

performance, there are three classes employed: *long*, *short* and *neutral*. $\bar{R}$ can potentially have different values for different calibrated stocks allowing for "big hit" ability (the ability of a method to forecast effectively the movements with big relative magnitude) introduced in Hartzmark (1991).

$$\begin{cases} R_t > \bar{R},\ y_t = \{long\} \\ R_t < -\bar{R},\ y_t = \{short\} \\ -\bar{R} \leq R_t \leq \bar{R},\ y_t = \{neutral\} \end{cases} \tag{9}$$

Given the magnitude of weekly stock returns in the learning sample, $\bar{R}$ was selected for each stock independently during calibration from the grid $[0\%, 3\%]$ with the step of 0.5%.

Although CART chooses variables for the learning sample automatically when building a decision tree, there is always a possibility for spurious links between dependent and independent variables. That is the main reason to consider multiple possible input specifications for the learning sample. Unlike Brennan et al. (1999), where preliminary regression analysis of available data was supposed to find the most significant variables to be included in the tree(s), in this study the optimal specification for each stock was obtained from a calibration procedure, which is described below.

Two different specifications were considered. The first one resembles the ideas of fundamental analysis (Fama and French, 1992; Sorensen et al., 1999; Brennan et al., 1999) and therefore is based on variables of fundamental nature – these are listed in the upper part of Table 2. According to the first specification, the learning sample consists of four variables: $\frac{CF_t}{P_t}$, $\frac{EPS_t}{P_t}$, $\frac{\Delta_{12}EPS_t}{P_t}$ and ROE_t (t is the current time period).

Depending on the stability of a distribution and the level of noise of the learning sample over time, retaining the old observations in the learning sample may potentially result in a deteriorated forecasting power of the model (Tam and Kiang, 1992); therefore, the second degree of freedom for calibration is the type of the learning sample, which can either have the fixed size over time (sliding window) or, when such setup provides inadequate calibration results (see below), include

Indicator	Type	Frequency	Comments
Sales/P	Fundamental	1 week	Sales to Price Ratio
CF/P	Fundamental	1 week	Cash Flow to Price Ratio
EPS/P	Fundamental	1 week	Earnings per Share to Price Ratio
Δ_{12}EPS/P	Fundamental	1 week	3-Month Change in EPS to Price Ratio
ROE	Fundamental	1 week	Return on Equity
Momentum	Technical	1 week	$M_t = P_t - P_{t-T},\ T = 20$
Stochastic	Technical	1 week	$\frac{P_t - P_L}{P_H - P_L},\ P_H = max(P_t),\ P_L = min(P_t)$
MA/P	Technical	1 week	$MA(T) = \frac{\sum_{i=t-T}^{t} P_i}{T},\ T = 12$
MACD	Technical	1 week	$(1 - \frac{n_1}{n_2})\{MA(n_1) - MA(n_2 - n_1)\}$ $n_1 = 12,\ n_2 = 26$
MA St. Error	Technical	1 week	Standard deviation of MA
ROC/P	Technical	1 week	$ROC_t = \frac{P_t}{P_{t-T}},\ T = 10$
TRIX	Technical	1 week	Triple exponentially smoothed MA
R_{t-1}	Technical	1 week	$R_{t-1} = \frac{P_t - P_{t-1}}{P_{t-1}}$, P_t – current stock price

Table 2: List of available variables as potential input factors for learning samples. All variables are available for each of 15 analyzed companies. The current time period is indicated by t

each new available observation with the following step.

For each stock independently the adequacy of calibration was assessed primarily based on the expected annualized yield – the higher the yield, the better the specification is assumed. To avoid the potential spuriousness of calibration results, the *activity ratio* indicator (the percentage of active operations during calibration for a given stock) was employed in the following way. First, the activity ratio has to exceed 40% in order for a specification to be considered reliable. Competing specifications (with the similar amount of yielded expected return) were selected in favor of those with the highest activity ratio. Additionally, the hit ratio (the proportion of correct active directional forecasts during the calibration) of a reliable specification had to exceed 45%.

If the first specification failed to provide adequate calibration results with both types of the learning sample, i.e. when the calibrated profit was negative for any setup or when the activity ratio or hit ratio constraints were violated, the second specification was considered. The second specification therefore implies the situation when the sole use of fundamental variables is not enough to explain

Stock	Specification	$\bar{R}$	Learning sample
ADS	fund. and tech.	0.5%	sliding window
ALT	fundamental	1.0%	sliding window
ALV	fundamental	1.0%	sliding window
BAS	fund. and tech.	0.5%	expanding
BAY	fundamental	0.5%	sliding window
BMW	fundamental	1.0%	sliding window
DCX	fundamental	1.0%	sliding window
EOA	N/A	N/A	N/A
LHA	N/A	N/A	N/A
LIN	fundamental	0.5%	sliding window
MAN	N/A	N/A	N/A
SAP	N/A	N/A	N/A
SCH	fundamental	0.5%	expanding
SIE	N/A	N/A	N/A
TUI	fund. and tech.	0.5%	expanding

Table 3: Calibration results for BNS tree pruning, N/A indicates situations when none of the inputs were able to produce positive calibration yield

the movements of the next period stock return adequately, therefore the variable set is expanded by available technical factors (Neftci, 1991; Sullivan et al., 1999) like ROC, TRIX or Stochastic listed in Table 2. According to the second specification, the learning sample consists of all 13 available variables.

