Neural Network Based Virtual Reality Spaces for Visual Data Mining of Cancer Data: An Unsupervised Perspective

Enrique Romero[1], Julio J. Valdés[2], and Alan J. Barton[2]

[1] Dept. of Languages and Information Systems, Polytechnic University of Catalonia
c/ Jordi Girona, 1-3, 08034 Barcelona, Spain
eromero@lsi.upc.edu

[2] National Research Council Canada
M50 1200 Montreal Rd, Ottawa, ON K1A 0R6, Canada
julio.valdes@nrc-cnrc.gc.ca,
alan.barton@nrc-cnrc.gc.ca

Abstract. Unsupervised neural networks are used for constructing virtual reality spaces for visual data mining of gene expression cancer data. Datasets representative of three of the most important types of cancer considered in modern medicine (liver, lung and stomach) are considered in the study. They 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 virtual 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.

1 Introduction

According to the World Health Organization (WHO) 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

F. Sandoval et al. (Eds.): IWANN 2007, LNCS 4507, pp. 1020–1027, 2007.

useful, and ultimately *understandable patterns* in data. Researchers need to *understand* their data rapidly and with greater ease. Further, the increasing complexity of the data analysis procedures makes it more difficult for the user to extract useful information out of the results given by the various techniques applied. Visual techniques are, therefore, very appealing. 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 decision support in the oncology area. In this paper a neural network based approach is used for constructing virtual reality spaces for exploring gene expression cancer data. Three datasets resulting from microarray experiments are used in the paper, representative of three of the most important types of cancer considered in medicine: liver, lung and stomach cancer.

2 Neural Networks for the Construction of Virtual Reality Spaces

Virtual reality spaces for the visual representation of information systems [1] and relational structures were introduced in [2,3]. A *virtual reality space* is the tuple $\Upsilon =< \underline{O}, G, B, \Re^m, g_o, l, g_r, b, r >$, 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; $\Re^m \subset \mathbb{R}^m$ is a *metric space* of dimension m (euclidean or not) which will be the actual virtual reality geometric space. The other elements are mappings: $g_o : O \rightarrow G$, $l : O \rightarrow \Re^m$, $g_r : \Gamma^v \rightarrow G$, $b : O \rightarrow B$.

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 [4]. In the unsupervised case, the function f mapping the original space to the virtual reality (geometric) space $\mathbb{R}^m$ can be constructed as to maximize some metric/non-metric structure preservation criteria as is typical in multidimensional scaling [5], or minimize some error measure of information loss [6]. A typical error measure is:

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

where δ_{ij} is a dissimilarity measure between any two objects i, j in the original space, and $\zeta_{i^v j^v}$ is another dissimilarity measure defined on objects i^v, j^v in the virtual reality space (the images of i, j under f). 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 doesn't need to be recomputed. Neural networks are natural candidates for constructing explicit representations due to their general function 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 virtual reality space and one or more hidden layers. The classical way of training the SAMANN network is described in [7]. 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, the weights can only be updated after pairs of examples are presented to the network.

3 Cancer Data Sets Description

Three microarray gene expression cancer databases were selected. They are representative of some of the leading causes of cancer death in the world and share the typical features of these kind of data: a small number of samples (in the order of tens), described in terms of a very large number of attributes (in the order of tens of thousands).

3.1 Lung Cancer Data

Gene expressions were compared in [8] for severely emphysematous lung tissue (from smokers at lung volume reduction surgery) and normal or mildly emphysematous lung tissue (from smokers undergoing resection of pulmonary nodules). The original database contained 30 samples (18 severe emphysema, 12 mild or no emphysema), with $22,283$ attributes. Genes with large detection P-values were filtered out, leading to a data set with $9,336$ genes, that were used for subsequent analysis. Nine classification algorithms were used to identify a group of genes whose expression in the lung distinguished severe emphysema from mild or no emphysema. First, model selection was performed for every algorithm by leave-one-out cross-validation, and the gene list corresponding to the best model was saved. The genes reported by at least four classification algorithms (102 genes) were chosen for further analysis. With these genes, a two-dimensional hierarchical clustering using Pearson's correlation was performed that distinguished between severe emphysema and mild or no emphysema. Other genes were also identified that may be causally involved in the pathogenesis of the emphysema. The data was obtained from `http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=737`.

