Learning Prediction of Time Series. A Theoretical and Empirical Comparison of CBR with some other Approaches*

Gholamreza Nakhaeizadeh
Daimler-Benz AG, Research and Technology
Postfach 2360, D-89013 Ulm, Germany
reza@fuzi.uucp

Abstract

Case-based Reasoning (CBR) is a rather new research area in Artificial Intelligence. The concept of K-Nearest Neighbours (KNN) that can be considered as a subarea of CBR traced back, however, to early fifties and during the last years it is deeply investigated by the statistical community. In dealing with the task "learning prediction of time series", besides the KNN-approach, the Statistician have investigated other approaches. Recently, neural networks and symbolic machine learning approaches are applied to performing this task as well. Although learning prediction of time series is a very important task, there is no comprehensive study in the literature which compares the performance of CBR with the performance of the other alternative approaches. The aim of this paper is to contribute to this debate.

Introduction

Besides the information about the past values of the time series itself, one can also use other information based on the exogenous indicators which have a significant impact on the development of the time series. K-Nearest-Neighbours and regression analysis can be mentioned as examples for such procedures. Recently, the attention is also focused on the application of Neural Networks. Some of symbolic machine learning algorithms based on ID3-concept can be used to predict the development of time series as well (Merkel and Nakhaeizadeh (1992)). It should be mentioned that although CBR-based approaches have found several applications for examples in classification, planning and design (see Althoff et al. (1992), Veloso (1992)), very little attention has been paid to the application of CBR to time series prediction. An exception is the work of Quinlan (1993) which applies both CBR-based and model based learning approaches to the prediction task. The CBR-approach used by Quinlan deals with the Instance-Based Learning (IBL) investigated by Aha et al. (1991).

*This is an extended abstract of a large paper to appear in the proceedings of the EWCBR(93), University of Kaiserslautern, Germany.

The above facts show that several alternative approaches can be applied to the prediction of time series. The aim of this study is to evaluate, firstly, these alternative approaches from a theoretical point of view and, secondly, to compare their performance in dealing with real-world prediction problems arise in industry and commerce. We will also refer to some results achieved within an Esprit-Project funded by the European Community.

A Short Description of the Applied Alternative Approaches

1.Statistical approaches

Denoting Y_t as a time series in period t, a linear regression model can be described by the equation

$$Y_t = a + \sum_{i=1}^{n} b_i X_{it}$$

In the above equation, X_{it} denotes the value of exogenous variable X_i in the period t. The value Y_{t+1} in the period $t+1$ can be predicted simply as:

$$\hat{Y}_{t+1} = \hat{a} + \sum_{i=1}^{n} \hat{b}_i X_{i(t+1)}$$

where $\hat{a}$ and $\hat{b}_i$ are the estimations for a and b_i and can be calculated using least-squares or maximum-likelihood method. Of course, one can use instead of a linear regression a nonlinear model as well. In this case, the parameters a and b_i can be estimated using numerical procedures. The regression analysis is theoretically well investigated and it is very simple to apply. One disadvantage of this method is the problem of model selection. A lot of other statistical approaches have the same disadvantage as well. The other problem is that the calculation of $\hat{Y}_{t+1}$ is only possible when all $X_{i(t+1)}$ are known for the period $t+1$ in advance, which is in praxis not always the case.

Concerning the Box-Jenkins approach, one can describe an ARMA (autoregressive moving average) model as:

$$Y_t + \alpha_1 Y_{t-1} + ... + \alpha_p Y_{t-p} = \epsilon_t + \beta_1 \epsilon_{t-1} + ... + \beta_q \epsilon_{t-q}$$

where ϵ_t are independent normal distributed random variables.

If the parameters α or β are zero, the above model will be reduced to a MA (moving average) or AR (autoregressive) process, respectively.