Finally, two BNS parameters need to be fixed as well. Similar to Breiman et al. (1987), an empirical rule of thumb was employed: set $\bar{p}$ to 75% and $\bar{n}$ to 10% of the size of the learning sample, refer to Osei-Bryson (2004) for a description of the so called discriminatory power measure, which is defined via $\bar{p}$, and Bramer (2002) for a discussion on size cutoff, which is in fact just a different name of $\bar{n}$. If a particular application requires more precision for $\bar{p}$ and $\bar{n}$, these two parameters can be calibrated analogously to $\bar{R}$.

If after all tested combinations during the calibration all specifications were considered inadequate for a given stock, this stock was excluded from the portfolio, see Section 6 for more details on portfolio creation.

Stock	Specification	$\bar{R}$	Learning sample
ADS	fund. and tech.	0.5%	sliding window
ALT	fundamental	0.5%	sliding window
ALV	fundamental	0.5%	sliding window
BAS	fund. and tech.	0.5%	expanding
BAY	fundamental	0.5%	sliding window
BMW	N/A	N/A	N/A
DCX	N/A	N/A	N/A
EOA	N/A	N/A	N/A
LHA	N/A	N/A	N/A
LIN	N/A	N/A	N/A
MAN	N/A	N/A	N/A
SAP	N/A	N/A	N/A
SCH	fundamental	0.5%	expanding
SIE	N/A	N/A	N/A
TUI	N/A	N/A	N/A

Table 4: Calibration results for cost-complexity tree pruning, N/A indicates situations when none of the inputs were able to produce positive calibration yield

The available market data were employed in the following way. The first 53 observations (or roughly one year) were allocated to the learning period. The next 25 points (or roughly half a year) comprised the test set for calibration. Finally, the rest 93 points (or a little less than two years) were used for validation. The size of the sliding window, when applicable, was set to the length of the learning period – 53 observations.

Such calibration was performed independently for BNS and the cost-complexity approaches of tree pruning. Tenfold cross-validation and 1-SE rule (Breiman et al., 1987) were employed to find optimal cost-complexity trees. In case when the resulting optimal tree was underparameterized (consisted of the single root node after pruning), 0-SE rule (Breiman et al., 1987) was employed instead.

6 XETRA DAX Stocks Backtesting

As it can be seen from Table 3 and Table 4, for BNS 10 out of 15 stocks (66.7%) showed positive performance at the test set and only 6 out of 15 (40%) – for the cost-complexity tree pruning.

If an open position was recommended for an arbitrary stock, it was then closed at the end of each period – no reinvesting was allowed. Transaction costs in the amount of 10 b.p. were accounted for every active operation.

Two various recursive portfolios – based on BNS and cost-complexity approach recommendations – were created. Their positions were updated weekly. Both portfolios were equally-weighted – this weighting scheme, firstly, comes to diversify the portfolios and reduce the risk of returns and, secondly, because there are no explicit reasons to prefer one stock to another (Amenc et al., 2003). According to Table 3 and Table 4, the first portfolio to backtest contained 10 stocks while the second one – 6 stocks.

Figure 7 depicts portfolio's weekly returns when BNS was used for tree pruning. Its annualized return is 17.17% while the Sharpe ratio is 1.26 for the risk-free rate of 4.5%. The hit ratio of this portfolio is 59%. However, one may notice that the vast majority of wrong classifications coincides with the relatively small values of stock price returns, therefore resulting in substantial profit and the high Sharpe ratio.

While the hit ratio of the second portfolio, which was built by employing the traditional cost-complexity approach, is close to the first one – 54%, the financial performance is far more different, refer to Figure 8 for details. Although it manages to produce the positive annualized profit – 2.87%, its returns are obviously more volatile resulting in the Sharpe ratio of only -0.09.

BNS exhibited superior performance comparing with the cost-complexity approach, however, another indirect comparison is also possible. Some of the studies mentioned in Section 1 employed decision trees for stock picking and reported the corresponding results. Although the markets and

