Expert Systems with Applications 39 (2012) 13193–13201

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Data and knowledge visualization with virtual reality spaces, neural networks and rough sets: Application to cancer and geophysical prospecting data

Julio J. Valdés [a], Enrique Romero [b,*], Alan J. Barton [a]

[a] Information and Communication Technologies, National Research Council, Canada
[b] Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Spain

ARTICLE INFO

Keywords:
Data and knowledge visualization
Visual data mining
Virtual reality
Data projection
Neural networks
Rough sets

ABSTRACT

Visual data mining with virtual reality spaces is used for the representation of data and symbolic knowledge. High quality structure-preserving and maximally discriminative visual representations can be obtained using a combination of neural networks (SAMANN and NDA) and rough sets techniques, so that a proper subsequent analysis can be made. The approach is illustrated with two types of data: for gene expression cancer data, an improvement in classification performance with respect to the original spaces was obtained; for geophysical prospecting data for cave detection, a cavity was successfully predicted.

1. Introduction

Knowledge discovery is the non-trivial process of identifying valid, novel, potentially useful, and ultimately *understandable patterns* in data (Fayyad, Piatesky-Shapiro, & Smyth, 1996). In general, data under study may be described in terms of collections of *heterogeneous* properties, typically composed of properties represented by nominal, ordinal or real-valued (scalar) variables, as well as by others of a more complex nature, like images, time-series, etc. In addition, the data comes with different degrees of precision, uncertainty and information completeness (missing data is quite common). Patterns to discover are also of different kinds (geometrical, logical, behavioral, etc.).

Technological advancements in recent years are enabling the collection of large amounts of data in many fields. For example, in the field of Bioinformatics, high-throughput microarray gene expression experiments are possible, leading to an information explosion. The increasing rates of data generation require the development of data mining procedures facilitating the in-depth *understanding* of the internal structure of data more rapidly and intuitively.

The complexity of many data analysis procedures makes it more difficult for the user to extract useful information out of results. Classical data mining and analysis methods are sometimes difficult to use, the output of many procedures may be large and time consuming to analyze, and often their interpretation requires special expertise. Moreover, some methods are based on assumptions about the data which limit their application, specially for the purpose of exploration, comparison, hypothesis formation, etc. that are typical of the first stages of scientific investigation.

The role of visualization techniques in the knowledge discovery process is well known. The human brain still outperforms the computer in understanding complex geometric patterns, thus making the graphical representation of complex and abstract information directly appealing. A virtual reality (VR) technique for visual data mining on heterogeneous, imprecise and incomplete information systems was introduced in Valdés (2002b, 2003). Several reasons make VR a suitable paradigm for visual data mining: different representation models according to human perception preferences can be chosen, it allows *immersion*, it creates a *living* experience, it is *broad and deep*, and for using VR the user needs no mathematical knowledge and no special skills.

The purpose of this paper is twofold. First, to explore the construction of high quality VR spaces for visual data mining using a combination of neural networks and rough sets techniques. Second, to use the high quality constructed VR spaces for classification tasks. The whole process is, in turn, divided into two steps. In the first one, both data and symbolic knowledge are transformed into VR spaces where their structure and properties can be visually inspected and quickly understood. In the second step, a proper subsequent analysis is made within the constructed VR spaces with the aim of obtaining good classification results. This latter analysis depends on the problem at hand. The approach is illustrated with two types of data: gene expression cancer data and geophysical prospecting data for cave detection.

Three two-class gene expression cancer data sets were selected, representative of three of the most important types of cancer in modern medicine: liver, stomach and lung. They are composed of samples from normal and tumor tissues, described in terms of tens of thousands of variables, related to the gene expression intensities

* Corresponding author.
E-mail address: eromero@lsi.upc.edu (E. Romero).

0957-4174/$ - see front matter Crown © 2012 and Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.05.082

measured in microarray experiments. In the first step, neural networks for Sammon's projection (SAMANN) (Jain & Mao, 1992; Mao & Jain, 1995) are used for unsupervised structure-preserving mapping to low-dimensional feature spaces, where the corresponding VR spaces are constructed. Despite the very high dimensionality of the original patterns, high quality visual representations in the form of structure-preserving virtual spaces are obtained for every data set, which enables the differentiation of cancerous and non-cancerous tissues: the projected 3D spaces are polarized with two distribution modes, each one corresponding to a different class. In the second step, linear Support Vector Machines are constructed in the respective projected spaces, leading to an improvement in classification performance with respect to the original spaces.

A case of geophysical prospecting for underground caves is also studied. It is not the typical two-class presence/absence problem because only one class is known with certainty. In contrast, this is a problem with partially defined classes: the existence of a cave beneath a measurement station is either known for sure or *unknown*. In the first step, SAMANN and Non-linear Discriminant Analysis (NDA) networks (Mao & Jain, 1993, 1995; Webb & Lowe, 1990) are constructed. SAMANN networks are used for unsupervised mapping to low-dimensional feature spaces, obtaining high quality structure-preserving visual representations. NDA networks are used for supervised mapping to low-dimensional feature spaces where objects belonging to different classes are maximally differentiated. In the second step, the VR spaces and the NDA results allow the derivation of fuzzy cave membership function and the prediction of unknown objects to the cave class. In one of the areas with higher values, a borehole drilled actually hit a cavity. Rough sets methods are applied for evaluating the information content of the original descriptor variables and for the extraction of symbolic rules from the data. The general properties of the symbolic knowledge can be found with greater ease in the virtual reality space, and the structures of the knowledge base and the data were found to be very similar.

2. Virtual reality spaces for visual data mining

Information systems were introduced in Pawlak (1991). They have the form $S = \langle U, A\rangle$ where U is a non-empty finite set called the *universe* and A is a non-empty finite set of *attributes*, such that each $a \in A$ has a domain V_a and an evaluation function f_a. The V_a are not required to be finite. More generally, *heterogeneous* and *incomplete* information systems should be considered (Valdés, 2002a).

