Appears in "Neural Networks in the Capital Markets",
Proceedings of the Third International Conference (London, October 1995),
A. Refenes, Y. Abu-Mostafa, J. Moody, and A. Weigend, eds., World Scientific, London, 1996.

Trading with Committees: A Comparative Study

Steve Rehfuss, Lizhong Wu and John Moody
Computer Science Dept., Oregon Graduate Institute
Portland, OR 97291-1000, USA
E-mail: stever@cse.ogi.edu, lwu@cse.ogi.edu, moody@cse.ogi.edu

ABSTRACT

Combining experts by averaging their forecasts can be useful for prediction in environments where the individual experts are noisy. In decision-making situations such as trading systems, another possibility is to use *voting committees*, where committee members first decide individually, and then the individual *decisions* are used to produce the final decision. We compare combining forecasts with three types of voting using a trading system developed for the INFFC competition.

1. Introduction

Our basic approach to prediction is to use a *combination of forecasts*, or *committee of experts*[4]. For pure prediction or regression, a set of networks (linear and nonlinear) are trained, and a subset with good prediction performances are selected; the *combined prediction* is the (unweighted) average of the individual predictions. Unweighted averages are used as they have been shown to be generally more robust, especially in non-stationary environments[3,2]

For making decisions, however, it may be useful to use a *voting committee*, where a number of predictors are used individually to make trading decisions, and then the individual decisions are merged in some way to produce a final decision. We consider three such voting schemes.

2. Data

We compare these forms of committees using a trading system developed for the International Nonlinear Financial Forecasting Competition (INFFC). The training data consist of around 80000 ticks of a slightly manipulated commodity futures series. Only training data was given to participants, who then

submitted a working system for evaluation on a (non-distributed) continuation, so the test data is unknown to us.

3. Adaptation

A critical feature of our trading system is adaptation to the nonstationarity of the series [Moody, Levin, Rehfuss, 1993], [Moody, Wu, 1994]. This takes several forms:

- **Retraining.**

  Models are retrained over time. For certain models, especially linear ones, this retraining happens at every tick. For models where the retraining cost is larger, retraining happens periodically.

- **Exponential Decay.**

  Information decays over time. Where possible, e.g. fitting of linear predictors, input information is decayed exponentially with a time constant of about 10^4 ticks. In some situations, e.g. neural net predictors, this is not appropriate and is not done. In addition to any decay, models are fit using only inputs lying in a fixed window of the last several thousand ticks. The effect on performance of the decay constant can be seen in figure 1.

