Virtual reality spaces: visual data mining with a hybrid computational intelligence tool

Valdés, Julio; Barton, Alan

https://doi.org/10.4224/8913600

Report (National Research Council of Canada. Radio and Electrical Engineering Division. ERB); no. ERB-1137, 2006-04

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*Virtual Reality Spaces: Visual Data Mining with a Hybrid Computational Intelligence Tool **

Valdés, J., and Barton, A.
April 2006

* published as NRC/ERB-1137. 44 pages. April 2006. NRC 48501.

Canada

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

ERB-1137

NRC·CNRC

Virtual Reality Spaces: Visual Data Mining with a Hybrid Computational Intelligence Tool

Valdés, J., and Barton, A.
April 2006

Canada

NRC 48501

Virtual Reality Spaces: Visual Data Mining with a Hybrid Computational Intelligence Tool

J.J. Valdés[a] and A.J. Barton[a]

[a]National Research Council Canada
Institute for Information Technology
Bldg. M-50, 1200 Montreal Rd., Ottawa, Ontario, K1A 0R6
Canada

The information explosion requires the development of alternative data mining procedures that speed up the process of scientific discovery. The improved in-depth understanding and ease of interpretability of the internal structure of data by investigators allows focussing on the most important issues, which is crucial for the identification of valid, novel, potentially useful, and understandable patterns (regularities, oddities, surprises, etc). Computational visualization techniques are used to explore, in an immersive fashion, inherent data structure in both an unsupervised and supervised manner. Supervision is provided via *i)* domain knowledge contained in the data, and *ii)* unsupervised data mining procedures, such as fuzzy clustering, etc.

The Virtual Reality (VR) approach for large heterogeneous, incomplete and imprecise (fuzzy) information is introduced for the problem of visualizing and analyzing general forms of data. The method is based on mappings between a heterogeneous space representing the data, and a homogeneous virtual reality space. This VR-based visual data mining technique allows the incorporation of the unmatched geometric capabilities of the human brain into the knowledge discovery process. Traditional means of interpretation would require more time and effort in order to achieve the same level of deep understanding of complex high dimensional data as the proposed technique.

This hybrid approach has been applied successfully to a wide variety of real-world domains including astronomy, genomics, and geology, providing useful insights.

1. Introduction

There are many computational techniques that may be used within a knowledge discovery process and applied to complex real world data sets in order to combat the information explosion problem. These various computations may lead to a plethora of data mining results, which, along with the data, need to be analyzed and properly understood. Fortunately, humans are capable of visually perceiving large quantities of information at very high input rates (by far outperforming modern computers). Therefore, one approach to aid the discovery of knowledge (concepts/relationships/etc) from within large data sets (possibly containing space or time dependencies) and data mining results obtained from

computer procedures, is to orient knowledge representations towards this vast human capacity for visual perception; effectively attempting to include the best of computer (calculation) capabilities with the best of human perception (cognitive) abilities.

Scanning and sensor technologies and their applications are continuously increasing in importance for a larger number of domains. General examples include *i)* collecting measurements from an entity (such as airplane/truck/train/gene/etc.) and *ii)* simulation of entities. Today the application areas are diversifying rapidly, including fields like archeology, art, and many others. One common denominator of all of these applications is that either there is a *recreation* of reality (eg. a model of a machine, a dinosaur skull from a museum, etc.), or the *creation* of some reality, which is plausible in some conceptual world (eg. monsters in a computer game, or a conjectured organic molecule, not yet discovered in Nature). In these cases, despite the nature of the modelled objects being either real or imaginary, they are *conceptually compatible* with a 3D world.

However, domains exist in which the virtual objects represent abstract concepts whose nature is not related with a physical space. Examples of these kinds of abstract objects: *i)* in logics, database theory and data mining are mathematical notions like *semantic systems* and *information systems* [18], [29], and *ii)* in artificial intelligence are notions like *knowledge bases*, which are typically composed of decision rules. These rules are generated either by human experts or by machine learning techniques, like inductive methods [30], rough set algorithms [29], GUHA methods [18], association rules [1], etc). In all cases, the abstract nature of the virtual objects, their complexity and their large numbers makes it difficult to understand and interpret their properties. However, these objects are important from the point of view of *knowledge discovery*; understood as the non-trivial process of identifying valid, novel, potentially useful, and ultimately *understandable patterns* in data [10].

The purpose of this paper is: *i)* to demonstrate the use of a hybrid evolutionary computation based technique that may be used for visual exploration within a data mining process, *ii)* to use the hybrid technique in the exploration of complex heterogeneous (generalized) data sets containing imprecise and uncertain information in order to aid the discovery of rarities, anomalies, exceptions, and other kinds of knowledge (such as anomalous decision rules) through the presentation of the inherent structure of the (generalized) data, *iii)* to visually aid the interpretation of data mining results (such as rough sets, k-means or fuzzy clustering or other machine learning algorithms) with the additional benefit of incorporating the human brain more directly into the summarization process, *iv)* to explore the construction of VR spaces for data representation from *a)* a supervised perspective, by using a variant (w.r.t. the classical approach) of nonlinear discriminant analysis neural networks (NDA networks), *b)* an unsupervised perspective, using evolutionary computation based techniques such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO), etc. *v)* to perform case studies using geology, astronomy and bioinformatics (genes, medicine – e.g. breast cancer, leukemia, etc.) data in order to demonstrate some of the real world uses for which the hybrid technique may be applied. For example, gene expression data has very high dimension (in the order of thousands).

2. Virtual Reality as a Data Mining Tool

Classical data mining and analysis methods may possibly be *i)* difficult to use, *ii)* verbose in the quantity of output produced when many procedures are investigated; leading to a time consuming process of analysis, which may require special expertise for interpretation, and *iii)* 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.

Several reasons make Virtual Reality (VR) a suitable paradigm: VR is *flexible*, in the sense that it allows the choice of different representation models to better accommodate different human perception preferences. In other words, allows the construction of different virtual worlds representing *the same* underlying information, but with different look and feel. Thus, the user can choose the particular representation that is most appealing to her. VR allows *immersion*. That is, the user can navigate inside the data, interact with the objects in the world, change scales, perspectives, etc. VR creates a *living* experience. The user is not merely a passive observer or an outsider, but an actor in the world, in fact, part of the information itself. 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 or portions of the world. Of no less importance is the fact that in order to interact with a Virtual World, no mathematical knowledge is required, but only minimal computer skills.

A virtual reality, visual, data mining technique extending the concept of 3D modelling to relational structures was introduced [35], [37], (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 [36] (i.e. objects under study are described in terms of collections of *heterogeneous* properties).

2.1. The Heterogeneous Space

More and more, theoretical and applied domains are attempting to solve complex problems, where objects are described in terms of a large collection of *heterogeneous* properties. Consider, for example, the case of patient description where, besides the symptoms represented by simple nominal, ordinal or real-valued variables, there are other important features, like images (e.g. X-rays), time-series (e.g. ECG, EKG), documents (e.g. doctor's comments or evaluations), graphs (e.g. the structure of the blood vessels in a region of interest), videos (e.g. the diffusion within the blood stream of a given substance), and many others. Further, in a broad sense, the notion of data is not restricted to databases, but also includes logical relations and other forms of structured knowledge, with different degrees of precision, *uncertainty* and completion (*missing data* is quite common).

Formally, consider an *information system* $S =< U, A >$ where U and A are non-empty finite sets, called the *universe* and the set of *attributes* respectively, such that each $a \in A$ has a domain V_a and an evaluation function f_a assigns to each $u \in U$ an element $f_a(u) \in V_a$ (i.e. $f_a(u) : U \rightarrow V_a$) (here the V_a are not required to be finite). An example of such a system is shown in Fig-1. There are attributes with domains of different kinds (nominal, ordinal, ratio, fuzzy, images, time-series and graphs), and also containing missing values.

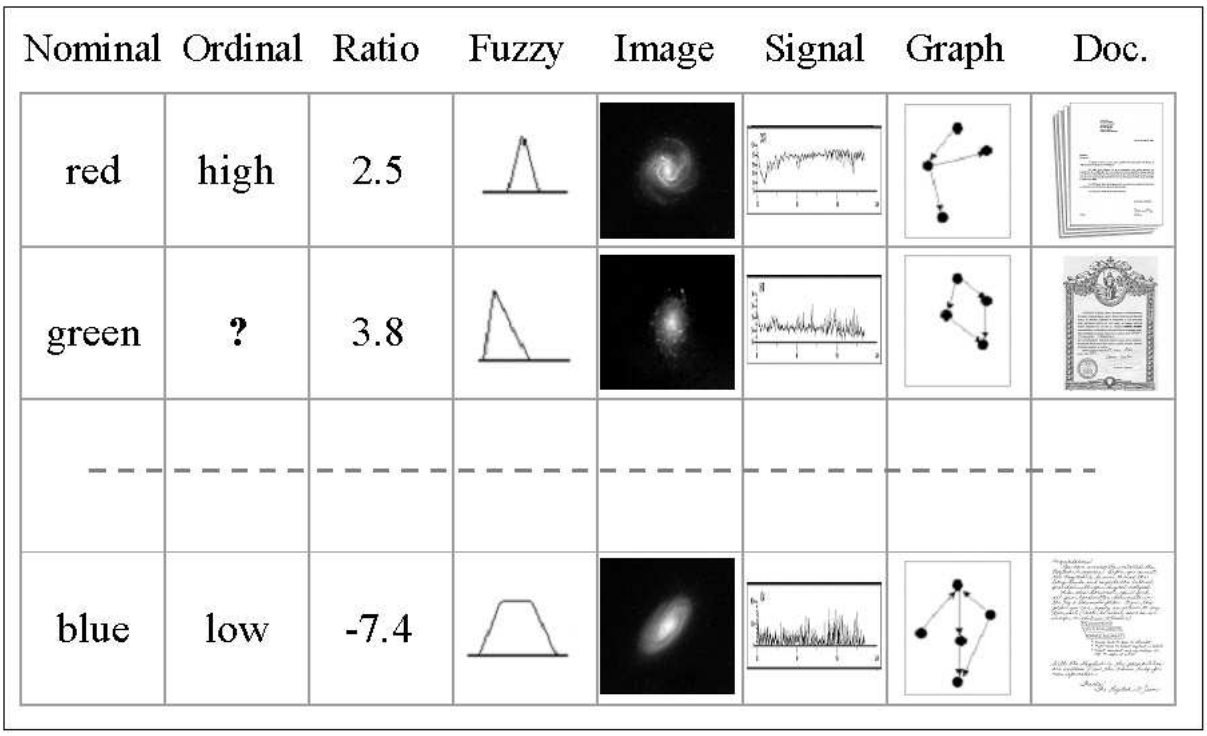

Figure 1. An example of a heterogeneous database. Nominal, ordinal, ratio, fuzzy, image, signal, graph, and document data are mixed. The symbol ? denotes a missing value.

In the present case, *heterogeneous and incomplete information systems* will be considered as follows. Let ? be a special symbol having two basic properties: *i*) if $? \in \Omega$ (Ω being an arbitrary set) and f is any unary function defined on Ω, then $f(?) = ?$, and *ii*) ? is an incomparable element w.r.t any ordering relation defined on Ω.
A heterogeneous domain is defined as a Cartesian product of a collection of *source sets* (Ψ_i): $\hat{\mathcal{H}}^n = \Psi_1 \times \ldots \times \Psi_n$, where $n > 0$ is the number of *information sources* to consider. Projections and cylindric extensions are defined in the usual way.
As an example, consider the case of a heterogeneous domain where objects are characterized by attributes given by continuous crisp quantities, discrete features, fuzzy features, graphs and digital images. Let $\mathbb{R}$ be the reals with the usual ordering, and $\mathcal{R} \subseteq \mathbb{R}$. Now define $\hat{\mathcal{R}} = \mathcal{R} \cup \{?\}$ to be a source set and extend the ordering relation to a partial order accordingly. For example, $\hat{\mathcal{R}}$ may model point measurements of some variable, possibly with missing values (e.g. temperature readings). Now let $\mathbb{N}$ be the set of natural numbers and consider a family of n_r sets ($n_r \in \mathbb{N}^+ = \mathbb{N} - \{0\}$) given by $\hat{\mathcal{R}}^{n_r} = \hat{\mathcal{R}}_1 \times \ldots \times \hat{\mathcal{R}}_{n_r}$ (n_r times) where each $\hat{\mathcal{R}}_j$ ($0 \leq j \leq n_r$) is constructed as $\hat{\mathcal{R}}$, and define $\hat{\mathcal{R}}^0 = \phi$ (the empty set). Now let $\mathcal{O}_j$, $1 \leq j \leq n_o \in \mathbb{N}^+$ be a family of finite sets with cardinalities k_j^o respectively, composed by arbitrary elements, such that each set has a fully ordering relation $\leq_{\mathcal{O}_j}$. Construct the sets $\hat{\mathcal{O}}_j = \mathcal{O}_j \cup \{?\}$, and for each of them define a partial ordering $\hat{\leq}_{\mathcal{O}_j}$ by extending $\leq_{\mathcal{O}_j}$ according to the definition of ?. Analogously construct

the set $\hat{\mathcal{O}}^{n_o} = \hat{\mathcal{O}}_1 \times \ldots \times \hat{\mathcal{O}}_{n_o}$ (n_o times and $\hat{\mathcal{O}}^0 = \phi$). For the special case of nominal variables, let $\mathcal{N}_j$, $1 \leq j \leq n_m$ ($n_m \in \mathbb{N}^+$) be a family of finite sets with cardinalities $k_j^m \in \mathbb{N}^+$ composed by arbitrary elements but such that no ordering relation is defined on any of the $\mathcal{N}_j$ sets. Now construct the sets $\hat{\mathcal{N}}_j = \mathcal{N}_j \cup \{?\}$, and define $\hat{\mathcal{N}}^{n_m} = \hat{\mathcal{N}}_1 \times \ldots \times \hat{\mathcal{N}}_{n_m}$, ($n_m$ times and $\hat{\mathcal{N}}^0 = \phi$). Sets $\hat{\mathcal{O}}^{n_o}$, $\hat{\mathcal{N}}^{n_m}$ may represent the case of n_o ordinal variables and n_m nominal variables respectively (according to statistical terminology). Similarly, a collection of n_f extended fuzzy sets $\hat{\mathcal{F}}_j$ ($1 \leq j \leq n_f$) , n_g extended graphs $\hat{\mathcal{G}}_j$ ($1 \leq j \leq n_g$) and n_i extended digital images $\hat{\mathcal{I}}_j$ ($1 \leq j \leq n_i$), can be used for constructing the corresponding cartesian products given by $\hat{\mathcal{F}}^{n_f}$, $\hat{\mathcal{G}}^{n_g}$ and $\hat{\mathcal{I}}^{n_i}$.
The heterogeneous domain is given by $\hat{\mathcal{H}}^n = \hat{\mathcal{R}}^{n_r} \times \hat{\mathcal{O}}^{n_o} \times \hat{\mathcal{N}}^{n_m} \times \hat{\mathcal{F}}^{n_f} \times \hat{\mathcal{G}}^{n_g} \times \hat{\mathcal{I}}^{n_i}$. Elements of this domain will be objects $o \in \hat{\mathcal{H}}^n$ given by tuples of length $n = n_r + n_o + n_m + n_f + n_g + n_i$, with $n > 0$ (for exclude the empty set). Other kinds of heterogeneous domains can be constructed in the same way, using the appropriate source sets. Furthermore, more general information systems are those in which the universe is endowed with a set of relations of different arities. Let $t =< t_1, \ldots, t_p >$ be a sequence of p natural integers, called *type*, and $\underline{Y} =< Y, \gamma_1, \ldots, \gamma_p >$ a relational structure as defined in [18], where Y is a non-empty domain of objects and the $\Gamma = \{\gamma_i\}$ ($i = 1, \ldots, p$) are different relations of various arities defined on Y (according to t). The extended information system will be $\hat{S} =< U, A, \Gamma >$, endowed with the relational system $\underline{U} =< U, \Gamma >$.