A *virtual reality* (VR) space for the visual representation of information systems (Valdés, 2002b, 2003), is defined as $\Upsilon = \langle \underline{O}, G, B, \Re^m, g_o, l, g_r, b, r\rangle$. $\underline{O}$ is a relational structure composed by objects and relations ($\underline{O} = \langle O, \Gamma^v\rangle$, $\Gamma^v = \langle \gamma_1^v, \ldots, \gamma_q^v\rangle$, $q \in \mathbb{N}^+$ and the $o \in O$ are objects), G is a non-empty set of *geometries* representing the different objects and relations. 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 (the actual VR 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 selecting which of the original relations will be represented in the virtual world. The representation of an information system $\widehat{S}$ in a virtual world requires the specification of several sets and a collection of extra mappings: $\widehat{S}^v = \langle O, A^v, \Gamma^v\rangle$, $\underline{O}$ in Υ, which can be done in many ways. A desideratum for $\widehat{S}^v$ is to keep as many properties from $\widehat{S}$ as possible. Thus, a 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 $f : \widehat{\mathcal{H}}^n \to \Re^m$. If $u = \langle f_{a_1}(u), \ldots, f_{a_n}(u)\rangle$ and $\xi(u) = o$, then $l(o) = f(\xi(\langle f_{a_1}(u), \ldots, f_{a_n}(u)\rangle)) = \langle f_{a_1^v}(o), \ldots, f_{a_m^v}(o)\rangle$ ($f_{a_i^v}$ are the evaluation functions of A^v).

Humans perceive much information through vision, in large quantities and at very high input rates. The human brain is extremely well qualified for the fast understanding of complex visual patterns, and still outperforms computers. Several reasons make VR a suitable paradigm: (i) it is *flexible* (it allows the choice of different representation models to better suit human perception preferences), (ii) allows *immersion* (the user can navigate inside the data, and interact with the objects in the world), (iii) creates a *living* experience (the user is not merely a passive observer, but an actor in the world) and (iv) VR is *broad and deep* (the user may see the VR world as a whole, and/or concentrate on specific details of the world). Of no less importance is the fact that in order to interact with a virtual world, only minimal skills are required (Simoff, Bhlen, & Mazeika, 2008).

3. Neural networks for the construction of virtual reality spaces

The typical *desiderata* for the visual representation of data and knowledge can be formulated in terms of minimizing information loss, maximizing structure preservation, maximizing class separability, or their combination, which leads to single or multi-objective optimization problems. In many cases, these concepts can be expressed deterministically using continuous functions with well defined partial derivatives. This is the realm of classical optimization where there is a plethora of methods with well known properties. In the case of heterogeneous information the situation is more complex and other techniques are required (see, for example (Valdés, 2004; Valdés & Barton, 2005)).

In the unsupervised case, the function f mapping the original space to the VR (geometric) space $\mathbb{R}^m$ can be constructed as to maximize some metric/non-metric structure preservation criteria (Lee & Verleysen, 2007) as is typical in multidimensional scaling (Borg & Lingoes, 1987), or minimize some error measure of information loss (Sammon, 1969). A typical error measure is:

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

where δ_{ij} is a dissimilarity measure between any two objects i, j in the original space, and ζ_{ij} is another dissimilarity measure defined on objects i, j in the VR space (the images of i, j under f). Usually, the mappings f obtained using approaches of this kind are *implicit* because the images of the objects in the new space are computed directly. However, a functional representation of f is highly desirable, specially in cases where more samples are expected *a posteriori* and need to be placed within the space. With an implicit representation, the space has to be computed every time that a new sample is added to the data set, whereas with an explicit representation the mapping can be used to compute directly the image of the new sample. As long as the incoming objects can be considered as belonging to the same population of samples used for constructing the mapping function, the space does not need to be recomputed. Neural networks are natural candidates for constructing explicit representations due to their universal function approximation property. If proper training methods are used, neural networks can learn structure-preserving mappings of high dimensional samples into lower dimensional spaces suitable for visualization (2D, 3D). Such an example is the SAMANN network. This is a feedforward network and its architecture consists of an input layer with as many neurons as descriptor attributes, an output layer with as many neurons as the dimension of the VR space and one or more hidden layers. The classical way of training the SAMANN network is described in Jain and Mao (1992) and Mao and Jain (1995). It consists of a gradient descent method where

the derivatives of the Sammon error are computed in a similar way to the classical backpropagation algorithm. Different from the backpropagation algorithm, the weights can only be updated after a pair of examples is presented to the network. As previously mentioned, the advantage of using SAMANN networks is that, since the mapping f between the original and the VR space is *explicit*, a new sample can be easily transformed and visualized in the VR space. The same networks could be used as non-linear feature generators in a preprocessing step for other data mining procedures. A recent application of SAMANN networks for data visualization can be found at Dzemyda, Marcinkevičius, and Medvedev (2011).

In the supervised case, a natural choice for representing the f mapping is an NDA neural network (Mao & Jain, 1993, 1995; Webb & Lowe, 1990). The NDA network is also feedforward with an input layer with as many neurons as descriptor attributes, an output layer with as many neurons as classes contain the decision attribute, a last hidden layer with a number of neurons equal to the dimension of the VR space and optionally other hidden layers. The NDA network is trained in a standard way (Mao & Jain, 1993, 1995) to minimize the mean squared error.

4. Support vector machines for classification

Support vector machines (SVMs) for classification can be described as follows (Vapnik, 1995): the input vectors are mapped into a (usually high-dimensional) inner product space through some non-linear mapping ϕ, chosen *a priori*. In this space (the *feature space*), an optimal separating hyperplane is constructed. By using a (positive definite) kernel function $K(u,v)$ the mapping becomes implicit, since the inner product defining the hyperplane can be evaluated as $\langle \phi(u), \phi(v) \rangle = K(u,v)$ for every two vectors $u, v \in \mathbb{R}^N$. In the SVM framework, an optimal hyperplane means a hyperplane with maximal normalized margin for the examples of every class (the normalized margin is the minimum distance to the hyperplane). When the data set is separable by a hyperplane (either in the input space or in the feature space), the maximal normalized margin hyperplane is called the hard margin hyperplane. When the data set is not separable by a hyperplane (neither in the input space nor in the feature space), some tolerance to noise is introduced in the model, associated to a parameter C that allows to control the trade-off between the margin and the errors in the data set. By setting $C = \infty$, the hard margin hyperplane is obtained.

