Expert Systems with Applications 31 (2006) 329–336

Expert Systems with Applications

www.elsevier.com/locate/eswa

A new hybrid data mining technique using a regression case based reasoning: Application to financial forecasting

Se-Hak Chun [a,*,1], Yoon-Joo Park [b,2]

[a] Department of Business Administration, Seoul National University of Technology, 172 Gongneung 2-dong, Nowon-gu, Seoul 139-743, South Korea
[b] Korea Advanced Institute of Science and Technology, 207-43 Cheongryangri-dong, Dongdaemoon-gu, Seoul 130-012, South Korea

Abstract

This paper proposes a regression case based reasoning (RCBR) which applies different weights to independent variables before finding similar cases. The traditional CBR model has focused on finding similar cases from a case base without considering the importance of independent variables. Thus, when extracting similar cases the traditional CBR has put same weights on each independent variable.

The proposed regression CBR (RCBR) finds a relative importance of independent variables from the relationship between independent variables and a dependent variable using a regression analysis and puts relative weights using regression coefficients on independent variables. Then, it selects nearest neighbor or similar cases using weighted independent variables through the traditional CBR machine and updates weights dynamically for the next target case and again performs the traditional CBR machine.

Keywords: Regression case based reasoning; Artificial intelligence; Data mining; Case based reasoning; Neural network; Statistical methods; Learning techniques

1. Introduction

This paper proposes a new hybrid learning technique, a regression CBR (RCBR) which applies different weights to independent variables before finding similar cases. The traditional CBR model has focused on finding similar cases from a case base without considering the importance of independent variables. Thus, when extracting similar cases the traditional CBR has put same weights on each independent variable.

The proposed regression CBR (RCBR) finds a relative importance of independent variables from the relationship between independent variables and a dependent variable using a regression analysis and applies relative weights using regression coefficients to independent variables. Then, it selects nearest neighbor or similar cases using weighted independent variables through the traditional CBR machine and updates weights dynamically for the next target case and again performs the traditional CBR machine. These concepts are investigated against the backdrop of a practical application involving the prediction of a stock market index.

The rest of this paper is organized into five sections. Section 2 reviews case based reasoning (CBR) as a knowledge discovery technique. Section 3 introduces a regression CBR. Section 4 presents the case study and reports the results. The case study intends to investigate the effect of a regression CBR on the predictive performance in forecasting a stock market index. Finally, the concluding remarks are presented in Section 6.

2. Case based reasoning

2.1. Prior knowledge through case reasoning

A learning system should make increasingly useful decisions as it accumulates experience. This is the express goal of the work in case-based reasoning (CBR).

Perhaps, the most important advantage of CBR is the affinity to human learning. People take account of observations and utilize them for future decision making. Often the extrapolation to new situations is ad hoc, as in modifying a set of evaluation criteria for the silicon-based computing industry into a similar one for the emerging vendors of photonic hardware. In other cases, the extrapolation is more formal and takes the form of

* This work was supported by Korea Research Foundation Grants (KRF-2004-003-B00073).

E-mail addresses: shchun@snut.ac.kr (S.-H. Chun), mirage20@kgsm.kaist.ac.kr (Y.-J. Park).
[1] Corresponding author. Tel.: +82 2 970 6487; fax: +82 2 973 1349.
[2] Tel.: +82 2 958 3697; fax: +82 2 958 3604.

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.09.053

inductive propositions such as formulas, principles, laws, or rules of thumb.

Related to the affinity of CBR to human learning is the ease of enhancing system performance. More generally, the knowledge in a particular domain can be stored in formats which are conventional for that domain. For instance, a knowledge base for balancing stocks, bonds, and other instruments in an investment portfolio can store the information about previous financial strategies in the cognitive format used by human analysts.

This is in contrast to other knowledge-level representations such as production rules, in which the system developer is required to extricate the pertinent decision rules used by a human. In general, the problem of knowledge extraction is further compounded by the fact that people often perform admirably in various domains without using-or even being aware of the existence of-any such decision rules.

The CBR methodology can be effective even if the knowledge base is imperfect. Certain techniques of automated learning, such as explanation-based learning, work well only if a strong domain theory exists. In contrast, case reasoning can use many examples to overcome the gaps in a weak domain theory while still taking advantage of the domain theory (Porter, Bareiss, & Holte, 1990). CBR can also be used when the descriptions of the cases, as well as the domain theory, are incomplete. A further advantage of CBR is the relative ease of combining techniques with other approaches such as production rules (Golding & Rosenbloom, 1991). An example of such compatibility is a system which uses case reasoning to solve problems whenever possible; otherwise it resorts to heuristics to decompose a problem into a simple one.

2.2. Retrieving precedents

Case reasoning requires the retrieval of past experience in the form of cases. In this task, two types of difficulties can arise. The *matching* problem refers to the task of associating a new problem to pertinent prior cases. A key issue lies in retrieving prior cases which are similar to the new problem in substantive rather than superficial ways. This relates in part to the issue of *indexing*, which deals with the organization of the case base.

2.3. Matching problem

Problem solving in any arena is dedicated to the attainment of a goal. To this end, the decision maker must find prior cases which resolve the specified or comparable objectives, rather than those that match only surface features having little impact on the effectiveness of the solution. For instance, two portfolios may be of similar dollar amount and contain a number of shares in common; but one is directed toward high income while the other seeks stable growth. Consequently, a CBR system must search through the base of previous cases by first attempting to find solutions that meet the primary design goals, and subsequently examine them against secondary objectives.

The matching problem can be addressed in a number of ways. The default scheme is to perform an exhaustive search through the case base each time a new problem arises. However, system performance can be degraded by such a tedious approach.

A more systematic way is for a human to identify the relevant prior cases. Unfortunately, this technique requires continuing human intervention if a system is to improve its performance over time.

To automate the task of matching in CBR, previous cases can be organized in some fashion to enable the rapid identification of potentially relevant cases. To this end, previous problems and solutions can be indexed by their key attributes and the features which distinguish them from other cases.

2.4. Indexing problem

The indexing problem refers to the task of storing cases for effective and efficient retrieval. In terms of efficacy, the subissues are *accuracy*—finding only relevant cases—and *completeness*—identifying all relevant cases.

In general the prior cases retrieved by case reasoning will match the required solution only imperfectly. In particular, the source cases may fail to fulfill some of the requisite objectives. At this point, an analogy can be formed between the functionality of the precedent solutions and the goals of the current problem. The prior solutions may then be modified to eliminate or circumvent the limitations. Then a process of iterative refinement can be employed to adapt an old solution to the new problem context. Whether or not analogy is used, an organization may be imposed on the case base through the use of clustering techniques (Kim & Novick, 1993). In this way, a target case may be readily accommodated into an existing case base.

2.5. Case based reasoning and composite neighbors

Conventional methods of prediction based on discrete logic usually seek the single best instance, or a weighted combination of a small number of neighbors in the observational space. For instance, the rule of thumb in case based reasoning (CBR) is to seek the nearest neighbor to a target case. In an analogous way, certain algorithms in neural networks seek a fixed number of the closest neighbors; this approach is illustrated by the use of self-organizing maps for pattern recognition tasks (Kohonen, 1984).

An intelligent learning algorithm should therefore 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, 1995; Kim & Chun, 1998; Chun & Kim, 2003, 2004a,b; Chun, Kim, & Kim, 2002; Chun & Park, 2005). The procedure for case reasoning using composite neighbors and Euclidean distance method is presented in