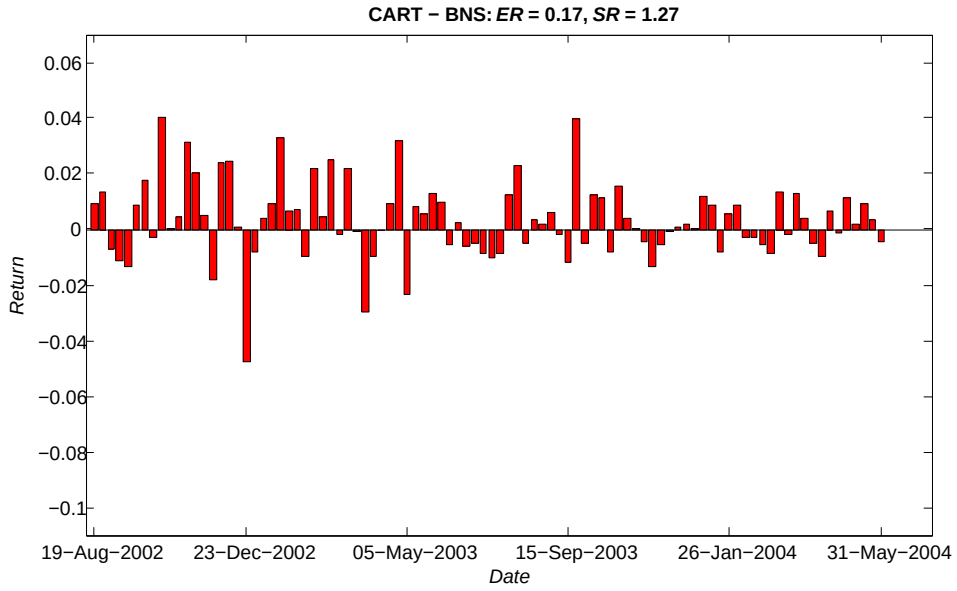

Figure 7: Equally weighted portfolio of stocks performance when BNS is employed for tree pruning, ER – annualized expected return, SR – the Sharpe ratio

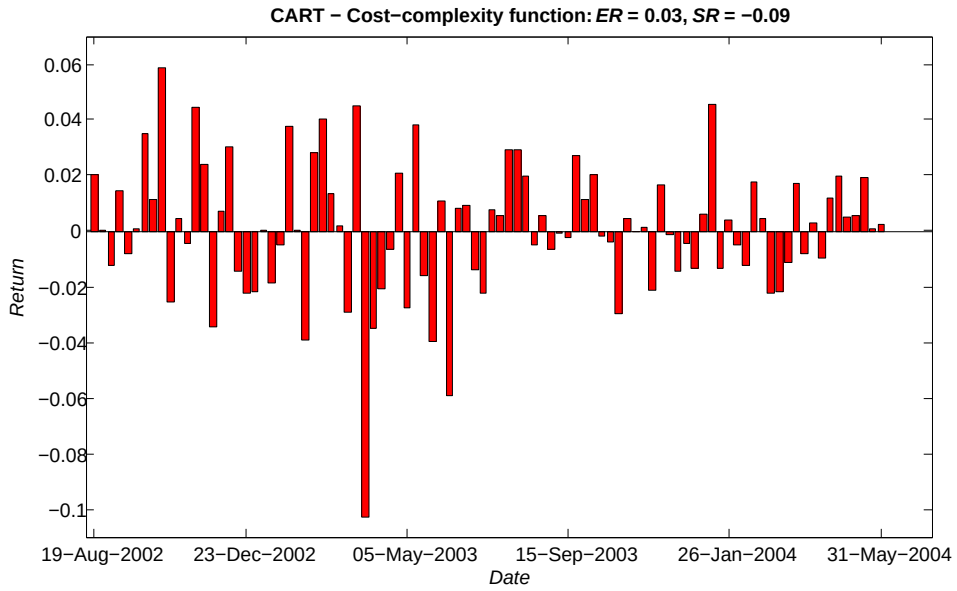

Figure 8: Equally weighted portfolio of stocks performance when the traditional cost-complexity approach is employed for tree pruning, ER – annualized expected return, SR – the Sharpe ratio

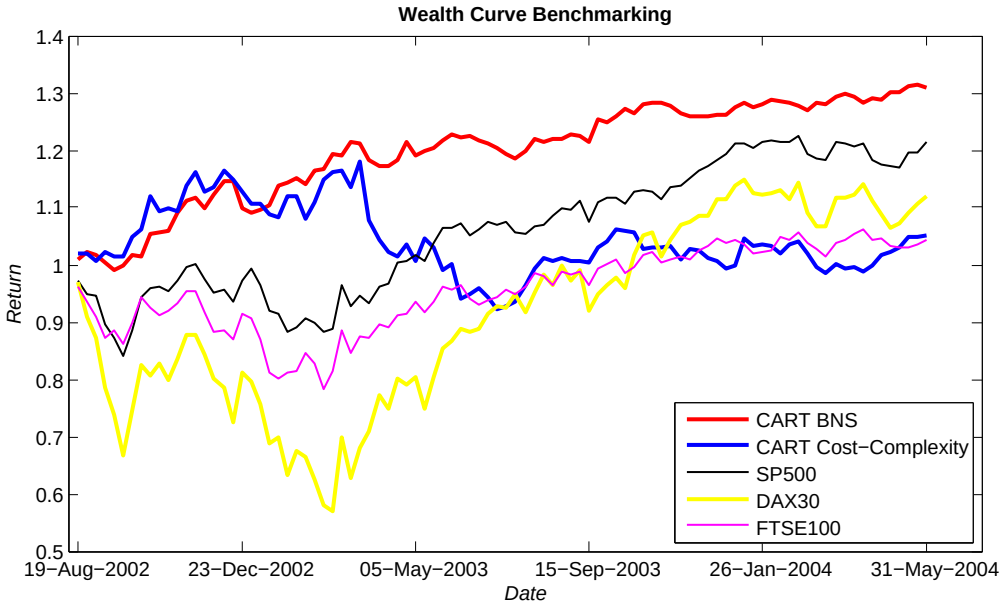

Figure 9: Wealth curves for two active CART strategies and three passive investment strategies

the time periods are different, it may still be interesting to compare these results in terms of relative returns and their risks. In Seshadri (2003) the three-class (*overweight*, *underweight*, *neutral*) recommendations for stocks from the S&P500 universe were provided. As at August 6, 2003, the model has returned 14.6% (annualized) with a corresponding Sharpe ratio of 1.5.