2.2. The Virtual Reality Space

A *virtual reality space* is a structure composed by different sets and functions defined as $\Upsilon =< \underline{O}, G, B, \Re^m, g_o, l, g_r, b, r >$. $\underline{O}$ is a relational structure defined as above ($\underline{O} =< O, \Gamma^v >$, $\Gamma^v =< \gamma_1^v, \ldots, \gamma_q^v >$, $q \in \mathbb{N}^+$ and the $o \in O$ are objects), G is a non-empty set of *geometries* representing the different objects and relations (the *empty* or *invisible* geometry is a possible one). B is a non-empty set of *behaviors* (i.e. ways in which the objects from the virtual world will express themselves: movement, response to stimulus, etc.). $\Re^m \subset \mathbb{R}^m$ is a *metric space* of dimension m (euclidean or not) which will be the actual virtual reality geometric space. The other elements are mappings: $g_o : O \to G$, $l : O \to \Re^m$, $g_r : \Gamma^v \to G$, $b : O \to B$, r is a collection of characteristic functions for Γ^v, $(r_1, \ldots, r_q)$ s.t. $r_i : \gamma_i^{vt_i} \to \{0, 1\}$, according to the type t associated with Γ^v.

The representation of an extended information system $\hat{S}$ in a virtual world implies the construction of another $\hat{S}^v =< O, A^v, \Gamma^v >, \underline{O}$ in Υ, which requires the specification of several sets and a collection of extra mappings (w.r.t. those required for Υ). Clearly, it can be done in many ways other than the one described here. A desideratum for $\hat{S}^v$ is to keep as many properties from $\hat{S}$ as possible. Thus, an obvious requirement is that U and O are in one-to-one correspondence (with a mapping $\xi : U \to O$). The structural link is given by a mapping $f : \hat{\mathcal{H}}^n \to \Re^m$. If $u =< f_{a_1}(u), \ldots, f_{a_n}(u) >$ and $\xi(u) = o$, then $l(o) = f(\xi(< f_{a_1}(u), \ldots, f_{a_n}(u) >)) =< f_{a_1^v}(o), \ldots, f_{a_m^v}(o) >$ ($f_{a_i^v}$ are the evaluation functions of A^v). This gives *semantics* to the pair $< g_o(o), l(o) >$ (it determines important properties like geometry, visibility and location).
It is natural to require that $\Gamma^v \subseteq \Gamma$ (possibly empty), thus having a virtual world portraying selected relations from the information system, represented according to the choices made for G and g_r.

2.3. The Direct and Inverse Transforms

As mentioned, f 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 to consider and this is the case when the location and adjacency relationships between the objects O in Υ should give an indication about the *similarity relationships* [7] between the objects U in the original heterogeneous space $\hat{\mathcal{H}}^n$, as given by A [36]. Other interpretations about internal structure are related with the linear/non-linear separability of class membership relations defined on the data [19]. In this sense, f can be constructed as to maximize some metric/non-metric structure preservation criteria as has been done for decades in multidimensional scaling [25], [5], or minimize some error measure of information loss [33], [19]. 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), two examples of error measures frequently used are:

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

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

The f mappings obtained using approaches of this kind are only *implicit*, as no functional representations are found. Moreover, its usefulness is restricted to the final errors obtained in the complex optimization process. However, explicit mappings can be obtained from these solutions using neural network or genetic programming techniques. An explicit f is useful for both practical and theoretical reasons. On one hand, in dynamic data sets (e.g. systems being monitored or databases formed incrementally from continuous processes) an explicit direct transform f will speed up the incremental update of the virtual reality information system S^v. On the other hand, it can give semantics to the attributes of the virtual reality space A^v, 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.

The existence of an *inverse transformation* f^{-1} from Υ back to $\hat{\mathcal{H}}^n$ is, in many cases, worth considering. If a sense is made of patterns of objects in Υ in terms of abstract concepts, and new conjectured objects or relations are conceived, it is natural to ask what kind of previously unseen or undiscovered objects or relations they would correspond to in $\hat{\mathcal{H}}^n$. The existence of ξ^{-1} is guaranteed by its one-to-one correspondence property, thus, in order to have a complete reversion of the process, only f^{-1} is required. Several approaches can be followed for finding the inverse transformation and they depend on

the kind of information system involved. Neural networks are obvious choices because of their flexibility and general function approximation property. In particular, counterpropagation networks are very appealing due to their ability of learning direct and inverse transforms simultaneously (Fig-2). Also pairs of feed-forward networks [11] and hybrid networks with heterogeneous neurons [36] can be used, trained with the results of implicit mapping techniques as input-target pairs.

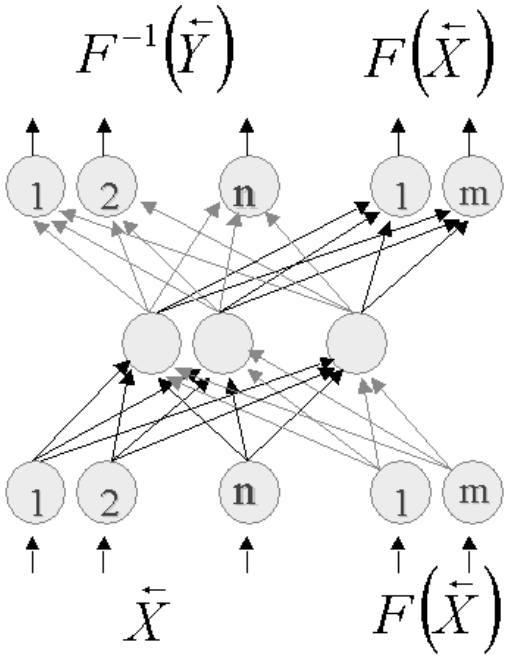

Figure 2. Simultaneous direct and inverse transformation with a counterpropagation neural network, useful for bridging the original (heterogeneous) and the constructed (virtual reality) spaces.

2.4. The Problem of Large Datasets

Regardless of the criteria followed when computing a virtual reality space, complex optimization procedures are applied involving the estimation of the image of the data objects. The objective function surface becomes more complex and convoluted with the increase of the dimensionality of the parameter space, and local extrema entrapment is typical. Even if all of the difficulties related with the amount of memory and the numeric computation involved are put aside (note that a dissimilarity matrix grows quadratically with the number of objects), the graphical representation of millions or possibly billions of objects in a screen with the current computer technologies, is neither feasible, nor practical. On the other hand, assuming that it would be possible, the amount of information presented to the user will be overwhelming, and will obscure, rather than clarify, the

presence of meaningful or interesting patterns. The approach followed, is to study the properties of the dataset ($\mathbf{X}$), possibly huge, in order to extract a subset of a sufficiently smaller cardinality which will either retain as much structural information as possible, or guarantee its preservation up to a predefined threshold. In this approach only the non-redundant objects up to a predefined degree are preserved, thus producing a kernel or core representation of the original dataset. If a similarity measure S is chosen as a redundancy criterium, and a similarity threshold T_s is set forth as a parameter, it is possible to construct a set $\mathbf{L} \subseteq \mathbf{X}$, such that $\forall x \in \mathbf{X}, \exists\ l \in \mathbf{L}, S(x,l) \geq T_s$ (Fig-3). There are efficient algorithms which can generate $\mathbf{L}$-sets at different T_s-levels, and this parameter will determine both the cardinality of the resulting $\mathbf{L}$-set, as well as its semantics. According to this approach, a VR representation of a large or huge dataset is obtained by first extracting an $\mathbf{L}$-set according to a suitable similarity threshold, and then computing its VR space. Since each of the data objects is represented by a sufficiently similar l-object (lower bounded by T_s), the VR space is compliant with the similarity structure of the whole dataset $\mathbf{X}$ at that level.

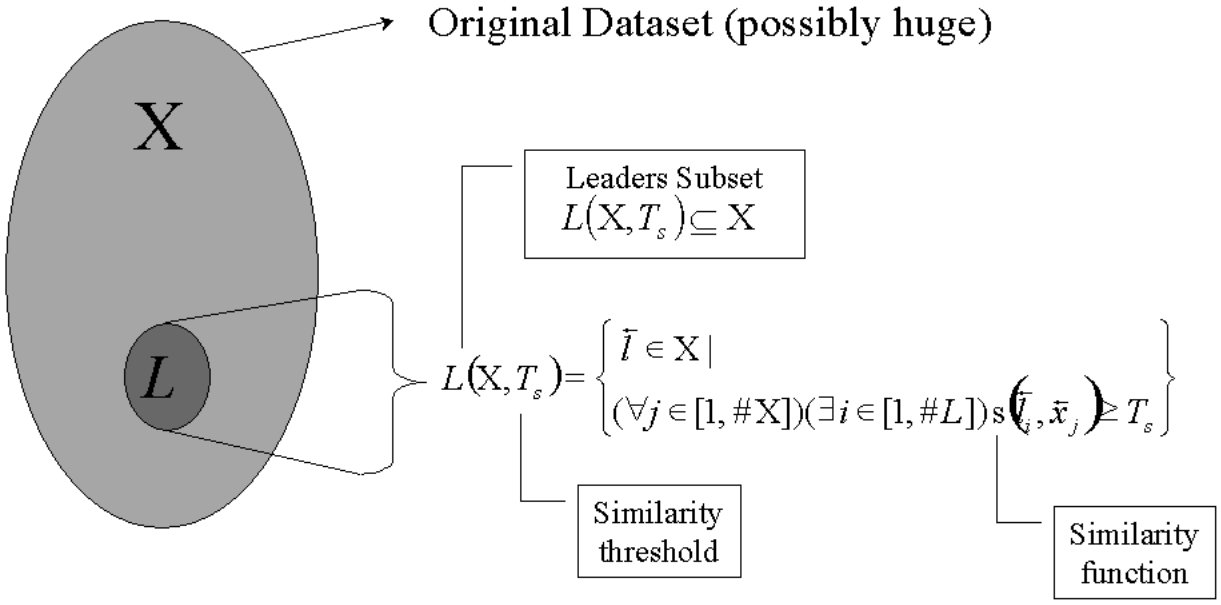

Figure 3. Relation between a dataset $\mathbf{X}$ and its corresponding $\mathbf{L}$-subset at the T_s-similarity level (# denotes set cardinality).

2.5. Unsupervised Virtual Reality Construction: Use of Evolutionary Computation (EC)

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 [21], [22], [20]. 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 [35], [37]. 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 [40], [38].

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 classical methods with well known properties. However, factors like the complexity of the objective function(s) chosen (e.g. f in the definition of the VR space), the dimensionality of the problem (an integer multiple of the data size), and the intrinsic properties of the data themselves, condition highly complex multidimensional error surfaces in the parameter space, with the risk of getting trapped in local extrema. The need to achieve high quality visual representations using virtual reality spaces implies the use of algorithms which explore the search space both globally and locally. A hybrid approach combining evolutionary computation methods or simulated annealing with classical optimization techniques, like Powell, Fletcher-Reeves, Davidon-Fletcher-Powell, Newton, and others is natural. A simple and straightforward way is to start the optimization process with a global optimizer, and at some point, use the current best solution (or the k-best solutions) as initial approximation for a classical optimization technique. A more elaborated strategy would be to alternate cycles of global and local search in parallel. In this respect, EC techniques like genetic algorithms (GA), evolution strategies (ES) and particle swarm optimization (PSO) have a great potential because of their global search capabilities.

Preliminary experiments showed that genetic algorithms and evolution strategies were more affected by the *curse of dimensionality* than particle swarm optimization. This topic deserves a dedicated investigation because the construction of a virtual reality space of dimension m representing N objects implies the optimization of a function depending on $m \cdot N$ parameters. For the typical case of a 3D representation, it means three times the data base size, measured in terms of data objects. The size of current databases and knowledge bases range from hundreds to thousands or even million of objects; posing a major problem.

