Multi-objective Evolutionary Optimization for Visual Data Mining with Virtual Reality Spaces: Application to Alzheimer Gene Expressions. *

Julio J. Valdés
National Research Council Canada
Institute for Information Technology
M50, 1200 Montreal Rd.
Ottawa, ON K1A 0R6
julio.valdes@nrc-cnrc.gc.ca

Alan J. Barton
National Research Council Canada
Institute for Information Technology
M50, 1200 Montreal Rd.
Ottawa, ON K1A 0R6
alan.barton@nrc-cnrc.gc.ca

ABSTRACT

This paper introduces a multi-objective optimization approach to the problem of computing virtual reality spaces for the visual representation of relational structures (e.g. databases), symbolic knowledge and others, in the context of visual data mining and knowledge discovery. Procedures based on evolutionary computation are discussed. In particular, the NSGA-II algorithm is used as a framework for an instance of this methodology; simultaneously minimizing Sammon's error for dissimilarity measures, and mean cross-validation error on a k-nn pattern classifier. The proposed approach is illustrated with an example from genomics (in particular, Alzheimer's disease) by constructing virtual reality spaces resulting from multi-objective optimization. Selected solutions along the Pareto front approximation are used as nonlinearly transformed features for new spaces that compromise similarity structure preservation (from an unsupervised perspective) and class separability (from a supervised pattern recognition perspective), simultaneously. The possibility of spanning a range of solutions between these two important goals, is a benefit for the knowledge discovery and data understanding process. The quality of the set of discovered solutions is superior to the ones obtained separately, from the point of view of visual data mining.

Categories and Subject Descriptors

I.2.m [**Artificial Intelligence**]: Miscellaneous; I.5.m [**Pattern Recognition**]: Miscellaneous; J.3 [**Computer Applications**]: LIFE AND MEDICAL SCIENCES; H.5.m [**Information Systems**]: Miscellaneous

General Terms

Algorithms, Experimentation

*This research was conducted within the scope of the BioMine Project in the Integrated Reasoning Group (National Research Council Canada, Institute for Information Technology).

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.
ACM 1-59593-186-4/06/0007.

Keywords

visual data mining, virtual reality spaces, multi-objective optimization, genetic algorithms, NSGA-II algorithm, k-nn classification, cross-validation error, similarity structure preservation, non-linear mapping, Sammon error, Alzheimer disease, genomics

1. INTRODUCTION

Knowledge discovery is the non-trivial process of identifying valid, novel, potentially useful, and ultimately *understandable patterns* in data [9], and the role of visualization techniques in the knowledge discovery process is well known. Data and patterns are concepts which should be considered in a broad sense. There are different kinds of data (relational, graphical, symbolic, etc.), and patterns of different kinds (geometrical, logical, etc.). The increasing rates of data generation require the development of procedures facilitating the *understanding* of the internal structure of data more rapidly and intuitively. Moreover, the increasing complexity of the data analysis procedures makes it more difficult for the user (not necessarily a mathematician or data mining expert), to extract useful information out of the results generated by the various techniques. This makes graphical representation directly appealing.

Several reasons make Virtual Reality (VR) a suitable paradigm: Virtual Reality is *flexible*, it allows the construction of different virtual worlds representing *the same* underlying information, but with a different look and feel. VR allows *immersion*, that is, the user can navigate inside the data, interact with the objects in the world. VR creates a *living* experience. The user is not merely a passive observer but an actor in the world. VR is *broad and deep*. The user may see the VR world as a whole, and/or concentrate the focus of attention on specific details of the world. Of no less importance is the fact that in order to interact with a Virtual World, no mathematical knowledge is required, and the user only needs minimal computer skills. A virtual reality technique for visual data mining on heterogeneous, imprecise and incomplete information systems was introduced in [23, 24].

These VR spaces are obtained by transforming the original set of attributes describing the objects, often defining a heterogeneous high dimensional space, into another space of small dimension (typically 2-4) and intuitive metric (e.g. Euclidean). The operation almost always involves a non-linear transformation of the set of original attributes; implying some information loss. There are basically two kinds of spaces sought: *i)* spaces preserving the structure of the objects as determined by the original set of attributes (one objective measure to minimize in order to achieve this goal could be similarity information loss), and *ii*) spaces preserving the

distribution of an existing class or decision attribute defined over the set of objects (one objective measure to minimize in order to achieve this goal could be classification error). The complexity of the data, the unknown adequacy of the set of descriptor attributes, their relevance, noise, and many other factors imply that they do not necessarily relate with sufficient accuracy to the class or decision attribute. Therefore, both kinds of spaces are usually conflicting. They are also different from the point of view of the nonlinear transformations defining them. This situation creates problems during visualization, and confuses the human interpreter, because the same set of objects has a different distribution over the two spaces. Clearly, it would be much better to construct spaces where both criteria could be simultaneously partially or fully satisfied which leads to a multiobjective problem formulation.

Evolutionary multiobjective optimization (EMO) provides an alternative to classical multiobjective optimization techniques due to its population-based nature, which allows the creation of a set of non-dominated solutions in a single run. Moreover, the presence of noise in the data, the presence of large search spaces and other factors make EMO an interesting approach which has proven to be effective in other real world domains. We propose to introduce an EMO approach in visual data mining using virtual reality.

