A neural networks-based approach in phytotaxonomic research

Marina Gorunescu[1] **& Florin Gorunescu**[2]

[1] Department of Computer Science, University of Craiova, 13 Al. I. Cuza St., 200585 Craiova, Romania, e-mail: mgorun@inf.ucv.ro

[2] Department of Mathematics, Biostatistics and Computer Science, University of Medicine and Pharmacy of Craiova, 2-4 Petru Rares St., 200349, Craiova, Romania, e-mail: fgorun@rdslink.ro ; gorun@umfcv.ro

Received: October 27, 2006 ▹ Accepted: December 12, 2006

Abstract. Modern plant taxonomic classification has progressed steadily since the 18th century modified by the advancement in morphology, evolution, and genetics. This paper introduces a novel computational technique, based on the machine learning paradigm, which can accurately differentiate between different species of plants. Discrimination is based on an artificial neural network trained to learn by examples and using numerically quantified features of plants. This neural computing-based approach represents an efficient tool in the computer-aided taxonomy. An illustrative application concerning the classification of *Iris* flowers is also presented.

Key words: classification, Iris flower, neural networks, plant taxonomy

Introduction

Taxonomy (Greek verb *τασσεῖν* or *tassein* = "to classify" and *νόμος* or *nomos* = "law, science") is the method by which scientists, conservationists and naturalists classify and organize the vast diversity of living things in an effort to understand the evolutionary relationships between them. Modern taxonomy originated in the mid-1700s when Swedish-born Carolus Linnaeus (Carl von Linné) published his famous *Systema Naturae* (first edition – 1735, eleven pages only), outlining his new and revolutionary method for classifying and especially for naming living organisms. Biosystematics, as a study of the diversity of life and the relationships among living things in time, uses taxonomy (or often, scientific classification) as a primary tool in understanding organisms.

Plant taxonomy, one of the main branches of taxonomy, basically represents the science that finds, describes, classifies and names plants (Stace 1992). It originated in the earliest classifications of plants made by the Greek philosophers, such as Aristotle and Theophrastus, and got a scientific characteristic due to Linnaeus' *Species Plantarum* (1753). Modern plant classification involves chemical and morphological analyses and, ultimately, DNA-based techniques.

Artificial intelligence (AI) is a branch of Computer Science concerned with making computers behave like humans (the term was coined in 1956 by John McCarthy at the Massachusetts Institute of Technology – MIT). The goal of AI is the development of paradigms or algorithms that require machines to perform cognitive tasks, at which humans are currently better. Machine learning (ML), one of the broad subfields of AI, is concerned with the development of algorithms and techniques that allow computers to "learn". One of the main topics covered by ML is represented by the artificial neural networks or, simply, neural networks (NN), also known as neural computing, a methodology originating in the mid-1900s and introduced by the neurophysiologist W. McCulloch and the logician W. Pitts (1943), attempting to imitate the way a human brain works by creating connections between processing elements, the computer equivalent of neurons. NN, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. In summary, the main disadvantages of using NN are the following: (a) generally they have no design theory or unique solution, (b) they cannot be generally guaranteed to converge on their global minimum, or occasionally to even converge at all, (c) they are too slow for practical use in large-scale issues. Furthermore, the problem of training NN to perform well is a sensitive issue, especially for cases where only very limited numbers of training samples are available (i.e. the network overfits the training data and fails to capture the true statistical process generating the data). To overcome such a problem the computational learning theory is used.

Despite their ability to learn by example, which makes them very flexible and powerful tools in knowledge discovery, they have not been intensively applied in taxonomy yet. Among the very few contributions in this domain, we can mention the plant seed classification (Goodacre & al. 1996), taxonomic discrimination of higher plants (Kim & al. 2004), or species identification (Clark 2003). Accordingly, the aim of this paper is to develop a NN-based methodology, capitalizing on their ability to learn from data with or without a teacher and thus making of them invaluable tools in classification or pattern recognition. In order to illustrate the practical use of such a technique in taxonomical research, we have applied it to the *Iris* database (Fischer 1936).

