Online Ensembles for Financial Trading

Jorge Barbosa[1] and Luis Torgo[2]

[1] MADSAD/FEP, University of Porto,
R. Dr. Roberto Frias, 4200-464 Porto, Portugal
jorgebarbosa@iol.pt
[2] LIACC-FEP, University of Porto, R. de Ceuta, 118, 6., 4050-190 Porto, Portugal
ltorgo@liacc.up.pt,
WWW home page: http://www.liacc.up.pt/~ltorgo

Abstract. This paper describes an application of online ensembles to financial trading. The work is motivated by the research hypothesis that it is possible to improve the performance of a large set of individual models by using a weighted average of their predictions. Moreover, we explore the use of statistics that characterize the recent performance of models as a means to obtain the weights in the ensemble prediction. The motivation for this weighting schema lies on the observation that, on our application, the performance ranking of the models varies along the time, i.e. a model that is considered the "best" at a time t will frequently loose this position in the future. This work tries to explore this diverse and dynamic behavior of models in order to achieve an improved performance by means of an online ensemble. The results of our experiments on a large set of experimental configurations provide good indications regarding the validity of the dynamic weighting schema we propose. In effect, the resulting ensembles are able to capture a much larger number of trading opportunities with a signal accuracy similar to the best individual models that take part on the ensemble.

1 Introduction

Financial markets are highly dynamic and complex systems that have long been the object of different modelling approaches with the goal of trying to anticipate their future behavior so that trading actions can be carried out in a profitable way. The complexity of the task as well as some theoretical research, have lead many to consider it an impossible task. Without entering this never ending argumentation, in this paper we try to experimentally verify a hypothesis based on empirical evidence that confirms the difficulty of the modelling task. In effect, we have tried many different modelling approaches on several financial time series and we have observed strong fluctuations on the predictive performance of these models. Figure 1 illustrates this observation by showing the performance of a set of models on the task of predicting the 1-day future returns of the Microsoft stock. As we can observe the ranking of models, according to the used performance measure (NMSE measured on the previous 15 days), varies a lot,

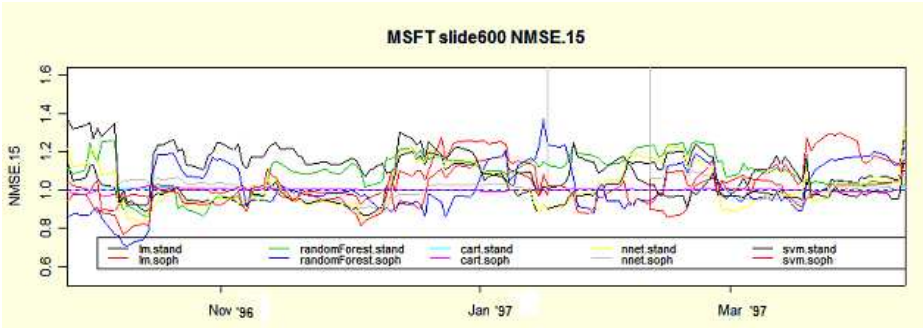

Fig. 1. The performance of different models on a particular financial time series.

which shows that at any given time t, the model that is predicting better can be different.

Similar patterns of performance where observed on other time series and using several other experimental setups. In effect, we have carried out a very large set of experiments varying: the modelling techniques (regression trees, support vector machines, random forests, etc.); the input variables used in the models (only lagged values of the returns, technical indicators, etc.); the way the past data was used (different sliding windows, growing windows, etc.). Still, the goal of this paper is not the selection of the best modelling approach. The starting point of this work are the predictions of this large set of approaches that are regarded (from the perspective of this paper) as black boxes. Provided these models behave differently (i.e. we can observe effects like those shown in Figure 1), we have a good setup for experimentally testing our hypothesis.

Our working hypothesis is that through the combination of the different predictions we can overcome the limitations that some models show at some stages. In order to achieve this, we claim that the combination should have dynamic weights so that it is adaptable to the current ranking of the models. This means that we will use weights that are a function of the recent past performance of the respective models. This way we obtain some sort of dynamic online ensembles, that keep adjusting the ensemble to the current pattern of performance exhibited by the individual models.