3.2 Stomach Cancer Data

A study of genes that are differentially expressed in cancerous and noncancerous human gastric tissues was performed in [9]. The original database contained 30 samples (22 tumor, 8 normal) that were analyzed by oligonucleotide microarray, obtaining the expression profiles for $6,936$ genes ($7,129$ attributes). Using the $6,272$ genes that passed a prefilter procedure, cancerous and noncancerous tissues were successfully distinguished with a two-dimensional hierarchical clustering using Pearson's correlation. However, the clustering results used most of the genes on the array. To identify the genes that were differentially expressed between cancer and noncancerous tissues, a Mann-Whitney's U test was applied to the data. As a result of this analysis, 162 and 129 genes showed a higher expression in cancerous and noncancerous tissues, respectively. In addition, several genes associated with lymph node metastasis and histological classification (intestinal, diffuse) were identified. The data was obtained from `http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=1210`.

3.3 Liver Cancer Data

Zebrafish liver tumors were analyzed and compared with human liver tumors in [10]. First, liver tumors in zebrafish were generated by treating them with carcinogens. Then, the expression profiles of zebrafish liver tumors were compared with those of zebrafish normal liver tissues using a Wilcoxon rank-sum test. The original database had 20 samples (10 normal, 10 tumor) and $16,512$ attributes. As a result of this comparison, a zebrafish liver tumor differentially expressed gene set consisting of $2,315$ gene features was obtained. This data set was used for comparison with human tumors. The results suggest that the molecular similarities between zebrafish and human liver tumors are greater than the molecular similarities between other types of tumors (stomach, lung and prostate). The data was obtained from `http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=2220`.

Table 1. Statistics of the best $1,000$ SAMANN networks obtained

	Sammon Error			
Data Set	Minimum	Maximum	Mean	Std.Dev.
Stomach Cancer	0.062950	0.077452	0.072862	0.003346
Lung Cancer	0.079242	0.107842	0.094693	0.006978
Liver Cancer	0.039905	0.055640	0.049857	0.003621

4 Experimental Settings

Data preprocessing. For stomach and lung data, each gene was scaled to mean zero and standard deviation one (original data were not normalized). For liver data, no transformation was performed (original data were $\log_2$ ratios).

Model training. For every data set, SAMANN networks were constructed to map the original data to a 3-D virtual reality space. The activation functions used were sinusoidal for the first hidden layer and hyperbolic tangent for the rest. A collection of models was obtained by varying some of the network controlling parameters: number of units in the first hidden layer (two different values), weights ranges in the first hidden layer (three different values), learning rates (three different values), momentum (three different values), number of pairs presented to the network at every iteration (three different values), number of iterations (three different values) and random seeds (four different values), for a total of $1,944$ SAMANN networks for every data set.

Computing environment. All of the experiments were conducted on a Condor pool (`http://www.cs.wisc.edu/condor/`) located at the Institute for Information Technology, National Research Council Canada.

5 Results

For every data set, we constructed the histograms of the Sammon error for the obtained networks. All of the empirical distributions were positively skewed (with the mode on the lower error side), which is a good behavior. In addition, the general error ranges were small. In table 1 some statistics of the experiments are presented: minimum, maximum, mean and standard deviation for the best (i.e., with smallest Sammon error) $1,000$ networks.