To fix notation, consider the classification task given by a data set $X = \{(x_1,y_1),\ldots,(x_L,y_L)\}$, where each instance x_i belongs to the input space $\mathbb{R}^N$, and $y_i \in \{-1,+1\}$. Using Lagrangian and Kuhn–Tucker theory, the maximal margin hyperplane, for a binary classification problem given by a data set is a linear combination of simple functions depending on the data:

$$f_{SVM}(x) = b + \sum_{i=1}^{L} y_i \alpha_i K(x_i, x) \tag{2}$$

where the vector $(\alpha_i)_{i=1}^{L}$ is the (1-norm soft margin) solution of the following constrained convex optimization problem:

$$\begin{aligned} &\text{Maximize}_{\alpha} \quad \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{L} y_i \alpha_i y_j \alpha_j K(x_i, x_j) \\ &\text{subject to} \quad \sum_{i=1}^{L} y_i \alpha_i = 0 \quad \text{(bias constraint)} \\ &\qquad\qquad 0 \leqslant \alpha_i \leqslant C \quad i = 1 \ldots L. \end{aligned} \tag{3}$$

The points x_i with $\alpha_i > 0$ (active constraints) are named *support vectors*. The most usual non-linear kernel functions $K(u,v)$ are Gaussian, polynomial or wavelet kernels (Ozer, Chen, & Cirpan, 2011).

5. Symbolic knowledge via rough sets and its representation with virtual reality

The rough set theory (Pawlak, 1991) bears on the assumption that in order to define a set, some knowledge about the elements of the data set is needed, 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 knowledge is understood to be the ability of characterizing all classes of the classification (Peters, Lingras, Ślęzak, & Yao, 2012).

A decision table is any information system of the form $\mathbf{S} = \langle U, A \rangle$ where $A = A' \cup \{d\}$, A' are the condition attributes and d is the decision attribute. The lower approximation of a concept consists of all objects, which surely belong to the concept, whereas the upper approximation consists of all objects, which possibly belong to the concept. For any $B \subseteq A$ an equivalence relation $IND(B)$ defined as $IND(B) = \{(x,x') \in U^2 | \forall a \in B,\ f_a(x) = f_a(x')\}$, is associated. 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 $\mathbf{S}$ is denoted $RED(A)$ (reduct computation is NP-hard, and several heuristics have been proposed (Thangavel & Pethalakshmi, 2009; Wróblewski, 2001)). Reduction of knowledge consists of removing superfluous partitions such that the set of elementary categories in the information system is preserved, in particular, with respect to those categories induced by the decision attribute. Minimum reducts (those with a small number of attributes) are extremely important, as decision rules can be constructed from them (Bazan, Skowron, & Synak, 1994). The algorithms for computing reducts and rules used in this paper are those of the Rosetta system (Øhrn & Komorowski, 1997).

6. Experiments with gene expression cancer data sets

According to the World Health Organization, cancer is a leading cause of death worldwide (http://www.who.int/cancer/en). From a total of 58 million deaths in 2005, cancer accounts for 7.6 million (or 13%) of all deaths. The main types of cancer leading to overall cancer mortality are (i) lung (1.3 million deaths/year), (ii) stomach (almost 1 million deaths/year), (iii) liver (662,000 deaths/year), (iv) colon (655,000 deaths/year) and (v) breast (502,000 deaths/year).

6.1. Data sets description

Three microarray gene expression cancer databases were selected, representative of three of the most important types of cancer in the world: liver, stomach and lung cancer. They share the typical features of these kind of data: a small number of samples (in the order of tens) described in terms of a very large number of attributes (in the order of tens of thousands), related to the gene expression intensities measured in microarray experiments.

6.1.1. Liver Cancer

The data were those used in Lam et al. (2006), where zebrafish liver tumors were analyzed and compared with human liver tumors. First, liver tumors in zebrafish were generated by treating them with carcinogens. Then, the expression profiles of zebrafish liver tumors were compared with those of zebrafish normal liver tissues using a Wilcoxon rank-sum test. The original database had 20 samples (10 normal, 10 tumor) and 16,512 attributes. As a result of this comparison, a zebrafish liver tumor differentially expressed gene set consisting of 2315 gene features was obtained. This data set was used for comparison with human tumors. The results suggest that the molecular similarities between zebrafish and

human liver tumors are greater than the molecular similarities between other types of tumors (stomach, lung and prostate). The data can be found at http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=2220.

6.1.2. Stomach cancer

The data were those used in Hippo et al. (2002), where a study of genes that are differentially expressed in cancerous and noncancerous human gastric tissues was performed. The original database contained 30 samples (22 tumor, 8 normal) that were analyzed by oligonucleotide microarray, obtaining the expression profiles for 6936 genes (7129 attributes). Using the 6272 genes that passed a prefilter procedure, cancerous and noncancerous tissues were successfully distinguished with a two-dimensional hierarchical clustering using Pearson's correlation. However, the clustering results used most of the genes on the array. To identify the genes that were differentially expressed between cancer and noncancerous tissues, a Mann–Whitney's U test was applied to the data. As a result of this analysis, 162 and 129 genes showed a higher expression in cancerous and noncancerous tissues, respectively. In addition, several genes associated with lymph node metastasis and histological classification (intestinal, diffuse) were identified. The data can be found at http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=1210.

6.1.3. Lung cancer

The data were those used in Spira et al. (2004), where gene expressions of severely emphysematous lung tissue (from smokers at lung volume reduction surgery) and normal or mildly emphysematous lung tissue (from smokers undergoing resection of pulmonary nodules) were compared. The original database contained 30 samples (18 severe emphysema, 12 mild or no emphysema), with 22,283 attributes. Genes with large detection P-values were filtered out, leading to a data set with 9336 genes, that were used for subsequent analysis. Nine classification algorithms were used to identify a group of genes whose expression in the lung distinguished severe emphysema from mild or no emphysema. First, model selection was performed for every algorithm by leave-one-out cross-validation, and the gene list corresponding to the best model was saved. The genes reported by at least four classification algorithms (102 genes) were chosen for further analysis. With these genes, a two-dimensional hierarchical clustering using Pearson's correlation was performed that distinguished between severe emphysema and mild or no emphysema. Other genes were also identified that may be causally involved in the pathogenesis of the emphysema. The data can be found at http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=737.

