Virtual Reality Visual Data Mining with Nonlinear Discriminant Neural Networks: Application to Leukemia and Alzheimer Gene Expression Data

Valdés, Julio; Barton, Alan

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Virtual Reality Visual Data Mining with Nonlinear Discriminant Neural Networks: Application to Leukemia and Alzheimer Gene Expression Data *

Valdés, J., and Barton, A.
August 2005

* published in the Proceedings of the International Joint Conference on Neural Networks (IJCNN'05). Montreal, Québec, Canada. August 1- 4, 2005. NRC 48124.

Canada

Virtual Reality Visual Data Mining with Nonlinear Discriminant Neural Networks: Application to Leukemia and Alzheimer Gene Expression Data.

Julio J. Valdés and Alan J. Barton
National Research Council
Institute for Information Technology
M50, 1200 Montreal Rd. Ottawa, ON K1A 0R6
Canada
E-mail: julio.valdes@nrc-cnrc.gc.ca
alan.barton@nrc-cnrc.gc.ca

***Abstract*— A hybrid stochastic-deterministic approach for solving NDA problems on very high dimensional biological data is investigated. It is based on networks trained with a combination of simulated annealing and conjugate gradient within a broad scale, high throughput computing data mining environment. High quality networks from the point of view of both discrimination and generalization capabilities are discovered. The NDA mappings generated by these networks, together with unsupervised representations of the data, lead to a deeper understanding of complex high dimensional data like Leukemia and Alzheimer gene expression microarray experiments.**

I. Introduction

Humans are capable of perceiving vast quantities of sensor information at very high input rates through vision systems in the brain that outperform current computer capabilities. Including such human processing power in the knowledge discovery process is one way to help uncover the important concepts from large amounts of (possibly space or time dependant) information contained in *i)* real world data sets *ii)* results obtained from computer procedures.

Several reasons make Virtual Reality (VR) a suitable paradigm: is *flexible*: it allows the choice of different representation models to better accommodate different human perception preferences; allows *immersion*: the user can navigate inside the data, interact with the objects in the world, change scales, perspectives, etc.; creates a *living* experience: the user is not merely a passive observer or an outsider, but an actor in the world; is *broad and deep*: the user may see the VR world as a whole, and/or concentrate the focus of attention on specific details or portions of the world; is *user friendly*: no specialized knowledge is required.

Constructing VR spaces for visual data mining of databases and symbolic knowledge requires the solution of a multivariate data projection problem, which in turn can be performed according to different criteria (unsupervised, or supervised). Neural networks are natural choices for feature extraction and multivariate data projection [6], [7], [5]. VR spaces constructed using an underlying unsupervised paradigm have proven to be successful tools for understanding both data and knowledge structures, from a visual data mining perspective [10], [12]. In particular, very good results have been obtained with this technique (unsupervised mode) in both the analysis of gene expression data, and in the evaluation of the results obtained by other data mining algorithms [14], [13].

The purpose of this paper is: *i)* to explore the construction of VR spaces for *visual* data mining from a supervised perspective, by using nonlinear discriminant analysis neural networks (NDA), in order to generate a *few new* nonlinear features (e.g. 3), with good class membership approximation capability; therefore, conventional classifiers would not be appropriate. These new features will be used as a base for an Euclidean space based virtual world suitable for data visualization, *ii)* to introduce a variant of NDA with respect to the classical approach. Instead of using feedforward networks trained with backpropagation, hybrid stochastic-deterministic networks are used, which works with a combination of simulated annealing and conjugate gradient. These kinds of networks are more robust, and less prone to local extrema entrapment. As part of a broader data mining objective, thousands of NDA's are constructed and evaluated in a High Throughput Computing environment, and *iii)* to study the behavior of these networks in processing gene expression data. This kind of data has a great importance in bioinformatics and medicine, and presents great challenges because of its complexity due to its high dimensionality (in the order of thousands).

II. Virtual Reality Representation Of Relational Structures

A virtual reality, visual, data mining technique extending the concept of 3D modelling to relational structures was introduced [10], [12], (see also `http://www.hybridstrategies.com`). It is oriented to the understanding of large heterogeneous, incomplete and imprecise data, as well as symbolic knowledge. The notion of data is not restricted to databases, but includes logical relations and other forms of both structured and non-structured knowledge. In this approach, the data objects are considered as tuples from a heterogeneous space [11]. An example of a heterogeneous database is shown in Fig.1.

