Composite Neighbors for Case Based Prediction: Structural Effects on Stock Price Forecasting

Steven H. Kim and Dae Suk Kang
Graduate School of Management
Korea Advanced Institute of Science and Technology
Seoul, Korea

Abstract

Learning methodologies such as neural networks or genetic algorithms usually require long training times. Case based reasoning, however, attains peak performance swiftly and is often appropriate for learning even with small data sets. Previous work has shown that an extended case reasoning methodology can yield superior performance in the task of predicting financial data series. This paper examines the impact of reasoning procedures on stock price prediction. The following characteristics are evaluated: size of input vector, multiplicity of neighboring states, and a scaling factor for growth. The concepts are illustrated in the context of predicting the price of an individual price.

Key words:
Case-based reasoning, knowledge discovery

INTRODUCTION

Case based reasoning offers a number of advantages for predicting time series data (Kim, 1994a, 1995a). However, the impact of various parameters on system performance has not yet been extensively investigated. This paper is an effort in that direction.

The effect of the size of the input vector and the number of neighbors is examined, as well as a scaling factor for secular growth. These parameters are investigated through a series of 841 computational experiments.

METHODOLOGY

In a larger sense, CBR is difficult to avoid in practical contexts. Decisions are made against the light of past experience, whether such information is encoded into a machine-compatible format or is available only informally. In addition, each new problem and the attendant solution constitute a case. As a result, case-base reasoning is unavoidable in experiential modes of learning whether or not the techniques are labeled as "CBR".

An intelligent learning algorithm should take account of a "virtual" or composite neighbor whose parameters are defined by some weighted combination of actual neighbors in the case base. In this way, the algorithm can utilize the knowledge reflected in a larger subset of the case base than the immediate collection of proximal neighbors (Kim, 1995b).

Case-based Reasoning. In general, a CBR cycle consists of the following four processes as shown in Figure 1 (Aamodt and Plaza, 1996).

1. Retrieve the most similar case or cases.
2. Reuse the information and knowledge in that case to solve the problem.
3. Revise the proposed solution.
4. Retain the parts of the example which may be useful for future problem solving.

As with any computational technique, the use of effective data structures is a vital aspect of implementing case based reasoning. In procedural terms, the key aspects are the indexing of old cases and the retrieval of "neighbors" which resemble the target case (Kim and Novick, 1993; etc.).

Composite Neighbor. Conventional methods of prediction based on declarative reasoning usually seek the nearest neighbor in the observational space. In real life, our predictions are guided not only by the proximity of neighboring cases, but also by their

density. We may take account of a "virtual" or composite neighbor whose parameters are defined by some weighted combination of actual neighbors in the case base. The key to the composite approach lies in the determination of the most effective set of weights to use in order to construct the virtual neighbor. In addition to utilizing the knowledge inherent in a larger set of neighbors, an advantage of the weighted indexing approach lies in the practicality of a probabilistic forecast. With this approach, a spectrum of forecasts may be provided in conjunction with their associated probabilities (Kim, 1995b).

Probabilistic Prediction. Probabilistic prediction refers to the estimation of the future value $s(t+\lambda)$ of a process $s(t)$ in terms of its past $s(t-\tau)$ for $\lambda > 0$ and $\tau > 0$. Many phenomena in process prediction and control exhibit quasi-random characteristics. In such stochastic domains, the system may be modeled by including probability distributions over state transitions (Kim, 1994b).

CASE STUDY

The case study involves the prediction of IBM stock on each market day. The raw input consisted of 590 trading days, from 30 August 1993 to 8 March 1996. The input data are the daily closing prices. The goal of the predictive system at time t is to forecast the closing price on the next day, t + 1, using information available up to the current date t. The conceptual architecture for the case study is illustrated in Figure 2.

Different types of scaling were used to construct the goal variable to be predicted. One is the conventional unit scaling employed in previous work (Kim and Kang, 1996). Another approach is the use of a growth ratio to take into account the secular change in the variable of interest. This approach is clarified in Figure 3.

RESULTS

As shown in Figure 4, the main influence on predictive accuracy was the number of neighbors rather than the size of the input vector. Using a composite input vector of size 3 in conjunction with 3 neighbors, the predictions are plotted together with actual values in Figure 5. The discrepancy is depicted more clearly in Figure 6.

In addition, the impact of varying the scaling factor is shown in Table 1. The results indicate that the growth scaling factor improved predictive accuracy.