Particle swarm optimization (PSO) is a population-based stochastic search process, modeled after the social behavior of bird flocks and similar animal collectives [23,24]. The algorithm maintains a population of particles, where each particle represents a potential solution to an optimization problem. In the context of PSO, a swarm refers to a number of potential solutions to the optimization problem, where each potential solution is referred to as a particle. Each particle i maintains information concerning its current position and velocity, as well as its best location overall. These elements are modified as the process evolves, and different strategies have been proposed for updating them, which consider a variety of elements like the intrinsic information (history) of the particle, *cognitive* and *social* factors, the effect of the *neighborhood*, etc, formalized in different ways. The swarm

model used has the form proposed in [43]

$$\begin{aligned} \nu_{id}^{k+1} &= \omega \cdot \nu_{id}^{k} + \phi_1 \cdot (p_{id}^{k} - x_{id}^{k}) + \phi_2 \cdot (p_{gd}^{k} - x_{id}^{k}) \\ x_{id}^{k+1} &= x_{id}^{k} + \nu_{id}^{k+1} \\ \phi_i &= b_i \cdot r_i + d_i, \qquad i = 1, 2 \end{aligned} \tag{3}$$

where ν_{id}^{k+1} is the velocity component along dimension d for particle i at iteration $k+1$, and x_{id}^{k+1} its location; b_1 and b_2 are positive constants both equal to 1.5; r_1 and r_2 are random numbers uniformly distributed in the range $(0, 1)$; d_1 and d_2 are positive constants both equal to 0.5, to cooperate with b_1 and b_2 in order to confine ϕ_1 and ϕ_1 within the interval $(0.5, 2)$; ω is an inertia weight.

2.6. Supervised Virtual Reality Construction: Use of Nonlinear Discriminant Analysis (NDA) Neural Networks

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 [7]). From an unsupervised perspective, the role of l could be to maximize some metric/non-metric structure preservation criteria [5], 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 [37]. Hybrid requirements are also possible.

In the supervised case, a natural choice for representing the l mapping is an NDA neural network [41], [21], [22], [20]. 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-4.

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. A variant in the construction of NDA networks was

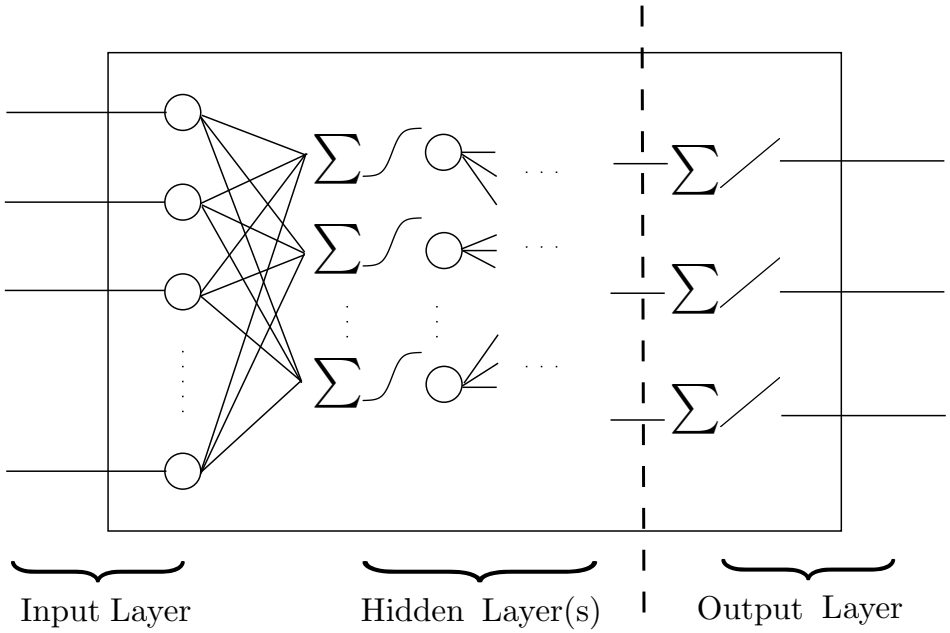

Figure 4. Network Architecture in which the NDA network is learned. $\int$ means nonlinear activation, $/$ linear activation , and Σ aggregation

introduced by using hybrid stochastic-deterministic feed forward networks (SD-FFNN) [39]. The SD-FFNN is a hybrid model where training is based on a combination of simulated annealing with the powerful minima seeking conjugate gradient [27], 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.

2.7. Use of Convex Hulls and α-shapes in the VR Space

Let the set $S \subset \mathbb{R}^d$ contain n points. It may be of interest to know the shape formed by these n points. This is quite a vague notion and there are probably many possible interpretations [9], the α-shape being one of them. The α-shape of S degenerates to the point-set S when $\alpha \to 0$ and the α-shape for $\alpha \to \infty$ is the convex hull of S. Intuitively for small dimensions ($d = 2, 3$), one can think of a convex hull as a piece of elastic cellophane tightly wrapped around all of the points. Such a cellophane boundary ∂S_α at a particular

α value is, for any value $0 \leq \alpha \leq \infty$, a subset of the Delaunay triangulation of S, leading to the conclusion that only faces of the Delaunay triangulation are candidates of the α-shape. [9]

In general, the α-shape, convex hull, etc. may be used to more abstractly represent the shape of groups of objects (having particular properties) in the VR space constructed in either a supervised or unsupervised fashion. For example, one class of objects (e.g. a set of galaxies from the same class) may be more abstractly characterized as a convex hull rather than (or in addition to) individual geometries (such as spheres/cubes/etc) in order to demonstrate class structure of a particular class and/or w.r.t. other abstract characterizations of class structures (e.g. a different set of galaxies all having the same class, but different from that previously mentioned) within a particular VR space representation. In essence, the introduction of a convex hull as a summarization technique for a set of objects, may itself become a geometric object in the VR space.

2.8. Use of Crisp Clustering Method Results as a Source of Heterogeneous Data

Clustering with classical partition methods constructs crisp (non overlapping) subpopulations of objects or attributes. Three algorithms were used in this study: *i)* the Leader algorithm [16] indirectly within the scope of the VR representation, *ii)* Forgy's k-means [2] and *iii)* rough k-means [26].

The leader algorithm operates with a dissimilarity or similarity measure and a preset threshold. A single pass is made through the data objects, assigning each object to the first cluster whose leader (i.e. representative) is close enough to the current object w.r.t. the specified measure and threshold. If no such matching leader is found, then the algorithm will set the current object to be a new leader; forming a new cluster. This technique is fast; however, it has several negative properties. For example, *i)* the first data object always defines a cluster and therefore, appears as a leader, *ii)* the partition formed is not invariant under a permutation of the data objects, and *iii)* the algorithm is biased, as the first clusters tend to be larger than the later ones since they get first chance at "absorbing" each object as it is allocated. Variants of this algorithm with the purpose of reducing bias include: *a)* reversing the order of presentation of a data object to the list of currently formed leaders, and *b)* selecting the absolute best leader found (thus making the object presentation order irrelevant).

The k-means algorithm is actually a family of techniques, where a dissimilarity or similarity measure is supplied, together with an initial partition of the data (e.g. initial partition strategies include: random, the first k objects, k-seed elements, etc). The goal is to alter cluster membership so as to obtain a better partition w.r.t. the measure. Different variants (Forgy's, Jancey's, convergent, and MacQueen's [2]) very often give different partition results. For the purposes of this study, only Forgy's k-means was used.

The classical Forgy's k-means algorithm consists of the following steps: *i)* begin with any desired initial configuration. Go to *ii)* if beginning with a set of seed objects, or go to *iii)* if beginning with a partition of the dataset. *ii)* allocate each object to the cluster with the nearest (most similar) seed object (centroid). The seed objects remain fixed for a full cycle through the entire dataset. *iii)* Compute new centroids of the clusters. *iv)* alternate *ii)* and *iii)* until the process converges (that is, until no objects change their

cluster membership). In Jancey's variant, the first set of cluster seed objects is either given or computed as the centroids of clusters in the initial partition. At all succeeding stages each new seed point is found by reflecting the old one through the new centroid for the cluster. MacQueen's method is composed of the following steps: *i)* take the first k data units as clusters of one member each. *ii)* assign each of the remaining objects to the cluster with the nearest (most similar) centroid. After each assignment, recompute the centroid of the gaining cluster. *iii)* after all objects have been assigned in step *ii)*, take the existing cluster centroids as fixed points and make one more pass through the dataset assigned each object to the nearest (most similar) seed object. A so called convergent k-means is defined by the following steps: *i)* begin with an initial partition like in Forgy's and Jancey's methods (or the output of MacQueen's method). *ii)* take each object in sequence and compute the distances (similarities) to all cluster centroids; if the nearest (most similar) is not that of the object's parent cluster, reassign the object and update the centroids of the losing and gaining clusters. *iii)* repeat steps *ii)* and *iii)* until convergence is achieved (that is, until there is no change in cluster membership).

The leader and the k-means algorithms were used with a similarity measure rather than with a distance. In particular Gower's general coefficient was used [14], where the similarity between objects i and j is given by $S_{ij} = \sum_{k=1}^{p} s_{ijk} / \sum_{k=1}^{p} w_{ijk}$ 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. For quantitative attributes (like the ones of the dataset used in the paper), the scores s_{ijk} are assigned as $s_{ijk} = 1 - |X_{ik} - X_{jk}|/R_k$, where X_{ik} is the value of attribute k for object i (similarly for object j), and R_k is the range of attribute k.

2.9. Use of Fuzzy Clustering Method Results as a Source of Heterogeneous Data

The purpose of unsupervised classification is to construct subgroups or clusters based on the similarity structure between the data objects. This is determined by the attributes used for characterizing the objects, and by a given formal criterium for evaluating the similarity (or dissimilarity). The classical idea of crisp clustering was extended to that of a fuzzy partition by [32], and later on investigated by many others [4], [15], [12], [34]. In a fuzzy partition of n objects into K clusters, the state of clustering is by a nxK matrix $U = (u_{ik})$ where $u_{ik}\epsilon[0,1]$, $i = 1, \cdots, n$; $k = 1, \cdots, K$, and the requirement that $\Sigma_{k=1}^{K} u_{ik} = 1$. The u_{ik} represent the memberships of each data object w.r.t each cluster. Memberships close to unity signify a high degree of similarity between the object and a cluster while memberships close to zero imply little similarity. This approach generalizes the classical crisp partition clustering, as an object may belong entirely to a single cluster or enjoy partial membership in several fuzzy clusters. This is typical for hybrid objects which can not be appropriately described by the classical hard partition clustering.

When constructing fuzzy partitions, a measure of goodness of clustering is given by a sum of generalized within-class dispersion:

$$J_m = \sum_{k=1}^{K} \sum_{i=1}^{n} (u_{ik})^m d(\overleftarrow{x}_i, \overleftarrow{v}_k)^2 \tag{4}$$

where $\overleftarrow{x_i}$ is a vector representing data object i, $\overleftarrow{v_i}$ is a vector representing the centroid of

class k, d is a norm, and the exponent m represents a degree of fuzziness of the cluster. Usual norms are Euclidean, but others could be used as well.

Obtaining a good fuzzy partition imply minimizing 4. The classical algorithm proceeds by obtaining successive approximations by first estimating the centroids $\overleftarrow{v}_{ka} = \frac{\sum_{i=1}^{n} (u_{ik})^m \overleftarrow{x}_{ia}}{\sum_{i=1}^{n} (u_{ik})^m}$, where $a = 1, \cdots, p$ (p is the number of attributes of the data objects). Then, the memberships are approximated according to $u_{ik} = \left[\sum_{j=1}^{K} \left(\frac{d(\overleftarrow{x}_i, \overleftarrow{v}_k)}{d(\overleftarrow{x}_j, \overleftarrow{x}_k)} \right)^{1/m-1} \right]^{-1}$

The problem of the optimality of J_m is a difficult one. The obtained solution might represent a local or a global optimum for the corresponding problem, and usually other measures of cluster validity are used in practice to complement 4. Among them are the partition coefficient F_c, and the entropy H_c of the partition U, given by

$$F_c(U) = \sum_{k=1}^{K} \sum_{i=1}^{n} \frac{(u_{ik})^2}{n} \tag{5}$$

$$H_c(U) = -\sum_{k=1}^{K} \sum_{i=1}^{n} \frac{u_{ik} \ln(u_{ik})}{n} \tag{6}$$

2.10. Use of Rough Sets Results as a Source of Heterogeneous Data

The Rough Set Theory [29] bears on the assumption that in order to define a set, some knowledge about the elements of the data set is needed. This is in contrast to the classical approach where a set is uniquely defined by its elements. In the Rough Set Theory, some elements may be indiscernible from the point of view of the available information and it turns out that vagueness and uncertainty are strongly related to indiscernibility. Within this theory, knowledge is understood to be the ability of characterizing all classes of the classification. More specifically, an information system is a pair $\mathbf{A} = (U, A)$ where U is a non-empty finite set called the universe and A is a non-empty finite set of attributes such that $a : U \rightarrow V_a$ for every $a \in A$. The set V_a is called the value set of a. For example, a decision table is any information system of the form $\mathbf{A} = (U, A \cup \{d\})$, where $d \in A$ is the decision attribute and the elements of A are the condition attributes. For any $B \subseteq A$ an equivalence relation $IND(B)$ defined as $IND(B) = \{(x, x') \in U^2 | \forall a \in B, a(x) = a(x')\}$, is associated. In the Rough Set Theory a pair of precise concepts (called lower and upper approximations) replaces each vague concept; the lower approximation of a concept consists of all objects, which surely belong to the concept, whereas the upper approximation of the concept consists of all objects, which possibly belong to the concept. A *reduct* is a minimal set of attributes $B \subseteq A$ such that $IND(B) = IND(A)$ (i.e. a minimal attribute subset that preserves the partitioning of the universe). The set of all reducts of an information system $\mathbf{A}$ is denoted $RED(A)$. Reduction of knowledge consists of removing superfluous partitions such that the set of elementary categories in the information system is preserved, in particular, w.r.t. those categories induced by the decision attribute. In particular, minimum reducts (those with a small number of attributes), are extremely important, as decision rules can be constructed from them [3].

However, the problem of reduct computation is NP-hard, and several heuristics have been proposed [42].

3. Virtual Reality Examples

Many different possible perspectives of a (particular) data set exist that portray different kinds of information oriented towards the deeper understanding of a particular problem under consideration. For this study, the particular properties and characteristics of a number of data sets are enumerated through visual representations.