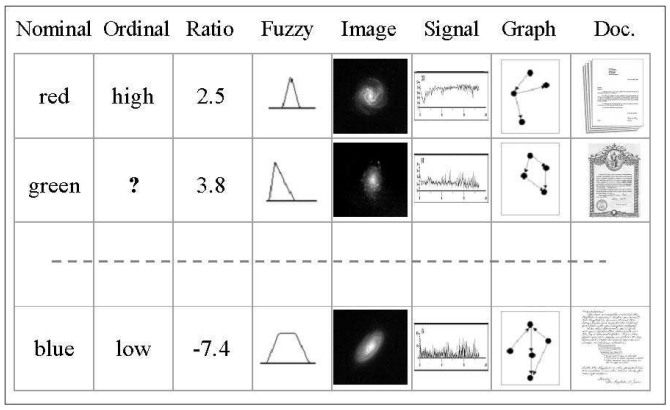

Fig. 1. An example of a heterogeneous database. Nominal, ordinal, ratio, fuzzy, image, signal, graph, and document data are mixed. The symbol ? denotes a missing value.

Different information sources are associated with the attributes, relations and functions, and these sources are associated with the nature of what is observed (e.g. point measurements, signals, documents, images, etc). They are described by mathematical sets (of the appropriate kind) called source sets (Ψ_i), constructed according to the nature of the information source to represent (e.g. point measurements of continuous variables by subsets of the reals in the appropriate ranges, structural information by directed graphs, etc). Source sets also account for incomplete information. A heterogeneous domain is a Cartesian product of a collection of source sets: $\hat{\mathcal{H}}^n = \Psi_1 \times \cdots \times \Psi_n$, where $n > 0$ is the number of information sources to consider. For example, in a domain where objects are described by attributes like continuous crisp quantities, discrete features, fuzzy features, time-series, images, and graphs (missing values are allowed). They can be represented as Cartesian products of subsets of real numbers($\hat{R}$), nominal ($\hat{N}$) or ordinal sets($\hat{O}$), fuzzy sets($\hat{F}$), sets of images ($\hat{I}$) , sets of time series ($\hat{S}$) and sets of graphs ($\hat{G}$), respectively (all extended for allow missing values). The heterogeneous domain is $\hat{\mathcal{H}}^n = \hat{N}^{n_N} \times \hat{O}^{n_O} \times \hat{R}^{n_R} \times \hat{F}^{n_F} \times \hat{I}^{n_I} \times \hat{S}^{n_S} \times \hat{G}^{n_G}$, where n_N is the number of nominal sets, n_O of ordinal sets, n_R of real-valued sets , n_F of fuzzy sets , n_I of image-valued sets, n_S of time-series sets, and n_G of graph-valued sets, respectively ($n = n_N + n_O + n_R + n_F + n_I + n_S + n_G$).

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 ($\underline{O} =< O, \Gamma^v >$, the 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. $\Re^m \subset \mathbf{R}^m$ is a *metric space* of dimension m (euclidean or not) which will be the actual virtual reality geometric space. The other elements are mappings: $g_o : O \rightarrow G$, $l : O \rightarrow \Re^m$, $g_r : \Gamma^v \rightarrow G$, $b : O \rightarrow B$.

Of particular importance is the mapping l. If the objects are in a heterogeneous space, $l : \hat{\mathcal{H}}^n \rightarrow \Re^m$. Several desiderata can be considered for building a VR-space. One may be to preserve one or more properties from the original space as much as possible (for example, the similarity structure of the data [2]). From an unsupervised perspective, the role of l could be to maximize some metric/non-metric structure preservation criteria [1], or minimize some measure of information loss. From a supervised point of view l could be chosen as to emphasize some measure of class separability over the objects in O [12]. Hybrid requirements are also possible.

III. Nonlinear Discriminant Neural Networks

In the supervised case, a natural choice for representing the l mapping is an NDA neural network [15], [6], [7], [5]. One strong reason is the nature of the class relationships in complex, high dimensional problems like gene expression data, where objects are described in terms of several thousands of genes, and classes are often either only separable with nonlinear boundaries, or not separable at all. Another is the generalization capabilities of neural networks which will allow the classification of new incoming objects, and their immediate placement within the created VR spaces. Of no less importance is that when learning the mapping, the neural network hidden layers create new nonlinear features for the mapped objects, such that they are separated into classes by the output layer. However, these nonlinear features could be used independently with other data mining algorithms. The typical architecture of such networks is shown in Fig.2

