

A Financial Data Mining Model for Stock Picking

Un Modello di “Financial Data Mining” per la selezione di portafogli finanziari

Claudio Conversano

Dipartimento di Matematica e Statistica, Università di Napoli Federico II
Via Cintia, Monte S. Angelo, I-80126, Napoli, Italy, conversa@unina.it

Abstract¹: Nell’ambito della problematica del *benchmarking* per dati finanziari si presenta una strategia integrata per la selezione di portafogli di attività finanziarie. All’uopo, si impiegano le procedure di partizione ricorsiva per l’individuazione di attività finanziarie di tipo *growth* ed i modelli semiparametrici di tipo *GAM-MM* per la stima del peso di ciascuna attività in portafoglio. L’approccio proposto permette di considerare simultaneamente diverse tipologie di dati, di tipo *cross-section* e longitudinale, e di costruire portafogli misti. Essa risulta particolarmente utile in contesti di *financial data mining*, allorquando si opera con *database* alimentati in tempo reale da un *data provider*. L’affidabilità della procedura proposta è valutata attraverso un’applicazione su dati relativi al mercato europeo provenienti dal sistema *Bloomberg*.

Keywords: Benchmarking, Recursive Partitioning, GAM-MM, Bagging.

1. The Problem

Nowadays, for companies managing Investments Funds it is crucial to succeed in monitoring stocks composing their portfolios and those composing financial indexes. Furthermore, these companies are legally confined to link the yield of their portfolios to some strategic or operative benchmarks, i.e. portfolios of financial indexes. Consequently, the challenge is to track or outperform a benchmark portfolio with an appropriate *stock picking process*, able to detect strong performers activities as well as the moment when to buy or sell them (*market timing*).

To deal with this problem, we present a strategy for the identification of a *managing portfolio* able to outperform a *benchmark portfolio*. It consists in combining suitable tree-structures with the semiparametric Generalized Additive Multi-Mixture Models. The first are used for the identification of stocks having growth potentialities, while the second for the definition of the weight of each stock in the managing portfolio. In this framework, we use two different types of data: in the tree growing step, we use data strictly related to the specific company or institution (like price-book ratio, earnings estimates, excess return, modified duration, etc.), whereas, in the modeling step, we use time series data concerning daily returns of each stock. Our challenge can be considered as a typical financial data mining application, because it involves the identification of relevant information from large (financial) databases, storing real-time data supplied by a financial data provider. These data usually include all the information concerning the different stocks composing the financial indexes included in the benchmark portfolio.

¹ This paper was financially supported by MURST funds (PRIN00; coordinator: Prof. N.C. Lauro).

2. The Proposed Strategy

We deal with a database storing information on n stocks composing the financial indexes defining the benchmark portfolio. These indexes are related to equities, bonds as well as derivatives and structured options. With respect to the type of information in the database we distinguish between two types of datasets, namely the *returns dataset* and the *intrinsic dataset*. The first provides information on trends in the stock price (such as time series data of daily returns of each stock), whereas the second provides information on the intrinsic features of each stock (as, for instance, company structure indicators in the equity case, or specification of the settlements in the bond case, etc.).

In order to define an *outperforming managing portfolio* composed of a restricted number of stocks, we associate the two main steps of the portfolio management activity, namely the *stock picking process* and the *portfolio composition process*, with the following two-stage strategy.

Partitioning Stage. An exploratory tree growing procedure is applied for each intrinsic dataset in order to discover unknown patterns in the data on the basis of a suitable response criterion (i.e., industry sector for equities, issuer for bonds). The aim is to identify a partition of the set of stocks as well as typological variables characterizing each group in the partition. In this way, it is possible to select those stocks whose book-value does not completely express their growing potentiality. We mainly follow the CART methodology of Breiman et. al (1984) to grow binary trees. At the end, the selected stocks for the different datasets are used in the next stage as predictors and concur to the composition of the managing portfolio.