The main assumption in the ARMA model is that the time series Y_t is stationary. A time series is stationary if its mean and variance remain unchanged with time. For a lot of real world time series, this assumption is not valid. In such cases, the time series should be transformed for example by taking successive differences so long as necessary to make the resulting series stationary. In this case, the original series is called an integrated ARMA process, i.e. an ARIMA process. Although the Box-Jenkins approach has some advantages, one needs a lot of experience to be able to apply it efficiently (see Henery and Nakhaeizadeh (1993)).

2. Symbolic Machine Learning and Neural Networks

Most of the symbolic machine learning algorithms are more appropriate to perform the classification tasks. Regarding the fact that in a prediction the target variable is, generally, continuous-valued, most of the symbolic machine learning algorithms can not be applied to prediction, directly. Exceptions are the ID3-type algorithms CART and NEWID which can handle continuous-valued classes as well. Of course, it is possible to discretize every continuous-valued target variable and reduce a prediction task to a classification one, but this would be connected with information loss. The algorithms like NEWID and CART can handle the continuous-valued classes, directly, and without discretization. They generate a predictor in the form of a regression tree that can be transformed to production rules. Furthermore, these learning algorithms apply a single attribute at each level of the tree and this is in contrast to the most statistical and neural learning algorithms which consider all attributes to make a decision. The main structure of regression trees will be discussed below (See Breiman et al. (1984) for more detail).

Like the classical regression analysis, rgression trees try to detect the causal dependency between a target variable Y that should be predicted and some other features $X_i, i = 1, 2, ..n$ which can have an significant impact on the target variable. In contrast to the regression analysis, the number of possible prediction values for the target feature is, however, known and is equal to the number of the terminal nodes of the regression tree.

A regression tree consists of different subtrees. A typical subtree consists of a parent node N and two children nodes N_1 and N_2. Suppose that we have used the attribute A and the threshold α to construct this subtree. In building such subtrees the following questions arise:

1. How can the attribute A and the threshold α be selected?
2. Which values should be assigned to the children nodes N_1 and N_2?
3. Is it necessary to split further in the children nodes N_1 and N_2?

We begin with the answer of the second question. Suppose that we have selected the attribute A and the threshold α and according to their values we have assigned the whole cases available in the parent node N to the children nodes N_1 and N_2 and, for example, q cases $C_1, C_2, ..., C_q$ which have the target variable values $Y_1, Y_2, ..., Y_q$ are assigned to the node N_1. The prediction value assigned to the node N_1 is just the average value of all $Y_1, Y_2, ..., Y_q$, namely:

$$\bar{Y_q} = \frac{1}{q} \sum_{i=1}^{q} Y_i$$

which minimizes

$$F = \frac{1}{q} \sum_{i=1}^{q} (Y_i - \bar{Y_q})^2$$

Regarding the question one, we discuss, firstly, how an optimum threshold α can be selected. Suppose that the cases which are assigned to the parent node N have the values $A_1, ..., A_M$, concerning the attribute A. Regarding these values in an increasing order leads to $A_{(1)}, ..., A_{(M)}$, where $A_{(1)}$ is the smallest and $A_{(M)}$ the largest value. A threshold value α_i can be defined as :

$$\alpha_i = \frac{A_{(i)} + A_{(i+1)}}{2} \quad (i = 1 \ldots (M-1)).$$

Using α_i, one can divide the cases assigned to the parent node N into two subgroups which will be assigned to the nodes N_1 and N_2, respectively. The cases for which the attribute values $A_{(1)}, ..., A_{(M)}$ are less or equal to α_i will be assigned to the node N_1; other cases to the node N_2. In this way, it is possible to define $M-1$ threshold α, from them the optimum one should be selected. Regarding the definition of F mentioned above, one can calculate F_{N_1} and F_{N_2} using the corrosponding Y-values of the nodes N_1 and N_2. The optimum threshold α^* is the threshold that minimize:

$$L = F_{N_1} + F_{N_2}$$

This procedure will be repeated for all attributes. The attribute which minimizes L will then be selected as splitting attribute of the cases which are assigned to the parent node N.

