Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers & Operations Research 32 (2005) 2583–2594
computers & operations research
www.elsevier.com/locate/dsw

Classification with incomplete survey data: a Hopfield neural network approach

Shouhong Wang*

Department of Marketing/Business Information Systems, Charlton College of Business, University of Massachusetts Dartmouth, North Dartmouth, MA 02747-2300, USA

Available online 20 May 2004

Abstract

Survey data are often incomplete. Classification with incomplete survey data is a new subject. This study proposes a Hopfield neural network based model of classification for incomplete survey data. Using this model, an incomplete pattern is translated into fuzzy patterns. These fuzzy patterns, along with patterns without missing values, are then used as the exemplar set for teaching the Hopfield neural network. The classifier also retains information of fuzzy class membership for each exemplar pattern. When presenting a test sample, the neural network would find an exemplar that best matches the test pattern and give the classification result. Compared with other classification techniques, the proposed method can utilize more information provided by the data with missing values, and reveal the risk of the classification result on the individual observation basis.

Keywords: Incomplete data; Survey data; Classification; Hopfield neural network; Fuzzy sets; Uncertainty

1. Introduction

Classification (or discriminant analysis, pattern recognition) is important for management decision making [1]. Among many statistical and non-traditional techniques of classification, neural networks have been widely used for classification [2–4]. One of the advantages of the neural network method is that it requires few restrict assumptions about the data, compared with other methods.

As classification has been a major approach in data mining [5,6], neural networks play even more significant role in classification. In data mining, the large number of dimensionality and the huge

* Tel.: +1-508-999-8579; fax: +1-508-999-8646.
E-mail address: swang@umassd.edu (S. Wang).

0305-0548/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cor.2004.03.018

volume of data make neural networks competitive in classification due to their imperviousness of "the curse of dimensionality" and low computational cost.

Standard neural networks, however, are not able to process input data with missing values. On the other hand, in data mining, it is a rare case where the data set contains entries for all of the variables for each pattern. Although there have been a large amount of work on fuzzy neural networks (e.g., [7,8]), the fuzzy neural network models reported in the literature do not address the problem of incomplete data. This study is to add an important advantage to the neural network classification techniques for data mining on survey data (e.g., marketing survey and questionnaire) by proposing a Hopfield neural network based model that can effectively deal with missing data.

In this paper, the type of survey data is assumed to be discrete. This is because business online survey data are discrete, with few exemptions. Also any continuous data can be approximately represented by discrete data for the classification purpose. Without loss of generality, two-group classification problems are investigated. The model described in this paper can be generalized to more than two groups classification problems, considering that c group classification problems can be solved approximately using $(c-1)$ or $c(c-1)/2$ discriminant functions in general [9,10].

Next, Section 2 describes a myth that commonly exists in the classification literature, and explains the motivation of this study. Section 3 presents a brief overview of the Hopfield neural networks. Section 4 discusses the issue of incomplete data and provides a fuzzy sets model for translating discrete patterns with missing values into fuzzy patterns. Section 5 proposes the model that performs classification based on all patterns including those fuzzy patterns. Section 6 shows experimental results on a real-data case using the proposed classification model for incomplete data. Finally, Section 7 gives conclusions.

2. A myth in the classification area

In the classification literature, there has been a myth that high correct-classification rates (or low misclassification rates) for real-world data sets are the measurement for assessing classifiers. This criterion overlooks two closely related facts, as discussed below.

First, for a given sample population, there is an optimal correct-classification rate. When the two data set classes overlap, the optimal rate must be less than 100%. If the distribution of the population is known, few non-statistical methods can compete with statistical classification methods that can achieve the optimal results. For a real-world data set, the distribution of the population is unidentified, and then the optimal rate is unknown. An experimental classification result on a real data set might be better than the actual (unknown) optimal rate when the test samples do not represent the data population. Such a flaw is difficult to detect because the sample size is always limited. Thus, in principle, any claim of a good classifier based on a high correct-classification rate for a real-world data set is questionable.

Second, theoretical optimal correct-classification rate does not make all sense in term of decision making on a particular observation. This problem can be exposed clearly through a general example. Suppose the optimal classification boundary separates the pattern space into two regions as shown in Fig. 1. The optimal classification reflects the statistical behaviour of the population, but reveals nothing about the behaviour of the individual sample. On the other hand, the object of a decision (or a judgement) is usually an individual observation (e.g., credit approval for a specific client). In