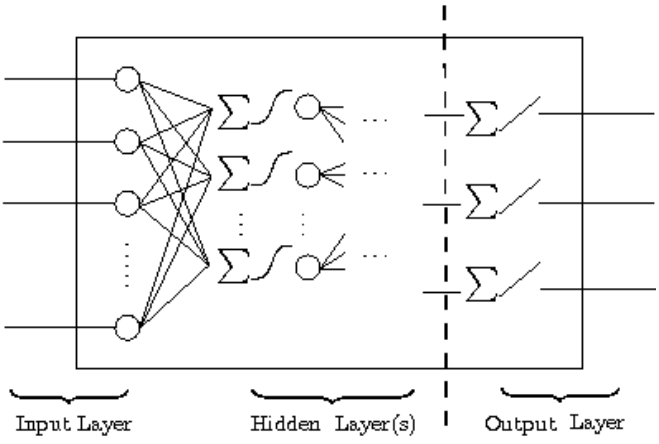

Fig. 2. Network Architecture in which the NDA network is learned. $\int$ means nonlinear activation, / linear activation , and Σ aggregation

This is a feedforward network with one or more hidden layers where the number of input nodes is set to the number of features of the data objects, and the number of neurons in the output layer to be the number of pattern classes. The number of neurons in the last hidden layer to m; the dimensionality of the projected space (for a VR space this is typically 3). From input layer to the last hidden layer, the network implements a nonlinear projection from the original n-dimensional space to an m-dimensional space. If the entire network can correctly classify a linearly-nonseparable data set, this projection actually converts the linearly-nonseparable data to separable data. The backpropagation learning algorithm is

used to train the feedforward network with two hidden layers in a collection of epochs, such that in each, all the patterns in the training data set are seen *once*, in a random order.

This classical approach to building NDA networks suffers from the well known problem of local extrema entrapment. In this paper a variant in the construction of NDA networks is introduced by using hybrid stochastic-deterministic feed forward networks (SD-FFNN). The SD-FFNN is a hybrid model where training is based on a combination of simulated annealing with the powerful minima seeking conjugate gradient [8], which improves the likelihood of finding good extrema while containing enough determinism. The global search capabilities of simulated annealing and the improved local search properties of the conjugate gradient reduces the risk of entrapment, and the chances of finding a set of neuron weights with better properties than what is found by the inherent steepest descent implied by pure backpropagation.

In the SD-FFNN network, simulated annealing (SA) is used in two separate, independent ways. First it is used for initializing (at high temperature with the weights centered at zero), in order to find a good initial approximation for the conjugate gradient (CG). Once it has reached a local minimum, SA is used again, this time at lower temperature, in order to try to evade what might be a local minimum, but this time with the weights centered at the values found by CG.

IV. Application to Leukemia and Alzheimer Gene Expression Data

Gene expression is the process by which a gene's coded information is translated into the structures present and operating in the cell (either proteins or RNAs). Current technologies measures the level of gene expression of tissue samples for a particular set of targeted genes. In this study, the following datasets from research into the corresponding diseases were used:

- Leukemia gene expression data for 7129 genes.
- Leukemia gene expression data for 7 selected genes.
- Alzheimer gene expression data for 9600 genes.
- Alzheimer gene expression data for 4 selected genes.

A. Experimental settings

For simplicity, and due to the restrictions imposed by printed materials, in all of the VR spaces constructed no behavior was associated to the objects; the only relation included was class membership (expressed as grey level). Geometries were spheres, and only in some cases, cubes; with the dimension of the space fixed at 3. For each of the data sets, NDA networks with one input layer, two hidden, and an output layer were used in all of the experiments. In contradistinction with the classical NDA training as a classification network, the training here was oriented to learn a remapped characteristic function of the classes associated with the datasets, where membership was set to 0.9 and non-membership to -0.9 in order to maximize performance w.r.t the hyperbolic tangent bahavior. The Mean Squared Error between the network outputs and the modified characteristic function of the classes was the error measure used. A total of 1600 NDA networks were computed for each of the four datasets processed, and the computations were performed in a Condor pool (`http://www.cs.wisc.edu/condor/`). The experimental settings used for the NDA networks are shown in Table.I.

