ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 1-8

An introduction to learning in web domains

M. Diligenti*, M. Gori*, M. Maggini*, F. Scarselli*, and A.C. Tsoi◇

* Dipartimento di Ingegneria dell'Informazione, Università di Siena, Siena (Italy)
◇ Office of Pro Vice-Chancellor, University of Wollongong, Wollongong (Australia)

Abstract. Artificial neural networks have been the subject of massive investigation in the last twenty years. Theoretical studies on architectural and learning issues and experimental evidence are now clearly indicating their potential capabilities and their limitations. In a different, apparently unrelated field, the problem of ranking Web pages for information retrieval has been studied giving rise to solutions based on a dynamical systems, which very much reminds typical neural network dynamics.
In this paper, we introduce the notion of learning in web domains, which represent an abstraction of the Web, giving insights on the way neural networks and Web page scoring systems can be bridged. Architectural and learning issues are discussed beginning from the theory of adaptive computation on structured domains.

1 Introduction

A strong limitation of most connectionist-based models is they are not well-suited for capturing topological features, which often play a crucial role in decision-making. This limitation is due to the flat data representation currently adopted also in other machine learning approaches, where links amongst samples of the training set are not typically taken into account. For instance, in some fields like pattern recognition, the application of widely-disseminated multilayer networks requires the dissipation of learning capabilities to incorporate translation and rotation invariance. As a consequence, Backprop cannot focus only on the actual detection of the distinguishing features of the patterns, thus limiting its capabilities. This limitation has been recognized by many people and, recently, the development of more general learning models capable of dealing with structured domains has been the subject of detailed investigation [8]. Basically, instead of processing flat representation, new models have been conceived which operate on graphs, thus extending the notion of learning environment to a collection of graphs.

Interestingly, the computation of the rank of Web pages proposed in [4], follows a computational scheme which very much reminds us the one adopted in neural networks for graphical domains [7]. The model adopted for computing the page rank is different with respect to the developed neural models in

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 1-8

structured domains in that it does not learn parameters, but nicely operates on a unique domain, namely the Web, by a computation which is guaranteed to converge thanks to the special linear structure of the model.

In this paper we propose a novel view of learning in web domains, which are abstractions of the Web in that they are a unique, typically huge, graph over which a function is defined. The aim of web learning is to infer that function relying on the knowledge of a subgraph of the web. Notice that the Web is just an example of a web domain, which could be the appropriate abstraction for different problems, especially in pattern recognition. Unlike the developed learning models for structured domains, in this paper we are interested in the case in which data are embedded into a unique graph, the web, not in a collection. We give a general view of the notion of web supervised learning in the framework of function optimization. We propose three examples of this general framework. First, we show that the models of learning in structured domains [7] are a special case of web learning. Second, we briefly review the learning of page rank on the Web as recently proposed in [10] and, finally, we show some intriguing links with graph spectral analysis as recently used in pattern recognition [5].

2 Web domains and node ranking

The theory of multi-dimensional systems [3] can be given a nice extension in the case in which, like the Web, the domain of the function becomes a graph. Instead of the traditional grid on which functions are defined, the domain can be generalized to a graph, whose labelled nodes contain vectors of real numbers. Hence in our view, a web domain can be formalized as follows:

Definition 2.1 *Let V be the set of vertices and consider and a subset $G \subset V \times V$, which represents a directed graph. Let $\ell : V \to I\!R^m : v \to \boldsymbol{u}_v$ be a function which associates a real vector to any node of the graph. The corresponding image $\ell(G)$ of function $\ell(\cdot)$ is referred to as a* WEB DOMAIN.