It is impossible to illustrate appropriately the look, feel and immersion of a virtual reality 3D environment within the limits imposed by printed paper. Grey level screen snapshots from different application examples are presented only to give a rough idea. The design of the virtual reality spaces was kept simple in terms of the geometries used (G), and in particular, behaviors were excluded ($B = \phi$ in Υ). In all cases the snapshots were simplified w.r.t the information included in the corresponding Υs to avoid information overload. The direct transform between the original space and Υ was found by minimizing Sammon error, with ζ_{ij} given by the euclidean distance in Υ and $\delta_{ij} = (1 - \hat{s}_{ij})/\hat{s}_{ij}$, where $\hat{s}_{ij}$ is Gower's similarity [14].

3.1. Dataset Descriptions

3.1.1. Iris Data

The classical pattern recognition Iris Plants Database from Fisher, 1936. See `ftp://ftp.ics.uci.edu/pub/machine-learning-databases/iris/iris.names`. The data set contains 4 numeric attributes and 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. In particular, the attributes are: *1)* sepal length (cm), *2)* sepal width (cm), *3)* petal length (cm), *4)* petal width (cm), and the classes are *i)* Iris Setosa *ii)* Iris Versicolour, and *iii)* Iris Virginica

3.1.2. Geology Data

In a geological prospecting research project, the problem was to detect the presence of a known underground cave in a test area based on a set of 5 geophysical field measurements and the topographic information. This is a problem with *partially defined* classes (present/unknown, instead of present/absent), as measurements made outside of the area of the known cave don't mean the absence of cave in those places.

3.1.3. Astronomy Data

The science of astronomy has experienced unprecedented progress in the last years. The advances in computer, communication, and observation technologies have increased by many orders of magnitude the quantity and quality of astronomic data. For example, the Sloan Digital Sky Survey (SDSS) will systematically map one-quarter of the entire sky, producing a detailed image of it and determining the positions and absolute brightnesses of more than 100 million celestial objects. It will also measure the distance to a million of the nearest galaxies, giving a three-dimensional picture of the universe through a volume one hundred times larger than that explored to date. The Sky Survey will also record the distances to 100,000 quasars, the most distant objects known, giving an un-

precedented hint at the distribution of matter to the edge of the visible universe. Apache Point Observatory, in Sunspot, New Mexico, is the site of the SDSS telescopes, and is operated by the Astrophysical Research Consortium (ARC), a not-for-profit consortium of seven research institutions whose mission is to develop and operate astronomical research facilities for scientists affiliated with the member institutions and their collaborators (`http://www.sdss.org/`).

Stephanie Juneau, Department of Astronomy, University of Montreal and Dr. Luc Simard, Herzberg Institute of Astrophysics (HIA), National Research Council Canada (NRC) provided a data set based on SDSS data and their own contributions, to the authors. In particular, a morphological analysis of the galaxies was performed by Dr. Luc Simard. The provided data consists of photometry in the *u,g,r,i* and *z* bands for spectroscopy of $374,767$ galaxies covering wavelengths from 3800 to 9200 Angstroms with a target magnitude limit of Petrosian $r < 17.77$. A subsample of $174,947$ galaxies was selected such that it was flux-limited ($14 < r < 18$) with a median redshift of about 0.1. The parameters were chosen to incorporate *i)* intrinsic properties of the galaxies *a)* absolute rather than apparent magnities, *b)* distance scales in kpc rather than arcsec (however, need to choose a cosm. model) and *ii)* physical relevance *a)* morphological properties are linked to galaxy type, and *b)* Observer-dependent quantities (e.g. galaxy inclination) are disregarded. In particular, the 11 atributes are: *1)* g-band absolute magnitude, *2)* r-band absolute magnitude, *3)* bulge fraction in g-band, *4)* bulge fraction in r-band, *5)* bulge effective radius in g-band, *6)* bulge effective radius in r-band, *7)* disk scale length in g-band, *8)* disk scale length in r-band, *9)* half-light radius of the galaxy in g-band, *10)* half-light radius of the galaxy in r-band, and *11)* spectroscopic redshift – cosmic epoch (lookback time).

3.1.4. Alzheimer Data

Alzheimer's disease (AD) is a chronic, progressive, debilitating condition which, along with other neurodegenerative diseases, represents the largest area of unmet need in modern medicine [40]. There is now renewed hope that genomics technologies, particularly gene expression profiling, can contribute significantly to the understanding of the disease. 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 [6], [17], [40], and others. However, in such data sets, patient samples are characterized by thousands of attributes representing the expression intensities of the different genes chosen in the framework of the experiment. They exhibit extremely complex patterns of dependencies, redundancies, noise, etc, making the process of understanding the meaning, role, and importance of the different genes, very difficult. In particular, a study of gene expression Alzheimer's data from a data mining perspective is presented in [40].

The data set was provided by Dr. P.R. Walker from the National Research Council of Canada's Institute for Biological Sciences and is composed of a total of 4 clinically diagnosed AD patients and 5 normal patients of similar age with a total of 23 samples taken from them, each characterized by 9600 genes. In [40] a simple screening algorithm was used with the purpose of finding individual relevant genes from the point of view of their ability to differentiate the class of samples having Alzheimer's disease from the normal ones. The idea of the procedure is to analyze each gene individually and determine

the threshold intensity value which dichotomizes the range of intensity values of the analyzed gene in order to maximize the conditional probability of the class. After the screening process, four genes were individually able to partition the data with perfect coincidence between the known classes and those induced by the dichotomization using the threshold values found. Accordingly, a new data set was defined containing all of the objects, but described in terms of only the four best genes found.

3.1.5. 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 [13] 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.

3.1.6. Breast Cancer Data

A public data set was chosen at random, with a slight bias towards larger numbers of samples, from the Gene Expression Omnibus (GEO) (See `http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=360`) in order to conduct a series of, successively more complex, and complementary, visualizations. The breast cancer data selected [8] consists of 24 core biopsies taken from patients found to be resistant (greater than 25% residual tumor volume, of which there are 14) or sensitive (less than 25% residual tumor volume, of which there are 10) to docetaxel treatment. The number of genes (probes) placed onto (and measured from) the microarry is $12,626$.

3.2. Example VR Space including Convex Hulls using Geology data

The Geology data was used in Fig-5, which shows the extended information system in Υ, where Γ^v on O is given by convex hulls defining the boundary of the corresponding classes. The one belonging to the Cave class, also contains unknown objects, thus making them likely candidates to have caves beneath their locations. Objects far from the Cave class hull are less likely to be located in cave areas. In fact, posterior drilling in the area confirmed these predictions.

3.3. Example of Inverse Transformation using Iris data

The use of the inverse transformation with the classical Iris data is illustrated in Fig-6. Iris Setosa and Iris Versicolour are two species well differentiated according to the sample objects. However, if there would be an *unknown hybrid* specimen between the two, its location in Υ should be somewhere in between the two corresponding clouds (objects 24 and 58 from the two classes are shown as reference. The inverse transform was computed with a neural network trained over the implicit direct transform and applied

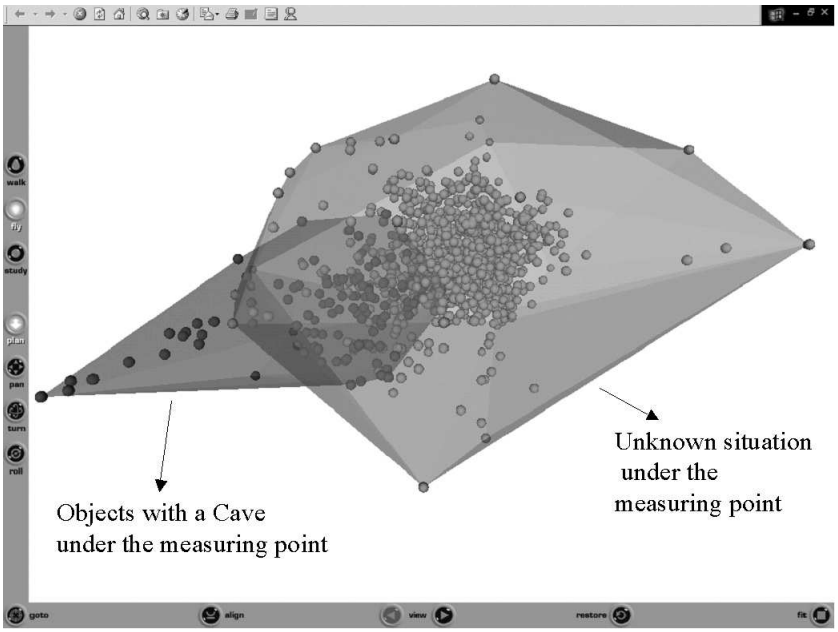

Figure 5. Geophysical prospecting for caves. The arrows and the text labels were added for explanatory purposes (see text). Γ^v in Υ is given by convex hulls associated with the classes. Several objects from the unknown class are contained within the hull of the Cave class.

to the theoretically conjectured object. The reconstructed original attributes were: petal length = 5.0, petal width = 2.9 sepal length = 2.4, sepal width = 0.7. It is easy to imagine the potential applications of this method in many domains.

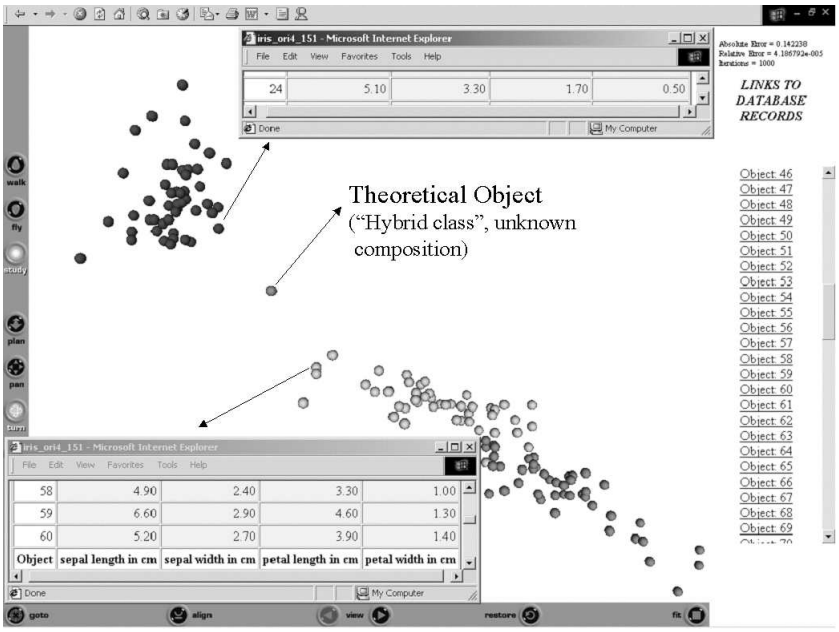

Figure 6. Iris data. Spheres are data objects and grey tones represent different classes. The arrows and the text label at the center of the screen were added for explanatory purposes (see text). The two embedded web pages are part of the virtual world and show the data objects in the original space.

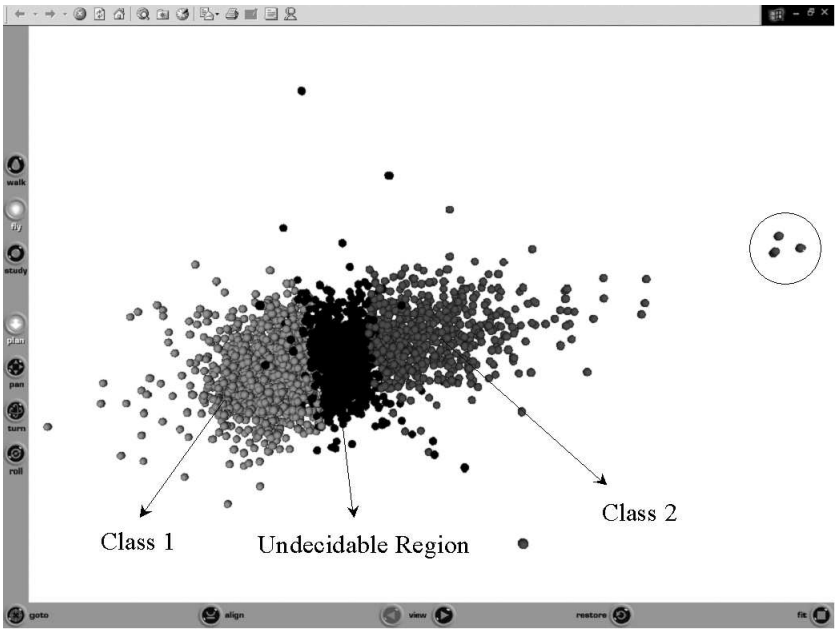

Figure 7. Rough K-means clustering on Gene Expression data showing the undecidable (rough) region between two classes. The arrows, the text labels and the circle (extreme right) were added for explanatory purposes (see text).

3.4. Example of a VR Space Containing Data Mining (Rough Clustering) Results

For genomic research in neurology, time-varying expression data for 2611 genes in 8 times were measured. Fig-7 shows the representation in Υ of the information system as well as the result of a previous rough k-means clustering experiment [26]. Besides showing that there is no *major* class structure in this data, the undecidable region between the two enforced classes is perfectly appreciated in Υ. The rough clustering parameters were $k = 2, \omega_{lower} = 0.9, \omega_{upper} = 0.1$ and $threshold = 1$. The small cluster encircled at the upper right, contains a set of genes actually discovered when examining the virtual reality world and was considered very interesting by the domain experts. This pattern is clearly identifiable in Υ but remained undiscovered since it was masked by the clustering procedure (its objects were assigned to the nearby bigger and unnatural cluster).