Material and methods

NN represent a Computer Science discipline concerned with non-programmed adaptive information processing systems that develop associations between objects and response to their environment. The basic unit of any NN is represented by an artificial neuron, which captures the essence of the biological neural model, synthetically displayed in Fig. 1.

Basically, the neuron receives a certain number of inputs x_i and sums them to produce an output. Usually the sums of each node are weighted (the weight parameters w_i), and the sum is passed through the *activation function*, to produce the output of the neuron. The synthetic scheme of an artificial neuron is illustrated in Fig. 2.

There are two phases in neural information processing: the *training phase* and the *using phase*. In the training phase, a training dataset is used to determine the weight parameters w_i that define the neural model. This trained neural model will be used later in the using phase to process real test patterns and yield classification results.

The *perceptron* (Rosenblatt 1962) is the simplest form of NN, also called a *single-layer network*, used for the classification of linearly separable patterns (i.e. patterns that lie on opposite sides of a hyperplane) only (perceptron convergence theorem – Rosenblatt 1962). Basically, it consists of a single neuron with adjustable (synaptic) weights. Since NN with a single layer of weights has very limited capabilities in solving complex classification problems, we have to consider networks consisting of successive layers of adaptive weights, that is a *multi-layer perceptron* (MLP) (Rumelhart & al. 1986a, b). A synthetic scheme of a multi-layered network (two hidden layers) is presented in Fig. 3.

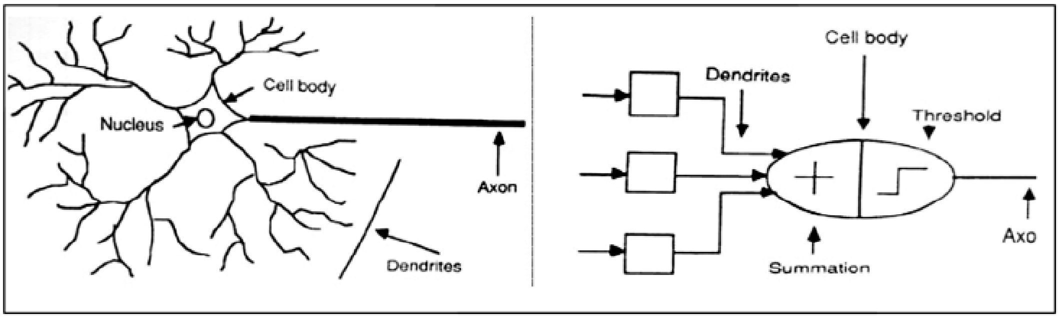

Fig. 1. The synthetic model of a biological neuron.

A MLP has three distinctive characteristics making it capable, at least theoretically, to represent a wide range of computable functions:

- The model of each neuron in the network usually includes a *smooth* (i.e. differentiable everywhere) nonlinear activation function, as opposed to Rosenblatt's perceptron, generalizing the input-output relation of the network;
- The network contains one or more layers of hidden neurons that are not part of the input or output of the network, enabling the network to learn complex tasks by extracting progressively more meaningful features from the input data;
- The network exhibits a high degree of connectivity between neurons.

Note that networks with just two layers are capable of approximating any continuous function.

The architecture of NN (number of neurons and topology of connections) can have a significant impact on its performance in any particular application. Various techniques have been developed for optimizing the architecture, in some cases as part of the network training process itself. Techniques such as exhaustive search through a restricted class of network architecture, pruning algorithms, or network committee and mixture of experts, are commonly adopted in practice. Finding an appropriate architecture for a given application by using techniques of these forms requires a significant computational effort, which is a time- consuming process. An alternative approach is to consider automated optimization procedures by using, for instance, evolutionary computation (e.g. genetic algorithms, genetic programming).

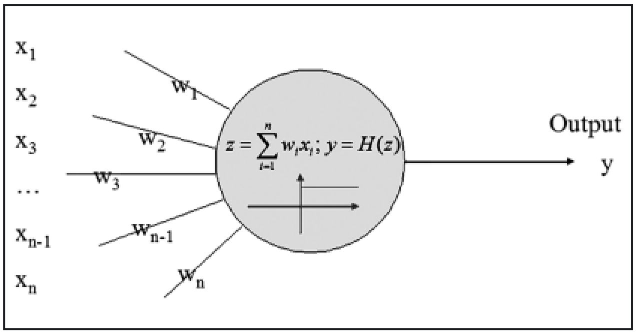

Fig. 2. The synthetic scheme of an artificial neuron.