The purpose of this paper is to explore the construction of high quality VR spaces for visual data mining using a multi-objective optimization technique; in particular, optimization based on genetic algorithms. This approach provides both a solution for the previously discussed problem, and the possibility of obtaining a set of spaces in which the different objectives are expressed in different degrees, with the proviso that no other spaces could improve any of the considered criteria individually (if spaces are constructed using the solutions along the Pareto front). This strategy clearly represents a conceptual improvement in comparison with spaces computed from the solutions obtained by single-objective optimization algorithms in which the objective function is a weighted composition involving different criteria.

This approach is applied to a real world problem: namely, the representation of a very high dimensional dataset from the domain of genomics, consisting of microarray gene expression data from samples of patients with and without Alzheimer's disease.

2. VIRTUAL REALITY REPRESENTATION OF RELATIONAL STRUCTURES

A virtual reality, visual, data mining technique extending the concept of 3D modelling to relational structures was introduced [23], [24], (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 [22].

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

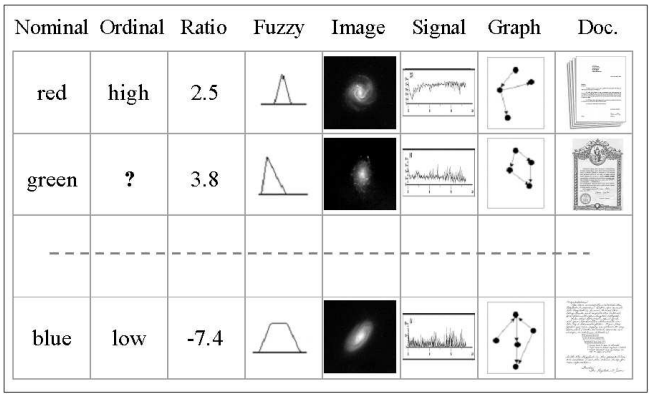


Figure 1: An example of a heterogeneous database. Nominal, ordinal, ratio, fuzzy, image, signal, graph, and document data are mixed. The symbol ? denotes a missing value.

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 to 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 ($\underline{O} =< O, \Gamma^v >$, 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: $g_o : O \to G, l : O \to \Re^m, g_r : \Gamma^v \to G, b : O \to B$.

Of particular importance is the mapping l. If the objects are in a heterogeneous space, $l : \hat{\mathcal{H}}^n \to \Re^m$. 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 [4]). From an unsupervised perspective, the role of l could be to maximize some metric/non-metric structure preservation criteria [2], 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 [24].

2.1 Structure preservation: An unsupervised perspective

As mentioned, l plays an important role in giving semantics to the virtual world, and there are many ways in which such a mapping can be defined. To a great extent it depends on which features from the original information system need to be highlighted. In particular, internal structure is one of the most important ones to consider and this is the case when the location and adjacency relationships between the objects O in Υ should give an indication of the *similarity relationships* [4] between the objects U in the original het-

erogeneous space $\hat{\mathcal{H}}^n$, as given by the set of attributes [22]. Other interpretations about internal structure are related with the properties of the space w.r.t. the linear/non-linear separability of class membership relations [13]. On the other hand, l can be constructed to maximize some metric/non-metric structure preservation criteria as has been done for decades in multidimensional scaling [14], [2], or minimize some error measure of information loss [20]. For example, if δ_{ij} is a dissimilarity measure between any two $i, j \in U$ ($i, j \in [1, N]$, where N is the number of objects), and $\zeta_{i^v j^v}$ is another dissimilarity measure defined on objects $i^v, j^v \in O$ from Υ ($i^v = \xi(i), j^v = \xi(j)$, they are in one-to-one correspondence). Examples of error measures frequently used are:

$$\text{S stress} = \sqrt{\frac{\sum_{i<j} (\delta_{ij}^2 - \zeta_{ij}^2)^2}{\sum_{i<j} \delta_{ij}^4}}, \quad (1)$$

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

$$\text{Quadratic Loss} = \sum_{i<j} (\delta_{ij} - \zeta_{ij})^2 \quad (3)$$

Typically, classical algorithms have been used for directly optimizing these measures, like Steepest descent, Conjugate gradient, Fletcher-Reeves, Powell, Levenberg-Marquardt, and others. The l mappings obtained using approaches of this kind are only *implicit*, as no functional representations are found. Moreover, their usefulness is restricted to the final errors obtained in the optimization process. However, explicit mappings can be obtained from these solutions using neural network or genetic programming techniques. An explicit l is useful for both practical and theoretical reasons. On one hand, in dynamic data sets (e.g. systems being monitored or data bases formed incrementally from continuous processes) an explicit direct transform l will speed up the incremental update of the virtual reality information system. On another hand, it can give semantics to the attributes of the virtual reality space, thus acting as a dimensionality reducer/new attributes constructor.
The possibilities derived from this approach are practically unlimited, since the number of different similarity, dissimilarity and distance functions definable for the different kinds of source sets is immense. Moreover, similarities and distances can be transformed into dissimilarities according to a wide variety of schemes, thus providing a rich framework where one can find appropriate measures able to detect interrelationships hidden in the data, better suited to both its internal structure and external criteria. In particular, for heterogeneous data involving mixtures of nominal and ratio variables, the Gower similarity measure [11] has proven to be suitable.