There are a few assumptions behind this hypothesis. First of all, we are assuming that we can devise a good characterization of the recent past performance of the models and more over that this statistic is a good indicator (i.e. can serve as a proxy) of the near future performance of the models. Secondly, we must assume that there is always diversity among the models in terms of the performance at any time t. Thirdly, we must also assume that there will always be some model performing well at any time t, otherwise the resulting combination could not perform well either.

Obviously, the above assumptions are what we could consider an ideal setup for the hypothesis to be verified in practice. Several of these assumptions are

difficult to meet in a real world complex problem like financial trading, left alone proving that they always hold. The work we present here can be regarded as a first attempt to experimentally test our hypothesis. Namely, we propose a statistic for describing the past performance of the models and then evaluate an online ensemble based on a weighing schema that is a function of this statistic on a set of real world financial time series.

2 The Dynamic Weighting Schema

Traders do not look for models that achieve a good average predictive accuracy. The reason lies on the fact that the most common return on a stock price is around zero. As such, predicting well on average usually means being very good at predicting these zero-like future returns. However, these predictions are useless for a trader as no one can earn money with such short variations on prices due to the transactions costs. As such, traders are more interested in models that predict well the larger but rare variations [3]. Based on these arguments we have decided not to use a standard statistic of prediction error, like for instance the Normalized Mean Squared Error (NMSE), as an indicator of the past performance of a model. Instead, we have used a measure that is more useful for trading. We have used two thresholds on the predicted return that determine when buy (sell) signals would be issued by an hypothetical trader. These thresholds can be seen as creating 3 bins on the range of returns. In effect, if the predicted future return $\hat{R}$ is above (below) a threshold $\alpha(\mu)$ we generate a **buy** (**sell**) signal, otherwise we have a **hold** signal. This process creates a discretized target variable, the predicted signal.

Using this discretization process we can then calculate statistics regarding the signals predicted by the models. Namely, we have calculate the *Precision* of the predictions as the proportion of buy (sell) signals that are correct (i.e. correspond to returns that really overcame the thresholds). We have also calculated the *Recall* as the proportion of true signals (i.e. real returns that were above (below) the threshold) that are signaled as such by the models. Finally, we have combined these two measures into a single statistic of performance using the F-measure [2].

The F-measure calculated on the previous 15 days was the measure of past performance that we have used to obtain the weights of each model in the ensemble. Namely, the combined prediction at any time t was obtained by,

$$\hat{R}_{t+1,ens} = \frac{\sum_{k=1}^{S} \hat{R}_{t+1,k} \times F.15_k}{\sum_{k=1}^{S} F.15_k} \tag{1}$$

where $F.15_k$ is the value of the F-measure calculated using the predictions of model k for the time window $[t-15..t]$.

3 Experiments and Results

We have carried out a very large set of experiments designed to test several instantiations of our working hypothesis [1]. Due to space restrictions we will

Table 1. The results for the DELL stock.

	±1%		±1.5%		±2%	
	Prec	Rec	Prec	Rec	Prec	Rec
lm.stand	0.405	0.077	**0.472**	0.017	0.364	0.006
lm.soph	**0.453**	0.173	0.460	0.045	0.519	0.014
randomForest.stand	0.434	0.227	0.402	0.080	0.328	0.026
randomForest.soph	0.420	0.321	0.399	0.154	0.345	0.075
cart.stand	0.444	0.004	0.444	0.004	**0.600**	0.003
cart.soph	0.231	0.004	0.154	0.004	0.154	0.004
nnet.stand	0.388	0.246	0.353	0.128	0.289	0.075
nnet.soph	**0.453**	0.063	0.360	0.037	0.323	0.030
svm.stand	0.391	0.380	0.357	0.190	0.326	0.075
svm.soph	0.415	0.360	0.397	0.189	0.373	0.097
Ensemble	0.438	**0.451**	0.354	**0.287**	0.296	**0.210**

limit our description to the ensembles formed by a weighted combination of predictions using Equation (1).

In our experiments we have tried three different setups in terms of thresholds on returns for generating trading signals, namely, $\pm 1.0\%, \pm 1.5\%$ and $\pm\, 2\%$.

We have compared our dynamic online ensembles with several individual models that participated on the ensemble. We present the results in terms of *Precision* and *Recall*, as different traders may require different tradeoffs between these two conflicting statistics. We have carried out this and other experiments on 9 different financial time series containing daily data for more than 10 years. The experiments were carried out over this large time period of time using a sliding window approach. Due to space limitations we can only show the results for one stock, DELL, which involves around 10 years of testing (approximately 2500 daily predictions). The reader is referred to [1] for full details.