6.2. Experimental settings

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

Model training. For every data set, one-hidden layer SAMANN networks were constructed to map the original data to a 3D VR space. The Euclidean distance was the dissimilarity measure used in the original (δ_{ij}) and the VR (ζ_{ij}) spaces. The activation functions were sinusoidal for the hidden layer and hyperbolic tangent for the output layer. A collection of models was obtained by varying some of the network controlling parameters (Table 1), for a total of 1944 SAMANN networks for every data set.

Table 1
Parameters used for the SAMANN networks with cancer data sets.

Parameters	Liver
Neurons in the hidden layer	{10,30}
Weights ranges	{(10,5),(10,10),(15,5)}
Learning rates	{(.1,.01),(.2,.01),(.2,.02)}
Momentum	{0,.5,.7}
Number of iterations	{200,500,1000}
Random seed	Four different values
Presented pairs at every iteration	(All,50,100)
Parameters	**Stomach**
Neurons in the hidden layer	{10,30}
Weights ranges	{(10,5),(10,10),(15,5)}
Learning rates	{(.5,.1),(1,.1),(2,.1)}
Momentum	{0,.5,.7}
Number of iterations	{200,500,1000}
Random seed	Four different values
Presented pairs at every iteration	(All,50,200)
Parameters	**Lung**
Neurons in the hidden layer	{10,20}
Weights ranges	{(50,5),(50,10),(60,5)}
Learning rates	{(.005,.0005),(.01,.001),(.02,.002)}
Momentum	{0,.5,.7}
Number of iterations	{200,500,1000}
Random seed	Four different values
Presented pairs at every iteration	(All,50,200)

6.3. Results

6.3.1. Visualization of SAMANN Results

For every data set, we constructed the histograms of the Sammon error for the obtained networks. The empirical distributions were positively skewed (with the mode on the lower error side), which is a good behavior. In addition, the general error ranges were small. In Table 2 some statistics of the experiments are presented: minimum, maximum, mean and standard deviation for the best (i.e., with smallest Sammon error) 1000 networks.

Clearly, it is impossible to represent a VR space on printed media (navigation, interaction, and world changes are all lost). Therefore, very simple geometries were used for objects and only snapshots of the virtual worlds are presented. For simplicity, in all of the VR-spaces presented $G = \{dark\ spheres, light\ spheres\}$, $B = \{static\}$ and the l function is based on the representation of f given by a SAMANN network. r is a single characteristic function for the relation $\mathcal{C}$ with the equivalent classes such that objects of one class will be represented as dark spheres and those of the other class by light ones.

Figs. 1–3 show the VR spaces corresponding to the best obtained networks for the liver, stomach and lung cancer data sets, respectively. Although the mapping was generated from an unsupervised perspective (i.e., without using the class labels), objects belonging to different classes were represented in the VR space differently for comparison purposes. Transparent membranes wrap the corresponding classes, so that the degree of class overlap can be easily observed. In addition, the wrapping allows one to look for particular samples with ambiguous diagnostic decisions.

The low values of the Sammon error indicate that the spaces preserved most of the distance structure of the data, therefore giving a good indication of the distribution in the original spaces. The three

Table 2
Statistics of the best 1000 SAMANN networks obtained for cancer data sets.

Data set	Sammon error			
	Minimum	Maximum	Mean	Std. dev.
Liver	0.03991	0.05564	0.04988	0.00362
Stomach	0.06295	0.07745	0.07286	0.00335
Lung	0.07924	0.10784	0.09469	0.00698

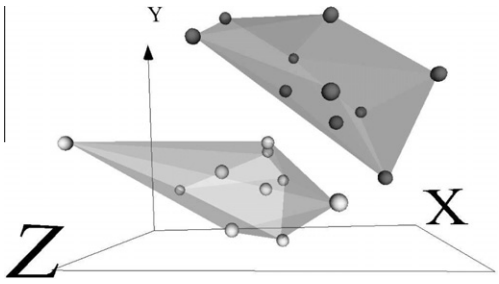

Fig. 1. VR space of the liver cancer data set (Sammon error = 0.03991, best out of 1944 experiments). The space was generated from an unsupervised perspective, but classes are displayed for comparison purposes. Dark spheres: normal, light spheres: cancerous samples.

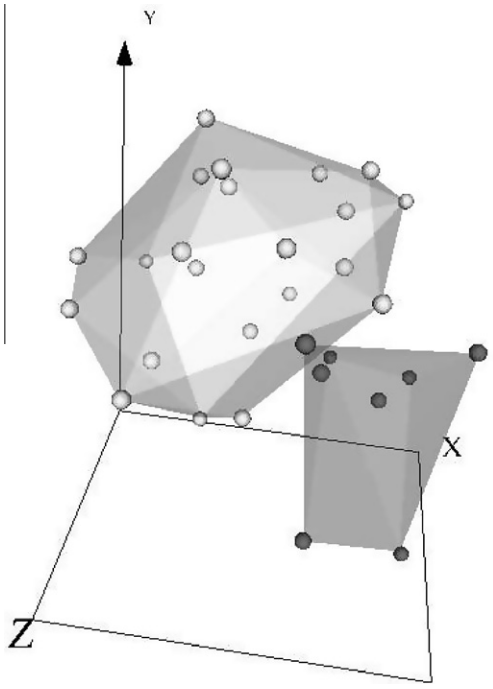

Fig. 2. VR space of the stomach cancer data set (Sammon error = 0.06295, best out of 1944 experiments). The space was generated from an unsupervised perspective, but classes are displayed for comparison purposes. Dark spheres: normal, light spheres: cancerous samples.

VR spaces are clearly polarized with two distribution modes, each one corresponding to a different class. Note, however, that classes are more clearly differentiated for the liver and stomach data sets than for the lung data set, where a certain level of overlap exists. The reason for this may be that mild and no emphysema were considered members of the same class (see section 6.1).

Since the distance between any two objects is an indication of their dissimilarity, a new point is more likely to belong to the same class of its nearest neighbors. In the same way, outliers can be readily identified, although they may result from the space deformation inevitably introduced by the dimensionality reduction.

6.3.2. Classification results with SVMs

Since the projected spaces are polarized with two distribution modes, each one corresponding to a different class, a linear classification model is suitable from a supervised perspective. In our case, SVMs were used to that end, as explained next.