Similarly, technological stocks were classified into three performance buckets in Sorensen et al. (1999), and for the period of 1996-1999 the model returned 19.62% with a corresponding Sharpe ratio of 1.23. Interestingly, while the recursive partitioning mechanism is described in this report, nothing is said about tree pruning that led to the achieved performance.

Figure 9 depicts the benchmarking of different strategies when an investor has alternative opportunities to invest into DAX index fund, FTSE100 index fund or SP500 index fund. While the markets are, of course, different (excluding XETRA DAX virtual index fund), this benchmarking accounts for some alternative types of passive investment.

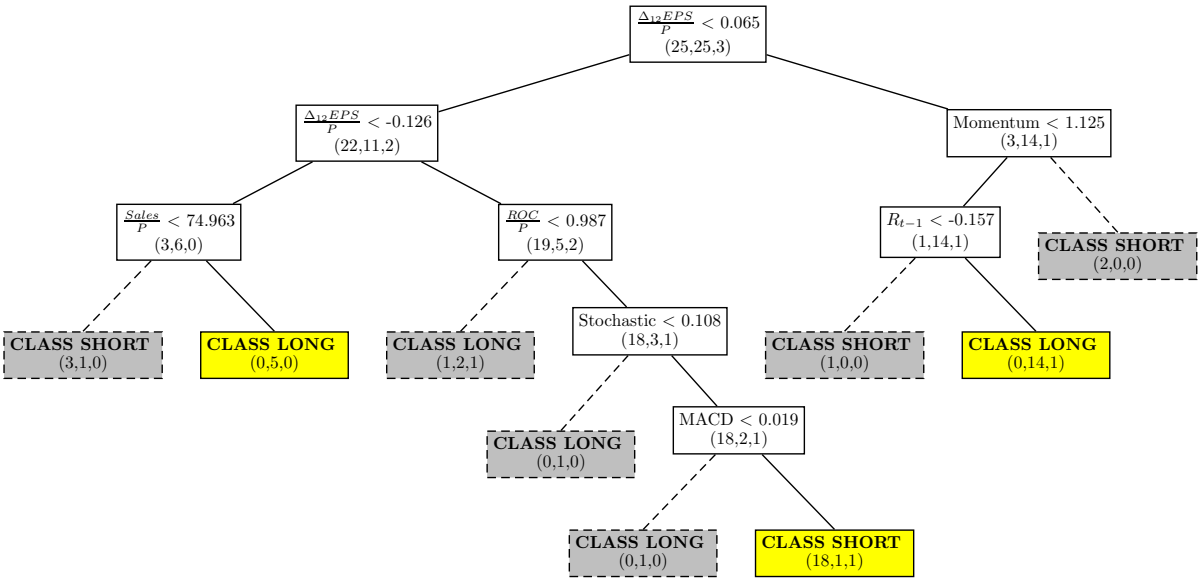

Figure 10: An example of the decision tree for ADS stock. Here $\bar{n} = 5$ and $\bar{p} = 0.75$. The numbers in parentheses reflect the number of observations for a given node belonging to classes *short*, *long* and *neutral* respectively. BNS-reliable nodes are marked with solid lines and yellow color, BNS-unreliable – dashed lines and grey color

To illustrate the difference in performance exhibited by the cost-complexity approach and BNS, Figure 10 shows the sample tree $T(7)$ for ADS stock. The root node – at the very top of the tree – is constituted by 25 observations of class *short*, 25 – of class *long* and 3 – of class *neutral*. The numbers in parentheses show the quantity of observations in a given node for respective three classes as in the root node. With $\bar{n} = 5$ for this case, many terminal nodes are considered unreliable since they contain fewer number of points of the dominating class. For instance, to reach the terminal node in the very bottom of the tree that has a recommendation *short* and is constituted by 18 observations of class *short*, one observation of class *long* and one – of class *neutral*, one would need to keep the majority of the tree's structure preserved. However, for the cost-complexity approach this single node may not have such a strong influence. Because other nodes in the close vicinity are quite impure, it may happen that the target node becomes pruned simply because the significant number of points from the test set falls into these impure neighbor nodes: the cost-complexity

function value for a subtree carrying the target node may become too high.

Finally, to test the statistical significance of the financial performance difference in the results exhibited by the traditional cost-complexity and the novel BNS tree pruning approach, the Diebold-Mariano test was employed (Diebold and Mariano, 1995). While the hit ratios of the compared portfolios are quite close, the main motivation for this test is to take into account the economic value of the forecasts and not just their directional accuracy. The null hypothesis H_0 of the Diebold-Mariano test is that the expected value of an arbitrary loss differential d is equal to zero:

$$H_0 : E\left(d\right) \equiv E\left[g\left(e^{BNS}\right) - g\left(e^{CC}\right)\right] = 0 \tag{10}$$

where $g(\cdot)$ is an arbitrary function and e^{BNS}, e^{CC} – vectors of forecast errors associated with BNS and cost-complexity portfolios.