Modeling Stage. The dependence relation between the returns of the financial index forming the benchmark portfolio and the returns of the set of stocks selected in the previous stage is modeled through the Generalized Additive Multi-Mixture Models (GAM-MM) introduced by Conversano et al. (2001a) in the framework of the supervised learning problem. GAM-MM have proven to be effective in data mining applications (Conversano et al., 2001,b) and outperformed the standard GAM approach of Hastie and Tibshirani (1990). In the specific case of the portfolio composition, we define for each predictor (i.e. each stock) a set of weights that penalizes the underperforming returns, and force all the predictors selected in the tree growing stage to enter in the model in order to estimate their weights in the managing portfolio.

3. GAM-MM for Portfolio Composition

In GAM-MM, the conditional expectation of the dependent variable given a set of predictors is expressed by a linear combination of mixtures of classifiers/regressors. We specify a set of regressors/classifiers $F = \{f_1, \dots, f_i, \dots, f_K\}$, including scatterplot smoothers, tree-based models, the linear estimator, etc. The idea is to combine additively suitable mixtures of classifiers associated to each X_j , namely:

$$g(\mu) = \alpha + \sum_{j=1}^d \sum_{i=1}^K \pi_{i|j} f_{ij}(X_j, \theta_{ij}) \quad (1)$$

where f_{ij} is the i -th classifier assigned to the j -th predictor; θ_{ij} is the vector of parameters of the i -th classifier assigned to the j -th predictor; π_{ij} is a mixing parameter, such that $0 \leq \pi_{ij} \leq 1$ and $\sum_i \pi_{ij} = 1$; α is the intercept term; g is an exponential link function relating the additive combination of mixtures of classifiers to the response variable. In (1), the mixing parameters derive from the scores obtained by each of the K classifiers with respect to the j -th predictor. Thus, the additive part of the model might include up to $d \times K$ terms. The estimation procedure involves two steps: the first concerns the estimation of the mixing parameters, and the second the model fitting. In the latter, an inner backfitting-like algorithm provides the estimation of the additive part of the model, and an outer local scoring procedure relates the previous estimates to the response variable.

For the specific case of portfolio composition, we propose a variant of the bagging procedure of Breiman (1996) to estimate the mixing parameters in (1). In particular, we adopt a consensus criterion such that the mixing parameters are given by the *voting* each regressor/classifier achieved over B bootstrap replications, namely: $\pi_{ij} = \sum_b \delta_{ij(b)} \omega_{bj}$,

$\delta_{ij(b)}$ being a dummy parameter equal to one if the i -th regressor/classifier is selected for the j -th predictor in the b -th bootstrap sample and zero otherwise. The selection is made according to an appropriate scoring measure that takes into account the tradeoff between model complexity and parsimony (see Conversano et. al, 2001a).

Note that this variant of the bagging procedure usually adopts a suitable sampling schema to draw cases in each bootstrap replication. Indeed, we penalize the underperforming return cases by assigning an higher weight to the cases where each stock overperformed the relative financial index. In this way, for each predictor (stock) we select the appropriate mixture of regressors/classifiers able to modeling in the best way the overperformance cases.