TABLE I
Experimental Settings used for the NDA networks

No. Neurons in Input Layer	$[7129, 7, 9600, 4]$
No. Neurons in First Hidden Layer	$[1, 2, 3, ..., 10]$
No. Neurons in Second Hidden Layer	3
No. Neurons in Output Layer	$[2]$
Aggregation Function	Scalar Product
Activation Function	Hyperbolic Tangent
Seed 1	$[1, 301, 601, 901]$
Seed 2	$[3, 303, 603, ..., 2703]$
Allowable MSE	$[0.004, 0.003, 0.002, 0.001]$
Maximum No. of Annealing Trials	15

For comparison purposes, unsupervised VR spaces using Sammon's original algorithm for computing the l mapping, but using a dissimilarity in the space of the original attributes (genes) given by $\delta_{ij} = (1 - \hat{s}_{ij})/\hat{s}_{ij}$, where $\hat{s}_{ij}$ is Gower's similarity coefficient [4] was used, with Euclidean distance set as the measure used as dissimilarity in the VR space.

B. Leukemia data

Cancer can potentially kill a human through disabling the normal function of tissues and/or organs. One such cancer is Leukemia, which originates in the bone marrow of humans. The cause of leukemia is not known.

For the study, 72 patients from [3] were used. They are separated into two groups, *i)* a training set containing 38 bone marrow samples: 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML), obtained from patients at the time of diagnosis, and *ii)* a testing set containing 34 samples (24 bone marrow and 10 peripheral blood samples), where 20 are ALL and 14 AML.

In this paper no explicit preprocessing of the data was performed, in order to not introduce bias and to be able to expose the behavior of the data processing strategy, the methods used, and their robustness. That is, no background subtraction, deletions, filtering, or averaging of samples/genes were applied.

In [13], a methodology was proposed for gene discovery from many noisy and potentially unrelated genes. It consists of two configurable learning stages. In Stage-I, a partition clustering algorithm is configured to either *i)* select a gene to represent a set of closely related genes (in terms of expression proximity), or *ii)* construct a synthetic gene by aggregating the properties of a set of genes. The representatives are then Stage-II processed in order to find the most discernibility preserving genes (i.e. the set of genes contained in the union of all discovered reducts). The learned knowledge may then be used for discretizing and classifying future leukemia samples.

Two sets of experiments were performed; 1600 for the original 7129-dimensional space and 1600 on the reduced (using the aforementioned methodology) 7-dimensional space, yielding 3200 leukemia experiments executed in a distributed computing environment using Condor.

The performance of each network on training and test sets is in Fig.3 for the 7129 case and Fig.4 for the 4 gene case.

Networks with a balanced training/test MSE and low test MSE are located around $(0.8, 0.9)$ in Fig.3, and around $(10^{-7}, 10^{-7})$ in Fig.4.

Table.II contains sample mean and ranges for the MSE for each of the experiments for both training and test set. The effect of reducing the number of genes per sample can be readily seen. The mean MSE on the test set has been reduced from 0.8762 down to 0.3711, a factor of over 2.3, the maximum MSE has been reduced by a factor of 2, and the minimum MSE has been reduced by a factor of over $38,000$.

An unsupervised 3D projection using Sammon's algorithm [9] (Fig.5) illustrates the complexity of this data. The ALL and AML classes appear completely interleaved for both the training and test sets. An NDA network result with balanced training/test as well as low test MSE, is shown in Fig.6. Only in a small region of the space do the two classes overlap (6 ALL and 2 AML samples), whereas the rest of the space contains well differentiated samples. Clearly, the NDA result substantially improves the unsupervised projection. When samples are described only in terms of the 7 selected genes, a Sammon projection Fig.7 shows a class differentiated structure. The NDA counterpart sharing the same training/test and test MSE properties (Fig.8), exhibits an even clearer differentiation.

TABLE II

MEAN AND RANGES OF MSE FOR LEUKEMIA DATA

	Training Set	Test Set
Exp-1	$\overline{x} = 0.00691$	$\overline{x} = 0.8762$
(all genes)	$[0.00000..0.08222]$	$[0.071471..1.84182]$
Exp-2	$\overline{x} = 0.00012$	$\overline{x} = 0.37114$
(7 genes)	$[1.67705\ 10^{-16}..0.00398]$	$[1.8445\ 10^{-07}..0.96931]$

C. Alzheimer data

Alzheimer's disease (AD) is an incurable, chronic, progressive, debilitating condition which, along with other neurodegenerative diseases, represents the largest area of unmet need in modern medicine [14]. In that study, a total of 4 clinically diagnosed AD patients and 5 normal patients of similar age were investigated. A total of 23 samples were taken from them, each characterized by 9600 genes. Despite such a high dimensionality in the original space, an unsupervised VR representation with low Sammon error (Fig.9), successfully portrays a structure in which Alzheimer's samples are clustered. They are wrapped by the class of normal samples, which appears more irregular. In the supervised case, due to