Since the aim of applying the Diebold-Mariano test here is to compare the expected economic values of two forecasts, function $g(\cdot)$ resembles the wealth curves from Figure 9:

$$\begin{cases} g(e_1) = 1 + e_1, \\ g(e_i) = g(e_{i-1}) + e_i, \quad 1 < i \leq N \end{cases} \tag{11}$$

where e_1 is the forecast error at the first time period, N – number of forecasts made (the length of the backtesting period).

Forecast errors are computed as the difference between the realized portfolio profit and any arbitrary benchmark – the resulting form of the loss differential d is invariant with respect to the choice of the benchmark as shown below. Given (11), if Π^{BNS} and Π^{CC} are vectors of values of the two respective portfolios and Π^{DAX} is the vector of values of some arbitrary DAX benchmark, then:

$$E\left(d\right) = E\left[\left(\Pi^{BNS} - \Pi^{DAX}\right) - \left(\Pi^{CC} - \Pi^{DAX}\right)\right] = E\left[\Pi^{BNS} - \Pi^{CC}\right] \tag{12}$$

and therefore the loss differential d is the difference between wealth curves for BNS and cost-complexity portfolios.

The test statistic is defined as

$$DM = \frac{\bar{d}}{\sqrt{2\pi \hat{f}_d(0)/N}} \tag{13}$$

where $\bar{d}$ is the sample mean of the the loss differential d, $\hat{f}_d(0)$ is a consistent estimate of spectral density of the loss differential at the zero frequency and N is the number of forecasts.

The variance $2\pi \hat{f}_d(0)$ was estimated using the Bartlett kernel with automatic bandwidth selection (Andrews, 1991; Newey and West, 1994). As a result, $DM = 13.14$ and the p-value $= 1.37 \cdot 10^{-38}$, which indicates that H_0 is rejected at the 0.1% confidence level. One may therefore conclude that the economic value associated with portfolio returns generated by BNS and cost-complexity decision tree pruning strategies are statistically significantly different in favor of BNS.

7 Conclusions

The new tree pruning technique introduced in this study – Best Node Strategy (BNS) – proved its high potential over the traditional approach based on the cost-complexity function for the analyzed XETRA DAX stocks data set. Backtesting has shown the superiority of BNS in terms of financial performance of the recursive equally weighted portfolio: the annualized profit of 17.17% vs 2.87% and the Sharpe ratio of 1.27 vs -0.09. Active stock management via BNS showed its higher efficiency also compared to selected passive investment strategies.

While the hit ratios of the both active strategies are quite close – 59% and 54% – and do not significantly deviate from 50%, the difference in economic value of both forecasts is undeniably significant according to the Diebold-Mariano test. At this point it is worth citing professional equity investment managers from *Schroders* (€189.4 billion under management as at December 31,

2007) commenting a very similar outcome (in their study, the backtested annualized return of a decision tree based trading strategy over the whole period is 12%): *"Although these hit rates do not seem significantly different from 50% (which is indicative of no skill in stock picking), this is very typical in financial applications and it would be rare to observe models with average hit rates in excess of 55%. Indeed, as the chart above illustrates, hit rates even slightly better than 50% can generate strong strategy outperformance in practice. [...] We would conclude from this analysis that the model is very successful at locating the key stock characteristics that identify future relative performance"* (Schroders, 2006).

With the proven reverse propagation property of BNS, it is easy to build the tree of an optimal size possessing much more flexible non-symmetric structure than its symmetric canonically pruned counterpart.

Appendix

Lemma 1. *Let's t_P be the parent node for t_L and t_R given some arbitrary split s. If the following inequalities hold:*

$$\begin{cases} i(t_P) > i(t_L) \\ i(t_P) \geq i(t_R) \end{cases} \tag{14}$$

and one of them holds as strict, for instance, for t_L, then it is true that

$$\Delta i(t_P, s) = i(t_P) - p_L i(t_L) - p_R i(t_R) > 0. \tag{15}$$

The reverse statement is also true.

Proof. The proof of the first part is straightforward and can be found in (Breiman et al., 1987). Let us prove the reverse part of the lemma. Using the link between p_L and p_R, one can get the

following inequality:

$$\begin{cases} \Delta i(t_P, s) = i(t_P) - p_L i(t_L) - p_R i(t_R) > 0 \\ p_L + p_R = 1,\ p_L \in (0;1),\ p_R \in (0;1) \end{cases} \Rightarrow\ i(t_P) > p_L i(t_L) + (1 - p_L) i(t_R) \tag{16}$$

Let us suppose that $i(t_P) < i(t_L)$ and to be more specific: $i(t_P) = p_L i(t_L) < i(t_L) \quad \forall p_L \in (0;1)$. Then $p_L i(t_L) > p_L i(t_L) + (1 - p_L) i(t_R)$ that is equivalent to $(1 - p_L) i(t_R) < 0 \ \Leftrightarrow\ i(t_R) < 0$, which is impossible by definition of $i(\cdot)$. Hence one can conclude that $i(t_P) \geq i(t_L)$.

Let us suppose now that $i(t_P) < i(t_R)$ and let $i(t_P) = (1 - p_L) i(t_R) < i(t_R) \quad \forall p_L \in (0;1)$. Then $(1-p_L) i(t_R) > p_L i(t_L) + (1-p_L) i(t_R) \ \Leftrightarrow\ i(t_L) < 0$, which is impossible. That is why $i(t_P) \geq i(t_R)$.