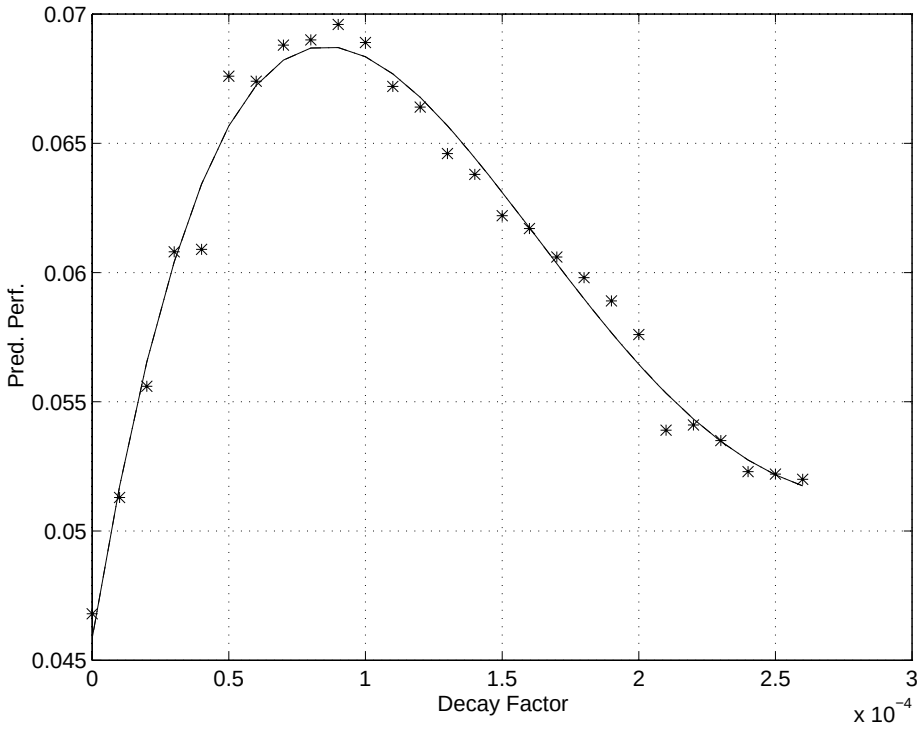


Figure 1: Prediction performances versus decay factors. The prediction performance is measured by the difference between the fractions of correct predictions and incorrect predictions. The points on the graph are the test results. The solid curve is a smooth curve fit to these points. Optimal prediction performance occurs when the decay factor is about equal to 10^{-4}. The test period is from tick 10,000 to tick 80,000 of the training data.

- **Outlier Detection.**

  Input variables are checked for being outliers, by comparison with their "recent" empirical distribution as determined over the last few thousand ticks. A value is classified as an outlier if it is more than some number of standard deviations from the mean. Trading is disallowed at ticks where there are outliers.

4. Trading

We compare combination of forecasts with three types of voting committees using a simple trading strategy.

- The combined forecast or voting committee is used to produce a trading signal, long, short, or neutral, as described below.

- If the current position is neutral, and there is a non-neutral trading signal, the signal is acted on, entering the market. If the current position is not neutral, and a fixed amount of time has passed since the market was entered, the current trading signal is acted on, taking the indicated position: short, long or neutral.

- No trades are made, and positions are held, whenever: the combined prediction appears to be an outlier, an input outlier has been detected, or at times when the statistics of the series are known to be "abnormal"[a]. The useful effect of this "invalidation" of ticks can be seen in figure 2.

In the *combination of forecasts* version of our system, the trading signal is produced as follows: the combined forecast is a prediction of the future return over an interval starting at the current time and having a certain fixed horizon. It is formed as the simple average of the return predictions of the individual predictors. If the (combined) predicted return deviates from its mean value by more than a fixed factor of its standard deviation, the *prediction threshold*, this is taken as a trading signal of the corresponding sign. Otherwise, the trading signal is taken to be "neutral". Both the mean and standard deviation are calculated over the last few thousand ticks.

In the *quantized predictor* version, the trading signal is instead produced as follows: each individual prediction of future return is used to produce a trading signal by comparing it to its local mean and standard deviation, exactly

[a] For the INFFC data, these times include the beginning and ending of the day, and days with abnormally low trading volume (estimated at the beginning of the day).