Clearly, it is impossible to represent a virtual reality space on printed media (navigation, interaction, and world changes are all lost). Therefore, very simple geometries were used for objects and only snapshots of the virtual worlds are presented. Figures 1, 2 and 3 show the virtual reality spaces corresponding to the best networks for the lung, stomach and liver cancer data sets respectively. The mapping was generated from an unsupervised perspective (i.e., without using the class labels) and objects from different

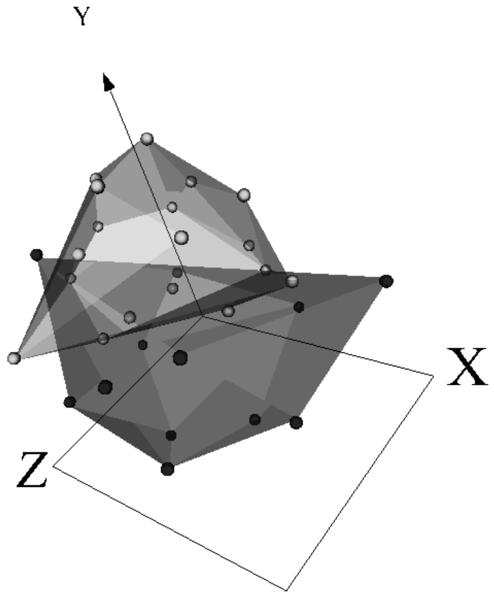

Fig. 1. VR space of the lung cancer data set (Sammon error = 0.079, best out of 1,944 experiments). The space was generated from an unsupervised perspective but the classes are displayed for comparison purposes. Dark spheres: severe emphysema, Light spheres: mild or no emphysema. The boundary between the classes in the VR space seem to be a low curvature surface.

classes were represented in the VR space differently for comparison purposes. Transparent membranes wrap the corresponding classes, so that the degree of class overlapping can be easily seen. In addition, it allows to look for particular samples with ambiguous diagnostic decisions.

The low values of the Sammon error indicate that the spaces preserved most of the distance structure of the data, therefore, giving a good idea about the distribution in the original spaces. The three virtual spaces are clearly polarized with two distribution modes, each one corresponding to a different class. Note, however, that classes are more clearly differentiated for the liver and stomach data sets than for the lung data set, where a certain level of overlapping exists. The reason for this may be that mild and no emphysema were considered members of the same class (see section 3).

The advantage of using SAMANN networks is that, since the mapping f between the original and the virtual space is *explicit*, a new sample can be easily transformed and visualized in the virtual space. Since the distance between any two objects is an indication of their dissimilarity, the new point is more likely to belong to the same class of its nearest neighbors. In the same way, outliers can be readily identified, although they may result from the space deformation inevitably introduced by the dimensionality reduction.

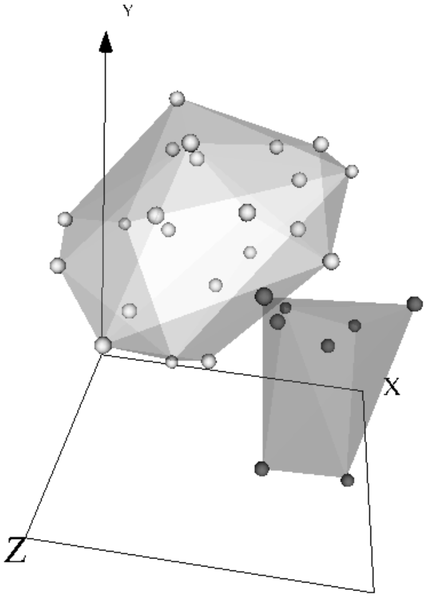

Fig. 2. VR space of the stomach cancer data set (Sammon error = 0.063, best out of $1,944$ experiments). The space was generated from an unsupervised perspective but the classes are displayed for comparison purposes. Dark spheres: normal, Light spheres: cancerous samples.