First, for every data set we obtained Sammon-projected data with dimension values ranging from 3 to 20. For every dimension, 2500 Newton minimization procedures were applied varying the initial point and the step size to obtain implicit representations of the original data. Similar to SAMANN networks, the Euclidean distance was the dissimilarity measure used in the original (δ_{ij}) and the VR (ζ_{ij}) spaces. The representations with smallest Sammon error were selected. Then, for every dimension, a leave-one-out cross-validation was performed to the projected data with hard margin linear SVMs ($C = \infty$).

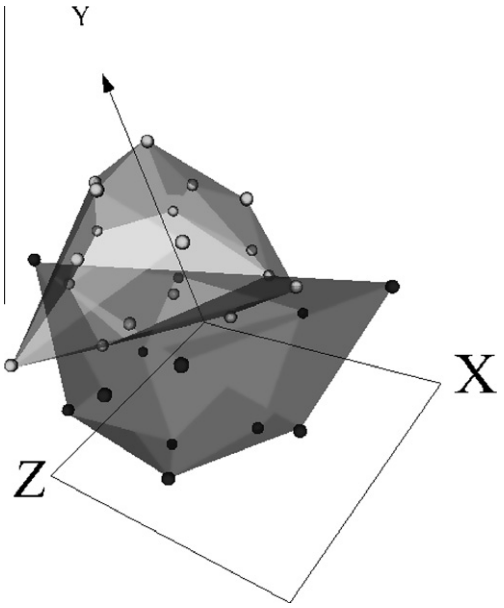

Fig. 3. VR space of the lung cancer data set (Sammon error = 0.07924, best out of 1944 experiments). The space was generated from an unsupervised perspective, but classes are displayed for comparison purposes. Dark spheres: severe emphysema, light spheres: mild or no emphysema. The boundary between the classes in the VR space seem to be a low curvature surface.

Table 3 shows the best leave-one-out cross-validation performances obtained for every projected data set, together with the results obtained with the original dimensions. As it can be seen, classification performance improves in the projected spaces (with a very low dimension).

7. Experiments with geophysical prospecting data

The proposed approach was applied to the detection of underground caves with geophysical data. Cave detection is a very important problem in civil and geological engineering. Sometimes the caves are opened to the surface, but typically they are buried and geophysical methods are required to detect them. This task is usually very complex.

7.1. Data set description

The studied area contained an accessible cave (see Fig. 8 right). Geophysical methods complemented with a topographic survey were used in the studied area with the purpose of finding their relation with subsurface phenomena (Valdés & Gil, 1982).

The set of geophysical methods included (1) the spontaneous electric potential (SP_{dry}) at the surface of the earth in the dry

Table 3
Best leave-one-out cross-validation results for linear SVMs ($C = \infty$) with cancer data sets for original (left) and projected (right) data.

Data set	Dimension	Test (%)	Dimension	Test (%)
Liver	16,512	90.0	13	100.0
Stomach	7129	100.0	4	100.0
Lung	22,283	73.3	11	90.0

season, (2) the vertical component of the electro-magnetic field in the VLF region of the spectrum, (3) the spontaneous electric potential in the rainy season (SP_{rain}), (4) the gamma ray intensity (*Rad*) and (5) the local topography (*Alt*). The raw data consist of these 5 fields (the attributes) on a spatial grid containing 1225 measurement stations (the data objects).

This is not the typical two-class presence/absence problem because only one class is known with certainty. In contrast, this is a problem with *partially* defined classes: the existence of a cave beneath a measurement station is either known for sure or *unknown*. Note, however, that this is not a one class problem, because two different classes exist. Since the classes are partially defined, a combination of unsupervised and supervised approaches is required.

7.2. Experimental settings

Data preprocessing. In order to eliminate the data distortion introduced by the different units of measure and to reduce the influence of noise and regional geological structures, a data preprocessing process was performed consisting of: *(i)* conversion of each physical field to standard scores. *(ii)* model each physical field f as composed of a trend, a signal and additive noise: $f(x,y) = t(x,y) + s(x,y) + n(x,y)$ where t is the trend, s is the signal, and n is the noise component. *(iii)* fit a least squares 2D linear trend $\hat{t}(x,y) = c_0 + c_1x + c_2y$ and obtain the residual: $\hat{r}(x,y) = f(x,y) - \hat{t}(x,y)$. *(iv)* convolve the residual with a low pass 2D filter to attenuate the noise component: $\hat{s}(x,y) = \sum_{k_1=-N}^{N}\sum_{k_2=-N}^{N} h(k_1,k_2)\hat{r}(x-k_1,y-k_2)$, where $\hat{s}(x,y)$ is the signal approximation, and $h(k_1,k_2)$ is the low-pass zero-phase shift digital filter. *(v)* recompute the standard scores and add a class attribute indicating whether there is a known cave below the corresponding measurement station or if its presence is unknown. The pre-processed data will be called *cave-prp-data*.

Reducts from the original data set. The *cave-prp-data* set was discretized using the boolean reasoning algorithm and the reducts were found by Johnson's algorithm (Øhrn & Komorowski, 1997). A single reduct was found, consisting of all of the 5 original variables, proving that no proper subset of these variables exactly preserves the discernibility relation of the original data. That is, no lower dimensional space based on the power set of the original variables is discernibility-preserving. Thus, lower dimensional spaces based on non-linear combinations must be constructed for visualization.

Model training. A collection of experiments was conducted with one-hidden layer SAMANN networks in order to select adequate models for the visualization. Two-hidden layer NDA networks were used to construct a space where objects belonging to different classes are maximally differentiated. The activation functions were sinusoidal for the first hidden layer and hyperbolic tangent for the rest of the layers. For the SAMANN networks, the Euclidean distance was the dissimilarity measure used in the original (δ_{ij}) and the VR (ζ_{ij}) spaces. A collection of models was obtained by varying some of the network controlling parameters (Table 4), for a total of 1260 for the NDA and 324 for the SAMANN networks, respectively.

Table 4
Parameters used for the SAMANN and NDA networks with the *cave-prp-data* set.