To answer the third question, one can use different criteria among them the number of the cases assigned to

each node. For example if the number of the cases assigned to N_1 is less than a given threshold T, further splitting in this node should be stopped. Another criterion may be defined using the empirical variance of the target variable. Suppose that we have used totally R cases to construct the regression tree with target values $Y_1, Y_2, ..., Y_R$. We can now define:

$$\bar{Y} = \frac{1}{R}\sum_{i=1}^{R} Y_i$$

$$F_R = \frac{1}{R}\sum_{i=1}^{R}(Y_i - \bar{Y})^2$$

Furthermore, suppose that β is a given fraction of F. Now, one can use β as a criterion and stop splitting in the node N_1 if $F_{N_1} < \beta$.

The above method can be applied only to continuous-valued features. In regression trees, in contrast to ID3-algorithm, the splitting procedure for the qualitative-valued features is a binary one. Suppose that B is a qualitative-valued attribute, $B_1, ..., B_l$ are the values which attribute B can accepted and $\mathcal{P}$ denotes all possible subsets of $B_1, ..., B_l$. The splitting values consist in this case of different pairs of the elements of $\mathcal{P}$ (of course the empty set and the set $B_1, ..., B_l$ itself will not be regarded). The rest of the procedure is just the same as the case of the continuous-valued attributes.

In the recent years, one can also see in literature some efforts put to apply Neural Networks to the prediction of time series. Although the development of Neural Networks at early stage was stimulated by modelling of learning process in human brain, the further development of this technology shows a very strong similarity with statistical approaches. There are some studies which compare the Neural Networks with some statistical procedures like nonlinear regression from a theoretical point of view. However, it should be mentioned that the ability of adaptive learning which characterizes the most of Neural Networks is not implemented in statistical procedures like regression analysis and Box-Jenkins approach.

Application of CBR to predicting of time series

As it is mentioned already KNN can be considered as a subarea of CBR that traces back to early fifties (see for example Fix & Hodges (1951)). An excellent review of research in this field in the last fourty years can be find in Dasarathy (1991). It should be mentioned that KNN presents only the statistical aspects of CBR. Regarding the fact that in particular the prediction of time series is a knowledge intensive task, the other aspects of CBR like knowledge acquisition and knowledge representation can improve the quality of forecasting by using other information sources which are not involved in the applied datasets. Using such possibilities, CBR can complete the statistical aspects of forecasting of time series in an efficient way and, in our opinion, there is a potential demand of research in this area.

Back to application of KNN to time series forecasting; denoting the target feature by Y and the features which can have a significant impact on Y by vector X (of course the elements of X can also be the historical values of Y itself), KNN can be used as below. Suppose that we have observed n cases $(X_1, Y_1), ...(X_n, Y_n)$, where $X_i \quad (i = 1 \ldots n)$ represents the vector involving the feature values of the case i

The aim in KNN-approach is to apply these cases and X_{n+1} as well to forecasting of Y_{n+1}. To perform this task, KNN finds in the historical data K cases with most similarity to X_{n+1}. Suppose that these are the cases number $1, 3$ and 5 $(K = 3)$ with target values Y_1, Y_3 and Y_5, respectively. The prediction value for X_{n+1} is than Y^* that is just a combination of these target values (such combination, for example, could be the mean, other weighted averages or just the median of the target values).

The problems of application of Neural Networks exist in application of CBR and KNN as well. Especially, finding the optimal length of the searched pattern and determining the number of considered patterns (K) needs again using a separate test dataset which reduces the number of available training cases.

There is a controversial discussion if specially KNN and generally CBR can be regarded at all as learning systems. The reason for this controversy is that the learning task in the most inductive systems generates, in contrast to CBR, a general concept which can later be used for predicting the class of unseen cases. On the other hand, it is true that in CBR one uses the information given by the cases. This information is applied, however, to measure a pre-defined distance function but it is not applied to find a general prediction concept which is the main part of inductive learning. Formalization of the relation between CBR and inductive concept learning is discussed by Jantke (1992).