CONCLUSION

Extended case based reasoning using composite neighbors offers a promising approach to swift, accurate learning. In the task of predicting a stock price, the number of neighbors was shown to be more important than the multiplicity of elementary cases in the input vector. Moreover, the use of a scaling factor in constructing the goal variable led to improved accuracy.

REFERENCES

Aamodt, Agnar and Enric Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches." *Artificial Intelligence Communications*, v. 7, http://www.iiia.csic.es/People/enric/AICom_Toc.html, 1996.

Dagum, P., et al., "Uncertain reasoning and forecasting.", International Journal of Forecasting, 1995 : 73-87.

Gazula, Srinivas and Mansur R. Kabuka, "Design of Supervised Classifiers Using Boolean Neural Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v.17, 1995: 1239-1246.

Kim, S. H. *Knowledge Systems through Prolog.* New York: Oxford Univ. Press, 1991.

Kim, S. H. "Learning Systems for Process Automation through Knowledge Integration." *Proc. Second World Congress on Expert Systems*, Lisbon, Portugal, 1994a.

Kim, S. H. *Learning and Coordination.* Dordrecht, Netherlands: Kluwer, 1994b.

Kim, S. H. and M. B. Novick. "Using Clustering Techniques to Support Case Reasoning." *International J. of Computer Applications in Technology*, v.6(2/3), 1993: 57-73.

Kim, S. H. "Integrated Knowledge Discovery for Investment prediction: Critical Issues and System Design Implications." Work Paper, Management Information Systems, KAIST, Seoul, Korea, 1995a.

Kim, S. H. "Knowledge Based Prediction of Nonlinear Phenomena." Work Paper, Management Information Systems, KAIST, Seoul, Korea, 1995b.

Kim, S. H. and Kang, D. S., "Implicit versus Explicit Learning for Forecasting: Case Study in Intraday Stock Index Prediction", Work Paper, Management Information Systems, KAIST, Seoul, Korea, 1996.

Kjærulff, Uffe. "dHugin: a computational system for

dynamic time-sliced Baysian networks.", *International Journal of Forecasting*, 1995: 89-111.

Kolodner, Janet. *Case-Based Reasoning*. Morgan Kaufmann Publishers, 1993.

Mott, Steve. "Insider Dealing Detection at the Toronto Stock Exchange." In Suran Goonatilake and Philip Treleaven. *Intelligent Systems for Finance and Business*. NY: Wiley, 1995.

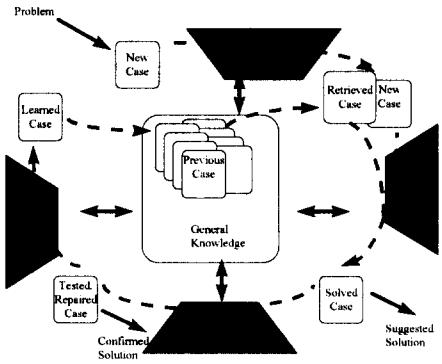

Figure 1. The CBR Cycle (Source: Aamodt and Plaza, 1996).

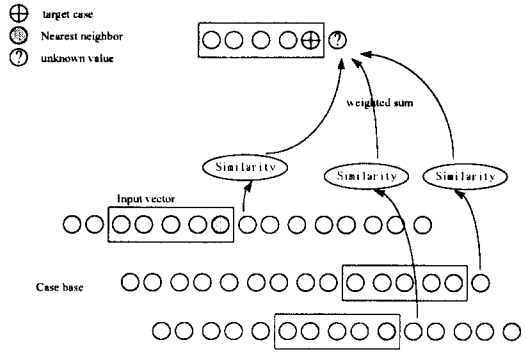

Figure 2. Conceptual representation of the predictive task in CBR.

j	time	Low	High	Close X_{t_j}	Next Close X_{t_j+1}
1	t_1	10	20	10	11
2	t_2	20	30	20	22
Target	t_*	30	40	30	?

Suppose neighbor weights are equal: $w_1 = w_2 = 0.5$.

Prototype value of the goal variable (tomorrow's close, X_{t_*+1}) is constructed using corresponding information from the neighbors. The predicted goal value

$$\hat{X}_{t_*+1} = \sum_{j=1}^{J} w_j \lambda_j X_{t_j+1}$$

where $\lambda_j \equiv X_{t_*} / X_{t_j}$ is the scaling factor for neighbor j.