Definition 2.2 *Let $\Omega = \ell(G)$ be a web domain. Given any node v, we can construct, on Ω, a* LOCAL MAP *based on the following dynamical system Σ*

$$\boldsymbol{x}_v = f(\boldsymbol{x}_{ch(v)}, \boldsymbol{u}_v; \Theta) \tag{1}$$

$$y_v = g(\boldsymbol{x}_v) \tag{2}$$

where $\boldsymbol{x} \in I\!R^n$, $\boldsymbol{y} \in I\!R$, and $\Theta \in I\!R^p$ is a (learning) parameter. The symbol $\boldsymbol{x}_{ch(v)}$ denotes an ordered list of states associated with the children of node v. The web domain Ω equipped with the dynamical system Σ, is referred to as a WEB, *and is denoted $\{\Omega, \Sigma\, \Theta\}$.*

Here we report noticeable examples of different form of web computation.

- RECURSIVE NEURAL NETWORKS
  Interestingly enough, if a web domain is properly partitioned in graphs,

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 1-8

we can easily see that the computation corresponding to Σ becomes the one introduced in [**?**] for the case of directed acyclic ordered graphs. The hypothesis of directed graphs makes it possible to carry out a forward computation, while the ordering of the children is required by the function arguments. This hypothesis has been recently removed in different ways (see. e.g. [2]), but removing the acyclic assumption is more critical and requires the relaxation to an equilibrium point. This is basically what happens in the computational scheme behind Google's PageRank.

- GOOGLE'S PAGERANK
  A remarkable example is given by Google's PageRank [4]. The basic idea is that of introducing a notion of page authority which is independent of the page content. Such an authority only emerges from the topological structure of the Web. In Google's PageRank, the authority reminds the notion of citation in scientific literature. In particular, the authority of a page p depends on the number of incoming hyperlinks (number of citations) and on the authority of the page q which cites p by a forward link. Moreover, selective citations from q to p are assumed to provide more contribution to the score of p than uniform citations. Hence, PageRank x_p of p is computed by taking into account the set of pages pa$[p]$ pointing to p. According to Brin and Page [4]:

$$x_p = d \sum_{q \in \mathrm{pa}[p]} \frac{x_q}{h_q} + (1-d) \ . \tag{3}$$

  Here $d \in (0,1)$ is a DAMPING FACTOR and h_q is the HUBNESS of q, that is the number [1] of hyperlinks outcoming from q. When stacking all the x_p into a vector $\boldsymbol{x}$, we get

$$\boldsymbol{x} = d\boldsymbol{W}\boldsymbol{x} + (1-d)\mathbb{1}_N \ , \tag{4}$$

  where $\mathbb{1}_N = [1, \ldots, 1]'$ and $\boldsymbol{W} = \{w_{i,j}\}$ — the TRANSITION MATRIX — is such that $w_{i,j} = 1/h_j$ if there is a hyperlink from j to i and $w_{i,j} = 0$, otherwise. Thus, $\boldsymbol{W}$ is a non–null matrix, where each column either sums to 1 or to 0. More precisely, the j–th column $\boldsymbol{W}_j$ is null if page j does not contain hyperlinks. Otherwise, $\boldsymbol{W}_j$ can be constructed by the normalization of the j–th row of the Web adjacency matrix. It can easily been seen that the solution of equation 4 can be obtained, as $t \to \infty$, as a fixed point of equation $\boldsymbol{x}(t+1) = d\boldsymbol{W}\boldsymbol{x}(t) + (1-d)\mathbb{1}_N$, which is asymptotically stable.

- SPECTRAL ANALYSIS
  Web computation is well-suited for attacking the complexity inherently associated with topological invariance in pattern recognition. Even though, many connectionist models possess strong learning capabilities, problems of invariance under translations and rotations are not inherently solved,

[1] In graph theory, this is also referred to as the outdegree of node q.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 1-8

and require neural networks to consume learning resources which could be more focussed on the recognition process. On the other hand, web computation focuses on graphical representations of the data, thus stressing topological properties. This is especially clear in the PageRank scheme, in which relevant pages are discovered on the Web, regardless of their "geographical location." The importance of emphasizing the topological properties of the data in pattern recognition has been widely advocated. Recently, Carcassoni and Hancock [5] have proposed a nice framework for extending classic approaches of graph matching which makes use of graph spectral properties. Interestingly, the web computation inherently captures topological feature, thus carrying out a sort of spectral analysis. If we consider graphs without information attached to the nodes, following PageRank, the solution of equation 4 is

$$\boldsymbol{x}(d) = (1-d) \cdot (\boldsymbol{I} - d\boldsymbol{W})^{-1} \mathbb{1}_N \,, \tag{5}$$