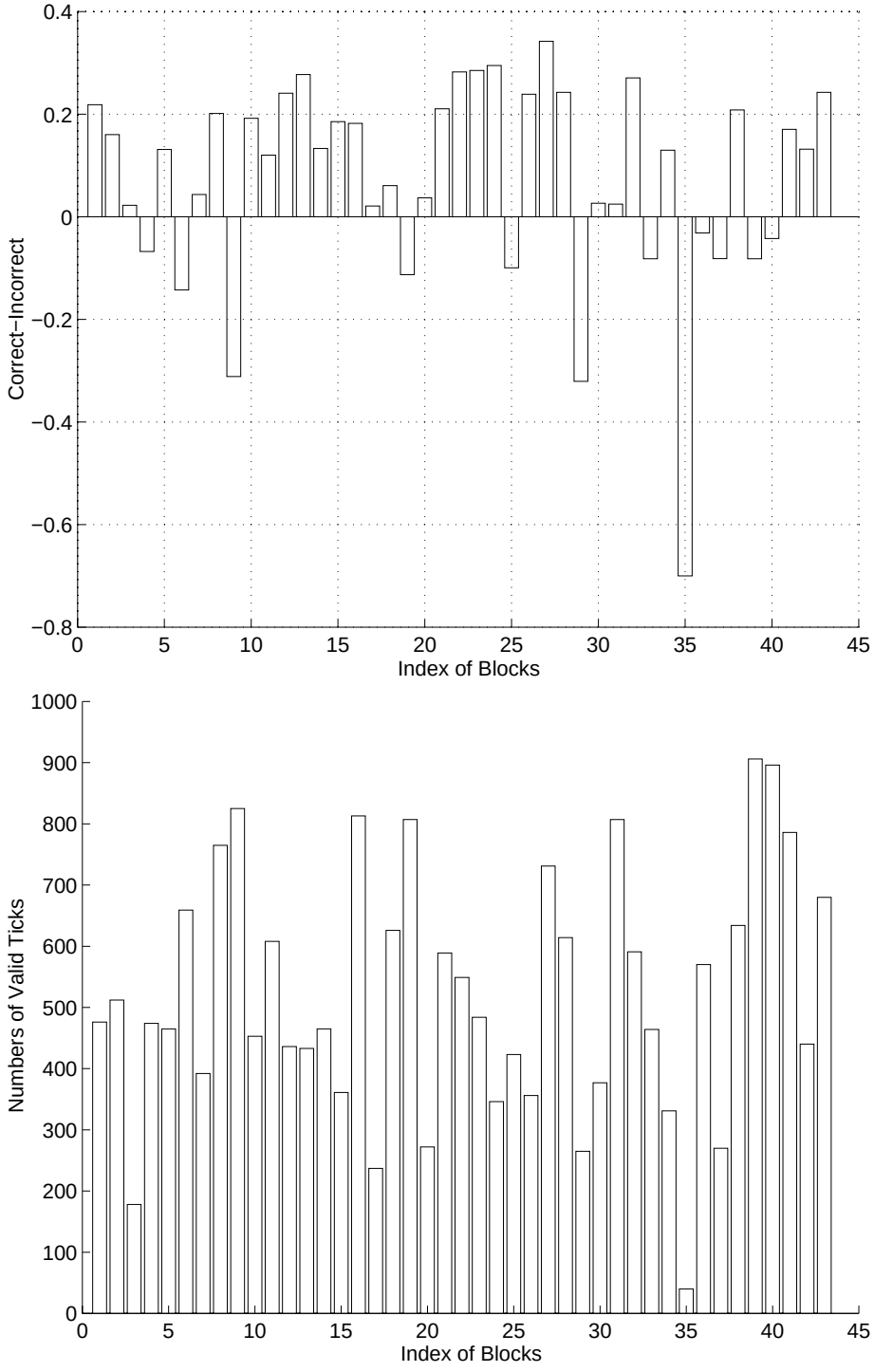


Figure 2: Difference of percentage correct and incorrect over a sequence of non-overlapping blocks. Each block consists of 1600 ticks. Overall, the INFFC training data from line 10,000 to 80,000 is divided into 43 blocks. There are 31 blocks for which the percentage of correct predictions is larger than that of incorrect predictions. The mean of the difference is +0.07 (7%). The lower panel plots the number of "valid" ticks of each block. A trading signal may be issued only on a valid tick. In general, the number of valid ticks is small when the prediction performance is poor, as in the 35th block in the graph, although not always (block 9).

