PERGAMON

Expert Systems with Applications 19 (2000) 117–123

Expert Systems
with Applications

www.elsevier.com/locate/eswa

The predictive accuracy of artificial neural networks and multiple regression in the case of skewed data: exploration of some issues

P.N. SubbaNarasimha[a,*], B. Arinze[b,1], M. Anandarajan[b,2]

[a]*Department of Management, G.R. Herberger College of Business, St. Cloud State University, St. Cloud, MN 56301, USA*
[b]*Department of Management, College of Business, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA*

Abstract

Business organizations can be viewed as information-processing units making decisions under varying conditions of uncertainty, complexity, and fuzziness in the causal links between performance and various organizational and environmental factors. The development and use of appropriate decision-making tools has, therefore, been an important activity of management researchers and practitioners. Artificial neural networks (ANNs) are turning out to be an important addition to an organization's decision-making tool kit. A host of studies has compared the efficacy of ANNs to that of multivariate statistical methods. Our paper contributes to this stream of research by comparing the relative performance of ANN and multiple regression when the data contain skewed variables. We report results for two separate data sets; one related to individual performance and the second to firm performance. The results are used to highlight some salient issues related to the use of ANN and multiple regression models in organizational decision-making. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Neural networks; Regression; Skewed data

1. Introduction

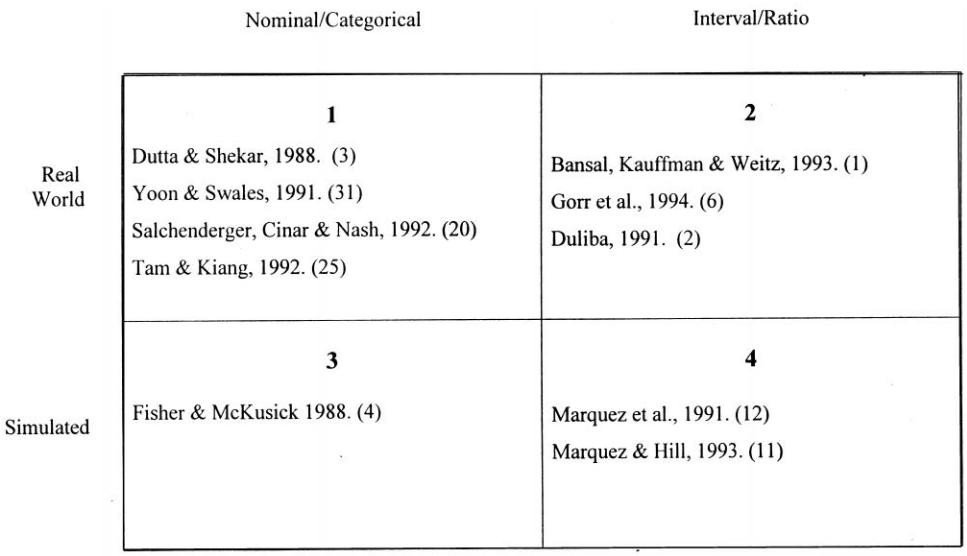
Increasing attention is being paid to the use of artificial neural networks (ANN) in managerial decision-making. As many managerial decision situations are fraught with variety, ambiguity and complexity (Mintzberg, Raishinghani & Theoret, 1976), ANNs are appealing as managerial decision-making aids precisely because of their expected effectiveness in such situations (Lippman, 1987). An important element in the life-cycle of an innovation (and ANN is an innovation in the field of organizational decision-making aids) is the establishment of the contingencies under and contexts in which the innovation is most effective. Hence, a stream of studies on ANNs has focused on delineating the boundaries of their usefulness (e.g. Duliba, 1991; Dutta & Shaker, 1988; Gorr, Nagin & Szczypula, 1994; Marquez & Hill, 1993; Marquez, Hill, Worthley & Remus, 1991; Sharda & Wilson, 1993). The studies reported in this paper add to this line of research. Using 'real-world' data we conduct an empirical investigation into the relative efficacy of ANNs and multiple regression when the sample data are not normally distributed, i.e. they are skewed. The effect of this contingency on the behavior of ANN and multiple regression models needs to be investigated given that reliability of multivariate statistical methods requires that data be multivariate normal.