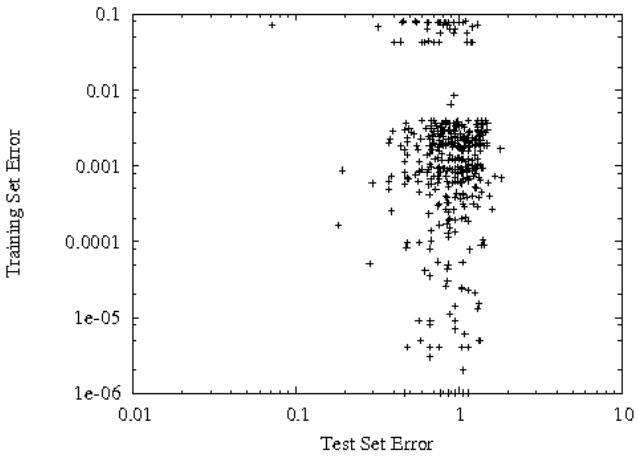

Fig. 3. Leukemia: Mean squared errors for 1600 runs, each with 38 train and 34 test samples, respectively. Samples described in terms of 7129 genes.

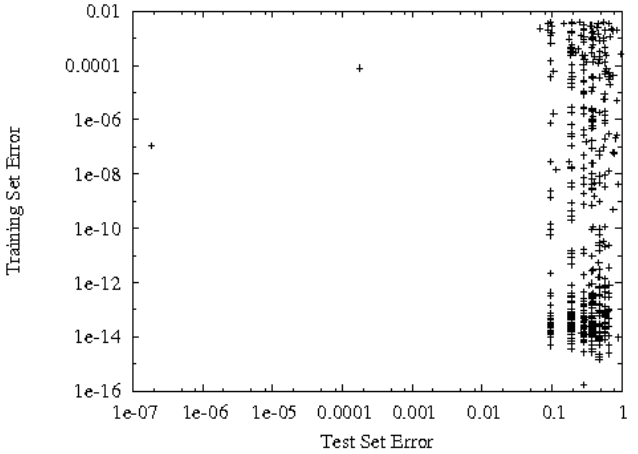

Fig. 4. Leukemia: Mean squared errors for 1600 runs, each with 38 train and 34 test samples, respectively. Samples described in terms of 7 selected genes.

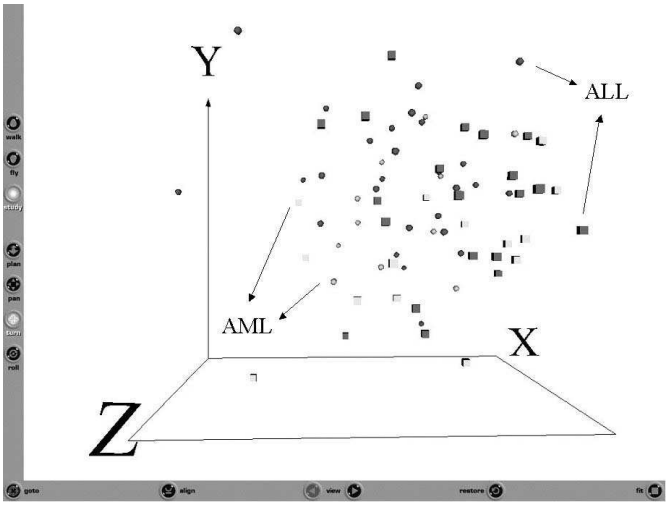

Fig. 5. Leukemia: Unsupervised (Sammon) representation of the original training and test data, in terms of 7129 genes. Dark objects= ALL, Light objects=AML. Spheres = training, Cubes = test. Sammon error = 0.143.

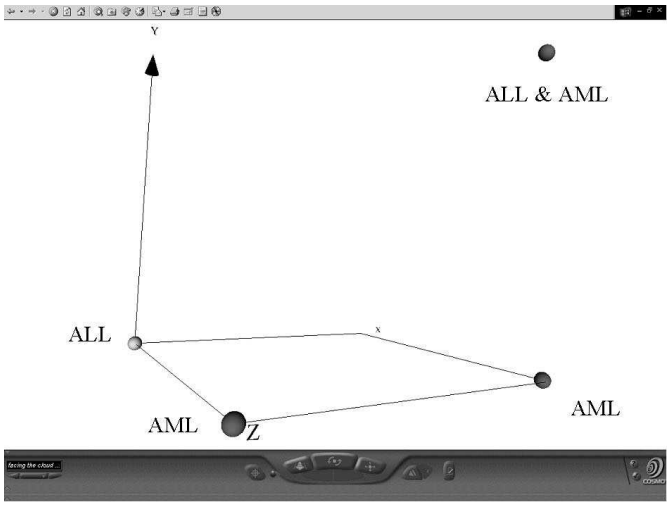

Fig. 6. Leukemia: Supervised (NDA) representation of the original training and test data, in terms of 7129 genes. Dark objects= ALL, Light objects=AML. Training error = 0.0710, test error = 0.0715.