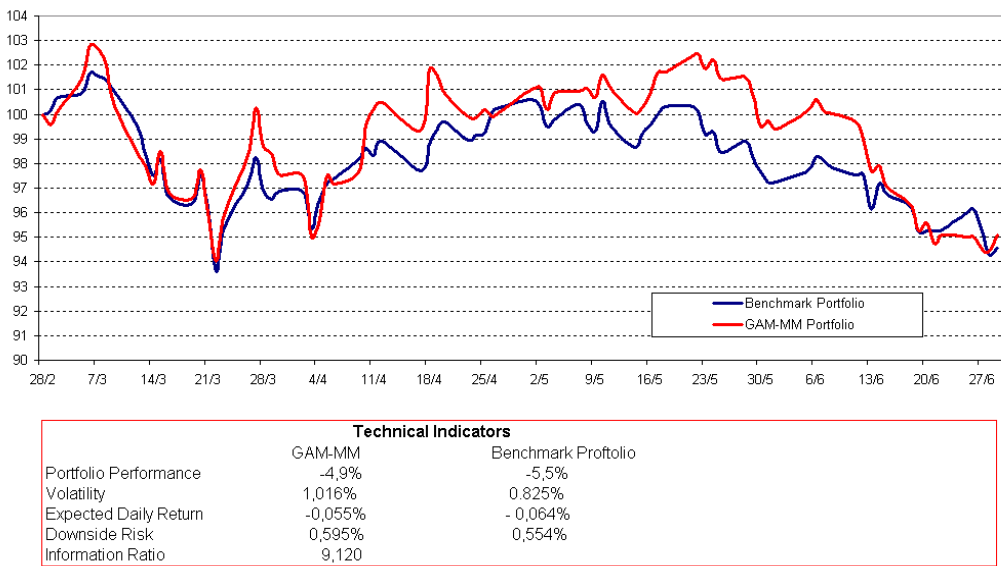
For the portfolio weights definition we forced all the predictors (i.e., the previously selected stocks) to enter in the model, and define the final weight of each stock as the normalized relative decrease in the residual sum of squares caused by the entrance of the corresponding predictor in the model. In particular, the ordering entrance of the predictors in the model is defined according to a measure introduced by Conversano (2001), that compares the model with s predictors with the one including $(s-1)$ predictors ($s = 2, \dots, d$), by considering simultaneously both the reduction in the residual sum of squares and the increase in the proportional number of parameters.

3. An Application on European Financial Data

We summarize the results of an application on a benchmark portfolio composed by two equally weighted European indexes: the *Merryll Lynch Emu Direct Government Euro* (used as benchmark for European Government Bonds) and the *Eurostoxx 50* (used as benchmark for European Equities). For these indexes and for all the stocks composing them, we considered about 50 company indicators (as intrinsic data) as well as the time series of stock returns (return data) observed daily from February 28 to June 30, 2001.

The proposed strategy was performed separately for the two European Indexes. The first stage led to the following selected stocks: a) for the bonds, the procedure pick 10 over the 268 stocks composing the index, that expressed good potentialities in terms of some typological variables (such as the coupon, the excess return, the effective yield and the effective modified duration); b) for the equities, the procedure pick 14 over the 50 stocks composing the index, that expressed good potentialities in terms of other typological variables (such as the beta, the return on asset, the trading volume and the price-earning ratio). The second stage provided the weights of the selected stocks in the managing portfolio. The final managing portfolio was obtained combining with an equal weight the managing portfolios related to the two European indexes. Figure 1 summarizes the results of some performance and risk indicators of the final managing portfolio evidencing its effectiveness with respect to the benchmark portfolio. Due to the lack of space, we refer to the full paper for all the other details.

Figure 1: *Performance and Risk of the managing portfolio with respect to the benchmark portfolio.*



References

- Breiman, L. (1996). Bagging predictors, *Machine Learning*, 26, 46-59.
- Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. *Classification and regression trees*, Belmont C.A. Wadsworth, 1984.
- Conversano C., (2001). “Generalized Additive Multi-Mixture Models: Scoring Models and Predictors”, in Klein B. e Korsholm (eds.) *New Trends in Statistical Modelling, Proceedings of the 16th International Workshop on Statistical Modelling*, University of Southern Denmark, pag. 103-110.
- Conversano C., Siciliano R. and Mola F., (2001a). Generalized additive multi-mixture models for data mining. *Computational Statistics and Data Analysis*, in press.
- Conversano C., Mola F. and Siciliano R, (2001b). Partitioning algorithms and combined model integration for data mining. *Computational Statistics*, 16: 323-339.
- Hastie, T., Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall, London.