The similarity between objects i and j is given by

$$S_{ij} = \sum_{k=1}^{p} s_{ijk} / \sum_{k=1}^{p} w_{ijk} \quad (4)$$

where the weight of the attribute (w_{ijk}) is set equal to 0 or 1 depending on whether the comparison is considered valid for attribute k. If $v_k(i), v_k(j)$ are the values of attribute k for objects i and j respectively, an invalid comparison occurs when at least one them is missing. In this situation w_{ijk} is set to 0.

For quantitative attributes (like the ones of the datasets used in the paper), the scores s_{ijk} are assigned as

$$s_{ijk} = 1 - |v_k(i) - v_k(j)| / R_k$$

where R_k is the range of attribute k. For nominal attributes

$$s_{ijk} = \begin{cases} 1 & \text{if } v_k(i) = v_k(j) \\ 0 & \text{otherwise} \end{cases}$$

This measure can be easily extended for ordinal, interval, and other kind of variables. Also, weighting schemes can be incorporated for considering differential importance of the descriptor variables.

2.2 Class Separability: A supervised perspective

In the supervised case, a natural choice for representing the l mapping is an NDA neural network [26], [16], [17], [12]. 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 capability of neural networks which allows the classification of new incoming objects, and their immediate placement within the created VR space. 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

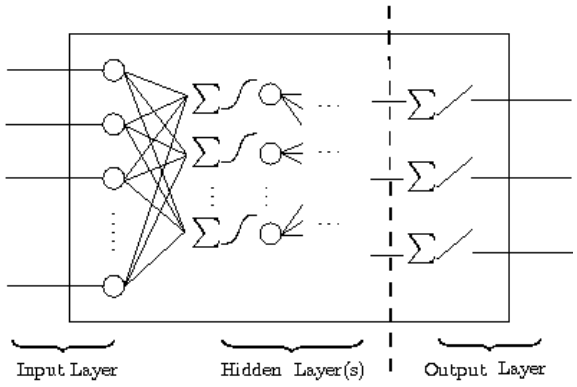


Figure 2: Network Architecture in which the NDA network is learned. $\int$ means nonlinear activation, / linear activation , and Σ aggregation

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 is m; the dimensionality of the projected space (for a VR space this is typically 3). From the 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. The construction of NDA networks can be done 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 conjugate gradient [18], which improves the likelihood of finding good extrema while containing enough determinism. Simulated annealing provides global search capabilities and conjugate gradient improved local search, reducing the risk of entrapment, and resulting in neuron weights with better properties than what is found by the inherent steepest descent implied by pure backpropagation. Alternatively, networks based on evolutionary algorithms can be used, or for instance, particle swarm optimization combined with classical optimization techniques [21].

2.3 The multi-objective approach: A hybrid perspective

It should be clear that a space with new features that satisfactorily preserves the similarity structure does not necessarily guarantee the maximization of class separability, since it results from the solution of an unsupervised problem (i.e. the decision attribute is not considered). Moreover, the relationship between the original descriptor variables and the class membership (expressed by the decision attribute) may be partial, total or poor. On the other hand, if classification is all that matters, then a set of nonlinear features may be found that successfully or acceptably classify the data, but at the cost of distorting the space considerably with respect to the one compliant with the similarity structure. In this case, the kind and amount of nonlinearity and distortion introduced may be so large that the data vector distribution in the two spaces may bear no resemblance at all. This makes the visual data mining process very difficult as the data objects have to be represented in two very different spaces, with different properties. In other words, the above discussed goals are usually in conflict and satisfying them separately, complicates the knowledge discovery process.

Therefore, the following is a very relevant question within the knowledge discovery process using visual data mining: "Are there alternative low dimensional feature spaces (possibly computed by non-linear transformations of the original descriptor variables), in which the class structure can be resolved as much as possible, while distorting the original similarity structure as little as possible?".

A multi-objective optimization approach to the problem of finding suitable nonlinear transformations for the representation of relational structures brings a new perspective to the problem.

2.3.1 Objective functions

In order to establish a formulation of the problem based on multi-objective optimization, a set of objective functions has to be specified, representing the corresponding criteria that must be simultaneouly satisfied by the solution. The minimization of a measure of similarity information loss between the original and the transformed spaces and a classification error measure over the objects in the new space can be used in a first approximation. Clearly, more requirements can be imposed on the solution by adding the corresponding objective functions. Following a principle of parsimony this paper will consider the use of only two criteria, namely, Sammon's error (Eq-2) for the unsupervised case and mean cross-validated classification error with a k-nearest neighbour pattern recognizer for the supervised case.