Function $\boldsymbol{x}(d)$ detects web topological features. Interestingly, the same analysis can be carried out in the case in which the input is a vector in $\mathbb{R}^m$. Notice that web computation also incorporates cellular neural networks, in which all the units process the input.

Generally speaking, web computation can either stress the topological component or the information attached to the nodes by function ℓ. For instance in PageRank, the Web computation only involves topological properties, regardless of the content of the Web pages. As a limit case, one can neglect the topological component and the web computation "collapse" to real-valued functions'. In pattern recognition, the vectors of real features and the graph topology can be jointly exploited [1], thus opening the doors to nice extensions of well-established paradigms.

3 Web learning

A given web $\{\Omega, \Sigma, \Theta\}$ exhibits a dynamics that is typically dependent on a set of parameters Θ, which makes it suitable for learning. Similar to artificial neural networks defined on traditional learning environments, one can construct learning theories which basically consist of adapting the parameters Θ. Both the supervised and unsupervised learning protocol can be conceived which, however, must take into account topological issues. Interestingly, the concepts to be discovered by learning from examples are now at node level and the nodes are inherently embedded into the web domain. For a concept to be meaningful, it has to be somehow related to the content and to the topological properties of the node. In this paper, we restrict the attention to supervised learning, where a teacher provides a real number to be attached to each node, so as to generate the LEARNING ENVIRONMENT $\{(v, t(v)), \quad v \in V_L \subset V\}$. Consequently, the degree of fitting of the learning environment is evaluated by minimizing the COST FUNCTION, generally subjected to constraints

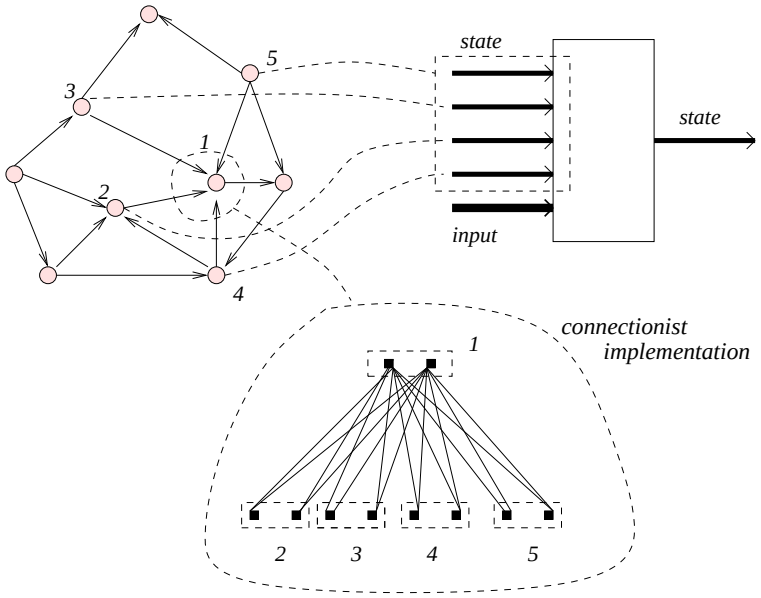

Figure 1: The local computation hypothesis in web domains and the connectionist implementation of function f.

$$\min_{\Theta} \sum_{v \in V_L} d(t(v) - y(v;\Theta)) \tag{6}$$

$$\phi(\boldsymbol{y}_L(\Theta)) = \mathbf{0} \tag{7}$$

where $d(\cdot)$ is a metrics on $\mathbb{R}$ and $\phi(\cdot)$ expresses a set of constraints on the values $y(v;\Theta)$. This problem arises when one wants to provide a score to Web pages on the basis of indications and constraints acquired by human experience, but there are plenty of applications in different field and, especially, in pattern recognition.

