Available online at www.sciencedirect.com
SCIENCE DIRECT®
Neurocomputing 64 (2005) 73–92
NEUROCOMPUTING
www.elsevier.com/locate/neucom

A preliminary empirical comparison of recursive neural networks and tree kernel methods on regression tasks for tree structured domains

Alessio Micheli[a], Filippo Portera[b], Alessandro Sperduti[b,*]

[a]*Dipartimento di Informatica, Università di Pisa, Pisa, Italia*
[b]*Dipartimento di Matematica Pura ed Applicata, Università degli Studi di Padova, Padova, Italia*

Available online 20 January 2005

Abstract

The aim of this paper is to start a comparison between recursive neural networks (RecNN) and kernel methods for structured data, specifically support vector regression (SVR) machine using a tree kernel, in the context of regression tasks for trees. Both the approaches can deal directly with a structured input representation and differ in the construction of the feature space from structured data. We present and discuss preliminary empirical results for specific regression tasks involving well-known quantitative structure-activity and quantitative structure-property relationship (QSAR/QSPR) problems, where both the approaches are able to achieve state-of-the-art results.

Keywords: Kernel methods; Kernels for structures; Recursive neural networks; Learning in structured domains

1. Introduction

In recent years, several researchers have started to consider the adaptive processing of structured data. This interest is motivated by two main reasons:

*Corresponding author.
E-mail addresses: micheli@di.unipi.it (A. Micheli), portera@math.unipd.it (F. Portera), sperduti@math.unipd.it (A. Sperduti).

0925-2312/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2004.11.013

(i) several very important computational problems in bioinformatics, chemistry, document classification and filtering (just to name a few) require the use of some machine learning procedure to be properly treated because their complexity does not allow a formal and precise definition of the problem and thus no algorithmic solution to the problem is known; (ii) in many of the above problems, the objects of interest are more naturally represented via structured representations of different sizes, such as sequences, strings, trees, directed or undirected graphs, which retain all the structural information relevant for solving the task. Within this area, there are two main streams of research which are relevant for the neural network community (for an overview see [15]): (a) Recurrent and recursive neural networks (RecNN) (e.g., see [11] for the basic theory and [1,2] for instances of recent developments on specific structures and applications in Bioinformatics); (b) kernel methods for structured data (e.g., see [12]).

Alternative approaches to structured domain learning have been proposed within the field of symbolic approaches to machine learning, such as ILP [18,25,10], and within the field of probabilistic approaches (e.g., [9]). Recent related work has also involved the combination of generative models and kernel methods for sequences or structures (see [17,26]).

The main aim of this work is to start a comparison between the two approaches described above for learning in domains constituted by trees, using the same basic assumptions for data representation. Specifically, here we discuss the differences between the two approaches and we report the results obtained for an (preliminary) empirical comparison of them on two representative regression tasks in the field of computational chemistry, namely, a quantitative structure-activity relationship (QSAR) problem and a quantitative structure-property relationship (QSPR) problem. The neural networks used for the comparison are recurrent cascade-correlation networks [28,29], while as kernel methods, we have used a support vector regression (SVR), with two different kernels: (i) a kernel based on string matching, where a string represents a tree; (ii) the tree kernel proposed in [7].

It should be stressed that, in the considered regression problems, recurrent cascade-correlation networks have already compared favorably with respect to state-of-the-art standard approaches used in the QSPR/QSAR field [4,20], and for this reason we do not repeat here the comparison with traditional approaches.

This paper is an expansion of the work presented in [24].

2. Regression of k-ary trees by Recursive NN and SVR with tree kernels

In this paper we focus on k-ary trees (in the following referred to as trees), which are rooted positional trees with finite out-degree k. In addition, we require that each node of a tree is associated to an element of a set L representing numerical or categorical variables, denoted as the label set. Examples of label set are given by a set of symbols, e.g. the alphabet used to generate a set of strings, or a set of real valued vectors which may represent, for example, the results of some measurement relating to the objects represented by each node of the tree. Let vert($\boldsymbol{t}$) be the set of vertexes