Let X be a set of N data records x_i, for $i \in [1, N]$, on p independent variables and a discrete dependent variable y (i.e. the class variable) with m possible values, s.t. $y_q = \{1, 2, ..., m\}, q \in [1, m]$. The i-th data record is a vector $\overleftarrow{x}_i$ which takes a value x_{ij} on the j-th independent variable. The k-nearest neighbor approach searches a set of training data records T (i.e., data records with known values for y) to find the k-nearest data records to $\overleftarrow{x}_i$. The proximity (or similarity) of $\overleftarrow{x}_i$ to a member $\overleftarrow{t}_k$ of T is defined by a distance (or similarity) calculated over the independent variables and can be defined by using a variety of measures. In the present case a normalized Euclidean distance is chosen:

$$d_{\overleftarrow{x}\,\overleftarrow{t}} = \sqrt{(1/p)\sum_{j=1}^{p}(x_{ij} - t_{kj})^2} \quad (5)$$

Let S_i be a set containing the k-nearest neighbors in T to $\overleftarrow{x}_i$, where $k \in [1, N]$ is predefined . Then the predicted value of y for $\overleftarrow{x}_i$, $\hat{y}_i$, is given by the value of y_q with the highest frequency within S_i, if such frequency is unique. Otherwise, the predicted value is undefined. This classical non-parametric pattern recognition classifier has been defined elsewhere [8], [10]. For each of the data objects in X there is a classification error w.r.t. the training set T if the predicted class variable does not coincide with its expected value for the corresponding object, or if the object is unclassifiable. The classification error associated with X w.r.t. T is the mean of the classification errors of the objects in X.

3. MULTI-OBJECTIVE OPTIMIZATION USING GENETIC ALGORITHMS

An evolutionary algorithm constructs a population of individuals, which evolve through time until stopping criteria is satisfied. At any particular time, the current population of individuals represent the current solutions to the input problem, with the final population representing the algorithm's resulting output solutions.

The genetic algorithm [1] is a particular evolutionary algorithm that permits particular sequences of operations on individuals of the current population in order to construct the next population in the series of evolving populations. The genetic algorithm requires each individual to have one measure of its fitness, which enables the genetic algorithm to select the fittest individuals for inclusion in the next population. For example, one operation is that of mating two individuals (parents) with the hope that useful pieces of genetic information contained within the chromosomes may be combined in such a way that child individuals (those individuals in the new population) may be fitter than their parents. Another genetic algorithm operation is that of mutation, whereby one individual is selected from the current population, and its chromosome representation is modified in some manner (e.g. probabilistically) in order to construct a new individual in the next population.

An enhancement to the traditional evolutionary algorithm, is to allow an individual to have more than one measure of fitness within a population. One way in which such an enhancement may be applied, is through the use of, for example, a weighted sum of more than one fitness value [3]. Multi-objective optimization, however, offers another possible way for enabling such an enhancement. In the latter case, the problem arises for the evolutionary algorithm to select individuals for inclusion in the next population, because a set of individuals contained in one population exhibits a Pareto Front[19] of best current individuals, rather than a single best individual. Most [3] multi-objective algorithms use the concept of dominance.

A solution $\overleftarrow{x}_{(1)}$ is said to dominate [3] a solution $\overleftarrow{x}_{(2)}$ for a set of m objective functions $< f_1(\overleftarrow{x}), f_2(\overleftarrow{x}), ..., f_m(\overleftarrow{x}) >$ if

1. $\overleftarrow{x}_{(1)}$ is not worse than $\overleftarrow{x}_{(2)}$ over all objectives.
   For example, $f_3(\overleftarrow{x}_{(1)}) \leq f_3(\overleftarrow{x}_{(2)})$ if $f_3(\overleftarrow{x})$ is a minimization objective.
2. $\overleftarrow{x}_{(1)}$ is strictly better than $\overleftarrow{x}_{(2)}$ in at least one objective. For example, $f_6(\overleftarrow{x}_{(1)}) > f_6(\overleftarrow{x}_{(2)})$ if $f_6(\overleftarrow{x})$ is a maximization objective.

One particular algorithm for multi-objective optimization is the elitist non-dominated sorting genetic algorithm (NSGA-II) [7], [6], [5], [3]. It has the features that it *i)* uses elitism, *ii)* uses an explicit diversity preserving mechanism, and *iii)* emphasizes the non-dominated solutions. The procedure is as follows: *i)* Create the child population using the usual genetic algorithm operations. *ii)* Combine parent and child populations into a merged population. *iii)* Sort the merged population according to the non-domination principle. *iv)* Identify a set of fronts in the merged population ($F_i, i = 1, 2, ...$). *v)* Add all complete fronts F_i, for $i = 1, 2, ..., k-1$ to the next population. *vi)* There may now be a front, F_k, that does not completely fit into the next population. So select individuals that are maximally separated from each other from the front F_k according to a crowding distance operator. *vii)* The next population has now been constructed, so continue with the genetic algorithm operations.

3.1 Implementation

The PGAPack library [15] is a general-purpose, data structure neutral, parallel genetic algorithm library. It is intended to provide most capabilities desired in a genetic algorithm library, in an integrated, seamless, and portable manner. Key features that are in PGAPack V1.0 include: *i)* Callable from Fortran or C, *ii)* Runs on uniprocessors, parallel computers, and workstation networks, *iii)* Binary-, integer-, real-, and character-valued native data types, *iv)* Full extensibility to support custom operators and new data types, *v)* Easy-to-use interface for novice and application users, *vi)* Multiple levels of access for expert users, *vii)* Parameterized population replacement, *viii)* Multiple crossover, mutation, and selection operators, *ix)* Easy integration of hill-climbing heuristics, *x)* Extensive debugging facilities, *xi)* Large set of example problems, and *xii)* Detailed users guide. The PGAPack library (`http://www-fp.mcs.anl.gov/CCST/research/reports_pre1998/comp_bio/stalk/pgapack.html` was extended to include Revision 1.1 (10 June 2005) of the NSGA-II algorithm, `http://www.iitk.ac.in/kangal/codes.shtml`, written in C with constraint handling.