It is worth mentioning that in most interesting real-world problems, $p << |V|$, since there are regularities amongst nodes on both content and topology. Basically, reasonable concepts to be learned require that one can infer properties on other nodes and, therefore, problems of overfitting suggest the adoption of an appropriate number of parameters.

Here we revisit the three class of models considered in the previous section to give some preliminary insights on learning.

- RECURSIVE NEURAL NETWORKS
  In recursive neural networks the learning of the parameters Θ is made possible by the hypothesis of dealing with directed ordered graphs. Similar to the case of sequences, we can unfold the network along the structure

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 1-8

and use BACKPROPAGATION THROUGH STRUCTURE [?]. The minimization of $\sum_{v \in V_L} d(t(v) - y(v;\Theta))$ assumes an appropriate sharing of parameters Θ which allows very good generalization to new examples in many interesting real-world problems [6]. Notice that the connectionist unfolding which produces feedforward networks in the case of directed acyclic graphs, gives rise to neural networks with cycles requiring a relaxation to an equilibrium point in the general case depicted in Fig 1.

- LEARNING THE PAGE RANK ON THE WEB
An example of learning in web domains has been recently proposed in [10] in which, apart from the Web pages whose PageRanks need to be modified, for the rest of the pages, we wish to minimize the modification of their PageRank. Unlike the case of recursive neural networks, in which the learning parameters are associated with the arcs, the chosen learning parameter Θ is associated with each node. Hence, the learning scheme is based on $\boldsymbol{x}_a = (1-d)\left(\boldsymbol{I} - d\boldsymbol{W}\right)^{-1}\boldsymbol{e} \doteq (1-d)\boldsymbol{M}\cdot\boldsymbol{e}$, where $\boldsymbol{e}$ is a $|V|$ - dimensional real vector, acting as a learning parameter. Let $\boldsymbol{x}_g$ be the PageRank of a set of Web pages and let $\boldsymbol{x}_a$ be the modified ranks of the same set when we apply the control. Then, we minimize

$$\min_{\boldsymbol{e}} J = \|\boldsymbol{x}_a - \boldsymbol{x}_g\|_2 \tag{8}$$

under the constraints:

$$\boldsymbol{x}_g = (1-d)\left(\boldsymbol{I} - d\boldsymbol{W}\right)^{-1}\mathbb{1} \tag{9}$$

$$\boldsymbol{x}_a = (1-d)\left(\boldsymbol{I} - d\boldsymbol{W}\right)^{-1}\boldsymbol{e} \tag{10}$$

$$\boldsymbol{B}\boldsymbol{x}_a \geq \mathbf{b} \tag{11}$$

the problem is converted to

Proposition 3.1 *The minimization stated by 8 can be converted to*

$$\begin{cases} \min_{\boldsymbol{e}} \;\; \boldsymbol{e}^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{e} - 2\,\mathbb{1}_n^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{e} \\ \\ \boldsymbol{B}\,\boldsymbol{M}\boldsymbol{e} \geq b \end{cases} \tag{12}$$

This proposition follows by straightforward algebra from 8 and 11.

Notice that (12) is a standard positive definite quadratic programming problem with an inequality constraint set which can be solved by common techniques [9]. The problem fits in the positive definite quadratic programming problem because M^TM is positive definite with an inequality constraint set.