Individual Predictor	Percentage of		
	Correct	Incorrect	Otherwise
1	52.12	45.66	2.22
2	51.92	45.85	2.22
3	52.22	45.56	2.22
4	52.01	45.77	2.22
5	52.23	45.54	2.22
6	52.02	45.76	2.22
7	51.99	45.78	2.22
8	51.84	45.94	2.22
m$\pm\sigma$	52.04$\pm$0.14	45.73$\pm$0.14	2.22$\pm$0
Combined	52.33	45.44	2.22

Table 1: Summary of combined predictor performance. The first 8 rows give the performance of the individual predictors, and the 9th row lists their means and standard deviations. The last row gives the performance of the combined forecast. The second, third and fourth columns give the percentage of time that the prediction and the target have the same sign ($x\hat{x} > 0$), different signs ($x\hat{x} < 0$), or are zero ($x\hat{x} = 0$), respectively. The percent difference of correct and incorrect predictions is about 7%. The results are based on ticks 10,000 to 80,000 of the training data.

as above for the combined predictor. The individual trading signals are then added, taking "long" as +1, "short" as -1, and "neutral" as 0. If the sum exceeds a certain positive threshold, the committee produces a "long" signal, if it falls below the negative of that threshold, the committee produces a "short" signal, otherwise a neutral signal is produced. Note that this system has two thresholds.

In the *bagging predictor* version[1], the individual votes are produced as above, and the final decision is whichever of "short", "long", or "neutral" gets the most votes.

In the *mixture* version, the trading signal is produced as in the quantized predictor scheme, the difference being that the committee members include not only the individual predictors, but also the combined forecast itself. In addition, the combined forecast is given two votes in the committee, as opposed to a single vote for each individual predictor.

The trading strategy is quite sensitive to the particular test interval and to the "hyperparameters", such as prediction threshold, prediction horizon, information decay factor, beginning and ending time of day and minimal daily volume for allowed trading. It may seem that this is a consequence of the rigid

rules for entering and leaving the market, but, in fact, all other strategies we have examined for the INFFC data have been at least as sensitive to these or other parameters. For the results below, unless specified otherwise, we selected hyperparameters maximizing the mean and median of Sharpe ratios computed over different blocks of the training set, while at the same time, minimizing the standard deviation of those Sharpe ratios.

5. Performance

The measure of trading system performance adopted for the INFFC was the *Sharpe ratio* computed over a test interval, defined as follows:

- at time t, the trading system produces a signal $s_{t+1} \in \{-1, 0, 1\}$, where -1 corresponds to taking a short position, 0 to taking a neutral one (leaving the market, not remaining unchanged), and 1 to going long.
- the change in the value of the position is then $g_t = s_{t-1} * (x_t - x_{t-1})$, where x_t is the (close) price series, and the transaction cost is $h_t = .001 * |s_t - s_{t-1}|$
- the "normalized monthly profit" at time T is defined as

$$p_T = \frac{\sum_{t=T-7200}^{T} g_t}{\frac{1}{7200}\sum_{t=T-7200}^{T} x_t} - \sum_{t=T-7200}^{T} h_t$$

  where 7200 is the number of minutes (not all containing ticks) in a 20 day month.

- sampling the p_T series monthly (every 7200 minutes) over a test interval gives a new series with mean μ_p and variance σ_p^2. The Sharpe ratio is defined as

$$SR = (\mu_p - 0.0025)/\sigma_p$$

  where 0.0025 is the mean monthly risk-free interest rate.

The Sharpe ratio rewards profit series that exceed the risk-free rate of interest after accounting for transaction costs, and that are steady from month to month.

We also measured results with the *Sterling ratio*. This is defined identically to the Sharpe ratio, except that instead of normalizing by the standard deviation σ_p, the *maximum drawdown* is used instead. The maximum drawdown is defined as the sum of the largest contiguous sequence of losses that occur in the time interval over which μ_p is computed. Use of the Sterling ratio gave the same qualitative results as the Sharpe ratio, and is not further reported here.