4. APPLICATION TO ALZHEIMER'S DISEASE (GENOMIC 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. Progress in understanding these diseases is hampered by their complexity, but there is now renewed hope that genomics technologies, particularly gene expression profiling, can have an impact. Genome-wide expression profiling of thousands of genes provides rich datasets that can be mined to extract information on the genes that best characterize the disease state [25]. A total of 4 clinically diagnosed AD patients and 5 "normal" patients of similar age were used in this study, comprising 12 AD and 11 normal samples, for a total of 23 samples. Each is characterized by a collection of 9600 attributes describing expression intensities of a corresponding number of genes. Details can be found in [25].

Each sample is a vector in a 9600 space, and therefore, direct inspection of the structure of this data, and of the relationship between the descriptor variables (the genes) and the type of sample (normal or Alzheimer), is impossible. Moreover, within the collection of genes there is a mixture of potentially relevant genes with others which are irrelevant, noisy, etc.

The need of simultaneously finding a visual representation (3D) respecting (as much as possible) the set of object interrelationships as defined by the 9600 original attributes, and the construction of a new feature space effectively differentiating the two classes of objects present, makes this problem suitable for a multi-objective optimization approach.

4.1 Experimental Settings

In the present case, there is a sample of $N = 23$ objects in a 9600-dimensional space (9600 genes describe each sample). All of the attributes (the gene expression intensity values) are real-valued. Therefore, the original domain is homogeneous, actually a particular case of the heterogeneous domains according to the formalism introduced in Section 2. $\hat{\mathcal{H}}^n = \hat{\mathbb{R}}^{n_R}$, where $n = n_R = 9600$. In the virtual reality space Υ, $m = 3$, and if continuous 3D spaces (for example, with Euclidean metric) are targeted, $l : \Re^{9600} \rightarrow \Re^3$. The result of the mapping l is an image of that set of original samples but in a 3-dimensional space. Accordingly, the number of attributes of the objects in the new space is $M = m = 3$. If the image set can be constructed (i.e. a set of $N = 23$ vectors of dimension $M = 3$), the ζ_{ij} terms in the error measures described by Eqs. $1, 2, 3$ can be evaluated for any pair of objects i, j. In particular, if Sammon error is chosen as an error measure (Eq. 2) and if an image is found such that this measure is minimized, then an implicit representation of the mapping l is obtained.

This problem can be described by a GA where each linear real-valued chromosome in the population represents a candidate image of the set of N objects in the VR space, with the chromosome elements being the coordinates of the objects in the VR space (a total of $N \cdot M = 23 \cdot 3 = 69$). The decoding scheme is simply decomposing the chromosome into chunks of M elements, such that the i-th chunk stands for the coordinates of the image of the corresponding object in the original sample (Fig-3). Thus, each chromosome represents the result of an implicit mapping $l : \Re^{9600} \rightarrow \Re^3$.

If the three attributes of the VR space are denoted as X, Y, Z, their relation with those of the original space is given by:

$$\begin{aligned} X &= \varphi_x(v_1, v_2, \cdots, v_{9600}) \\ Y &= \varphi_y(v_1, v_2, \cdots, v_{9600}) \\ Z &= \varphi_z(v_1, v_2, \cdots, v_{9600}) \end{aligned}$$

where $\{v_1, v_2, \cdots, v_{9600}\}$ are the original variables and $\varphi_x, \varphi_y, \varphi_z$ are the non-linear functions of the original variables defining the mapping l. Note that in this approach the explicit form of l is neither obtained nor needed. However, there are applications where an explicit l is required (they are developed elsewhere).

The collection of parameters describing the application of the NSGA-II algorithm is shown in Table-1.

It should be observed that a modest population size and number of generations were used, with a relatively high mutation probability in order to enable richer genetic diversity. Randomization of the set of data objects was applied in order to reduce the bias in the composition of the cross-validated folds by providing a more even class distribution between successive training and test subsets. The number of folds was set in consideration of the sample size.