The remaining step is to note that one of the two inequalities – $i(t_P) \geq i(t_L)$ or $i(t_P) \geq i(t_R)$ – must hold as strict because if $i(t_P) = i(t_L) = i(t_R)$, then $\Delta i(t_P, s) = 0$ that violates the conditions of the lemma. □

Lemma 2. *Let t_L and t_R be the two child nodes with t_P being the parent node and s – the relevant data split so that $\Delta i(t_P, s) > 0$. Let $S(t)$ be the dominating class of the node t. Then for the node $t \in \{t_L, t_R\}$ so that $S(t) = S(t_P)$ it is true that if $v(t) = 0$, then $v(t_P) = 0$ where $v(\cdot)$ is defined in (7) so that $\bar{n}$ and $\bar{p}$ do not violate (8).*

Proof.

1. Let us consider two sets of conditional probabilities $(p_1, p_2, \ldots, p_J)$ and $(p'_1, p'_2, \ldots, p'_J)$ where $p_i = p(i \,|\, t_P)$ and $p'_i = p(i \,|\, t)$, $t \in \{t_L, t_R\}$. Since the inequality $i(t_P) > i(t)$ holds as strict at least for one of the child nodes $\{t_L, t_R\}$ (Lemma 1), it follows that at least one of the values in the set $p(i \,|\, t)$ has changed compared to the set $p(i \,|\, t_P)$, refer to (Breiman et al., 1987) for the detailed description of the properties of an arbitrary impurity function.
2. Since $\sum_{i=1}^{J} p(i \,|\, t) = 1$, there exist at least one value of the conditional probability from $(p'_1, p'_2, \ldots, p'_J)$

that has *increased* compared to $(p_1, p_2, \ldots, p_J)$ and at least one – that has *decreased*, because the situation when each of the components $p(i|\,t) \geq 0$ changed their values in one direction is impossible.

3. For some class j let $p'_j = \max\limits_i p'_i = \max\limits_i p(i|\,t)$, i.e. the maximum value of the conditional probability from the second set. Then while there may exist an arbitrary number of components that increased or decreased their values when transferring from the first set of probabilities $p(i|\,t_P)$ to the second – $p(i|\,t)$, p'_j is the maximum value from the subset of values that have *increased*.

4. That is why $p_j \leq p'_j$ where $p_j = \max\limits_i p(i|\,t)$ and $p'_j = p(j|t_P)$.

   Since $j = \operatorname{argmax}\limits_i p\,(i\,|t\,)$, it follows that $S(t_P) = j$. It is given that $S(t) = S(t_P)$, therefore $S(t) = j$. Because $v(t) = 0$, it follows that $p(j|t) < \bar{p}$. However, it was proven that $p(j|t) \geq p(j|t_P)$. Therefore, $p(j|t_P) \leq p(j|t) < \bar{p}$. Hence $p(j|t_P) < \bar{p} \;\Rightarrow v(t_P) = 0$.

□

Theorem 1. *Let t_L and t_R be the two child nodes with t_P being the parent node and s – the relevant data split. Let t_L and t_R be terminal nodes in a tree $T(\frac{\bar{n}}{\bar{p}})$. Let $i(t)$ be the impurity function taking the form of the Gini index: $i(t) = 1 - \sum\limits_{j=1}^{J} p^2(j|t)$, J be the number of classes in the learning sample and $\Delta i(t_P, s) > 0$. Then if at least one of the child nodes is unreliable: $v(t) = 0$, then the parent node is also unreliable: $v(t_P) = 0$ where $v(\cdot)$ is defined in (7) so that $\bar{n}$ and $\bar{p}$ do not violate (8).*

Proof. Let $j^* = \operatorname{argmax}\limits_i p\,(i\,|t\,)$. One of the requirements for a node to be accounted as reliable is to show the significantly high probability of the dominating class: $p(j^*|t) \geq \bar{p}$. Since $\sum\limits_{i=1}^{J} p(i|\,t) = 1$, $0 \leq p(i|\,t) \leq 1$ and $i(t) = 1 - \sum\limits_{j=1}^{J} p^2(j|t)$, then the inequality $p(j^*|t) \geq \bar{p}$ implies the existence of the upper bound of the node impurity value – $\bar{\imath}$, so that

$$p(j^*|t) \geq \bar{p} \Leftrightarrow i(t) \leq \bar{\imath}$$

where

$$\bar{\imath} = \left[\begin{array}{l} 1 - \frac{\bar{p}^2}{J},\ \bar{p} = \frac{1}{J} \\ \frac{-J\bar{p}^2 + 2\bar{p} + J - 2}{J-1},\ \bar{p} > \frac{1}{J} \end{array} \right.$$

Since $v(t) = 0$, there are two possible configurations of the triplet $\{t_L, t_R, t_P\}$, where t_L and t_R are arbitrary child nodes and t_P – their parent node.

1. Both child nodes are unreliable: $v(t_L) = v(t_R) = 0$

   In this case $i(t) > \bar{\imath}$ where $t = \{t_L, t_R\}$ because $t \in \tilde{T}(\frac{\bar{n}}{\bar{p}})$. Since $\Delta i(t_P, s) > 0$, according to Lemma 1 it follows that $i(t_P) \geq i(t)$, and therefore $i(t_P) > \bar{\imath} \Rightarrow v(t_P) = 0$.