As already pointed out, however, the curse of dimensionality and the need to extract regularities, suggest to use a reduced number of parameters

for learning. Let us consider a partition $C_1, \ldots, C_k$ of Web pages. This clustering is used as a pre-processing step so as to reduce the amount of information and provide a regularization for the sub-sequent learning process. In [10], it is shown that equation 12 is reduced to a another quadratic programming problem operating on $\mathbb{R}^k$, where typically $k << |V|$. Experimental results are given which show that the clustering pre-processing is technically sound. If we consider that, in practice, humans are likely to provide also contradicting constraints, then it makes sense to relax their strength. As usual, this can be done by introducing the constraints as penalty functions. This relaxation has also been shown to be effective in practice [10]

- SPECTRAL ANALYSIS
  Following Carcassoni and Hancock [5], graph spectral analysis can be used on web domains. In addition to eigenvalues and eigenvectors, one can rely on random walk models, like PageRank, augmented with information, $\boldsymbol{u} \in \mathbb{R}^m$, within the nodes

$$\boldsymbol{x}(d) = (1-d) \cdot (\boldsymbol{I} - d\boldsymbol{W})^{-1} \boldsymbol{P}\boldsymbol{u} \ , \tag{13}$$

  In so doing, the matrix $\boldsymbol{P} \in \mathbb{R}^{|V|,m}$ can be used as a learning parameter, in addition to a sub-sequent learning taking place on the vectorial representation associated with the web [2].

4 Conclusion

One of the main motivations for proposing the reformulation of learning in web domains is that data often exhibit relationships, which are typically neglected. In particular, topological features, which are efficiently expressed by spectral analysis and random walk models, can hardly be learned by most traditional models.

In this paper we have introduced a new general framework for learning which is based on the concept of web. Beginning from present existing limitations of traditional learning schemes, we have stressed the importance of providing structured representation of the data. This is claimed by many people and is becoming the subject of studies which emphasize learning so as to exhibit robustness to noise [8]. The introduction of webs represents an extension of machine learning approaches based on collections of graphs, in that they represent a domain where the function to be learned is defined. In a sense, it reminds us the difference between learning from a collection of temporal sequences and on-line learning on a given signal, where there are no marks to denote sequences of the collection. We have reviewed learning approaches already proposed as particular cases of the proposed framework, with special emphasis on a recently proposed model for learning the page rank in the Web [10].

[2] Research towards this direction, however, is still in its infancy.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 1-8

References

[1] G. Adorni, S. Cagnoni, and M. Gori. Adaptive graphic pattern recognition: foundation and perspectives. In H. Bunke and A. Kandel, editors, *Hybrid methods in pattern recognition*, pages 23–59. World Scientific, 2002.

[2] M. Bianchini, M. Gori, and F. Scarselli. Processing directed acyclic graphs with recursive neural networks. *IEEE Transactions on Neural Networks*, pages 1464–1470, 2001.

[3] N. Bose. *Applied Multidimensional Systems Theory*. Van Nostrand Reinhold Co., N.Y, 1982.

[4] Sergey Brin and Lawrence Page. The anatomy of a large–scale hypertextual Web search engine. In *Proceedings of the 7th World Wide Web Conference (WWW7)*, 14–18 April 1998.

[5] M. Carcassoni and E.R. Hancock. Spectral correspondence for point pattern matching. *Pattern Recognition*, 36(1):193–204, 2003.

[6] P. Frasconi, C.Goller, M.Gori, A.Kuchler, and A.Sperduti. From sequences to data structures: Theory and applications. In J. Kolen and S.Kremer, editors, *A Field Guide to Recurrent Neural Networks*, pages 351–374. IEEE-Press, 2001.

[7] P. Frasconi, M. Gori, and A. Sperduti. A general framework for adaptive processing of data structures. *IEEE Transactions on Neural Networks*, 9(5):768–786, September 1998.

[8] P. Frasconi, M. Gori, and A. Sperduti. Guest editors' introduction: Special section on connectionist models for learning in structured domains. *IEEE Trans. on Knowlege and Data Engineering*, 13(3):145–147, March/April 2001.

[9] P. Gill, W. Murray, and W. Wright. *Practical optimization*. Academic Press, 1981.

[10] A. C. Tsoi, G. Morini, F. Scarselli, M. Hagenbuchner, and M. Maggini. Adaptive ranking of web pages. In *Proceedings of the the World Wide Web Conference (WWW12)*, 2003.