4.2 Results

The set of non-dominated solutions obtained by the NSGA-II algorithm is shown in the scatter plot of Fig-4, where the horizontal axis is the mean cross-validated knn error and the vertical axis the Sammon error. The approximate location of the Pareto front is defined by the convex polygon joining the solutions provided by chromosomes 0, 3, 2, 4, 1. Chromosome 0 defines a space with a perfect resolution of the supervised problem in terms of the Normal and Alzheimer classes (knn error $= 0$), but at the cost of a

Table 1: Experimental settings for computing the pareto-optimal solution approximations by the multi-objective genetic algorithm (PGAPack extended by NSGA-II).

population size	100
number of generations	200
chromosome length	69
ga seed	4001
objective functions should be	minimized
chromosome data representation	real
crossover probability	0.8
crossover type	uniform (prob. 0.6)
mutation probability	0.4
mutation type	gaussian
selection type	tournament
tournament probability	0.6
perform mutation and crossover	yes
population initialization	random, bounded
lower bound for initialization	0
upper bound for initialization	7
fitness values	raw objective values
stopping criteria	maximum iterations
restart ga during execution	no
parallel populations	no
number of objectives	2
number of constraints	0
pre-computed diss. matrix	Gower dissimilarity
evaluation functions	mean cross-validated error Sammon error
cross-validation (c.v.)	5 folds
randomize before c.v.	yes
knn seed	-101
k nearest neighbors	3
non-linear mapping measure	Sammon
dimension of the new space	3

severe distortion of the space. Whereas, chromosome 1 approximates a pure unsupervised solution (with low Sammon error). Its classification error is large indicating that few non-linear features preserving the similarity structure lacks classification power. This may be due to the large amount of attribute noise, redundancy, and irrelevancy within the set of 9600 original genes.

Clearly, it is impossible to represent virtual reality spaces on a static medium. However, a composition of snapshots of the VR spaces using the solutions along the Pareto front approximation is shown in Fig-5.a-5.e. For comparison Fig-5.f corresponds to an unsupervised single-objective solution obtained with deterministic optimization (Newton's method) using Sammon's error (Eq-2), Gower's similarity in the original space (Eq-4), and normalized Euclidean metric in the new space (Eq-5) was obtained in [25]. The error obtained was 0.1034 after 335 iterations. The error of this single objective solution is much better than the equivalent obtained with the multi-objective approach, but it should be considered that the reduced number of generations (in the latter case) as well as the modest population size, considerably reduces the search space.

A solution satisfying classification error as much as possible (actually with 0-error) is shown in Fig-5.a where both classes are not only completely separated, but linearly separated. If this space is compared with the MO solution most oriented towards similarity preservation (Fig-5.e) or with the pure single-objective solution of Fig-5.f, it is possible to see that according to the original variables, the two classes are not linearly separable. In fact, the Alzheimer class is surrounded by elements from the Normal class. In Fig-5.c, patterns from both Fig-5.a and Fig-5.e-f can be identified. The two classes are still separable (but less sharply, in this case by a nonlinear boundary), and also the Alzheimer class is closer and more mixed with the Normal class. If the spaces of Figs e and f (particularly the last one) represents an approximation to the 'natural' distribution of the data in the original 9600 dimensional space (a reasonable assumption supported by the low Sammon error of Fig-5.f, the distortion required for the space to achieve classification error =0 is large, as evidenced by Fig-5.a. Clearly, Fig-5.c shows a space less distorted and closer to that of Fig-5.f, but where the two classes are still clearly distinguishable. That is why visually, that space represents a compromise solution between the two goals and a tradeoff between the two objective functions. It should be remembered that the class information is not used at all for computing the space of Fig-5.f. Chromosome 2, according to Fig-4 and Fig-5.c, can be considered to be the best multi-objective compromised solution in which both error criteria are simultaneously as low as possible. It shows a reasonable class discrimination with a non-large similarity structure distortion, which is a very meaningful result.

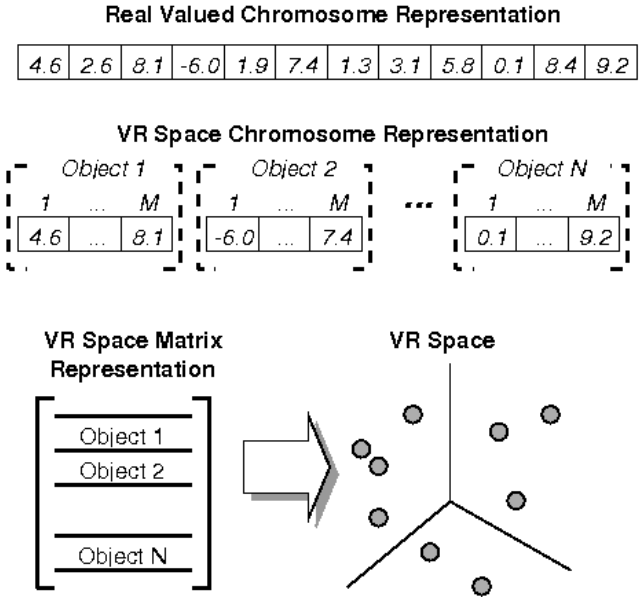


Figure 3: Multi-objective chromosome representation.

5. CONCLUSIONS

A multi-objective optimization approach was introduced for the problem of computing virtual reality spaces in the context of visual data mining and knowledge discovery applied to relational structures (e.g. databases). The multi-objective procedure was based on NSGA-II using two objective functions representative of unsupervised and supervised criteria (mean cross-validated knn error as a measure of missclassification, and Sammon error as a measure of similarity structure loss). This methodology was applied to the analysis of high dimensional genomic data collected in the framework of Alzheimer's disease research.