Regarding the above mentioned points, it is obvious that CBR can not be regarded as an inductive learning system like ID3, for example, or regression technique. But it should also be mentioned that the research in the field of machine learning is not limited just to inductive learning systems. Machine learning is a subarea of AI involving not only the inductive learning paradigm but a number of other learning approaches like deductive learning, explanation based learning, learning by analogy etc. In this connection CBR can be regarded, in our opinion, certainly as a subarea of machine learning

very close to learning by analogy.

Combined Approaches

Combining of different learning paradigms is known in machine learning as multistrategy learning. In ML-community, research in this area began in the end of eighties. It should be mentioned, however, that some works in the statistical community which use this approach traced back to the end of seventies (Dasarathy & Sheela (1979)). There are some efforts in the literature to combine CBR with other learning approaches (see for example Cardie (1993), Bamberg & Goos (1993)). Furthermore, the works in the Esprit-Project INRECA, funded by the European Community, is along this line of research.

Concerning the forecasting task, Quinlan (1993) combines the model based approaches with IBL. The main idea in his work can be described as follows. Suppose that X_1, X_2, ..., X_n are n attribute vectors correspond to the target values Y_1, Y_2, ..., Y_n and X_{n+1} is known as well and we want to predict the value of Y_{n+1}. Suppose also that X_i is one of the K-nearest-neighbours obtained for X_{n+1}. Before combining Y_i with the other target values to get Y^* as forecasting value for Y_{n+1}, Quinlan suggests a modification as :

$$\hat{Y_i} = Y_i - [M(X_i) - M(X_{n+1})],$$

where $\hat{Y_i}$ is the modified value of Y_i. After this modification, one uses instead of Y_i just the value $\hat{Y_i}$ to construct Y^*. The rest of the procedure remains unchanged. In the above relation M is an arbitrary model that can be used to predict the continuous-valued target Y. $M(X_i)$ and $M(X_{n+1})$ are the prediction values for cases i and $n+1$, respectively, using model M, where model M can be any learning approach able to handle continuous-valued classes.

Empirical Evaluation Results

There are some studies in literature which compare the performance of different statistical approaches using the time series data (Makridakis et al (1984)). But, there is no comprehensive study which includes the recent developed prediction approaches based on the AI-methodology like CBR, Neural Networks and Symbolic Machine Learning. An exception is the attempts put on this task within the Esprit-Project StatLog. In this Project three real time series datasets are applied to compare the performance of different learning algorithms. As it mentioned before, although a lot of learning algorithms can perform the classification task, they can not be applied to prediction, directly, because they can not handle the continuous-valued classes. It is, however, possible to consider the prediction task as classification by an appropriate discretization of the class values.

The first application used in the project StatLog deals with prediction of development of interest rates on successive trading days. The empirical results for this dataset are ambiguous. On one hand, some symbolic machine learning algorithms like CN2 deliver very precise predictions. On the other hand, the performance of the other machine learning algorithms like NEWID and C4.5 are very poor. CBR-type and Neural Networks algorithms are not evaluated for this dataset. The second and the third datasets are two versions of a real-world application which is in interest of the marketing department of Mercedes-Benz AG, Stuttgart. This application deals with the prediction of the number of registered cars and trucks in France. While the performance of Box-Jenkins method and NEWID are the best for this application, the prediction power of a CBR-type algorithm based on the KNN-concept is very poor. Other statistical and neural networks learning algorithms deliver an average performance (see Henery and Nakhaeizadeh (1993) for more detail).