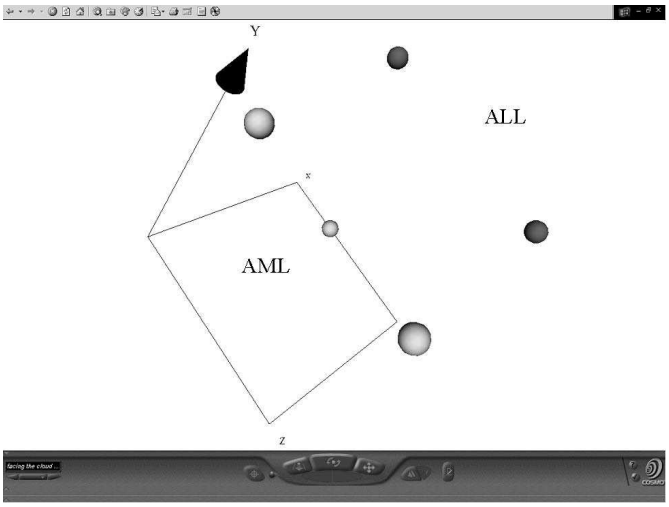

Fig. 8. Leukemia: Supervised (NDA) representation of the original training and test data, in terms of 7 selected genes. Dark objects= ALL, Light objects=AML. Training error = $1.1236\ 10^{-07}$, test error = $1.8445\ 10^{-07}$.

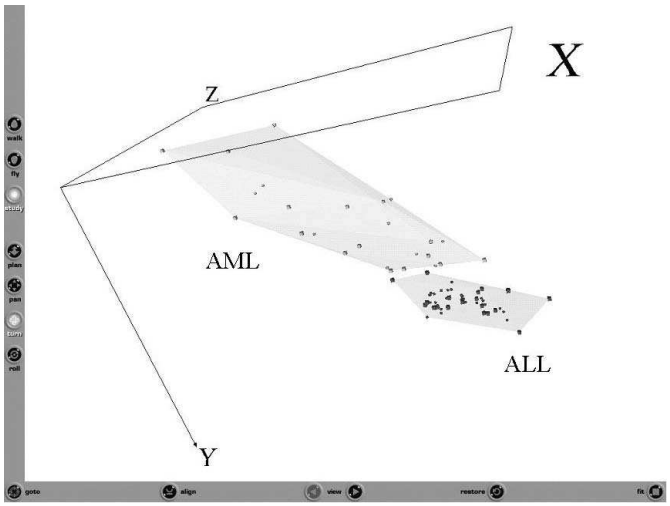

Fig. 7. Leukemia: Unsupervised (Sammon) representation of the original training and test data, in terms of 7 selected genes. Convex hulls wrap the classes. Dark objects= ALL, Light objects=AML. Spheres = training, Cubes = test. Sammon error = 0.103.

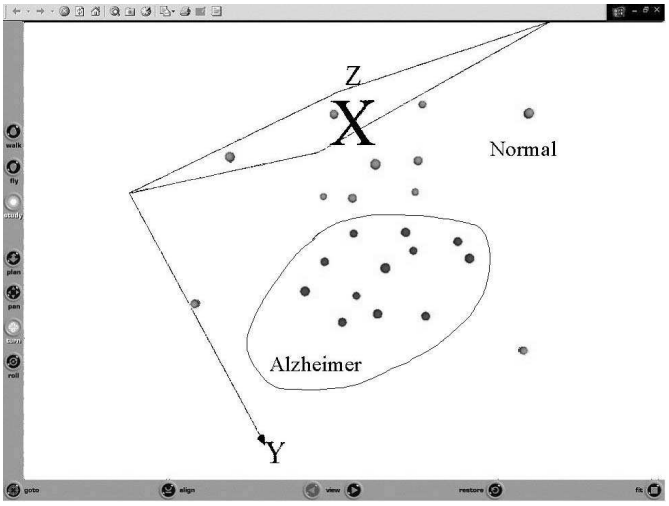

Fig. 9. Alzheimer: Unsupervised (Sammon) representation of the original training and test data, in terms of 9600 genes. Dark objects= ALL, Light objects=AML. A boundary delimiting the Alzheimer class was added for clarity. Sammon error = 0.103.

the small number of samples, the whole dataset was used when computing the NDA projections.