Table 1 shows the results of the individual models and of the ensemble for the DELL stock. The first observation we can make is that the results are generally poor. The individual methods achieve quite low values of *Recall* and usually less than 50% of *Precision*. A possible explanation for these poor results is the fact that all methods are obtained by optimizing a criterion (usually some form of mean squared error) that is an average error estimator and thus will not be optimal for predicting the extreme low and high returns [3].

Although we can observe a few high *Precision* scores these are usually obtained with too few trades (very low *Recall*) to make these strategies worth investing[3].

Regarding the results of our dynamic ensembles we can generally state that they are interesting. In effect, we have a much higher value of *Recall* with a *Precision* that is most of the times near the best individual scores. The results

[3] Although *Recall* and *Precision* do not directly translate into trading results, they still provide good indications as they are more related to trading decisions than pure prediction error metrics like mean squared error, for instance.

in terms of *Recall* get more impressive as we increase the thresholds, i.e. make the problem even harder because increasing the thresholds means that trading signals are even more rare and thus harder to capture by models that are not designed to be accurate at predicting rare values. Still, the *Precision* of the ensemble signals also suffers on these cases, which means that the resulting signals could hardly be considered for real trading. Nevertheless, the results with the thresholds set to $\pm 1\%$ can be considered worth exploring in terms of trading as we get both *Precision* and *Recall* around 45%. The best individual model has a similar *Precision* but with only 17.3% *Recall*.

The same general pattern of results were observed on the experiments with other stocks. In summary, this first set of experiments provides good indications towards the hypothesis of using indicators of the recent performance of models as dynamic weights in an online ensemble in the context of financial markets.

4 Conclusions

In this paper we have explored the possibility of using online ensembles to improve the performance of models in a financial trading application. Our proposal was motivated on the observation that the performance of a large set of individual models varied a lot along the testing period we have used. Based on this empirical observation we have proposed to use dynamic (calculated on moving windows) statistics of the performance of individual models as weights in an online ensemble.

The application we are addressing has some particularities, namely the increased interest on accurately predicting rare extreme values of the stock returns. This fact, lead us to transform the initial numerical prediction problem into a discretized version where we could focus our attention on the classes of interest (the high and low returns that lead to trading actions). As a follow up we have decided to use *Precision* and *Recall* to measure the performance of the models at accurately predicting these trading opportunities. Moreover, as statistical indicator of recent past performance we have used the F-measure that combines these two statistics.

The results of our initial approach to dynamic online ensembles in the context of financial trading are promising.We have observed an increased capacity of the ensembles in terms of signaling the trading opportunities. Moreover, this result was achieved without compromising the accuracy of these signals. This means that the resulting ensembles are able to capture much more trading opportunities than the individual models.

As main lessons learned from this application we can refer:

- Diverse behavior of models on predicting complex dynamic systems: in problems with so many unrecorded factors influencing the dynamics of a system is hard to find a predictive model that is good for all regimes. We have observed similar beahviors on other problems like for instance predicting algae blooms in freshwater systems.

– Predicting extreme values is a completly different problem: we have confirmed our previous observations [3] that the problem of predicting rare extreme values is very different from the standard regression setup, and this demands for specific measures of predictive performance.
– Dynamic online ensembles are a good means to fight instability of individual models: whenever the characteristics of a problem lead to a certain instability on models' performance, the use of dynamic online ensembles is a good approach to explore the advantages of the best models at each time step.

As future work we plan to extend our study of recent past performance statistics. We also plan to explore this hypothesis using base models that are better tuned towards the prediction of rare extreme values.

References

1. J. Barbosa. Metodos para lidar com mudancas de regime em series temporais financeiras - utilizacao de modelos multiplos na combinacao de previsoes (in portuguese). Master's thesis, Faculty of Economics, University of Porto, 2006.
2. C. Van Rijsbergen. *Information Retrieval*. Dept. of Computer Science, University of Glasgow, 2nd edition, 1979.
3. L. Torgo and R. Ribeiro. Predicting rare extreme values. In W. Ng, editor, *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'2006)*, number 3918 in Lecture Notes in Artificial Intelligence. Springer, 2006.