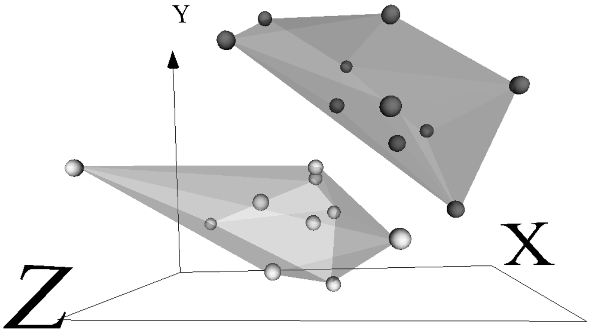

Fig. 3. VR space of the liver cancer data set (Sammon error = 0.040, best out of $1,944$ experiments). The space was generated from an unsupervised perspective but the classes are displayed for comparison purposes. Dark spheres: normal, Light spheres: cancerous samples.

6 Conclusions

High quality virtual reality spaces for visual data mining of typical examples of gene expression cancer data were obtained using unsupervised structure-preserving neural networks in a distributed computing data mining (grid) environment. These results show that a few nonlinear features can effectively capture the similarity structure of the data and also provide a good differentiation between the cancer and normal classes. However, in cases where the descriptor attributes are not directly related to class structure or where there are many noisy or irrelevant attributes the situation may not be as clear. In these cases, feature subset selection and other data mining procedures could be considered in a preprocessing stage.

Acknowledgments

This work was partially supported by the Consejo Interministerial de Ciencia y Tecnología (CICYT, Spain), under projects CGL2004-04702-C02-02 and TIN2006-08114. This research was conducted in the framework of the STATEMENT OF WORK between the National Research Council Canada (Institute for Information Technology, Integrated Reasoning Group) and the Soft Computing Group (Dept. of Languages and Information Systems), Polytechnic University of Catalonia, Spain.

References

1. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrecht (1991)
2. Valdés, J.J.: Virtual Reality Representation of Information Systems and Decision Rules: An Exploratory Tool for Understanding Data and Knowledge. In: International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (LNAI 2639), pp. 615–618 (2003)
3. Valdés, J.J.: Similarity-based Heterogeneous Neurons in the Context of General Observational Models. Neural Network World 12(5), 499–508 (2002)
4. Valdés, J.J.: Building Virtual Reality Spaces for Visual Data Mining with Hybrid Evolutionary-classical Optimization: Application to Microarray Gene Expression Data. IASTED International Joint Conference on Artificial Intelligence and Soft. Computing, pp. 161–166 (2004)
5. Borg, I., Lingoes, J.: Multidimensional Similarity Structure Analysis. Springer, Heidelberg (1987)
6. Sammon, J.W.: A Non-linear Mapping for Data Structure Analysis. IEEE Transactions on Computers C-18, 401–408 (1969)
7. Mao, J., Jain, A.K.: Artificial Neural Networks for Feature Extraction and Multivariate Data Projection. IEEE Transactions on Neural Networks 6, 296–317 (1995)
8. Spira, A., Beane, J., Pinto-Plata, V., Kadar, A., Liu, G., Shah, V., Celli, B., Brody, J.S.: Gene Expression Profiling of Human Lung Tissue from Smokers with Severe Emphysema. American Journal of Respiratory Cell. and Molecular Biology 31, 601–610 (2004)
9. Hippo, Y., Taniguchi, H., Tsutsumi, S., Machida, N., Chong, J.M., Fukayama, M., Kodama, T., Aburatani, H.: Global Gene Expression Analysis of Gastric Cancer by Oligonucleotide Microarrays. Cancer Research 62(1), 233–240 (2002)
10. Lam, S.H., Wu, Y.L., Vega, V.B., Miller, L.D., Spitsbergen, J., Tong, Y., Zhan, H., Govindarajan, K.R., Lee, S., Mathavan, S., Murthy, K.R.K., Buhler, D.R., Liu, E.T., Gong, Z.: Conservation of Gene Expression Signatures between Zebrafish and Human Tumors and Tumor Progression. Nature Biotechnology 24(1), 73–75 (2006)