For the sample described in terms of 9600 genes, even the output of the NDA network with the worst MSE ($2.4581\ 10^{-1}$) produced a total class differentiation Fig.10.

The data mining procedures applied in [14] reported a subset of 20 most relevant genes. From them, a subset of 4 were found to individually differentiate the classes with zero error. An unsupervised VR space constructed using Sammon's algorithm is shown in Fig.11 where the two classes are wrapped with their corresponding convex hulls. The quality of the representation is evidenced by both the value of the Sammon error (0.002), and the clear separation of the two classes. The effect of incorporating the class information into the analysis is shown in Fig.12, where the results of applying the worst NDA network are shown. Again, there is a total class differentiation.

V. Conclusions

The hybrid stochastic-deterministic approach used for solving NDA problems proves to be very effective at differentiating the classes existing in the very high dimensional biological data investigated. Broad scale, high throughput computing environments for data mining, enable the discovery of high quality NDA networks from the point of view of both discrimination and generalization capabilities. The joint use of NDA and unsupervised mappings provides more insight towards comprehensive interpretation and understanding of gene expression data. These are important features of these kinds of networks when applied to bioinformatic problems.

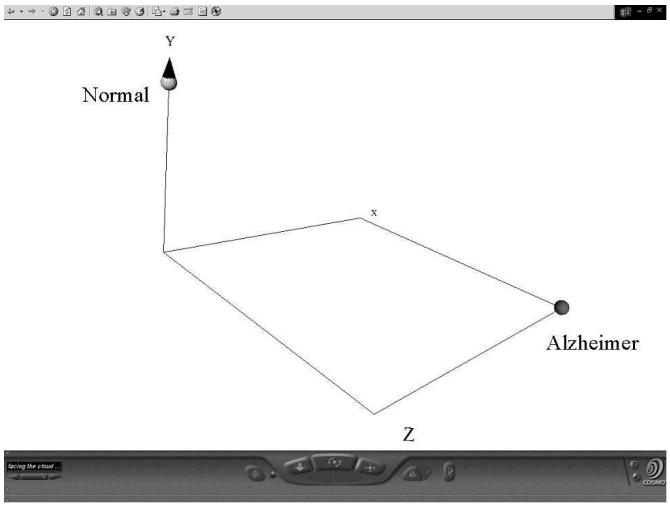

Fig. 10. Alzheimer: Supervised (NDA) representation of the original data, in terms of 9600 genes. Dark objects = Alzheimer, Light objects = Normal. Training error = $2.4581\ 10^{-1}$ (the worst of 1600 networks).

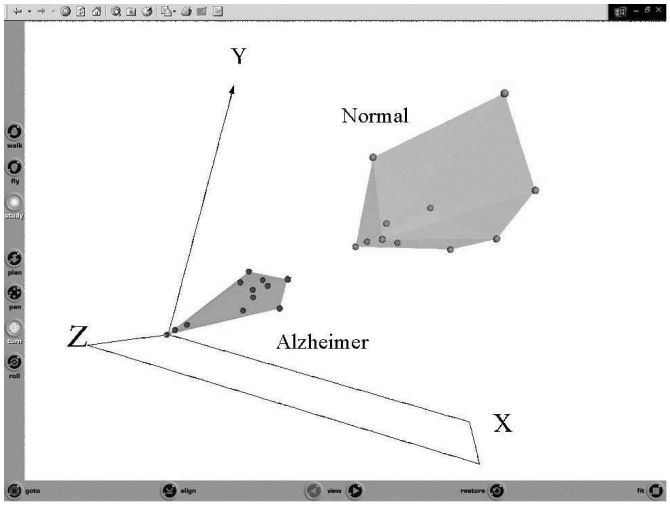

Fig. 11. Alzheimer: Unsupervised (Sammon) representation of the original training and test data, in terms of 4 selected genes. Convex hulls wrap the classes. Dark objects = Alzheimer, Light objects = Normal. Sammon error = 0.002.