A Pareto front approximation was recognizable from within the solutions provided by the final population. Selected solutions from along that approximation were used for the construction of a sequence of visualizations showing the progression from spaces with

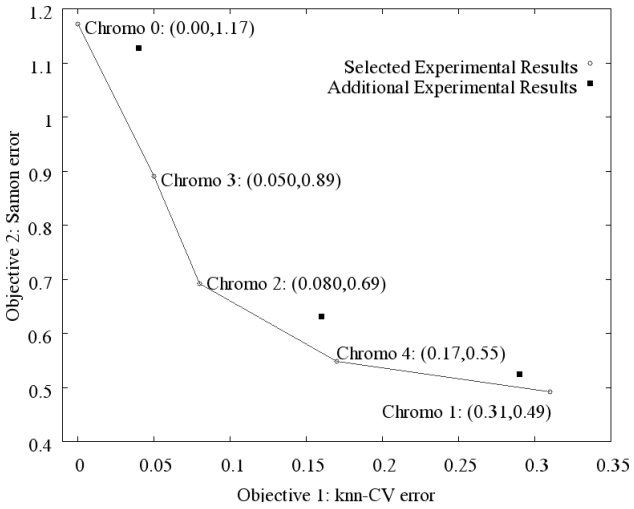


Figure 4: Set of 100 **multiobjective solutions. Those along the Pareto front approximation progressively span the extremes between minimum classification error and minimum dissimilarity loss. The errors for the two objective functions are shown in parenthesis.**

complete class separation and poor similarity preservation to spaces with reversed characteristics. A solution with a reasonable compromise between the two criteria was identified and clearly contained properties of both extreme solution spaces. This is the first investigation of virtual reality space construction using multi-objective optimization with genetic algorithms applied a specific real-world problem. Thus, these research results, although preliminary, showed large potential and further investigation is required.

6. ACKNOWLEDGMENTS

The authors would like to thank Robert Orchard from the Integrated Reasoning Group (National Research Council Canada, Institute for Information Technology) for his constructive criticism of the first draft of this paper.

7. REFERENCES

[1] T. Bäck, D. B. Fogel, and Z. Michalewicz. *Handbook of Evolutionary Computation*. Institute of Physics Publishing and Oxford Univ. Press, New York, Oxford, 1997.

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

[3] E. K. Burke and G. Kendall. *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*. Number 0-387-23460-8. Springer Science and Business Media, Inc., 233 Spring Street, New York, NY 10013, USA, 2005.

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

[5] K. Deb, S. Agarwal, and T. Meyarivan. A fast and elitist multi-objective genetic algorithm: Nsga-ii. In *IEEE Transaction on Evolutionary Computation*, volume 6 (2), pages 181–197, 2002.

[6] K. Deb, S. Agarwal, A. Pratap, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In *Proceedings of the Parallel Problem Solving from Nature VI Conference*, pages 849–858, Paris, France, 16-20 September 2000.

[7] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multi-objective genetic algorithm: Nsga-ii. Technical Report 2000001, Kanpur Genetic Algorithms Laboratory (KanGAL), Indian Institute of Technology Kanpur, 2000.

[8] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley New York, 1972.

[9] U. Fayyad, G. Piatesky-Shapiro, and P. Smyth. From data mining to knowledge discovery. In U. F. et al., editor, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press, 1996.

[10] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1972.

[11] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 1(27):857–871, 1973.

[12] A. K. Jain and J. Mao. Artificial neural networks for nonlinear projection of multivariate data. In *1992 IEEE joint Conf. on Neural Networks*, pages 335–340, Baltimore, MD, 1992.

[13] M. Jianchang and A. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans. On Neural Networks*, 6(2):1–27, 1995.

[14] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psichometrika*, 29:1–27, 1964.

[15] D. Levine. *Users Guide to the PGAPack Parallel Genetic Algorithm Library*. Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, January 1996.

[16] J. Mao and A. K. Jain. Discriminant analysis neural networks. In *1993 IEEE International Conference on Neural Networks*, pages 300–305, San Francisco, California, March 1993.

[17] J. Mao and A. K. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans. on Neural Networks*, 6:296–317, 1995.

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

[19] V. Pareto. *Cours D'Economie Politique*, volume I and II. F. Rouge, Lausanne, 1896.

[20] J. W. Sammon. A non-linear mapping for data structure analysis. *IEEE Trans. Computers*, C18:401–408, 1969.

[21] J. Valdés. Building virtual reality spaces for visual data mining with hybrid evolutionary-classical optimization: Application to microarray gene expression data. In *2004 IASTED International Joint Conference on Artificial Intelligence and Soft Computing, ASC'2004*, pages 161–166, Marbella, Spain, September 2004. IASTED, ACTA Press, Anaheim, USA.

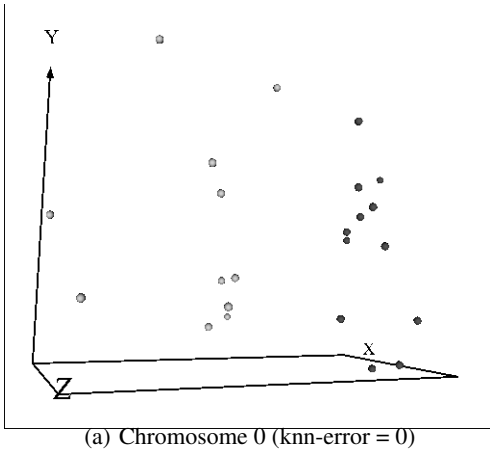
[22] J. J. Valdés. Similarity-based heterogeneous neurons in the context of general. *Neural Network World*, 12(5):499–508, 2002.

