Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 20 (2013) 433 – 438

Procedia Computer Science

Complex Adaptive Systems, Publication 3
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2013- Baltimore, MD

Knowledge Extraction from Survey Data Using Neural Networks

Imran Khan, Arun Kulkarni

The University of Texas at Tyler, Tyler, TX 75799, USA

Abstract

Surveys are an important tool for researchers. It is increasingly important to develop powerful means for analyzing such data and to extract knowledge that could help in decision-making. Survey attributes are typically discrete data measured on a Likert scale. The process of classification becomes complex if the number of survey attributes is large. Another major issue in Likert-Scale data is the uniqueness of tuples. A large number of unique tuples may result in a large number of patterns. The main focus of this paper is to propose an efficient knowledge extraction method that can extract knowledge in terms of rules. The proposed method consists of two phases. In the first phase, the network is trained and pruned. In the second phase, the decision tree is applied to extract rules from the trained network. Extracted rules are optimized to obtain a comprehensive and concise set of rules. In order to verify the effectiveness of the proposed method, it is applied to two sets of Likert scale survey data, and results show that the proposed method produces rule sets that are comparable with other knowledge extraction techniques in terms of the number of rules and accuracy.

Selection and peer-review under responsibility of Missouri University of Science and Technology

Keywords: Neural networks; Decision trees; Rule extraction; Knowledge discovery; Classification

1. Introduction

A survey is conducted to collect data from individuals to find out their behaviors, needs and opinions towards a specific area of interest. Survey responses are then transformed into usable information in order to improve or enhance that area. Survey data attributes can come in the forms of binary-valued (or binary-encoded), continuous data or discrete data measured on a Likert scale. All three forms of data attributes are used according to the survey requirements. Discrete data can be used as a measure on a Likert scale to provide some distinct advantages over the other two types of data attributes. It helps respondents choose an answer. For instance, some respondents may be too impatient to make fine judgments and to give their responses on a continuous scale. The options provided in a typical five-level Likert item are Strongly Disagree, Disagree, neither Agree nor Disagree, Agree and Strongly Agree. The collected data might be contaminated if the difficult or time consuming judgmental task is beyond the respondent's ability or tolerance. The use of a Likert scale has been proposed to alleviate these difficulties.

Classification and knowledge extraction from survey data is a very important step in the decision-making process. Based on this knowledge, decisions are taken to improve the area for which the survey was conducted. Classification of Likert-scale survey data depends on the number of attributes. Classification process may become more complex when the number of Likert scale options and attributes in the survey is large. In the case of a survey, these attributes or features are the questions. Another major issue in Likert-Scale data is the uniqueness of the tuples. Classification algorithms group data based on the patterns of the attributes. A large number of unique tuples may result in a large number of patterns. Due to a large number of patterns, the knowledge extraction process from these classifiers becomes complex, and often the outcome of knowledge extraction process may not be satisfactory. The main focus of this research is to classify Likert-scale survey data using a multi-layered feed forward (MLF) [1,2,3,4] neural network and to apply Artificial Neural Network Tree (ANNT) algorithm [5,6] to extract knowledge from trained neural network.

The method proposed in this research consists of two steps. The first step is to train and prune the MLP neural network using back propagation algorithm. The second step is to apply an ANNT algorithm to extract knowledge from the neural network in the form of rules and optimize them to obtain a comprehensive and concise set of rules.

Selection and peer-review under responsibility of Missouri University of Science and Technology
doi:10.1016/j.procs.2013.09.299

The proposed method was applied to two Likert scale surveys. The first survey was about the reading strategies of students. The name of the survey was "Metacognitive Awareness of Reading Strategies Inventory (MARSI)" [7]. The second data set is a teacher evaluation survey. The teacher evaluation survey is used to evaluate a teacher's performance and helped in decision making.

2. Method

Method to extract the knowledge from Likert scale survey data consists of two steps. The first step is to train and prune the neural network using a multi-layered back propagation algorithm. The second step is to apply an ANNT algorithm to extract rules from trained network. Responses of a Likert-scale survey are usually in a non-numeric form. For neural network training, responses were converted to the range of 1 to -1. The mapping shown in Table 1 was used.

Table 1. Normalization of Responses

Option	Option Value	Normalized Value
Option 1	1	-0.9
Option 2	2	-0.4
Option 3	3	-0.1
Option 4	4	0.4
Option 5	5	0.9

2.1 Neural network training and pruning

A MLF neural network consists of three layers (Figure 1). The first layer has *k* input neurons which send data via connection links to the second layer of *M* hidden neurons, and then via more connection links to the third layer of output neurons. The number of neurons in the input layer is usually based on the number of features in a data set. The second layer is also called the hidden layer. More complex systems will have multiple hidden layers of neurons. Given an input pattern x_i, $i \in \{1,2,\ldots k\}$, where k is the number of attributes in the data set, the activation value of each neuron *o* can be described by the following equation:

$$o_i = f\left(\sum_j (w_{ij} \,.\, x_j)\right) \tag{1}$$

where *f*(.) is the activation function. In this research, sigmoid function is used.

$$f(net) = \frac{1}{1 + exp^{-net}} \tag{2}$$

In order to calculate the change of weights, output vector **o** is compared with the target vector **d**, and the error between the two vectors is then propagated backward to obtain the change in weights Δq_{ij} that is used to update the weights. Δq_{ij} for weights between layers L_2L_3 is given by:

$$\Delta q_{ij} = \alpha \delta_i o_j \tag{3}$$

where α is a training rate coefficient (typically 0.01 to 1.0). o_j is the output of neuron *j* in layer L_3, and δ_i is given by

$$\delta_i = (d_i - o_i) o_i (1 - o_i) \tag{4}$$

where o_i represents the actual output, where as d_i represents the target output.