From an organizational effectiveness view, there is a need for studies that establish the pros and cons of any innovation. Corporations tend to have a pro-innovation bias-leading them to adopt 'promising' but untested innovations (Kimberly, 1985). For example, use of labels such as "Expert" and "Intelligent" has been shown to lead to complacency as well as unthinking dependence on such systems among users (Will, 1991). Further, innovation–adoption can be disruptive because of their organization-wide consequences (Sviola, 1990). Thus, there is clearly a need for systematic investigations of the contexts and contingencies affecting the predictive accuracy of ANN models.

A number of studies have focused on investigating the relative performance of statistical and ANN methods in forecasting. These studies can be differentiated from each other on two dimensions. The first is in data types. Some studies use actual data (i.e. data from real-world) and others have used simulated data. The second dimension relates to differences in measures. Studies have used measures that are

[*] Corresponding author. Tel.: +1-320-255-3823; fax: +1-320-654-5184.
E-mail addresses: psubba@stcloudstate.edu (P.N. SubbaNarasimha), arinzeob@drexel.edu (B. Arinze), murugan.anandarajan@drexel.edu (M. Anandarajan).
[1] Tel.: +1-215-895-1798.
[2] Tel.: +1-215-895-6212.

0957-4174/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0957-4174(00)00026-9



	Nominal/Categorical	Interval/Ratio
Real World	1 Dutta & Shekar, 1988. (3) Yoon & Swales, 1991. (31) Salchenderger, Cinar & Nash, 1992. (20) Tam & Kiang, 1992. (25)	2 Bansal, Kauffman & Weitz, 1993. (1) Gorr et al., 1994. (6) Duliba, 1991. (2)
Simulated	3 Fisher & McKusick 1988. (4)	4 Marquez et al., 1991. (12) Marquez & Hill, 1993. (11)

Fig. 1. Categorization of ANN efficacy studies.

either nominal/categorical or interval/ratio-scale. Fig. 1 provides some examples of studies in each of the cells.

There appears to be a preponderance of studies in cells one and four. If real-world data are used, usually the phenomenon is represented by a categorical variable (e.g. solvent/insolvent, bankrupt/not bankrupt, loan granted/denied, etc.). It is only when simulated data are used in the analyses that we observe the measures to be continuous/interval scale (cell 4). The two studies reported in this paper fall in cell two. Both the studies reported here use real-world data, and have variables that are measured in ratio-scale. While both studies are of organizational phenomena, they differ in their level of analysis. The first focuses on individual performance and the second on firm level performance.

In addition to differences in data type and variable measurement, our study differs from earlier ones by considering the additional dimension of variable skewness. Specifically, we investigate whether skewness in a sample's dependent variable affects the efficacy of ANN and multiple regression models. Analyses by Marquez et al. (1991) using simulated data, indicate that they do. The reliability of any statistically derived result is known to be strongly dependent on the degree to which the sample distribution is multivariate normal. Formulae for tests of statistical significance of regression coefficients are based on this assumption (Tabachnick & Fidell, 1983). To the extent that this assumption is violated, generalization of statistics-based results data beyond the sample will be highly suspect. Hence, data sets with skewed variables are particularly well suited for testing the relative efficacy of ANN and regression models.

A large proportion of studies support the use of ANN-based reasoning to deal with unstructured or semi-structured decision situations (e.g. Dutta & Shekar, 1988; Gallant, 1988; Yoon & Swales, 1991). These (and other) studies comparing ANN performance to that of multivariate statistical methods, have found ANNs to be better at prediction. However, Marquez et al. (1991) found in their simulation study that ANN-based systems perform better than regression techniques only when sample sizes are small and when variables are strongly correlated. Duliba (1991) found that an ANN model did not perform as well as regression when additional explanatory variables were introduced into the modeling. Gorr et al. (1994) noted that, for their data, although multiple regression was best overall there were no statistically significant differences in predictive accuracy across four different models. Further, they observe that

> Neither the stepwise regression nor the ANN benefited when additional model structures were incorporated (p. 31).

As skewness is a factor affecting the various models' behavior, we compare the performance of both multiple regression and ANN by deliberately choosing samples characterized by highly-skewed variables.

The paper is structured as follows: the two studies are reported next. Data set for the first study consists of a sample of MBA students where the focus is on predicting the students' graduating GPA. We conclude the paper by discussing the comparative performance of both ANN and regression models on the two data sets, suggesting guidelines for the use of ANNs for knowledge acquisition, and proposing future research directions.

2. Methodology

The same basic procedure was followed in each study: (1) use sampling without replacement of the full sample to