Besides the results we have achieved within the project StatLog, some other empirical works has be done by the Machine Learning Group at the Ressort Research and Technology of Daimler-Benz AG in Ulm. Besides the prediction of the number of cars and trucks for the other countries, we have evaluated different learning algorithms by using another real-world application which deals with the prediction of daily exchange rates of US-Dollar against D-Mark. The results show that the performance of CBR, Neural Networks and Symbolic Machine Learning algorithms are almost the same. But they are still too far from the accuracy rates which one can get for example by using classical chart analysis.

Conclusions

In this paper we have presented a theoretical and empirical comparison between CBR and other approaches which contribute to predicting the future development of time series. We have regarded KNN as a special case of CBR that can be used besides the inductive methods (symbolic and neural) to attack the forecasting problems. Up to now, our empirical results do not indicate the superiority of KNN to other approaches in dealing with prediction tasks. Combining of CBR with other approaches, however, seems promising and can improve the quality of forecasting. We will follow this line of research, specially, in dealing with forecasting of financial markets. On the other hand, forecasting of time series is a knowledge intensive task. In this connection, regarding the other aspects of CBR like knowledge representation and knowledge acquisition can contribute to improving the forecasting results. In our opinion, this is a second direction for further research in this area.

References

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms, Machine Learning 6, 1, 37-66.

Althoff, K. D., Wess, S. Bartsch-Spörl, B. and Janetzko, D. (Hrsg.) (1992). Proceedings of the Workshop: Ähnlichkeit von Fällen beim fallbasierten Schliessen. University of Kaiserslautern. Fachbereich Informatik.

Bamberger, S. K. and Goos, K. (1993). Integration of Case-based Reasoning and Inductive Learning Methods. Paper presented in EWCBR-93, University of Kaiserslautern, Germany.

Breiman, L., Friedman, J. H., Olshen, A. and Stone, C. J. (1984). Classification and Regression Trees. Wadsworth, Belmont.

Cardie, C. (1993). Using Decision Trees to improve Case-Based Learning. In: Proceedings of the Tenth International Conference on Machine Learning, 25-32. Morgan Kaufmann Publishers.

Dasarathy, B. V. (Ed.) (1991). Nearest Neighbor(NN) Norms. NN Pattern Classification Techniques. IEEE Computer Society Press.

Dasarathy, B. V. and Sheela, B. V. (1979). A composite Classifier System Design: Concepts and Methodology. In: Proceedings of the IEEE, Volume, 67, Number 5, 708-713.

Fix, E. and Hodges, J.L. (1951). Discriminatory Analysis, Nonparametric estimation: Consistency Properties. Report no 4, UASF School of Aviation Medicine, Texas.

Henery, R. and Nakhaeizadeh, G. (1993). Forecasting of Time Series. Mimeo, University of Strathclyde, Glasgow.

Jantke, K. P. (1992). Formalizations in Case-Based Reasoning. In : Althoff, K. D. Wess, S. Bartsch-Spörl, B. and Janetzko, D. (Hrsg.) (1992). Proceedings of the Workshop: Ähnlichkeit von Fällen beim fallbasierten Schliessen. University of Kaiserslautern. Fachbereich Informatik, 9-14.

Makridakis, S; Andersen, A; Carbone, R; Fildes, R; Hibon, M; Lewandowski, R; Newton, J; Parzen, E; and Winkler, R. (1984). The Accuracy of Extrapolation (Time Series) Methods: Results of a forecasting competition. In : Makridakis, S. (Ed.). The Forecasting Accuracy of Major Time Series Methods. Wiley & Sons. 103-166.

Merkel, A. and Nakhaeizadeh, G. (1992). Application of Artificial Intelligence Methods to Prediction of Financial Time Series. In: Gritzmann, P. et al. (Hrsg.). Operations Research 91, 557-559.

Quinlan, J. R. (1993). Combining instance-based and model-based learning. In: Proceedings of the Tenth International Conference on Machine Learning, 236-243. Morgan Kaufmann Publishers.

Veloso, M. M.(1992). Learning by Analogical Reasoning in General Problem Solving. Dissertation, Carnegie Mellon University.