3.5. Example of Joint Representation of Information Systems and Decision Rules

The extraction of symbolic relations from information systems in the form of decision rules is one of the most important techniques in data mining and the knowledge discovery process. In this case, the information systems are of the form $S =< U, A \bigcup \{d\} >$, $S_r =< R, A \bigcup \{d\} >$ (for the rules), where $\{d\}$ is the decision attribute (all attributes are nominal, i.e. for all $a \in A \bigcup \{d\}, V_a \subset \mathbb{N}^+$). Decision rules are of the form , $\bigwedge_{i=1}^{p}(A_{\tau_i} = v_{\eta_i}^{\tau_i}) \rightarrow (d = v^d{}_j)$, where the $A_{\tau_i} \subseteq A$, the $v_{\eta_i}^{\tau_i} \in V_{\tau_i}$ and $v^d{}_j \in V_d$ (i.e. the antecedent is an elementary

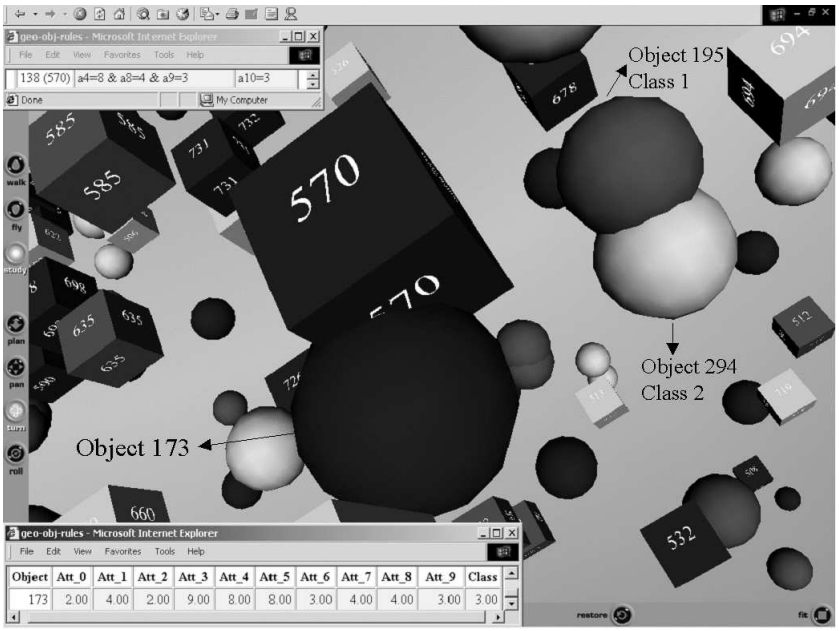

Figure 8. Simultaneous representation of data objects and decision rules (*geo* data set). Spheres are data objects and numbered cubes, decision rules. The arrows and the text labels they point to, were added for explanatory purposes (see text). The two embedded web pages are part of the virtual world and show the data objects and the rules in the original space.

conjunction of attribute-value pairs of length $p \in [1, card(A)]$, and the succedent is a pair associated with the decision attribute). The $\hat{s}_{ij}$ used for δ_{ij} in A, was given by: $\hat{s}_{ij} = \frac{1}{\sum_{a \in \breve{A}} \omega_{ij}} \sum_{a \in \breve{A}} (\omega_{ij} \cdot s_{ij})$, where: $\breve{A} = A^u$ if $i, j \in U$, A^r if $i, j \in R$ and $A^u \bigcap A^r$ if $i \in U$ and $j \in R$. The s, ω functions are defined as: $s_{ij} = 1$ if $f_a(i) = f_a(j)$ and 0 otherwise, $\omega_{ij} = 1$ if $f_a(i), f_a(j) \neq ?$, and 0 otherwise.

The example presented is the *geo* data set, distributed with the RSL library for rough set analysis [31]. It contains 432 objects and 11 attributes, all nominal. The last attribute was considered the decision attribute and the *very fast* strategy was used to generate the rules (328 were obtained and they classify the data set with 99% accuracy). The representation in Υ of $S \bigcup S_r$ is shown in Fig-8 ($G = \{spheres, cubes\}$ for objects and rules respectively). According to RSL results, Rule 570 is supported by object 173, and they appear very close in Υ. Also, objects 195 and 294 are very similar (although belonging to different classes) and they appear very close in Υ.

3.6. Example of Unsupervised VR Space creation using hybrid classical and Evolutionary Computation

In order to study the behavior of hybrid optimization when constructing virtual reality spaces for gene expression data, two kinds of experiments were made using two data

sets derived from Alzheimer Data. The experimental plan was: *i)* to apply the classical methods to the two data sets, and *ii)* for each of the data sets, to use the Particle Swarm Optimization (PSO) final result as an initial approximation for the classical methods, as a hybrid algorithm.

The classical techniques used in the study were: Powell, Fletcher-Reeves, and Davidon-Fletcher-Powell [28]. The Powell method does not require the partial derivatives of the objective function. In the first phase, the four classical optimization methods were applied to each data set, using 100 different random initial approximations for a total of 400 runs per data set. The same set of seeds was used in all experiments in order to ensure comparability.

From the point of view of constructing the virtual reality spaces, the f transform sought was one minimizing Sammon's error (1), with 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 [14]. The Euclidean distance was the measure used as ζ_{ij} in the VR space (Υ).

An implicit representation was computed via deterministic optimization with a gradient descent technique (Newton's method). Several solutions were found using different initial approximations, and for comparison purposes, the same seeds were used.

In this study, a set of experiments using PSO was set forth with the following parameters: number of particles = 100, maximum velocity = $\{0.1, 0.15, 0.20, 0.25, 0.30\}$, initial weight=$\{0.1, 0.2, 0.4, 0.6, 0.8, 0.9\}$, final weight = $\{0.1, 0.2, 0.4, 0.6, 0.8, 0.9\}$, and number of iterations = $\{1000, 2000, 4000, 8000\}$, for a total of 720 experiments. The results obtained with classical optimization (CO), PSO and the hybrid algorithm PSO+CO are shown in Table-1.

A snapshot of part of a VR space resulting from a hybrid algorithm applied to the original Alzheimer's data using all 9600 genes, is shown in Fig-9.

The samples corresponding to the Alzheimer's class are colored black, and the non-Alzheimer's with a light color. In this case it is clearly seen that the samples corresponding to the Alzheimer's class appear as more homogeneous and compact (i.e. more similar to each other) than those from the non-Alzheimer class. However, the Alzheimer class appears *wrapped* by the non-Alzheimer class, which is more irregular, indicating that the classes are not linearly separable.

In the case of the Alzheimer data with 9600 genes, the classical methods and the PSO gave very similar results. However, the error range for the hybrid algorithm is smaller than that of the classical methods. This indicates that high quality VR spaces can be obtained with relative independence of the classical method used by the hybrid algorithm (Table 1).

From the point of view of the PSO algorithm alone, the set of 100-best solutions is upper-bounded by an error of 0.0344337, which is better than many of those obtained with the classical methods . Within this set, 40% of the solutions were found using 4000 iterations, and the remaining 60% with 8000 iterations, indicating the overall computational effort involved. From the 100-best solutions, 20% of them were found with an increasing weight updating scheme, and in 60% with a decreasing one. In 20% of the solutions, the weight was constant. The distribution of maximum particle velocities within the 100-best was: $\{0.1 : 10\%, 0.15 : 10\%, 0.20 : 40\%, 0.25 : 40\%, 0.30 : 0\%\}$. It suggests that non-extreme velocities are better for this kind of problem.

Table 1
Experiments with Alzheimer data sets: Comparison of the classical methods with PSO and the hybrid algorithm formed by combining PSO with each of the classical methods. Data objects are described in terms of the original 9600 original genes, and in the four best discovered by a data mining procedure. Error ranges are shown in square brackets

	9600 Genes Sammon Error	4 Best Genes Sammon Error
Classical Method	[0.03437583, 0.03868285]	[0.0690399, 0.06936745]
Powell	[0.03437595, 0.03868285]	[0.06903999, 0.06933347]
Fletcher-Reeves	[0.03437584, 0.03694771]	[0.06903994, 0.06936745]
Davidon-Fletcher-Powell	[0.03437583, 0.03690736]	[0.0690399, 0.06936724]
PSO	0.034376	0.0690399
	9600 Genes Sammon Error	4 Best Genes Sammon Error
Hybrid Method	[0.03437583, 0.03437589]	[0.0690399, 0.0690399]
PSO + Powell	0.0343758948	0.0690398961
PSO + Fletcher-Reeves	0.0343758576	0.0690398961
PSO + Davidon-Fletcher-Powell	0.0343758315	0.0690398961

From the practical point of view, the results obtained by PSO and the classical methods can be considered equivalent. This suggests that the phase of global search could have been shortened without loosing performance in terms of the error measure. In this case, a reduction of the computational load of the hybrid algorithm could be achievable. In other words, the proportion between global and local search in this problem could be determined such that the same final solutions can be found at smaller computational cost. This topic deserves further investigation.

After a data mining process described in [40], a set of four most relevant genes was found. As in the previous case, a VR space representation was computed by the classical methods, PSO, and the hybrid algorithm. The settings for the PSO experiments were

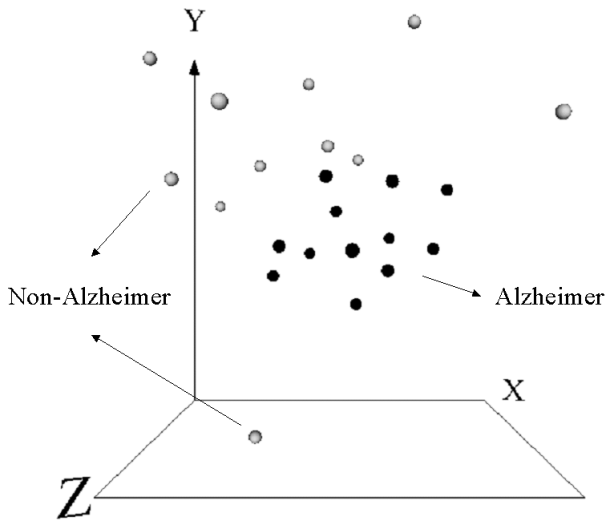

Figure 9. Alzheimer Data (9600 genes): VR space representing the data when object locations are resulting from a hybrid optimization using the best PSO result as initial approximation to Powell's method. X, Y, and Z are the axis of the VR space resulting from the nonlinear transformation of the original 9600 attributes (the genes), performed by the implicit function minimizing the error measure.

the same as with the Alzheimer data set using the original 9600 genes, for a total of 720 experiments. The results are shown in Table-1.

A snapshot of the VR space is shown in Fig-10. Now it appears completely polarized, with the Alzheimer and non-Alzheimer classes appearing as distant clouds occupying well separated half-spaces. The two classes are now well differentiated entities, also linearly separable.

The behavior of the classical methods, the PSO and the hybrid algorithms was almost identical, thus suggesting that the best solution found possibly corresponds to the global optimum. In this case, the structure of the dissimilarity matrix in the original 4D space had a decisive effect. As a result of the data mining procedure, the selected genes lead to a better distribution of the dissimilarity values within the matrix. This is a consequence of a better expressed class structure inherent to the data objects.

From the point of view of the PSO performance, the error level for the 100 best particles is 0.06904022, which is under those of the GD. Within this subset, 18% of the solutions were obtained with 2000 iterations, 43% with 4000 and the remaining 39% with 8000. Within the 100-best solutions, 40% were found to be obtained with an increasing weight updating scheme, and 60% with a decreasing one. None of the solutions in the subset were found using a constant weight. For this problem, the frequency of good solutions obtained with decreasing weight updating is larger than that obtained with an increasing scheme. This behavior differs from the one reported in [43] for several of DeJong's functions and deserves further studies. The distribution of maximum particle velocities within the 100-

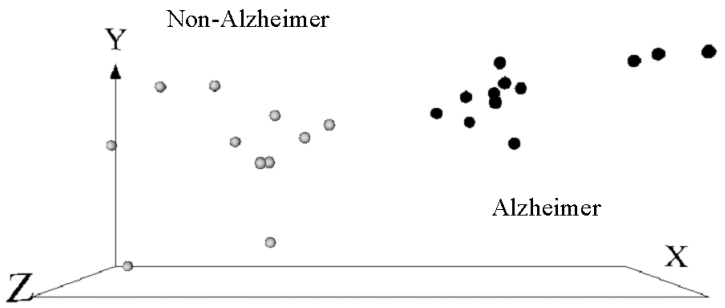

Figure 10. Alzheimer Data (4 best genes): VR space representing the data. The object locations are the result of Particle Swarm Optimization alone (in practical terms the classical methods couldn't improve the pure PSO result). X, Y, and Z are the axis of the VR space resulting from the error minimization.

best was: $\{0.1 : 20\%, 0.15 : 13\%, 0.20 : 27\%, 0.25 : 0\%, 0.30 : 40\%\}$. This behavior differs from the one found in the case of the 9600 genes data set. A multimodal frequency distribution is a curious behavior in this case, but the number of velocity values considered was very limited.

3.7. Example of VR Space creation using Nonlinear Discriminant Analysis

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.

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

Table 2
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

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/`) .

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 [14] was used, with Euclidean distance set as the measure used as dissimilarity in the VR space.

In [38], 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-11 for the 7129 case and Fig-12 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-11, and around $(10^{-7}, 10^{-7})$ in Fig-12.

Table-3 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 [33] (Fig-13) 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-14. 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-15 shows a class differentiated structure. The NDA counterpart sharing the same training/test and test MSE properties (Fig-16), exhibits an even clearer differentiation.

Table 3
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]

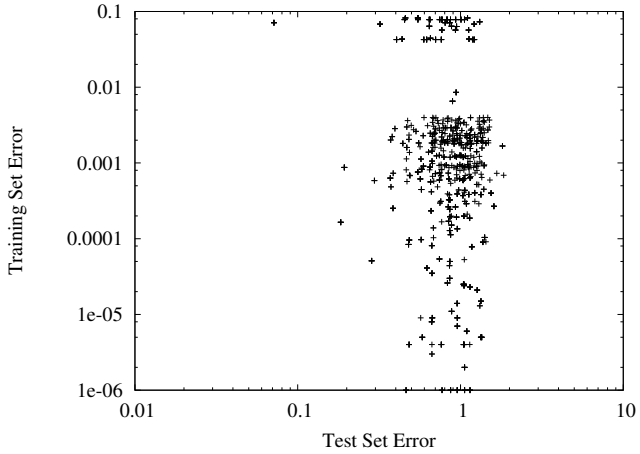

Figure 11. Leukemia: Mean squared errors for 1600 runs, each with 38 train and 34 test samples, respectively. Samples described in terms of 7129 genes.