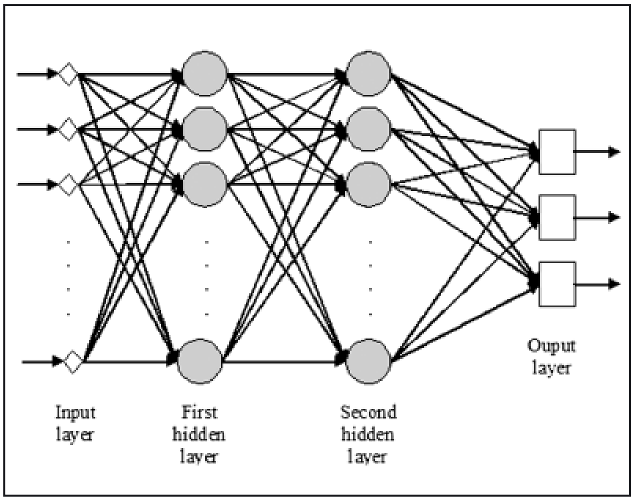

Fig. 3. The synthetic scheme of a multi-layered network (MLP).

In the classification problems area, a useful interpretation of network outputs is to estimate the probability of class membership, in which case the network is actually learning to estimate a probability density function (p.d.f.). This special case of NN, known as *probabilistic neural networks* (PNN) (Specht 1988), provides a general solution to pattern classification problems by following the probabilistic approach based on the Bayes decision theory. Basically, PNN represents a three-layer network, consisting of: (*a*) a *pattern layer*, (*b*) a *summation layer* and (*c*) a *decision layer*. The network paradigm uses a sum of small multivariate Gaussian distributions, centred at each training sample, that is:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}\sigma_i^p} \cdot \frac{1}{m_i} \cdot \sum_{j=1}^{m_i} \exp\left(-\frac{\left\| x - x_j \right\|^2}{2\sigma_i^2} \right),\ i = 1, 2, \ldots, q,$$

to estimate the p.d.f's $\boldsymbol{f}_\mathrm{i}(\boldsymbol{x})$ corresponding to each decision class Ω_i. The key factor in PNN is the way to determine the value of σ. Although the smoothing factor is chosen heuristically, there are different improvements to speed up the searching process (Gorunescu & a. 2005a, b, c).

In conclusion, in the training mode NN is trained to associate outputs with input patterns. When NN is used, it identifies the input pattern and tries to output the associated output pattern. If a pattern that has no output associated with it is given as an input, NN gives the output that corresponds to a taught input pattern that is least different from the given pattern. For more details concerning NN, see (Bishop 1995), (Haykin 1999) and (Zaknich 2003).

Taxonomic classification is an act of placing an object/instance into a set of categories based on the properties of the object/instance. Algorithmically, the process of classification involves:

- **Input:** A collection of records (*training set*). Each record contains a set of *attributes*; one of the attributes is the *class*.
- The findings of a *model* for class attribute as a function of the values of other attributes;
- A *test set* is used to determine the accuracy of the model. Usually, the given dataset is divided into training and test sets, with the training set used to build the model and the test set used to validate it;
- **Output:** previously unseen records should be assigned a class as accurately as possible.

While there are many methods for classification, the usual computer-aided techniques are the following: decision tree-based methods, rule-based methods, memory-based reasoning, neural networks-based methods, naïve Bayes and Bayesian belief networks, support vector machines.

Technically, let us consider the general case of the q-category classification problem, in which the states of nature are denoted by $\Omega_1, \Omega_2, \ldots, \Omega_q$. The goal is to determine the class (category) membership of a multivariate sample data (i.e. a p-dimensional random vector $\mathbf{x}$) into one of the q possible groups $\Omega_1, \Omega_2, \ldots, \Omega_q$. Concretely, each object/instance to be classified is represented by a p-dimensional vector, denoted $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, where the components $\mathbf{x}_i$ represent some of the most important characteristics (attributes) leading to the right classification. The choice of main attributes used in the classification process strongly depends on the human expertise in the respective domain. Let us remark that the data concerning the object attributes can be: numerical (quantitative), categorical (qualitative), ranks, percentages, rates, scores etc. For instance, we can handle continuous data, such as leaf length or width; nominal data, such as flower colour; ordinal data, such as small, medium, large (size); ratio, such as counts etc. Note that non-numerical data must be quantified in an appropriate form. Next, a distance measure in the feature space, in order to quantify the difference between objects, is needed to accomplish the classification process.