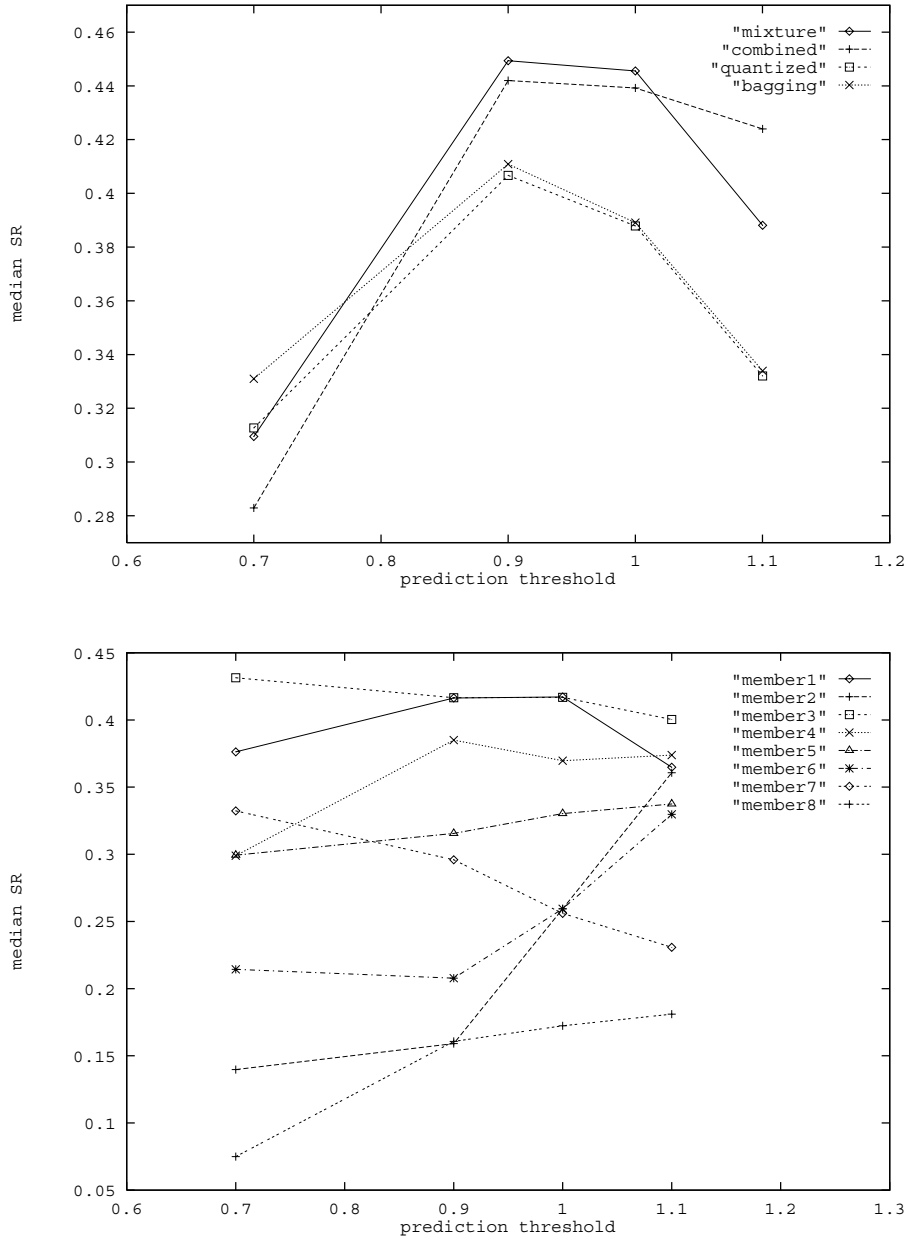
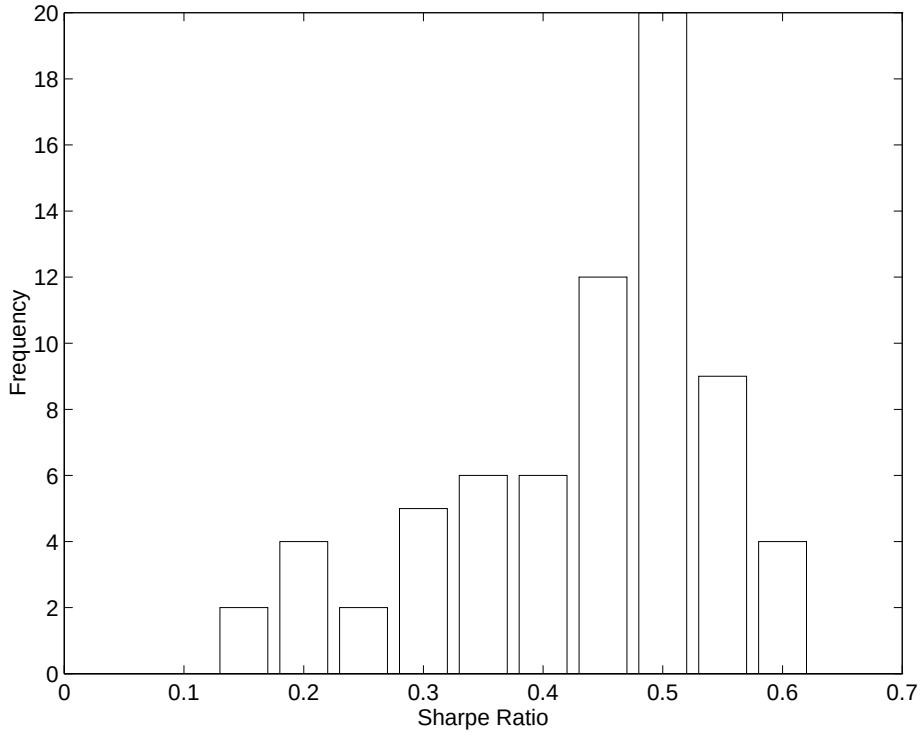