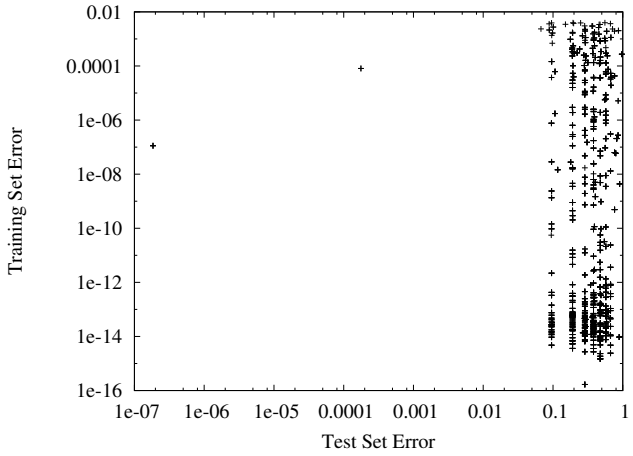

Figure 12. Leukemia: Mean squared errors for 1600 runs, each with 38 train and 34 test samples, respectively. Samples described in terms of 7 selected genes.

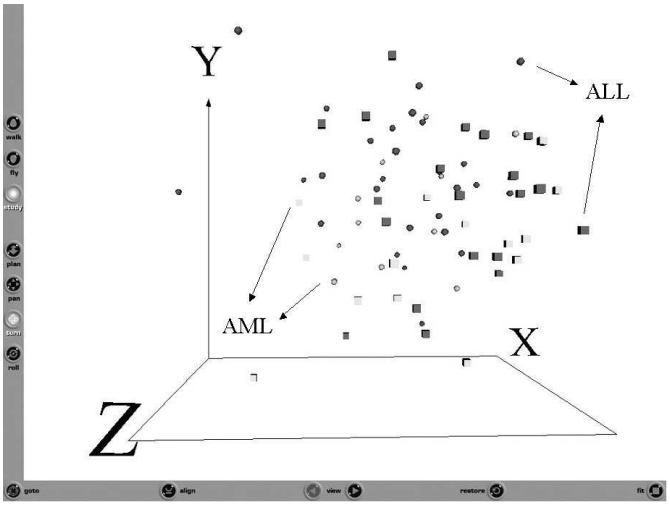

Figure 13. Leukemia: Unsupervised (Sammon) representation of the original training and test data, in terms of 7129 genes. Dark objects= ALL, Light objects=AML. Spheres = training, Cubes = test. Sammon error = 0.143.

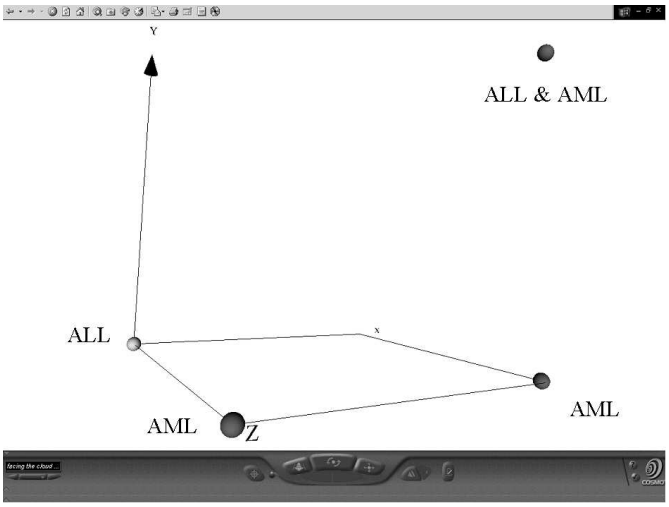

Figure 14. Leukemia: Supervised (NDA) representation of the original training and test data, in terms of 7129 genes. Dark objects= ALL, Light objects=AML. Training error = 0.0710, test error = 0.0715.

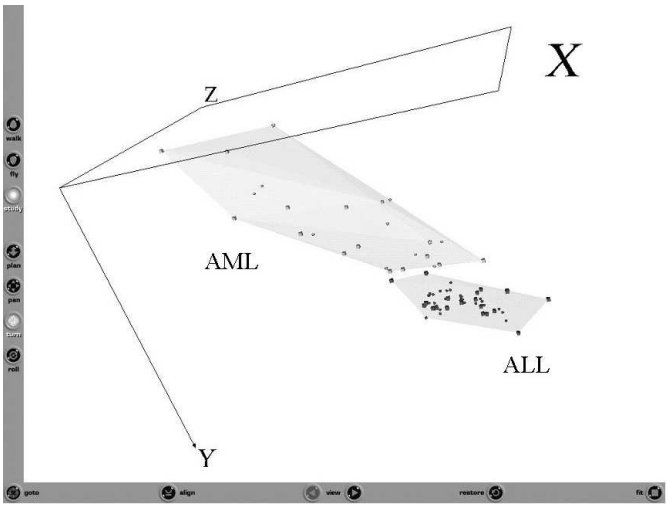

Figure 15. Leukemia: Unsupervised (Sammon) representation of the original training and test data, in terms of 7 selected genes. Convex hulls wrap the classes. Dark objects= ALL, Light objects=AML. Spheres = training, Cubes = test. Sammon error = 0.103.

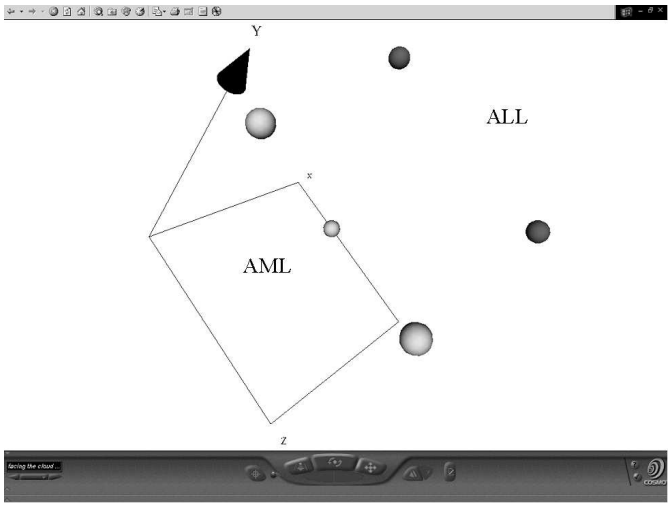

Figure 16. Leukemia: Supervised (NDA) representation of the original training and test data, in terms of 7 selected genes. Dark objects= ALL, Light objects=AML. Training error = 1.1236 10^{-07}, test error = 1.8445 10^{-07}.

Despite such a high dimensionality in the original space, an unsupervised VR representation with low Sammon error (Fig-17), 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 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-18.

The data mining procedures applied in [40] 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-19 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-20, where the results of applying the worst NDA network are shown. Again, there is a total class differentiation.

3.8. Example of Data Structure Visualization using Breast Cancer Data

A breast cancer data set has at least two immediate perspectives on its structure for a given point in time. It may be viewed from the point of view of the similarity between samples, or from the (transposed) point of view of the similarity between genes (probes). Fig-21 represents the 24 breast cancer samples in a 3-D virtual world. Each point in the 3-D space represents a breast cancer sample and was mapped from the original $12,625$ dimensional gene expression space. No supervisory information was used to construct the space (i.e. whether a sample is resistant or sensitive), but that information is available through a coloring of the spheres; where black spheres represent sensitive samples, and

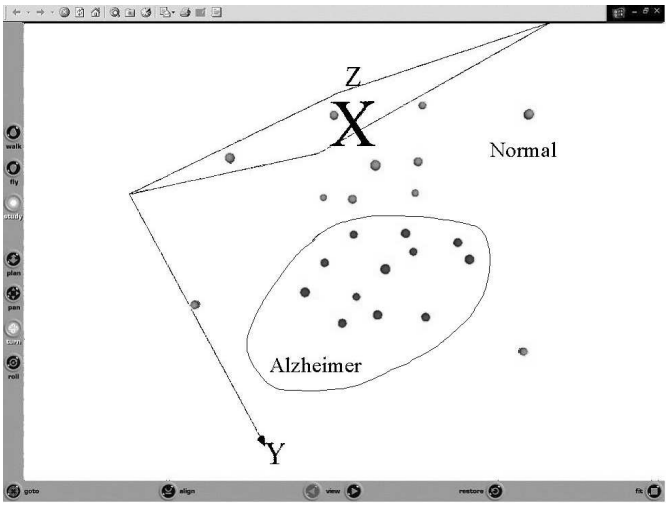

Figure 17. Alzheimer: Unsupervised (Sammon) representation of the original training and test data, in terms of 9600 genes. Dark objects= ALL, Light objects=AML. A boundary delimiting the Alzheimer class was added for clarity. Sammon error = 0.103.

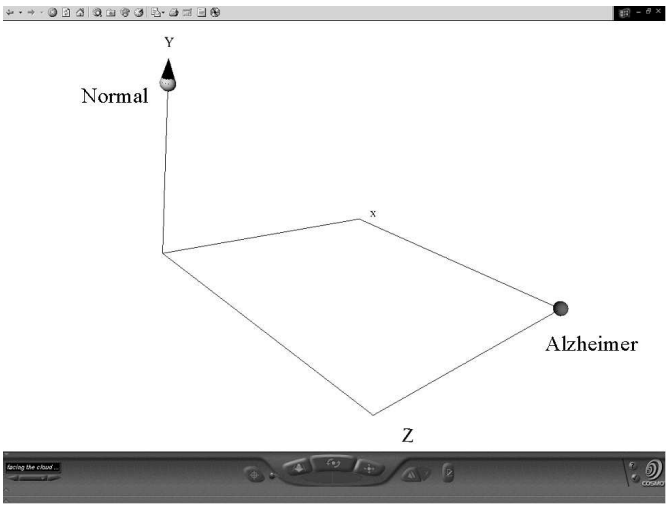

Figure 18. Alzheimer: Supervised (NDA) representation of the original data, in terms of 9600 genes. Dark objects = Alzheimer, Light objects = Normal. Training error = $2.4581\ 10^{-1}$ (the worst of 1600 networks).

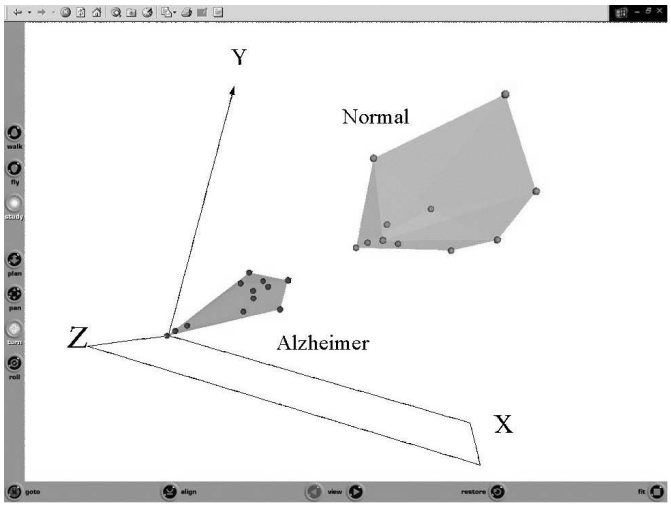

Figure 19. Alzheimer: Unsupervised (Sammon) representation of the original training and test data, in terms of 4 selected genes. Convex hulls wrap the classes. Dark objects = Alzheimer, Light objects = Normal. Sammon error = 0.002.

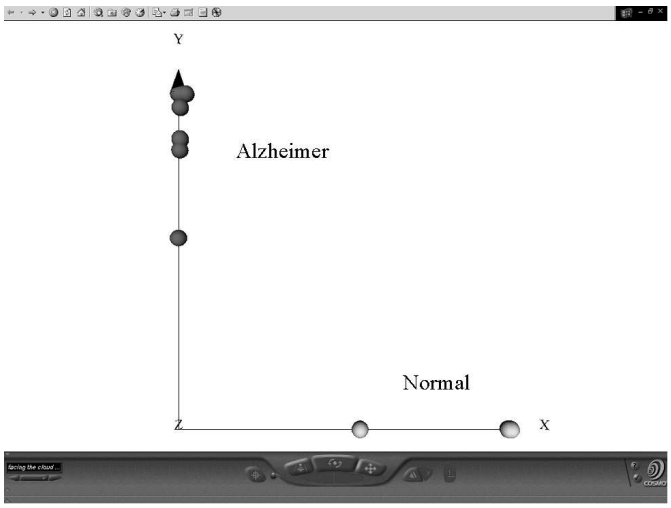

Figure 20. Alzheimer: Supervised (NDA) representation of the original data, in terms of 4 selected genes. Dark objects = Alzheimer, Light objects = Normal. Training error = 3.7416 10^{-3} (the worst of 1600 networks).

grey spheres represent resistant. Fig-22 demonstrates another perspective, with the introduction of geometries; where cubes represent resistant samples and spheres represent sensitive samples. It is interesting to notice that there is one resistant sample in between two sensitive samples, and one sensitive sample in between two resistant samples; suggesting that these samples may not have correct class labels.

For the other view point (that of viewing similarity between genes) Fig-23 plots 361 points in 3-D virtual space. Each point represents a set of points that are mapped to it from the original $12,625$ point gene space in 24 dimensions. For example, the large sphere in Fig-23 represents $10,973$ very similar (0.95 threshold) genes as measured over the 24 samples provided in the breast cancer data set downloaded. Whereas, the furthest point from that large sphere contains only the single gene 31962_at, meaning that it is very different in terms of similarity (when looking at all of the 24 samples) from most (as there are a few points near it) of the other $12,625$ genes.