Parameters	SAMANN
Neurons in the first hidden layer	{20,30,40}
Weights ranges	{(10,5),(10,10),(15,15)}
Learning rates	{(3.0,1.5),(2.0,1.0),(1.0,0.5)}
Momentum	{0,.1,.2}
Number of iterations	200
Random seed	Four different values
Presented pairs at every iteration	All
Parameters	**NDA**
Neurons in the first hidden layer	{20,30,40,50,60}
Weights ranges	({.1,.5,1,3,5,7,9},1.0)
Learning rates	.001,.001,.001
Momentum	{.1,.2,.3}
Number of iterations	{1000,2000,3000}
Random seed	Four different values

7.3. Results

7.3.1. Visualization of SAMANN Results

SAMANN networks mapped the original *cave-prp-data* 5-dimensional space to a 3D VR-space from an unsupervised perspective. The distribution of the Sammon error is shown in Fig. 4.

It is skewed towards the smaller errors end, which is a good behavior, with a mean of 0.0229 and a standard deviation of 0.0013 indicating that error values fluctuate within a narrow range. As an illustration, the VR-space corresponding to experiment 135 is shown in Fig. 5.

The low value of the Sammon error indicates that the space preserved most of the distance structure of the data, therefore, giving a good idea about the distribution in the original space. The space is clearly polarized with two distribution modes: one at the left hand side composed exclusively of cave objects, and another at the right hand side composed only of unknown objects. Since the distance between any two objects is an indication of their dissimilarity, objects of the unknown class closer to objects of the cave class are more likely to correspond to measurement stations having underground cavities than objects further away. In particular, those objects of the unknown class contained within the convex hull defined by the objects of the cave class are very interesting. It is also evident that only a smaller proportion of the objects of the unknown class are either contained, or close to the convex hull of the cave class, as expected from the typical lognormal-like distribution of many geological features.

A hierarchical clustering using Euclidean distance and Ward's method (Anderberg, 1973) (Fig. 6) clearly reveals the existence of two well defined clusters.

Their nature is explained by the 2×2 contingency table defined by the membership with respect to the cave/unknown classes vs. those corresponding to the two clusters emerging from the dendrogram. The table has a highly significant χ^2 value (165.872), indicating the high degree of association between the existing classes (specially the cave class) and the formed clusters. Cluster 2 corresponds to the cave class containing 120 of the 121 cave objects and 419 unknown objects (likely candidates to belong to the cave class). Clearly, those in cluster 1 correspond to locations less likely to have underground cavities beneath.

	Cluster 1	Cluster 2	Total
Unknown	685	419	1104
Cave	1	120	121

7.3.2. Visualization of NDA results and cave membership function

A structure preserving space is not necessarily class-discriminating and conversely. In a supervised situation, the information available from the decision attribute is used for constructing a space where objects belonging to different classes are maximally differentiated. NDA networks were used for that purpose. The distribution of the classification error for the cave class is shown in Fig. 4 (right) (the only determined class in the problem). The distribution exhibits a skewed-multimodal characteristic with the important modes shifted towards smaller error values (a good feature). Several networks have 0% classification error for the cave

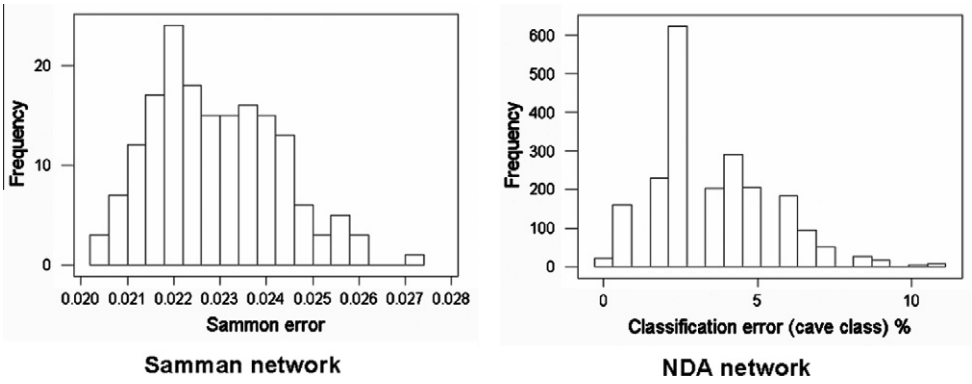

Fig. 4. Left: distribution of the SAMANN error (324 experiments) using SAMANN networks with the *cave-prp-data* set. Right: distribution of the classification error of the cave class (1260 experiments) using NDA networks.

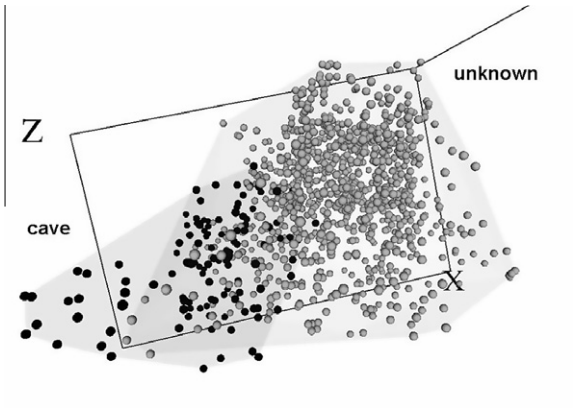

Fig. 5. VR-space of the *cave-prp-data* set corresponding to experiment 135 (Sammon error = 0.0208). Objects of the *cave* class are dark. Objects of the *unknown* class are light (this is for comparison purposes only, since the mapping generating the space is unsupervised). Transparent membranes wrap the corresponding classes.

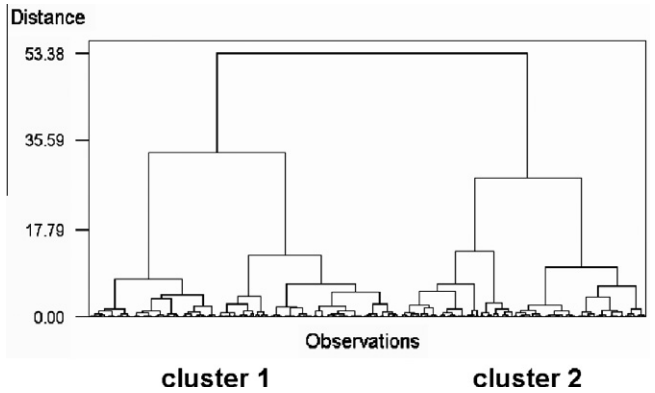

Fig. 6. Dendrogram of the objects in the VR-space of Fig. 5 for the *cave-prp-data* set (Ward's method using Euclidean distance).

class and a representative of them is the one found in experiment 174.