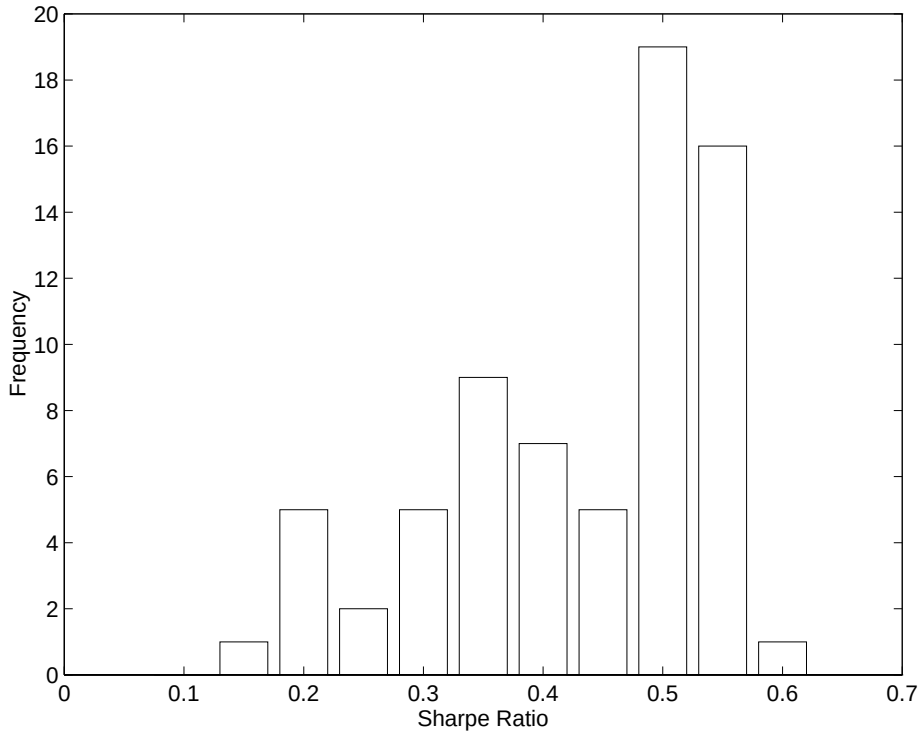