2. Only one of the child nodes is unreliable, for sake of simplicity let it be node t_R:

   Employing Lemma 1 once again, it is possible to conclude that $i(t_P) \geq i(t_L)$. Because the node t_L is pure, then $i(t_L) < \bar{\imath}$. However, it is not possible to say if $i(t_P) > \bar{\imath}$ or not.

   But for the node t_R the situation changes drastically. Again, $i(t_P) \geq i(t_R)$, but in this case $i(t_R) > \bar{\imath}$, so one can conclude that $i(t_P) > \bar{\imath} \Rightarrow v(t_P) = 0$.

Since it is given that $v(t) = 0$, the situation when both terminal nodes in the triplet are pure is impossible. This concludes the proof of the theorem.

If t_P is unreliable, the same set of arguments can be applied to this node because $n_{t_P} > n_t \geq \bar{n} \geq \frac{\bar{n}}{\bar{p}}$. Therefore, if a terminal node in $T(\frac{\bar{n}}{\bar{p}})$ is unreliable, each of its parent nodes is unreliable, too. $\square$

References

N. Amenc, P. Malaise, L. Martellini, and D. Sfeir. Portable alpha and portable beta strategies in the eurozone: Implementing active asset allocation decisions using equity index options and futures. *Edhec Business School*, 2003.

D. W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59, No. 3, 1991.

R. Balvers, T. Cosimano, and B. McDonald. Predicting stock returns in an efficient market. *The Journal of Finance*, 45, No. 4, 1990.

M. Bramer. Using J-pruning to reduce overfitting in classification trees. *Knowledge-Based Systems*, 15, 2002.

L. Breiman, H. J. Friedman, A. R. Olshen, and J.C. Stone. *Classification and regression trees*. The Wadsworth Statistics/Probability Series, 1987.

N. Brennan, P. Parameswaran, J. Gadaut, A. Luck, M. Dowle, S. Fagg, G. Brar, M. Turner, E. Flynn, D. Jessop, K. Yu, and A. McCutcheon. A method for selecting stocks within sectors. *Salomon Smith Barney Equity Research: Europe, Quantitative Strategy*, 1999.

J. Campbell and Y. Hamao. Predictable stock returns in the United States and Japan: A study of long-term capital market integration. *The Journal of Finance*, 47, No. 1, 1992.

N. Chen. Financial investment opportunities and the macroeconomy. *The Journal of Finance*, 46, No. 2, 1991.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, No. 3, 1995.

E. Fama. Efficient capital markets: II. *The Journal of Finance*, 46, No. 5, 1991.

E. Fama and K. French. Dividend yields and expected stock returns. *Journal of Financial Economics*, 22 Nm. 1:3–25, 1988a.

E. Fama and K. French. Permanent and temporary components of stock prices. *Journal of Political Economy*, 96 Nm. 2:246–73, 1988b.

E. Fama and K. French. The cross-section of expected stock returns. *The Journal of Finance*, 47, No. 2, 1992.

S. Feldman, D. F. Klein, and G. Honigfeld. The reliability of a decision tree technique applied to psychiatric diagnosis. *Biometrics*, 28, No. 3, 1972.

W. Ferson and C. Harvey. The variation in economic risk premia. *Journal of Political Economy*, 99:385–415, 1991.

G. Harman, P. Parameswaran, and M. Witt. Shares, bonds or cash? Asset allocation in the new economy using CART. *Salomon Smith Barney Equity Research: Australia, Quantitative Analysis*, 2000.

M. L. Hartzmark. Luck versus forecast ability: Determinants of trader performance in futures markets. *The Journal of Business*, 64, No. 1, 1991.

D. Keim and R. Stambaugh. Predicting returns in the stock and bond markets. *Journal of Financial Economics*, 17, 1986.

H. Kim and W.-Y. Loh. Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 96, No. 454, 2001.

I. Kolyshkina and R. Brookes. Data mining approaches to modelling insurance risk. *PricewaterhouseCoopers*, 2002.

S. N. Neftci. Naive trading rules in financial markets and Wiener-Kolmogorov prediction theory: A study of "technical analysis". *The Journal of Business*, 64, No. 4, 1991.

W. K. Newey and K. D. West. Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61, No. 4, 1994.

K.-M. Osei-Bryson. Evaluation of decision trees: a multi-criteria approach. *Computers & Operations Research*, 31, 2004.

Quantitative Equity Products Team at Schroders. Tree building at Schroders. *Schroder Investment Management Limited: Quantitative Equity Products*, 2006.

L. Seshadri. JPMorgan US quantitative factor model: August 2003 stock list. *JPMorgan Quantitative Equity and Derivatives*, 2003.

H. E. Sorensen, K. C. Ooi, and L. K. Miller. The decision tree approach to stock selection. *Salomon Smith Barney Equity Research: United States, Global Quantitative Research*, 1999.

H. J. Steadman, E. Silver, J. Monahan, P. S. Appelbaum, P. C. Robbins, E. P. Mulvey, T. Grisso, L. H. Roth, and S. Banks. A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior*, 24, No. 1, 2000.

