Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 70 (2006) 384–397
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Extracting rules from multilayer perceptrons in classification problems: A clustering-based approach

Eduardo R. Hruschka[a,*], Nelson F.F. Ebecken[b]

[a]Graduate Program in Computer Science, Catholic University of Santos (UniSantos), R. Carvalho de Mendonça, 144, CEP 11.070-906, Santos, SP, Brazil
[b]COPPE/Federal University of Rio de Janeiro, CEP 21.945-970, CP 68.506, Rio de Janeiro, Brazil

Received 8 April 2005; received in revised form 1 August 2005; accepted 31 December 2005
Communicated by A. Abraham
Available online 24 May 2006

Abstract

Multilayer perceptrons adjust their internal parameters performing vector mappings from the input to the output space. Although they may achieve high classification accuracy, the knowledge acquired by such neural networks is usually incomprehensible for humans. This fact is a major obstacle in data mining applications, in which ultimately understandable patterns (like classification rules) are very important. Therefore, many algorithms for rule extraction from neural networks have been developed. This work presents a method to extract rules from multilayer perceptrons trained in classification problems. The rule extraction algorithm basically consists of two steps. First, a clustering genetic algorithm is applied to find clusters of hidden unit activation values. Then, classification rules describing these clusters, in relation to the inputs, are generated. The proposed approach is experimentally evaluated in four datasets that are benchmarks for data mining applications and in a real-world meteorological dataset, leading to interesting results.

Keywords: Rule extraction from neural networks; Clustering; Genetic algorithms

1. Introduction

Neural networks have been successfully applied to solve data mining problems in several domains. In this sense, multilayer perceptrons (MPs) may achieve high classification accuracy, but the knowledge acquired by such neural networks is usually incomprehensible for humans [13]. This fact can be a major obstacle in data mining applications, in which human-interpretable patterns describing the data, like symbolic rules or other forms of knowledge structure, are important [37]. Therefore, many methods have been developed to alleviate the lack of explanation of neural network (NN) models.

Neural networks (NNs) learn by adjusting their connection weights, which somehow reflect the statistical properties of the data [17]. Thus, the knowledge acquired by a NN is codified on its connection weights, which in turn are associated to both its architecture and activation functions [2]. In this context, the process of knowledge acquisition from NNs usually implies the use of algorithms based on the values of either connection weights or hidden unit activations. The algorithms designed to perform such task are generally called *algorithms for rule extraction from neural networks*. The task of rule extraction from NNs is a computationally hard problem [23], and heuristics have been developed to overcome its combinatorial complexity [69]. In our work, a clustering genetic algorithm (CGA) is employed for rule extraction from MPs. The proposed method is based on the hidden unit activation values and consists of two main steps. First, the CGA is employed to find clusters of hidden unit activation values. Then, these clusters are translated into logical rules.

Andrews et al. [2] suggested a classification scheme for rule extraction algorithms. The proposed scheme is based on four aspects: (i) form and quality of the extracted rules; (ii) necessity of specific neural network training algorithms; (iii) complexity of the rule extraction algorithm; (iv) translucency of the neural network. According to this

*Corresponding author.
E-mail addresses: erh@unisantos.br (E.R. Hruschka), nelson@ntt.ufrj.br (N.F.F. Ebecken).

0925-2312/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2005.12.127

scheme, our method provides *If…Then* propositional rules and it does not require any specific MP training algorithm. In addition, it can be applied in classification problems involving discrete and continuous attributes. The rule extraction algorithm complexity is based on the employed CGA. As far as the *translucency* of the NN is concerned, there are three approaches: *decompositional*, *pedagogical and eclectic*. *Decompositional* approaches involve rule extraction at the level of hidden and output units, which are mapped in a binary form. *Pedagogical* approaches try to map inputs directly into outputs, using machine-learning techniques. In our work, hidden unit activation expressions are employed to get classification rules by means of a CGA. Thus, our approach can be classified as *eclectic*, because it is based on both decompositional and pedagogical approaches.