Figure 3: Performance of committees and individual members vs. prediction threshold. Panel 1 compares the different forms of committees, panel 2 shows the performance of the individual members. Each result is the median of a histogram of the Sharpe ratio of the trading system when started at different initial ticks.





	Mean	Std. Dev.	Min.	Max.	Median
combined	0.434	0.117	0.127	0.632	0.460
mixture	0.431	0.112	0.166	0.604	0.484

Figure 4: Histogram of Sharpe ratios computed with different initial ticks. Panel 1 is for a trading system based on combination of forecasts. Panel 2 is for a trading system based on a mixture committee. Each ratio is calculated over 40,000 ticks; the initial ticks range from 12,000 to 40,000. The lower table

Table 1 shows the performance of individual forecasts, compared to that of the combined forecast. As predicted returns are seldom exactly zero, we used as a figure of merit the number of times the sign of a non-zero return was correctly predicted. It can be seen that use of a combined forecast gives somewhat improved performance. This is slight, however, suggesting that the individual forecasts are highly correlated. The performance of the individual members is also shown in panel 2 of figure 3, where the median Sharpe ratio (described below) over the test interval is plotted as a function of the prediction threshold. In this metric, the responses of the individual members are seen to be quite different, in spite of their correlation.

The two panels of figure 4 compare using a mixture committee to using a combination of forecasts. Each result is presented as a histogram of the Sharpe ratio of the trading system when started at different initial ticks. This is necessary, due to the sensitivity of the trading system to the starting point. We see that the mixture committee distribution has higher (better) median and minimum values, but lower maximum and mean values. As well, it has a smaller variance. The mixture committee would seem to be preferred from "minimize maximum loss" and robustness points of view: minimum value is higher, mean plus one standard deviation is higher, median is higher. However, it is not clear that the differences are significant. Figure 3 plots the median of this distribution for the various kinds of committee and for the individual members, as a function of the prediction threshold. We see that the mixture committee is somewhat better in its optimal range of prediction threshold, that the pure voting and bagging committees are substantially worse than either the mixture or combining committees, and that the individual members are quite variable in their behavior, when measured by the Sharpe ratio, rather than by their percentage correctness as in table 1. In fact, the quantizing and bagging committees are both worse than the two best individual committee members, 1 and 3. This is not surprising, given highly correlated individual predictors[1].

6. Conclusions and Future Work

The experiments reported here are somewhat inconclusive. While voting and bagging committees perform poorly here, they can make use of predictors trained as classifiers rather than regressors, and this should be tried. No clear superiority of mixture versus combining committees is apparent, although mixture committes seem slightly preferable from a "minimize maximum loss" point of view. Nonetheless, combining experts into a committee is a useful way of

decreasing model variance in noisy situations. In decision-making situations, it may be useful to combine the *decisions* of the members, rather than their forecasts.

7. Acknowledgements

We wish to thank Hong Pi and Denni Rögnvaldsson for many useful discussions. We gratefully acknowledge support for this work under ARPA and ONR grants N00014-92-J-4062 and N0001-94-1-0071, and under NSF Postdoctoral Grant CDA-9309728.

8. References

1. Breiman, L. "Bagging Predictors", unpublished.
2. Clemen, Robert T, "Combining forecasts: a review and annotated bibliography", *Intl. J. Forecasting*, **5**, 1989, pp 559-583.
3. Granger, C. W. J., and P. Newbold, *Forecasting Economic Time Series*, 2nd ed., Academic Press, San Diego, California, 1986, pp 265 - 276.
4. Moody, J., A. Levin, and S. Rehfuss, "Predicting the U.S. Index of Industrial Production", *Neural Networks in the Capital Markets 93*, London, England, November 1993
5. Moody, J., and L. Wu, "Statistical Analysis and Forecasting of High Frequency Foreign Exchange Rates", *Neural Networks in the Capital Markets 94*, Pasadena, California, November 1994