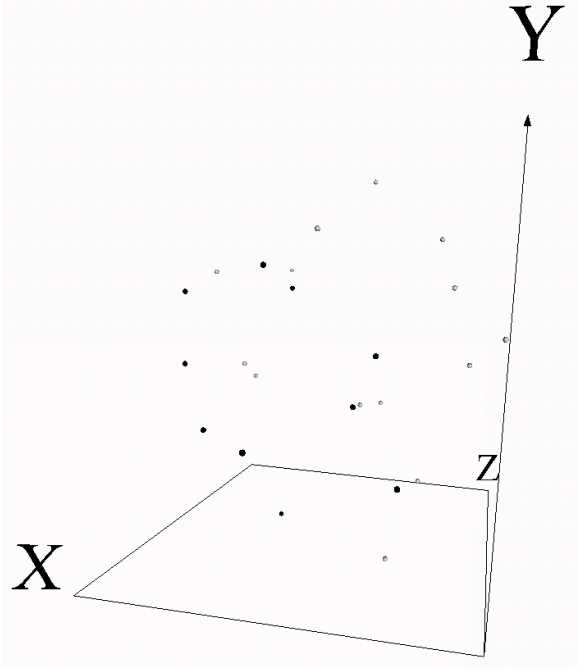

Figure 21. Visual representation (3 dimensions) of 24 breast cancer samples; each containing $12,625$ genes. Absolute Error $= 7.33 \cdot 10^{-2}$. Relative Mapping Error $= 1.22 \cdot 10^{-4}$

3.9. Example of Domain Knowledge Visualization using Breast Cancer Data

The breast cancer data has domain knowledge embedded in the form of sample classes. In particular, the two classes are *sensitive* and *resistant*. In order to more fully understand the class structure based on this domain knowledge, convex hulls can be added to the virtual representation as in Fig-24. The class structures now become much clearer. In particular, the distribution of samples within a class may be observed. For example,

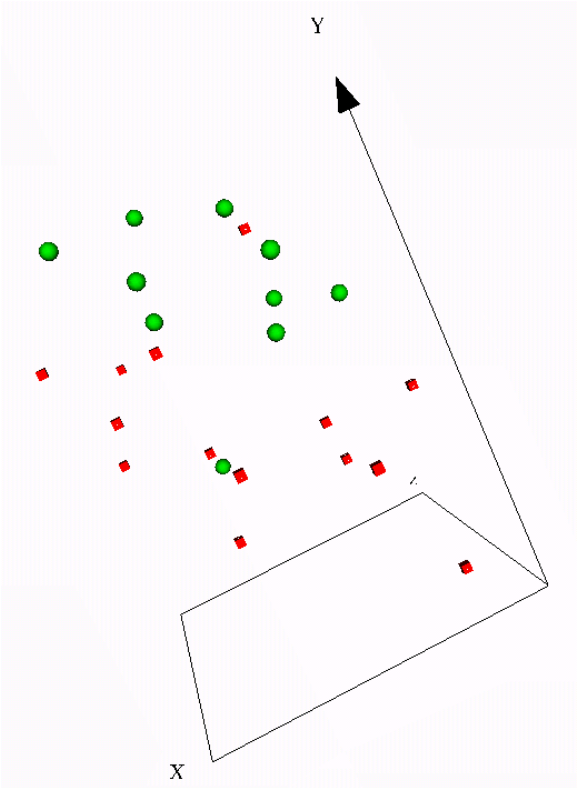

Figure 22. Visual representation (3 dimensions) of 24 breast cancer samples; each containing 12,625 genes. Spheres represent *sensitive* samples. Cubes represent *resistant* samples. Absolute Error = $2.59 \cdot 10^{-2}$. Relative Mapping Error = $2.36 \cdot 10^{-3}$

the larger *resistant* class contains samples that seem to be equally spaced throughout the convex hull, while the smaller *sensitive* class has a markedly different shape, both in terms of size, and distribution. At a more abstract level (that of class structure) the two classes can be observed to touch. When manipulated in the virtual environment (which is difficult to show on printed paper), it becomes clear that one sample from the *sensitive* class lies just inside that of the *resistant* classes' convex hull.

3.10. Example of Crisp Clustering Results Visualization using Breast Cancer Data

Two data mining clustering algorithms were selected in order to generate class information from the breast cancer data. The algorithms were not the subject of this study per se, but rather the results generated by the computational procedures were given to the visualization system in order to investigate the effectivity of the generated visual representations on understandability. In particular, the visualization system is used as a tool for the explicit demonstration of the clustering results and how those results relate to the underlying data structure. One investigation of the possibility of discovering a subset of genes that may be able to discriminate between the samples better than using the full 12,625 genes (as in this study) is [38].

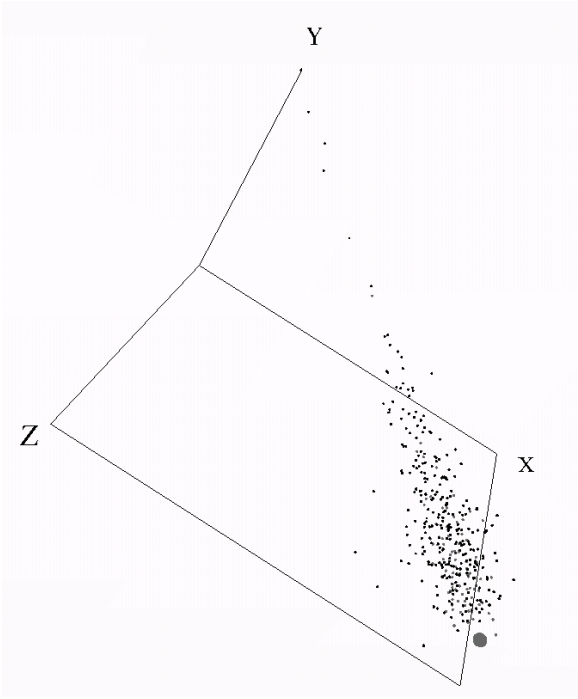

Figure 23. Visual representation (3 dimensions) of 12,625 breast cancer genes from 24 samples. Points (genes) reduced to 361 via a similarity threshold of 0.95. Large sphere represents 10,973 genes. Absolute Error = $9.90 \cdot 10^{-2}$. Relative Mapping Error = $1.39 \cdot 10^{-4}$

Forgy's k-means algorithm [2] was used to cluster the samples into 2 groups, simulating the situation when no class information is available. That is, the breast cancer sample labels were hidden from the clustering algorithm, forcing the algorithm to perform unsupervised clustering in order to discover the labels based on the data. Fig-25 can be seen to have 2 very distinct groups, as requested. Each group has the same number of individuals (12) and both are shaped approximately like ellipsoids. When the virtual world is explored, it is seen that for the upper class, the top sample point is sample GSM4903 (*sensitive*) while the bottom sample point (of the same class) is GSM4918 (*resistant*), indicating that for this data and algorithm parameters, incorrect classifications were made. Reinforcing the understanding that k-means assumes the data is based on hyperspheres when Euclidean distance is used, which is an assumption that the underlying data structure may not support.

A rough set k-means algorithm [26] was used to cluster the samples into 2 groups in a similar fashion as the Forgy's k-means algorithm. Fig-26 can be seen to have 2 classes, as requested. However, one class contains 5 objects and the other class contains the rest. When the small class is investigated, it, like the k-means case above, also does not contain samples from the same (either *resistant* or *sensitive*) class. It is interesting, that for this data set, and the particular algorithm parameters ($w_{lower} = 0.9$, $w_{upper} = 0.1$, $distanceThreshold = 1$) no boundary cases are reported.

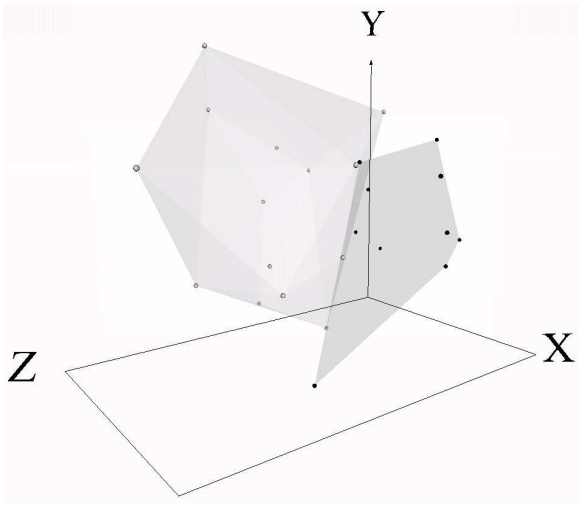

Figure 24. Visual representation (3 dimensions) of 24 breast cancer samples with 12,625 genes. Convex hulls wrap the *resistant*(size= 14) and *sensitive*(size= 10) classes. Absolute Error = $7.33 \cdot 10^{-2}$. Relative Mapping Error = $1.22 \cdot 10^{-4}$

3.11. Example of Fuzzy Clustering Results Visualization using Alzheimer Data

A snapshot of the representation of the resulting Alzheimer's data corresponding to a two-cluster solution with $m = 4$, and partition coefficient equal to 0.571556 is shown in Fig-27. In this representation the set of geometries of the virtual reality space was given by $G = sphere, cone, cube$. The spheres and the cones were used for representing the crisp relation defined by the decision class (Alzheimer vs. non-Alzheimer), whereas the cubes were used for indicating the location of the centroid objects of the two fuzzy classes. The colour used for displaying the images of each data object in the virtual reality space (or the grey level in the snapshots), was used for representing the membership matrix of the fuzzy partition U.

The crisp partition defining the Alzheimer and non-Alzheimer classes is associated with the centroids of the corresponding classes and are displayed with pure white in the case of the Alzheimer class, and pure black for the non-Alzheimer class. Thus, for each data object, its colour (grey level tone) was computed by a convex combination of the extreme colours black and white using the membership's value as its coefficients.

3.12. Example of Supervised Visualization using Breast Cancer Data

In order to take advantage of the domain knowledge contained within the breast cancer data set in terms of the class information (*resistant*, *sensitive*) a supervised visualization based on nonlinear discriminant analysis (NDA) [39] was performed. All 12,625 genes for each of the 24 samples along with supervisory class information were presented into the multi-layer feedforward neural network and its associated training mechanisms. The purpose was to use the complex neural network training (based on simulated annealing and conjugant gradient) in order to generate a 3 dimensional space (the neural network used as the mapping function) for the virtual world. The resultant supervised visualiza-

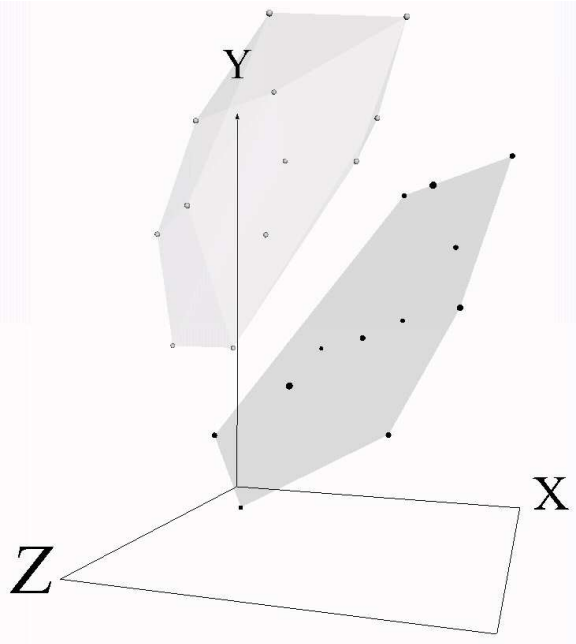

Figure 25. Visual representation (3 dimensions) of 24 breast cancer samples with 12,625 genes. Convex hulls wrap the *C1*(size= 12) and *C2*(size= 12) classes discovered by k-means. Absolute Error = $7.06 \cdot 10^{-2}$. Relative Mapping Error = $5.21 \cdot 10^{-5}$

tion is presented in Fig-28. Each sphere represents a set of breast cancer samples. In particular, 5 points are plotted in the visualization, which correspond to the 24 samples. The discrepancy occurs because some spheres represent more than one sample. For those multi-sample spheres, the colors presented in the figure, represent the class of the first sample that was placed into that sphere according to the leader clustering algorithm. All spheres except one contain homogeneous information. The exception is the lower right hand sphere, which contains objects 9 (*GSM4912*) and 18 (*GSM4914*). This exception indicates that the supervised algorithm tried to push the two classes as far apart as possible, but was not able to separate these two objects based on the combination of all 12,625 genes; an extremely high dimensional problem. This may indicate that these two samples are indistinguishable using all 12,625 genes, and that some form of preprocessing would be required (none was performed in this study).

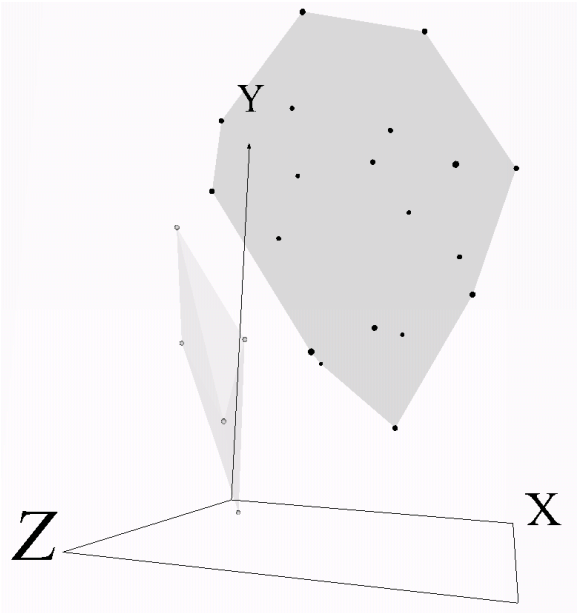

Figure 26. Visual representation (3 dimensions) of 24 breast cancer samples with 12,625 genes. Convex hulls wrap the *RC1*(size= 19) and *RC2*(size= 5) classes discovered by rough set based k-means. Absolute Error = $7.06 \cdot 10^{-2}$. Relative Mapping Error = $5.21 \cdot 10^{-5}$

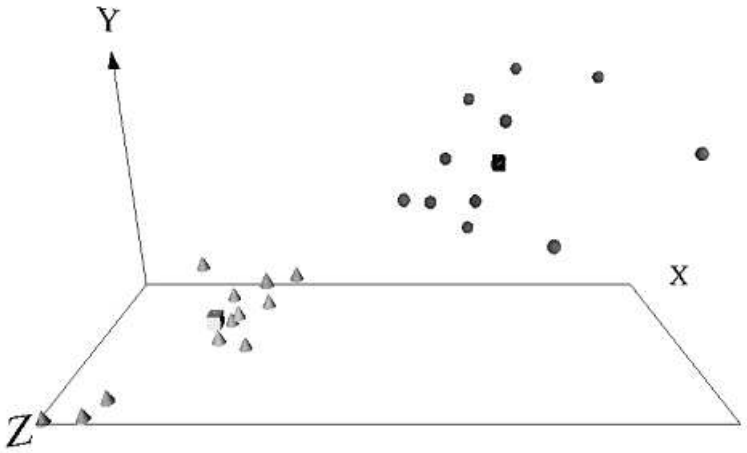

Figure 27. Snapshot of part of the virtual reality representation of the Alzheimer's data (with four selected genes). The cones represents the samples from the Alzheimer class, and the spheres the samples from the non-Alzheimer class. The cubes are the centroids of the corresponding classes (pure white for the Alzheimer class, and pure black for the non-Alzheimer). The grey level with which each object is represented is proportional to the fuzzy membership values w.r.t the two classes. The Sammon error of the overall space is 0.0651.