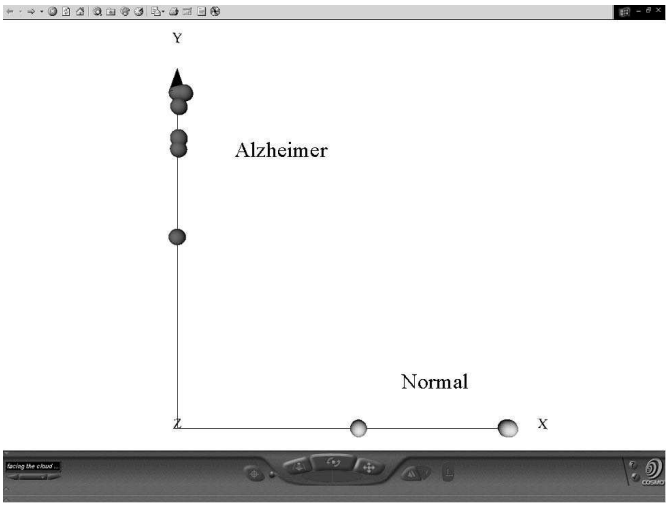

Fig. 12. Alzheimer: Supervised (NDA) representation of the original data, in terms of 4 selected genes. Dark objects = Alzheimer, Light objects = Normal. Training error = $3.7416\ 10^{-3}$ (the worst of 1600 networks).

These results are encouraging and further experimentation with other data sets are required.

ACKNOWLEDGMENT

The authors would like to thank Ratilal Haria, and Roger Impey (National Research Council Canada, NRC) for their assistance in using distributed computing facilities, as well as to Fazel Famili and Robert Orchard from the Integrated Reasoning Group (IIT-NRC). This research was conducted within the scope of the BioMine project (IIT-NRC).

REFERENCES

[1] I. Borg and J. Lingoes, *Multidimensional similarity structure analysis*, Springer-Verlag, 1987.

[2] J. L. Chandon and S. Pinson, *Analyse typologique. Théorie et applications* Masson, Paris, 1981.

[3] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science* vol. 286, pp. 531–537, 1999.

[4] J.C Gower, "A General Coefficient of Similarity and Some of its Properties". *Biometrics* vol. 1, No. 27 pp. 857–871, 1973.

[5] A. K. Jain and J. Mao , "Artificial Neural Networks for Nonlinear Projection of Multivariate Data," *Proceedings of the 1992 IEEE joint Conf. on Neural Networks*, Baltimore, MD, June. 1992, pp. 335–340.

[6] J. Mao and A. K. Jain, "Discriminant Analysis Neural Networks," *Proceedings of the 1993 IEEE International Conference on Neural Networks*, San Francisco, California, Mar. 1993, pp. 300–305.

[7] J. Mao and A. K. Jain, "Artificial Neural Networks for Feature Extraction and Multivariate Data Projection," *IEEE Trans. on Neural Networks* vol. 6, pp. 296–317, Mar. 1995.

[8] T. Masters, *Advanced Algorithms for Neural Networks*, John Wiley & Sons, 1993.

[9] J. W. Sammon, "A non-linear mapping for data structure analysis," *IEEE Trans. on Computers*, 1969, vol. C18, pp. 401–409.

[10] J. J. Valdés, "Virtual Reality Representation of Relational Systems and Decision Rules: An exploratory Tool for understanding Data Structure," *In Theory and Application of Relational Structures as Knowledge Instruments*, Meeting of the COST Action 274 (P. Hajek. Ed), Prague, Nov. 2002.

[11] J. J. Valdés, "Similarity-Based Heterogeneous Neurons in the Context of General Observational Models," *Neural Network World* vol. 12, no. 5, pp. 499–508, 2002.

[12] J. J. Valdés, "Virtual Reality Representation of Information Systems and Decision Rules: An Exploratory Tool for Understanding Data and Knowledge," *Lecture Notes in Artificial Intelligence LNAI 2639*, Springer-Verlag, 2003, pp. 615–618.

[13] J. J. Valdés and A. J. Barton. "Gene Discovery in Leukemia Revisited: A Computational Intelligence Perspective", *Proceedings of the 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, *Lecture Notes in Artificial Intelligence 3029*, Springer Verlag, 2004, pp. 118–127.

[14] P. R. Walker, B. Smith, Q. Y. Liu, A. Famili, J. J. Valdés, Z. Liu, and B. Lach, "Data Mining of Gene Expression Changes in Alzheimer Brain", Artificial Intelligence in Medicine, *Elsevier Science*. Jun. 2003.

[15] A. R. Webb and D. Lowe, "The Optimized Internal representation of a Multilayer Classifier", *Neural Networks* vol. 3, pp. 367-375, 1990.