A VR-space was built from a composition of the mapping function (φ) represented by that network, with a principal components transformation ($\mathcal{P}$) given by $f = (\varphi \circ \mathcal{P})$ (Fig. 7).

The intrinsic dimensionality of this space is very close to one, and its shape indicates an almost linear continuum within and between the two classes. Conceptually, the objects at the two extremes represent the maximum expression of a *cavehood* property, and its opposite, the maximum expression of being *solid rock*, in geological terms. In between there is a gradation of the *cavehood* property, which is actually a fuzzy concept. Let $o_m \in O$ be the object of the VR-space satisfying the property $((\varphi \circ \mathcal{P})(o_m))_{pc_1} \leqslant ((\varphi \circ \mathcal{P})(o))_{pc_1}$ for all $o \in O$ and let o_M be the object such that $d(o_m,o_M) >= d(o_m,o)$ for all $o \in O$, where d is the Euclidean distance and pc_1 is the first principal component. Then, a two dimensional membership function $\mu_c \in [0,1]$ for *caveness* can be constructed as $\mu_c(o) = (1 - (d(o_m,o)/d(o_m,o_M)))$. Note that although a supervised approach was used, this formulation is based only on the information about the known class. The distribution of μ within the investigated area is shown in Fig. 8 (left).

The behavior of μ depicts a very consistent and realistic geological pattern, where not only the known cave is correctly flagged with maximal membership values, but also defines a collection of halos around the known cave with progressively decreasing values. In addition, other smaller areas with medium to high values are indicated, suggesting locations where other underground cavities could be expected. In particular, a borehole drilled at a location within the white circle of Fig. 8 (left) actually hit a cavity.

7.3.3. Visualization of symbolic knowledge

Symbolic knowledge in the form of production rules was extracted from the *cave-prp-data* set using rough set techniques, as explained in Section 5. Structure preserving VR-spaces representing an information system with rules as objects can be constructed by minimizing the Sammon error (1). In this case the dissimilarity measure used for the original attributes was $\delta_{ij} = (1 - \hat{s}_{ij})/\hat{s}_{ij}$, where $\hat{s}_{ij}$ is Gower's similarity coefficient (Gower, 1971). The Euclidean distance was the measure used for ζ_{ij} in the VR space. A set of 345 rules were generated. Two representative examples are:

$$SP_{dry}([-0.16981,*)) \;\&\; VLF([-0.75462,*)) \;\&\; SP_{rain}([0.48744,*)) \;\&\; Rad([-0.21015,*)) \;\&\; Alt([0.00346,*)) \Rightarrow \textit{unknown}\ (123\ \textit{objects})$$

$$SP_{dry}([*,-1.50209)) \;\&\; VLF([*,-1.14882)) \;\&\; SP_{rain}([*,-0.46789)) \;\&\; Rad([*,-1.54413)) \;\&\; Alt([*,-1.22398)) \Rightarrow \textit{cave}\ (6\ \textit{objects})$$

Fig. 7. VR-space maximizing class separability for the 1225 objects in the *cave-prp-data* set according to the $(\varphi \circ \mathcal{P})$ function. The classification error of the cave class is 0%.

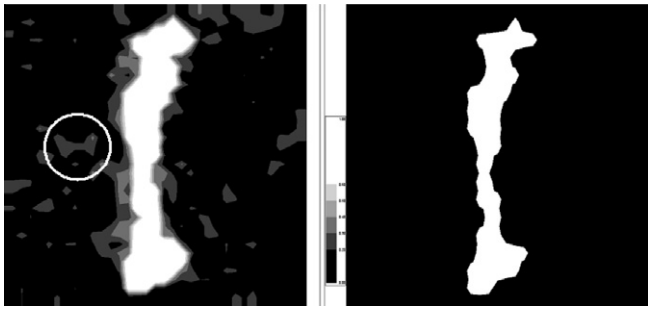

Fig. 8. Right: map of the known cave. Left: fuzzy membership function μ_c of the *cave* class computed from the VR-space obtained from the NDA network (Extreme values: white = 1, black = 0). The white circle indicates the area where a borehole hit a cavity, not opened to the surface.

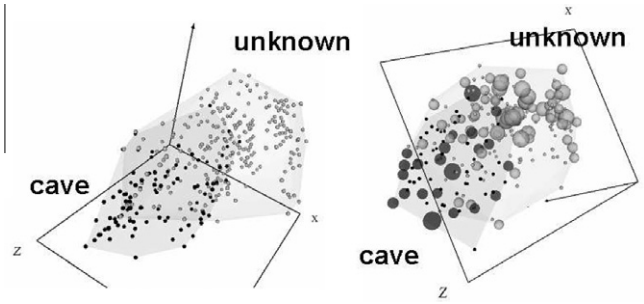

Fig. 9. Left: VR-space with a representation of the 345 rules for the *cave-prp-data* set. Right: VR-space with the 231 most representative rules (sizes are proportional to the amount of similar rules at a given location). Dark objects: rules concluding about the *cave* class. Light objects: rules concluding about the *unknown* class.

The approach described in Valdés (2002b, 2003) for the construction of VR-spaces representing symbolic knowledge in the form of production rules was applied and the corresponding space is shown in Fig. 9 (left). When compared with Fig. 5 it is clear that the structures of the knowledge base and the data are very similar. An even clearer distribution is obtained if the rule base is pre-processed with the Leader clustering algorithm (Hartigan, 1975) in order to select representatives for subsets of similar rules and work with a smaller information system.

Such a space is shown in Fig. 9 (right) where the relative size of an object at a particular location in the VR-space is proportional to the number of similar rules within its neighborhood (therefore, of data concentration in the original feature space).

This allows an easy identification of the most general rules from the more specific ones and also of knowledge granules. From the point of view of the distribution of the most important objects, the space is strongly polarized, allowing the identification of the rules describing the properties of the physical fields more accurately identifying the presence of underground caves and also the properties of the fields characterizing the areas most likely composed of solid rock. At the same time it allows the identification of the knowledge related with those objects of undetermined nature (i.e. from the undefined class).

8. Conclusions and future work

A combination of neural networks and rough set techniques was used for constructing virtual reality spaces for visual data mining suitable for representing data and symbolic knowledge. Good neural network models were found with the use of distributed computing techniques, that were used as mapping functions to produce high quality VR spaces where the properties of data and symbolic knowledge can be revealed.

