Neural Network–Based Visual Data Mining for Cancer Data

Enrique Romero
Technical University of Catalonia, Spain

Julio J. Valdés
National Research Council Canada, Canada

Alan J. Barton
National Research Council Canada, Canada

INTRODUCTION

According to the World Health Organization (http://www.who.int/cancer/en), cancer is a leading cause of death worldwide. From a total of 58 million deaths in 2005, cancer accounts for 7.6 million (or 13%) of all deaths. The main types of cancer leading to overall cancer mortality are *i)* Lung (1.3 million deaths/year), ii) Stomach (almost 1 million deaths/year), *iii)* Liver (662,000 deaths/year), *iv)* Colon (655,000 deaths/year) and *v)* Breast (502,000 deaths/year). Among men the most frequent cancer types worldwide are (in order of number of global deaths): lung, stomach, liver, colorectal, oesophagus and prostate, while among women (in order of number of global deaths) they are: breast, lung, stomach, colorectal and cervical.

Technological advancements in recent years are enabling the collection of large amounts of cancer related data. In particular, in the field of Bioinformatics, high-throughput microarray gene experiments are possible, leading to an information explosion. This requires the development of data mining procedures that speed up the process of scientific discovery, and the in-depth understanding of the internal structure of the data. This is crucial for the non-trivial process of identifying valid, novel, potentially useful, and ultimately *understandable patterns* in data (Fayyad, Piatesky-Shapiro & Smyth, 1996). Researchers need to *understand* their data rapidly and with greater ease. In general, objects under study are described in terms of collections of *heterogeneous* properties. It is typical for medical data to be composed of properties represented by nominal, ordinal or real-valued variables (scalar), as well as by others of a more complex nature, like images, time-series, etc. In addition, the information comes with different degrees of precision, uncertainty and information completeness (missing data is quite common).

Classical data mining and analysis methods are sometimes difficult to use, the output of many procedures may be large and time consuming to analyze, and often their interpretation requires special expertise. Moreover, some methods are based on assumptions about the data which limit their application, specially for the purpose of exploration, comparison, hypothesis formation, etc, typical of the first stages of scientific investigation. This makes graphical representation directly appealing. Humans perceive most of the information through vision, in large quantities and at very high input rates. The human brain is extremely well qualified for the fast understanding of complex visual patterns, and still outperforms the computer. Several reasons make Virtual Reality (VR) a suitable paradigm: *i)* it is *flexible* (it allows the choice of different representation models to better suit human perception preferences), *ii)* allows *immersion* (the user can navigate inside the data, and interact with the objects in the world), *iii)* creates a *living* experience (the user is not merely a passive observer, but an actor in the world) and *iv)* VR is *broad and deep* (the user may see the VR world as a whole, and/or concentrate on specific details of the world). Of no less importance is the fact that in order to interact with a virtual world, only minimal skills are required.

Visualization techniques may be very useful for medical decisión support in the oncology area. In this paper unsupervised neural networks are used for constructing VR spaces for visual data mining of gene expression cancer data. Three datasets are used in the paper, representative of three of the most important

types of cancer in modern medicine: liver, stomach and lung. The data sets are composed of samples from normal and tumor tissues, described in terms of tens of thousands of variables, which are the corresponding gene expression intensities measured in microarray experiments. Despite the very high dimensionality of the studied patterns, high quality visual representations in the form of structure-preserving VR spaces are obtained using SAMANN neural networks, which enables the differentiation of cancerous and noncancerous tissues. The same networks could be used as nonlinear feature generators in a preprocessing step for other data mining procedures.

NEURAL NETWORKS FOR THE CONSTRUCTION OF VIRTUAL REALITY SPACES

VR spaces for the visual representation of information systems (Pawlak, 1991) and relational structures were introduced in (Valdés, 2002a) (Valdés, 2003). A *VR space* is a tuple $\Omega =< \underline{O}, G, B, R^m, g_0, l, g_r, b, r >$, where $\underline{O}$ is a relational structure ($\underline{O} =< O, \Gamma^v >$), O is a finite set of objects, and Γ^v is a set of relations); G is a non-empty set of *geometries* representing the different objects and relations; B is a non-empty set of *behaviors* of the objects in the virtual world; $R^m \subseteq \Re^m$ is a *metric space* of dimension m (Euclidean or not) which will be the actual VR geometric space. The other elements are mappings: $g_0 : O \rightarrow G$, $l : O \rightarrow R^m$, $g_r : \Gamma^v \rightarrow G$ and $b : O \rightarrow B$.

The typical *desiderata* for the visual representation of data and knowledge can be formulated in terms of minimizing information loss, maximizing structure preservation, maximizing class separability, or their combination, which leads to single or multi-objective optimization problems. In many cases, these concepts can be expressed deterministically using continuous functions with well defined partial derivatives. This is the realm of classical optimization where there is a plethora of methods with well known properties. In the case of heterogeneous information the situation is more complex and other techniques are required (Valdés, 2002b) (Valdés, 2004) (Valdés & Barton, 2005). In the unsupervised case, the function f mapping the original space to the VR (geometric) space R^m can be constructed as to maximize some metric/non-metric structure preservation criteria as is typical in multidimensional scaling (Borg & Lingoes, 1987) or minimize some error measure of information loss (Sammon, 1969). A typical error measure is:

$$Sammon\ Error = \frac{1}{\sum_{i<j} \delta_{ij}} \frac{\sum_{i<j} (\delta_{ij} - \xi_{ij})^2}{\delta_{ij}}$$

where δ_{ij} is a dissimilarity measure between two objects i, j in the original space, and ξ_{ij} is another dissimilarity measure defined on objects i, j in the VR space (the images of i, j under f). Typical dissimilarity measures for δ_{ij} are the Euclidean distance or the dissimilarity based on Gower's similarity coefficient (Gower, 1971). The Euclidean distance is the usual measure for ξ_{ij} in the VR space.

Usually, the mappings f obtained using approaches of this kind are *implicit* because the images of the objects in the new space are computed directly. However, a functional representation of f is highly desirable, specially in cases where more samples are expected *a posteriori* and need to be placed within the space. With an implicit representation, the space has to be computed every time that a new sample is added to the set, whereas with an explicit representation, the mapping can be computed directly. As long as the incoming objects can be considered as belonging to the same population of samples used for constructing the mapping function, the space does not need to be recomputed. Neural networks are natural candidates for constructing explicit representations due to their general universal approximation property. If proper training methods are used, neural networks can learn structure preserving mappings of high dimensional samples into lower dimensional spaces suitable for visualization (2D, 3D). If visualization is not a requirement, spaces of smaller dimension than the original can be used as new features for noise reduction or other data mining methods. Such an example is the SAMANN network. This is a feedforward network and its architecture consists of an input layer with as many neurons as descriptor attributes, an output layer with as many neurons as the dimension of the VR space and one or more hidden layers. The classical way of training the SAMANN network is described in (Mao & Jain, 1995). It consists of a gradient descent method where the derivatives of the Sammon error are computed in a similar way to the classical backpropagation algorithm. Different from the backpropagation algorithm,