[23] J. J. Valdés. Virtual reality representation of relational systems and decision rules:. In P. Hajek, editor, *Theory and Application of Relational Structures as Knowledge Instruments*, Prague, Nov 2002. Meeting of the COST Action 274.

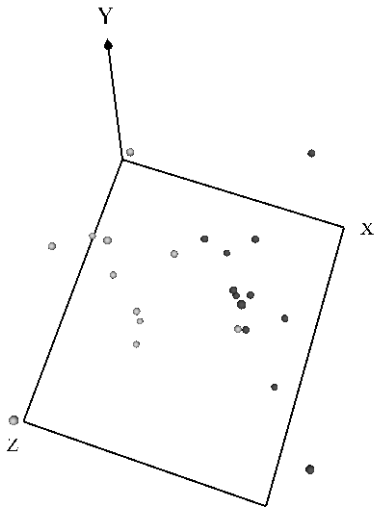
[24] J. J. Valdés. Virtual reality representation of information systems and decision rules:. In *Lecture Notes in Artificial Intelligence*, volume 2639 of *LNAI*, pages 615–618. Springer-Verlag, 2003.

[25] P. Walker, B. Smith, Y. Qing, F. Famili, J. J. Valdés, L. Ziying, and L. Boleslaw. Data mining of gene expression changes in alzheimer brain. *Artificial Intelligence in Medicine*, 31:137–154, 2004.

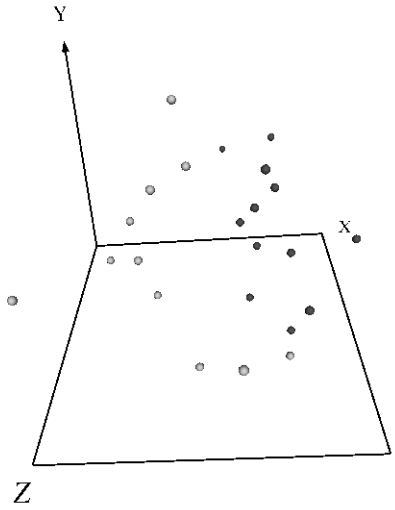
[26] A. R. Webb and D. Lowe. The optimized internal representation of a multilayer classifier. *Neural Networks*, 3:367–375, 1990.



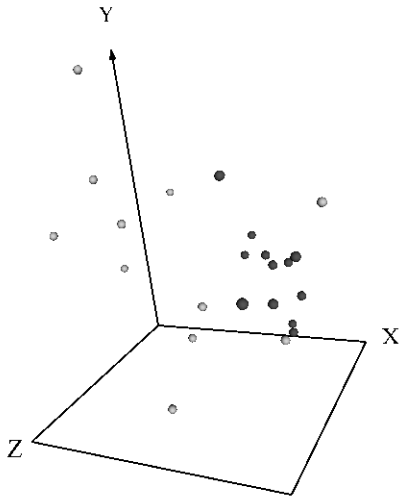
(a) Chromosome 0 (knn-error = 0)



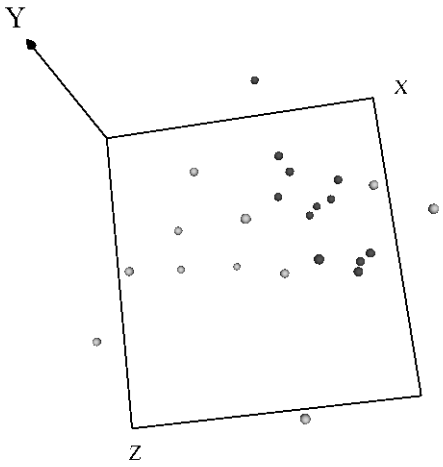
(b) Chromosome 3



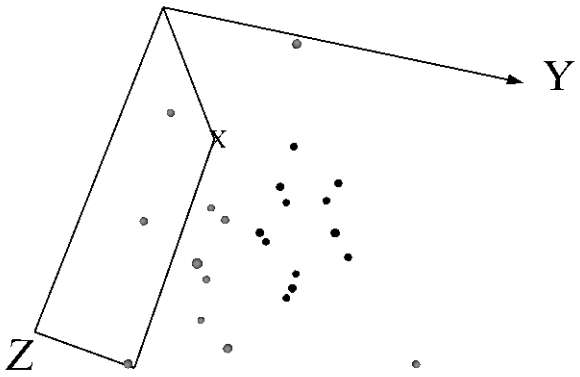
(c) Chromosome 2



(d) Chromosome 4



(e) Chromosome 1



(f) Walker et. al. 2004 (Sammon error = 0.1034)

Figure 5: Snapshots of vr-spaces computed with different solutions along the Pareto front approximation progressively spanning the extremes (minimum classification error; minimum dissimilarity loss). Light spheres = normal samples, dark spheres = Alzheimer samples.