For the example above:

$$\lambda_1 = 30/10 = 3, \lambda_2 = 30/20 = 1.5;$$

$$\hat{X}_{t_*+1} = (0.5)3(11)+(0.5)1.5(22) = 33$$

Note that $\lambda_j \equiv 1$ corresponds to conventional unit scaling. For this choice, the forecast vale is $(0.5)(11)+(0.5)(22) = 16.5$.

Figure 3. Definition and illustration of the growth scaling factor.

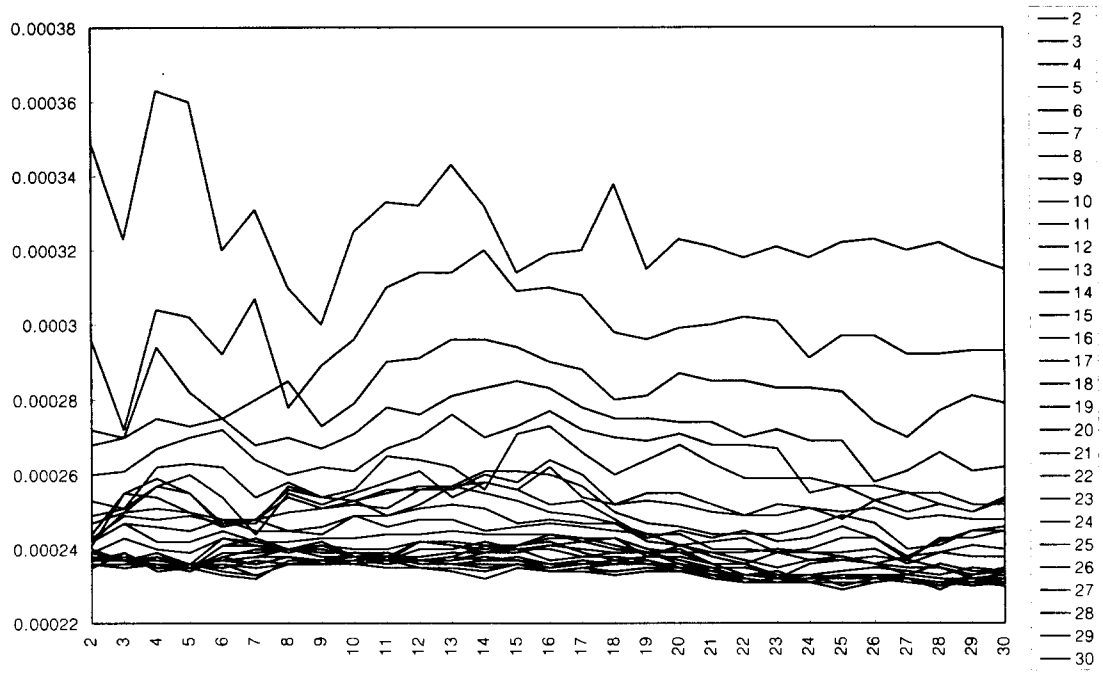

Figure 4. Overall MSE (Mean Squared Error) as a function of input vector size (N), using the number of neighbors (J) as a parameter.

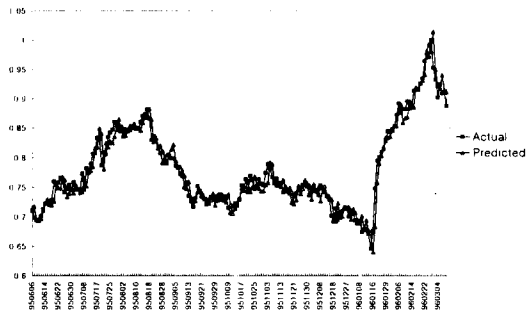

Figure 5. Plots of actual and predicted values (size of input vector = 3, # of neighbors = 3).

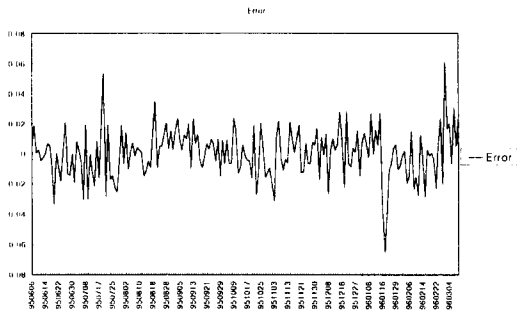

Figure 6. Residuals associated with Figure 5 (size of input vector = 3, # of neighbors = 3).

Scaling factor	MSE
unity	0.000371
growth ratio	0.000272

Table 1. Residuals associated with the use of alternative scaling factors.