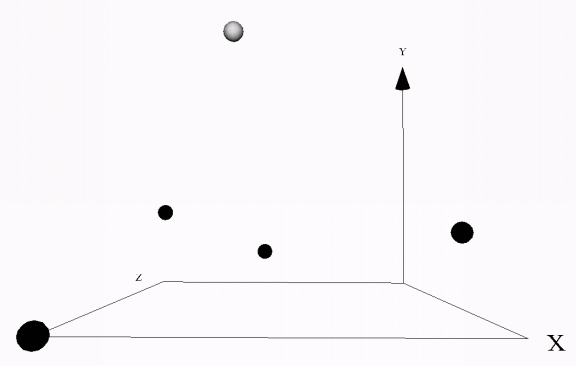

Figure 28. Supervised Visual representation (3 dimensions) of 24 breast cancer samples (5 explicitly shown) with $12,625$ genes as generated by NDA. All spheres are homogeneous except for lower right, containing one of both *sensitive* and *resistant* samples. Sphere colors indicate the class of the first object. Black is *resistant*; white is *sensitive*. Absolute Error = 2.999. Relative Mapping Error = $1.44 \cdot 10^{-9}$

3.13. Example of Unsupervised Visualization using Astronomy Data

An unsupervised VR space was created for a large data set in astronomy containing 174, 947 galaxies (Fig-29). There are some objects in the VR space that are substantially larger with respect to the other objects; indicating a greater number of galaxies represented. In particular, the largest sphere represents 16, 179 galaxies. It can also be observed that some objects appear to have an inherent clustering relationship, as can be seen in Fig-29 in terms of groups of objects in the VR space. Not only do groups of objects exist, but the distribution of the objects w.r.t. each other within the groups are non-random and not all of the same kind. For example, some groups seem to form a linear alignment, while others form a J-shape or ellipsoidal conglomeration. Furthermore, possible interesting objects lie on the periphery (outermost regions) of the VR space potentially indicative of anomalous, rare, surprising, or exceptional values.

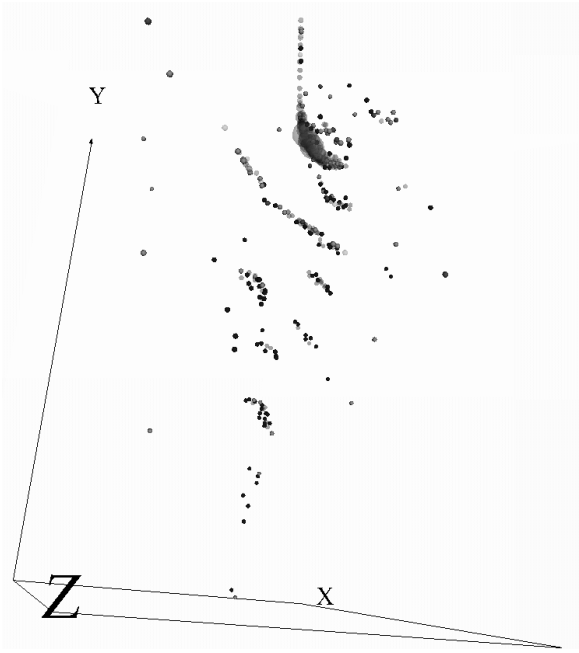

Figure 29. VR space containing 393 sphere objects was constructed in an unsupervised manner from 11 properties for each of 174, 947 galaxies. The largest sphere geometry is a representation for 16, 179 galaxies. Absolute Error $= 2.38 \cdot 10^{-5}$. Relative Mapping Error $= 9.96 \cdot 10^{-9}$

4. Conclusion

Complex and abstract entities not necessarily having physical 3D nature, like databases, symbolic knowledge, and data mining results may be used as heterogeneous data sources for the construction of a Virtual Reality space. The VR space representation enables the visualization, summarization and the understanding of the underlying data structure as demonstrated by examples from different, and potentially large, domains such as geology, astronomy and biomedical applications using gene expression data.

The hybrid VR space technique incorporates aspects from neural networks, evolutionary computation, (such as genetic algorithms, particle swarm optimization combined with classical optimization methods, etc), which may be applied to diverse data mining results visualization, as demonstrated using rough sets, crisp and fuzzy clusterings. The procedure may be constructed in either an unsupervised or supervised manner. The resulting VR space may contain geometries representing single or multiple entities, representing objects or abstract characterizations (summarizations) of class structural information such as that provided by the convex hull. The VR space may simultaneously display data and data mining results, for example, in the form of decision rules.

Further research is necessary in order to find the best ranges of parameters, the best evolutionary schemes, and more elaborated strategies for combining computational techniques with classical optimization methods, including the proportion between global and local search. Further studies focussing on the relationships between particular data mining strategies and visual representations would also be interesting to pursue.

5. Acknowledgments

The authors would like to thank Bob Orchard and Dr. F. Famili from the National Research Council (NRC) Canada's Institute for Information Technology's Integrated Reasoning Group for their support of this research, Dr. Luc Simard, NRC's Herzberg Institute of Astrophysics (HIA) and Stephanie Juneau, Department of Astronomy, University of Montreal for providing the SDSS data set along with Dr. Simard's morphological analysis of the astronomic data, and Dr. P.R. Walker from the NRC's Institute for Biological Sciences for providing the Alzheimer's gene expression data set.

REFERENCES

1. Agrawal, R, Mannila, H, Srikant, R. Toivinen, H, Verkamo, I.: Fast Discovery of Association Rules. [In] U.M. Fayyad et al. (eds.) *Advances in Knowledge Discovery and Data Mining*. AAAI Press, (1996), pp.307-328
2. Anderberg, M.: Cluster Analysis for Applications. Academic Press, (1973) 359pp.
3. Bazan, J.G., Skowron A., Synak, P: Dynamic Reducts as a Tool for Extracting Laws from Decision Tables. Proc. of the Symp. on Methodologies for Intelligent Systems. Charlotte, NC, Oct. 16-19 1994. Lecture Notes in Artificial Intelligence 869, Springer-Verlag (1994), 346–355.
4. J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, 1989.

5. Borg, I., Lingoes, J.: Multidimensional Similarity Structure Analysis. Springer-Verlag 1987.
6. Brown V. M, Ossadtchi A., Khan A.H., Cherry S.R., Leahy R.M. and Smith D.J., High-throughput imaging of brain gene expression. *Genome Res*, 2002, 12: 244-54.
7. Chandon, J.L., Pinson, S. : Analyse Typologique. Théorie et Applications. Masson, Paris, (1981)
8. Chang, J.C. et al. "Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer". Mechanisms of Disease. The Lancet. vol 362. August 2003.
9. H. Edelsbrunner and E. P. Mücke. Three-dimensional alpha shapes. Manuscript UIUCDCS-R-92-1734, Dept. Comput. Sci., Univ. Illinois, Urbana-Champaign, IL, 1992.
10. Fayyad, U., Piatesky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery. [In] U.M. Fayyad et al. (eds.) *Advances in Knowledge Discovery and Data Mining*. AAAI Press, (1996), pp. 1-34
11. Freeman, J.A., Skapura, D.M: Neural Networks: Algorithms, Applications, and Programming Techniques. Addison Wesley (1991)
12. I. Gath and A. Deva, "Unsupervised Optimal Fuzzy Clustering," IEEE Trans. Of Pat Anal. and Mach Intel, 1989, pp 773-781.
13. 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.
14. Gower, J.C.: A General Coefficient of Similarity and Some of its Properties. *Biometrics* Vol.1 No. 27 (1973) pp. 857-871
15. D.E. Gustafson and W.C.Kessel, "Fuzzy clustering with a covariance matrix," in IEEE Conference on Decision and Control, 1979, pp. 761-766.
16. Hartigan, J.: Clustering Algorithms. John Wiley & Sons, 351 pp, (1975).
17. Hata R., Masumura M., Akatsu H., Li F, Fujita H, Nagai Y, Yamamoto T, Okada H, Kosaka K., Sakanaka M. and Sawada T.: Up-regulation of calcineurin Abeta mRNA in the Alzheimer's disease brain: assessment by cDNA microarray. *Biochem Biophys Res Commun*, 2001, 284: 310-6.
18. Hajek, P., Havranek, T. : Mechanizing Hypothesis Formation. Springer Verlag (1978)
19. Jianchang, M., Jain, A. : Artificial Neural Networks for Feature Extraction and Multivariate Data Projection. *IEEE Trans. On Neural Networks*. Vol. 6, No. 2 (1995) pp. 296-317
20. 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.
21. 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.
22. 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.

23. J. Kennedy, R. C. Eberhart, Particle Swarm Optimization, *Proceedings of the IEEE International Joint Conference on Neural Networks*, Vol. 4, pp 19421948, 1995.
24. J. Kennedy, R. C. Eberhart, Y. Shi, *Swarm Intelligence*, (Morgan Kaufmann, 2002).
25. Kruskal, J.: Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psichometrika* Vol 29 1964 pp. 1-27
26. Lingras, P., Yao, Y. : Time Complexity of Rough Clustering: GAs versus K-Means. *Third. Int. Conf. on Rough Sets and Current Trends in Computing* RSCTC 2002. Malvern, PA, USA, Oct 14-17. Alpigini, Peters, Skowron, Zhong (Eds.) *Lecture Notes in Computer Science* (*Lecture Notes in Artificial Intelligence Series*) LNCS 2475, pp. 279-288. Springer-Verlag , 2002
27. T. Masters, *Advanced Algorithms for Neural Networks*, John Wiley & Sons, 1993.
28. W. Pres, B.P. Flannery, S.A. Teukolsky, W.T Vetterling, *Numeric Recipes in C.* (Cambridge, MA, Cambridge University Press, 1992).
29. Pawlak, Z. : Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrecht, Netherlands. (1991)
30. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Series in Machine Learning (1992)
31. Gawrys, M., Sienkiewicz, J. : Rough Set Library User's Manual (version 2.0). Inst. of Computer Science. Warsaw Univ. of Technology (1993)
32. E. Ruspini, "A New Approach to Clustering," Inform. Control, vol. 15, no. 1 pp. 22-32, April 1969.
33. J. W. Sammon, A non-linear mapping for data structure analysis. *IEEE Trans. Computers*, C-18, 401-408, (1969)
34. M. Sato, Y. Sato and L.C. Jain "Fuzzy Clustering Models and Applications," Physica Verlag, Heidelberg, New York 1997.
35. Valdés, J.J: Virtual Reality Representation of Relational Systems and Decision rules: an exploratory tool for understanding data structure. In TARSKI: Theory and Application of Relational Structures as Knowledge Instruments. Meeting of the COST Action 274, *Book of Abstracts*. Prague, Nov. 14-16, (2002)
36. Valdés, J.J: Similarity-based Heterogeneous Neurons in the Context of General Observational Models. *Neural Network World*. Vol 12., No. 5, (2002) pp. 499-508
37. J. J. Valdés, Virtual Reality Representation of Information Systems and Decision Rules: An Exploratory Tool for Understanding Data and Knowledge. *Proceedings of the 9-th Int. Conf. on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing* (Wang, Liu, Yao, Skowron, eds.). Chongqing, China, Oct 8-12, 2003. *Lecture Notes in Artificial Intelligence* LNAI 2639, pp. 615-618. Springer-Verlag, 2003.
38. 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 **LNAI 3029**, Springer Verlag, 2004, pp. 118–127.
39. J. J. Valdés and A. J. Barton. "Virtual Reality Visual Data Mining with Nonlinear Discriminant Neural Networks: Application to Alzheimer and Leukemia Gene Expression Data", *Proceedings of the International Joint Conference on Neural Networks 2005*, Lecture Notes in Artificial Intelligence, Springer Verlag, 2005, To appear.

40. P. R. Walker, B. Smith, Q. Y. Liu, F. Famili, J. J. Valdés, Z. Liu, B. Lach, Data Mining of Gene Expression Changes in Alzheimer Brain. *Int. Jour. of Artificial Intelligence in Medicine*, Elsevier Science 2003.
41. A. R. Webb and D. Lowe, "The Optimized Internal representation of a Multilayer Classifier", *Neural Networks* vol. 3, pp. 367-375, 1990.
42. Wróblewski, J: Ensembles of Classifiers Based on Approximate Reducts. Fundamenta Informaticae 47 IOS Press, (2001), 351–360.
43. Yong-ling Zheng, Long-hua Ma, Li-yan Zhang, Ji-xin Qian, Empirical Study of Particle Swarm Optimizer with an Increasing Inertia Weight. *Proceedings of the World Congress on Evolutionary Computation*. Canberra, Australia, Dec 8-12, 2003, IEEE Press, pp 221-226 pp.