To conclude, in plants taxonomy we need a set of plants with some significant features, coded/digitalized in order to be processed by computer, and a criterion to be used in classification.

In order to highlight the efficiency of this computer-aided taxonomy, we have considered the problem of classifying the *Iris* plant, taking into account three flower types (classes): *Setosa*, *Virginica*, *Versicolour*, and four attributes: sepal width and length and petal width and length. Note that the length and width of the petals/sepals varies not only between types but also within the same type of *Iris*. The chosen features adequately separate the three flower types. Data came from the UCI Machine-Learning Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html). The dataset consists of 150 *Iris* flowers in the following distribution: 50 *Setosa*, 50 *Virginica* and 50 *Versicolour*.

Results

We have used two main types of NN in the classification process: (*a*) the *multi-layer perceptron* (MLP) and (*b*) the *probabilistic neural network* (PNN).

In order to evaluate the classification efficiency, two main metrics have been computed: the testing error, along with the corresponding performance of the classifier. Moreover, NN can conduct a *sensitivity analysis* of its inputs, indicating which attributes are considered more important. The sensitivity analysis can give important insights into the usefulness of individual attributes. Concretely, since in general the attributes are not independent, the sensitivity analysis rates attributes according to the deterioration in modelling performance that occurs if that attribute is no longer available to the model. In conclusion, important attributes have a high error, indicating that the network performance deteriorates badly if they are not present. Thus, the sensitivity analysis identifies variables that can be safely ignored in subsequent analysis, and key variables that must always be retained.

The 10-fold cross-validation has been used as a testing method. Accordingly, the classification accuracy is computed 10 times, each time leaving out one of the sub-samples from the computations and using that subsample as a test sample for cross-validation, so that each subsample is used 9 times in the learning sample and just once as a test sample. The best results obtained using this computer-aided classification, running a number of 29 NN (15 MLP and 14 PNN), are depicted in Table 1 and prove the suitability and success of this methodology.

Table 1. The classification results obtained running NN on *Iris* flower.

NN type	Inputs	Hidden layers	TError	Performance
MLP	4	6	6.43 %	95.14 %
PNN	4	80	2.92 %	98.57 %

The number of inputs in the network represents the number of variables before pre-processing (each network includes its own pre- and post-processing layers). Note that the number of hidden layers, together with the number of input variables, defines the *complexity* of the network. The testing error (TError) is most relevant as an indicator of the ability of the NN to make predictions given new data. This is the root mean square (RMS) of the errors on each individual case, where the error on each individual case is measured by the network's error function. Performance, representing a measure of the success of the network, is actually measured on the testing subset. For NN used in the classification issues, performance is the percentage of cases correctly classified. From Table 1 follows that PNN is the best classifier in this concrete case. The sensitivity analysis showed that the hierarchy (descending order) of the four attributes is the following, regardless the NN type: #1 petal width, #2 petal length, #3 sepal width and #4 sepal length.

Discussion

The methodology developed by us makes possible the exploration and analysis by automatic means of large quantities of data related to plants characteristics, in order to obtain an optimal classification. The purpose of this paper is to demonstrate the suitability and ability of the NN methodology in classification problems regardless of the type of features, as long as these characteristics are (numerically) coded in an appropriate manner.

In the example presented in this paper, the petal and sepal dimensions are correlated. This means that it is possible to achieve a good degree of discrimination by using the petal/sepal length or width alone, and only a slightly better discrimination with both. Accordingly, in a concrete plant taxonomy issue, when selecting a large number of features, it is more than likely that there will be some correlation between them (identified by the sensitivity analysis). Consequently, it may be desirable to reduce the dimension of the feature vector by various statistical methods to achieve the smallest essential dimension without any loss of actual discriminating information.