The remainder of the paper is organized as follows. Section 2 situates the proposed method in the context of related work. Section 3 describes the CGA, which is applied to extract rules from MPs trained in classification problems. In Section 4, we present empirical results in four datasets that are benchmarks for data mining (Iris Plants, Wisconsin Breast Cancer, Australian Credit Approval and Pima Indians Diabetes) as well as in a real-world meteorological dataset. Finally, Section 5 concludes our work.

2. Related work

Several methods for rule extraction from NNs have been proposed in the literature, showing the increasing importance of this issue in several domains. Under this perspective, this section provides a brief description of several rule extraction methods. To do so, we follow a chronological order, considering the original work of each author. Then, we present our proposed method, comparing it with similar ones described in the literature.

In 1988, Gallant [19] proposed the first approach to understand the knowledge encoded by NNs. In [19], each NN node represents a conceptual entity, e.g. the input, hidden and output units represent symptoms, diseases, and treatments, respectively. This approach provides NN understanding, but requires available domain knowledge [46]. Also, since redundant rules can appear in the rule extraction process, this method presents the inconvenient necessity of establishing subjective criteria for choosing the appropriate rules [4]. In a later work, Gallant [20] developed the concepts relating NN learning with expert systems.

In 1992, McMillan et al. [43] described a neural network called *RuleNet*, which learns symbolic rules in array manipulation domains. The RuleNet was employed to natural language processing.

In 1993, Towell and Shavlik [68] showed how to use NNs for rule refinement. The developed algorithm was called SUBSET, which is based on the analysis of the weights that make a specific neuron active (considering binary activations). All methods published until 1992 are based on this concept, i.e. in a *decompositional* approach. Also in [68], the authors described the MofN algorithm, which provides more precise and understandable rules than the SUBSET. Thrun [67] described the *validity interval analysis* (VIA) approach, which is based on the propagation of activation values through the network, from the inputs to the outputs and vice versa. Activation ranges are used to prove rule correctness, and the procedure does not require any specific training method for back-propagation NNs. Fu [18] presented a hybrid system that combines symbolic and connectionist inferences. The initial NN architecture is dependent of available domain knowledge and a dataset is employed to refine the rules codified on the NN model. In this sense, the KT Algorithm [16] extracts the refined rules from the trained NN. Kane and Milgram [35] developed a rule extraction method based on numeric and logic-numeric units.

In 1994, Craven and Shavlik [7] explored the fact that trained NNs can be effectively questioned, answering questions about situations in which specific instances are described by a particular class. The rule extraction process is visualized as a learning task, whose main goal is to learn the function computed by the trained NN, providing conjunctive and MofN [68] rules. Sethi and Yoo [55] proposed a method for converting the connection weights in symbolic representations, and showed how to apply it to understand and approximate NNs.

In 1995, Alexander and Mozer [1] developed a rule extraction method, based on connection weights, that supposes activation functions showing approximately Boolean behavior. Andrews et al. [2] presented an important survey of the main approaches developed for rule extraction from NNs until 1995.

In 1996, Narazaki et al. [47] proposed a method to reorganize the knowledge distributed in NNs and to extract approximate fuzzy classification rules. The developed methodology is focused on the analysis of the function learned by a NN. Omlin and Giles [48] proposed a method to insert knowledge in recurrent NNs, showing that these are capable of reviewing the inserted rules. Sethi and Yoo [54] developed a rule extraction method based on the connection weights. Craven and Shavlik [6,61] proposed the TREPAN Algorithm, which does not require neither specific architectures nor particular NN learning algorithms, to extract symbolic representations from NNs. The authors use a classification tree to perform this task. Herrmann [26] described an algorithm to extract rules from logic NNs. Taha and Gosh developed three rule extraction algorithms [64,65] and also a methodology to evaluate the extracted rules [66]. Lu et al. [41] proposed an approach for rule extraction from NNs based on the clustering of hidden unit activation values.

In 1997, Setiono [56] described a method based both on the work of Lu et al. [41] and on an algorithm that eliminates NN redundant connections. Duch et al. [10] proposed two rule extraction algorithms that depend on a