For microarray gene expression cancer data sets, the obtained results show that a few non-linear features can effectively capture the similarity structure of the data and also provide a good differentiation between the cancer and normal classes. Linear support vector machines constructed in projected spaces lead to an improvement in classification performance. However, in cases where the descriptor attributes are not directly related to class structure or where there are many noisy or irrelevant attributes the situation may not be as clear. In these cases, feature subset selection and other data mining procedures could be considered in a preprocessing stage.

Problems with partially defined classes can be approached successfully by combining unsupervised and supervised techniques. A method for constructing membership functions in problems with partially defined classes is proposed which can be used as a forecasting tool, as illustrated with an example from geophysical prospecting. This approach can be extended to multiclass problems with partially defined classes.

Acknowledgments

This work was supported in part by the Ministerio de Ciencia e Innovación (MICINN), under project TIN2009-13895-C02-01. This research was conducted in the framework of the STATEMENT OF WORK between the National Research Council Canada (Institute for Information Technology, Integrated Reasoning Group) and the Soft Computing Group (Dept. de Llenguatges i Sistemes Informàtics), Universitat Politècnica de Catalunya, Spain.

References

Anderberg, M. (1973). *Cluster analysis for applications*. Academic Press.
Bazan, J. G., Skowron, A., & Synak, P. (1994). Dynamic reducts as a tool for extracting laws from decision tables. In *Symposium on methodologies for intelligent systems*. Lecture notes in artificial intelligence (Vol. 869, pp. 346–355).
Borg, I., & Lingoes, J. (1987). *Multidimensional similarity structure analysis*. Springer-Verlag.
Dzemyda, G., Marcinkevičius, & V., Medvedev, V. (2011). Large-scale multidimensional data visualization: A web service for data mining. In *European conference on towards a service-based internet*. Lecture notes in computer science (Vol. 6994, pp. 14–25).
Fayyad, U., Piatesky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In U. Fayyad et al. (Ed.), *Advances in knowledge discovery and data mining* (pp. 1–34). MIT Press.
Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27*, 857–871.
Hartigan, J. (1975). *Clustering algorithms*. John Wiley and Sons.
Hippo, Y., Taniguchi, H., Tsutsumi, S., Machida, N., Chong, J.-M., Fukayama, M., et al. (2002). Global gene expression analysis of gastric cancer by oligonucleotide microarrays. *Cancer Research, 62*(1), 233–240.
Jain, A.K., & Mao, J. (1992). Artificial neural networks for nonlinear projection of multivariate data. In *International joint conference on neural networks* (Vol. 2, pp. 335–340).
Lam, S. H., Wu, Y. L., Vega, V. B., Miller, L. D., Spitsbergen, J., Tong, Y., et al. (2006). Conservation of gene expression signatures between zebrafish and human tumors and tumor progression. *Nature Biotechnology, 24*(1), 73–75.
Lee, J. A., & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer.
Mao, J., & Jain, A. K. (1993). Discriminant analysis neural networks. In *1993 IEEE International conference on neural networks* (pp. 300–305).
Mao, J., & Jain, A. K. (1995). Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks, 6*, 296–317.
Øhrn, A., & Komorowski, J. (1997). Rosetta – A rough set toolkit for the analysis of data. In *International joint conference on information sciences* (Vol. 3, pp. 403–407).
Ozer, S., Chen, C. H., & Cirpan, H. A. (2011). A set of new Chebyshev Kernel functions for support vector machine pattern classification. *Pattern Recognition, 44*(7), 1435–1447.
Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*. Kluwer Academic Publishers.
Peters, G., Lingras, P., Ślęzak, D., & Yao, Y. (2012). *Rough sets: Selected methods and applications in management and engineering*. Springer.
Sammon, J. W. (1969). A non-linear mapping for data structure analysis. *IEEE Transactions on Computers, C-18*, 401–408.

Simoff, S. J., Bhlen, M. H., & Mazeika, A. (Eds.). (2008). *Visual data mining: Theory. Techniques and tools for visual analytics*. Springer.

Spira, A., Beane, J., Pinto-Plata, V., Kadar, A., Liu, G., Shah, V., et al. (2004). Gene expression profiling of human lung tissue from smokers with severe emphysema. *American Journal of Respiratory Cell and Molecular Biology, 31*, 601–610.

Thangavel, K., & Pethalakshmi, A. (2009). Dimensionality reduction based on rough set theory: A review. *Applied Soft Computing, 9*(1), 1–12.

Valdés, J., & Gil, J. L. (1982). Application of geophysical and geomathematical methods in the study of the Insunza Karstic Area (La Salud, La Habana). In *First international colloquium of physical-chemistry and karst hydrogeology in the Caribbean region* (pp. 376–384).

Valdés, J. J. (2002a). Similarity-based heterogeneous neurons in the context of general observational models. *Neural Network World, 12*(5), 499–507.

Valdés, J. J. (2002b). Virtual reality representation of relational systems and decision rules: An exploratory tool for understanding data structure. In *Theory and application of relational structures as knowledge instruments. Meeting of the COST action* (pp. 274).

Valdés, J. J. (2003). Virtual reality representation of information systems and decision rules: An exploratory tool for understanding data and knowledge. In *International conference on rough sets, fuzzy sets, data mining and granular computing*. Lecture notes in artificial intelligence (Vol. 2639, pp. 615–618).

Valdés, J. J. (2004). Building virtual reality spaces for visual data mining with hybrid evolutionary-classical optimization: Application to microarray gene expression data. In *IASTED international joint conference on artificial intelligence and soft computing* (pp. 161–166).

Valdés, J. J., & Barton, A. J. (2005). Virtual reality visual data mining with nonlinear discriminant neural networks: Application to leukemia and alzheimer gene expression data. In *International joint conference on neural networks* (pp. 2475–2480).

Vapnik, V. N. (1995). *The nature of statistical learning theory*. NY: Springer-Verlag.

Webb, A. R., & Lowe, D. (1990). The optimized internal representation of a multilayer classifier performs nonlinear discriminant analysis. *Neural Networks, 3*(4), 367–375.

Wróblewski, J. (2001). Ensembles of classifiers based on approximate reducts. *Fundamenta Informaticae, 47*(3-4), 351–360.