It is worthwhile to add that this methodology can be successfully used in other phytological domains, such as identification of plants diseases, for instance, based on symptoms and knowledge of previous cases.

NN are still in their infancy, but it is very likely that they will eventually and very soon have applications in almost all real systems, including phytology too. NN learn by examples so the details of how to recognize the plant are not needed. What is needed is a set of examples that are representative of all the variations of the plant concerned in the classification process. This approach was designed as a computer-assisted method for plant classification or other prediction tasks in the phytological domain, helping the researchers make an optimal decision.

Acknowledgements. The authors wish to thank Dr E. El-Darzi for his continued support and encouragement during their visit at Harrow School of Computer Science, University of Westminster, London, UK, as well as to the unknown reviewer for his/her valuable remarks.

This paper originates in the joint work of the International Exchange of Academic Staff and Research (1st August 2004–31st July 2005), University of Westminster, London, UK.

References

Bishop, C.M. 1995. Neural Networks for Pattern Recognition, Oxford Univ. Press, Oxford.

Clark, J.Y. 2003. Artificial neural networks for species identification by taxonomists. – BioSystems, **72**(1): 131-147.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. – Ann. Eugenics, **7**: 179-188.

Goodacre, R., Pygall, J. & Kell, D.B. 1996. Plant seed classification using pyrolysis mass spectrometry with unsupervised learning. The application of auto-associative and Kohonen artificial neural networks. – Chemometrics & Intelligent Laboratory Systems, **34**(1): 69-83.

Gorunescu, F., Gorunescu, M., El-Darzi, E. & Gorunescu, S. 2005a. An evolutionary computational approach to probabilistic neural network with application to hepatic cancer diagnosis. – In: **Cunningham, P. & Tsymbal, A.** (eds.), Proc. 18th IEEE Int. Symp. "Computer-Based Medical Systems", IEEE CBMS 2005, Dublin, Ireland. Pp. 461-466. IEEE Computer Science Press.

Gorunescu, F., Gorunescu, M., El-Darzi, E., Ene, M. & Gorunescu, S. 2005b. Statistical comparison of a probabilistic neural network approach in hepatic cancer diagnosis. – In: Proc. Eurocon2005, IEEE Int. Conf. "Computer as a tool", Belgrade, Serbia. Pp. 237-240. ©2005 IEEE.

Gorunescu, M., Gorunescu, F., Ene, M. & El-Darzi, E. 2005c. A heuristic approach in hepatic cancer diagnosis using a probabilistic neural network-based model. – In: **Janssen, J. &**

Lenca, P. (eds), Proc. 11th Applied Stochastic Models and Data Analysis –ASMDA 2005. Pp. 1016-1025. Brest, France.

Haykin, S. 1999. Neural Networks: a comprehensive foundation. Prentice Hall, NJ.

Kim, S.W., Ban, S.H., Chung, H.J., Choi, D.W., Choi, P.S., Yoo, O.J. & Liu, J.R. 2004. Taxonomic discrimination of higher plants by pyrolysis mass spectrometry. – Pl. Cell Rep., **22**(7): 246-250.

McCulloch, W.S. & Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. – Bull. Math. Biophys., **5**: 115-133.

Rosenblatt, **F.** 1962. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington DC: Spartan Books.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. 1986a. Learning internal representations by error propagation. – In: **Rumelhart, D.E., McClelland, J.L.** & **the PDP Research Group** (eds), Paralled Distributed Processing. Explorations in the Microstructure of Cognition. Vol. **1**, pp. 318-362. The MIT Press, Cambridge, MA.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. 1986b. Learning representations of back-propagation errors. – Nature (London), **323**: 533-536.

Specht, **D.F**. 1988. Probabilistic neural networks for classification, mapping, or associative memory. – In: IEEE Int.Conf. Neural Networks. Vol. **1**, pp. 525-532.

Stace, C.A. 1992. Plant Taxonomy and Biosystematics. 2nd ed. Cambridge Univ. Press, Cambridge.

Zaknich, A. 2003. Neural Networks for Intelligent Signal Processing. World Scientific Pub. Co Inc.