R. Sullivan, A. Timmermann, and H. White. Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance*, 54, No. 5, 1999.

K. Y. Tam and M. Y. Kiang. Managerial applications of neural networks: The case of bank failure predictions. *Management Science*, 38, No. 7, 1992.

SFB 649 Discussion Paper Series 2008

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Testing Monotonicity of Pricing Kernels" by Yuri Golubev, Wolfgang Härdle and Roman Timonfeev, January 2008.
002 "Adaptive pointwise estimation in time-inhomogeneous time-series models" by Pavel Cizek, Wolfgang Härdle and Vladimir Spokoiny, January 2008.
003 "The Bayesian Additive Classification Tree Applied to Credit Risk Modelling" by Junni L. Zhang and Wolfgang Härdle, January 2008.
004 "Independent Component Analysis Via Copula Techniques" by Ray-Bing Chen, Meihui Guo, Wolfgang Härdle and Shih-Feng Huang, January 2008.
005 "The Default Risk of Firms Examined with Smooth Support Vector Machines" by Wolfgang Härdle, Yuh-Jye Lee, Dorothea Schäfer and Yi-Ren Yeh, January 2008.
006 "Value-at-Risk and Expected Shortfall when there is long range dependence" by Wolfgang Härdle and Julius Mungo, Januray 2008.
007 "A Consistent Nonparametric Test for Causality in Quantile" by Kiho Jeong and Wolfgang Härdle, January 2008.
008 "Do Legal Standards Affect Ethical Concerns of Consumers?" by Dirk Engelmann and Dorothea Kübler, January 2008.
009 "Recursive Portfolio Selection with Decision Trees" by Anton Andriyashin, Wolfgang Härdle and Roman Timofeev, January 2008.
010 "Do Public Banks have a Competitive Advantage?" by Astrid Matthey, January 2008.
011 "Don't aim too high: the potential costs of high aspirations" by Astrid Matthey and Nadja Dwenger, January 2008.
012 "Visualizing exploratory factor analysis models" by Sigbert Klinke and Cornelia Wagner, January 2008.
013 "House Prices and Replacement Cost: A Micro-Level Analysis" by Rainer Schulz and Axel Werwatz, January 2008.
014 "Support Vector Regression Based GARCH Model with Application to Forecasting Volatility of Financial Returns" by Shiyi Chen, Kiho Jeong and Wolfgang Härdle, January 2008.
015 "Structural Constant Conditional Correlation" by Enzo Weber, January 2008.
016 "Estimating Investment Equations in Imperfect Capital Markets" by Silke Hüttel, Oliver Mußhoff, Martin Odening and Nataliya Zinych, January 2008.
017 "Adaptive Forecasting of the EURIBOR Swap Term Structure" by Oliver Blaskowitz and Helmut Herwatz, January 2008.
018 "Solving, Estimating and Selecting Nonlinear Dynamic Models without the Curse of Dimensionality" by Viktor Winschel and Markus Krätzig, February 2008.
019 "The Accuracy of Long-term Real Estate Valuations" by Rainer Schulz, Markus Staiber, Martin Wersing and Axel Werwatz, February 2008.
020 "The Impact of International Outsourcing on Labour Market Dynamics in Germany" by Ronald Bachmann and Sebastian Braun, February 2008.
021 "Preferences for Collective versus Individualised Wage Setting" by Tito Boeri and Michael C. Burda, February 2008.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

022 "Lumpy Labor Adjustment as a Propagation Mechanism of Business Cycles" by Fang Yao, February 2008.
023 "Family Management, Family Ownership and Downsizing: Evidence from S&P 500 Firms" by Jörn Hendrich Block, February 2008.
024 "Skill Specific Unemployment with Imperfect Substitution of Skills" by Runli Xie, March 2008.
025 "Price Adjustment to News with Uncertain Precision" by Nikolaus Hautsch, Dieter Hess and Christoph Müller, March 2008.
026 "Information and Beliefs in a Repeated Normal-form Game" by Dietmar Fehr, Dorothea Kübler and David Danz, March 2008.
027 "The Stochastic Fluctuation of the Quantile Regression Curve" by Wolfgang Härdle and Song Song, March 2008.
028 "Are stewardship and valuation usefulness compatible or alternative objectives of financial accounting?" by Joachim Gassen, March 2008.
029 "Genetic Codes of Mergers, Post Merger Technology Evolution and Why Mergers Fail" by Alexander Cuntz, April 2008.
030 "Using R, LaTeX and Wiki for an Arabic e-learning platform" by Taleb Ahmad, Wolfgang Härdle, Sigbert Klinke and Shafeeqah Al Awadhi, April 2008.
031 "Beyond the business cycle – factors driving aggregate mortality rates" by Katja Hanewald, April 2008.
032 "Against All Odds? National Sentiment and Wagering on European Football" by Sebastian Braun and Michael Kvasnicka, April 2008.
033 "Are CEOs in Family Firms Paid Like Bureaucrats? Evidence from Bayesian and Frequentist Analyses" by Jörn Hendrich Block, April 2008.
034 "JBendge: An Object-Oriented System for Solving, Estimating and Selecting Nonlinear Dynamic Models" by Viktor Winschel and Markus Krätzig, April 2008.
035 "Stock Picking via Nonsymmetrically Pruned Binary Decision Trees" by Anton Andriyashin, May 2008.

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".