

NRC Publications Archive Archives des publications du CNRC

Virtual Reality Representation of Information Systems and Decision Rules: An Exploratory Technique for Understanding Data Knowledge Structure

Valdés, Julio

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. /
La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

Publisher's version / Version de l'éditeur:

The 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC'2003) [Proceedings], 2003

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
<https://nrc-publications.canada.ca/eng/view/object/?id=56fb7f4a-6735-4674-bccf-386adede39fb>
<https://publications-cnrc.canada.ca/fra/voir/objet/?id=56fb7f4a-6735-4674-bccf-386adede39fb>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at
PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.



National Research
Council Canada

Conseil national
de recherches Canada

Institute for
Information Technology

Institut de technologie
de l'information

NRC-CNRC

*Virtual Reality Representation of Information Systems and Decision Rules: An Exploratory Technique for Understanding Data Knowledge Structure **

Valdés, J.
May 2003

* published in The 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing RSFDGrC'2003. Chongqing, China. May 26-29, 2003.
NRC 45817.

Copyright 2003 by
National Research Council of Canada

Permission is granted to quote short excerpts and to reproduce figures and tables from this report,
provided that the source of such material is fully acknowledged.

Canada

Virtual Reality Representation of Information Systems and Decision Rules: An exploratory technique for understanding data and knowledge structure.

Julio J. Valdés

National Research Council of Canada
Institute for Information Technology
1200 Montreal Road, Ottawa ON K1A 0R6, Canada
julio.valdes@nrc.ca

Abstract. This present paper introduces a virtual reality technique for visual data mining on heterogeneous information systems. The method is based on parametrized mappings between heterogeneous spaces with extended information systems and a virtual reality space. They can be also constructed for unions of heterogeneous and incomplete data sets together with knowledge bases composed by decision rules. This approach has been applied successfully to a wide variety of real-world domains and examples are presented from genomic research and geology.



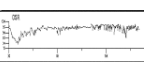



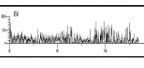

1 Introduction

In this paper a *Virtual Reality* approach is introduced for the problem of understanding heterogeneous, incomplete and imprecise data [9]. The notion of data is not restricted to classical data bases, but also to logical relations and other forms of structured knowledge. Examples are decision rules generated by inductive methods [7], rough set algorithms [6], and others. The role of visualization techniques in the knowledge discovery process is well known. Several reasons make Virtual Reality (VR) a suitable paradigm: it is *flexible* (allows the choice of different representation models according to human perception preferences), allows *immersion* (the user can navigate inside the data, interact with the objects, etc), creates a *living* experience, and is *broad and deep* (The user may see the whole world and/or concentrate on specific details). Moreover, the user needs no mathematical knowledge and only minimal computer skills.

2 The Virtual Reality Space

In the present case, *heterogeneous and incomplete information systems* will be considered [10]. They have the form $S = \langle U, A \rangle$ where U is the *universe* and A the set of *attributes*, such that each $a \in A$ has a domain V_a and an evaluation function f_a but here the V_a are not required to be finite (Table 1).

Table 1. An example of a heterogeneous data base. Attributes are from different domains(nominal, ordinal, ratio, fuzzy, images, time-series and graphs), also containing missing values (?).

A_1	A_2	A_3	A_4	A_5	A_6	A_7
yellow	?	2.5				
⋮	⋮	⋮	⋮	⋮	⋮	⋮
red	medium	?				

A heterogeneous domain is defined as a Cartesian product of a collection of *source sets* (Ψ_i): $\hat{\mathcal{H}}^n = \Psi_1 \times \dots \times \Psi_n$, where $n > 0$ is the number of *information sources* to consider. As an example, consider the case of a 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. 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 $\hat{\mathcal{R}}^{n_r} = \hat{\mathcal{R}}_1 \times \dots \times \hat{\mathcal{R}}_{n_r}$ (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, if \mathcal{O}_j is a family of ordinal source sets (with the corresponding ordering relation), \mathcal{N}_j a family of nominal variables, \mathcal{F}_j a collection of fuzzy sets, \mathcal{G}_j of graphs, and of digital images, \mathcal{I}_j , and the same procedure is applied, a heterogeneous domain is constructed as $\hat{\mathcal{H}}^n = \hat{\mathcal{R}}^{n_r} \times \hat{\mathcal{O}}^{n_o} \times \hat{\mathcal{N}}^{n_m} \times \hat{\mathcal{F}}^{n_f} \times \hat{\mathcal{G}}^{n_g} \times \hat{\mathcal{I}}^{n_i}$. Other kinds of heterogeneous domains can be constructed in the same way, using the appropriate source sets. In more general information systems the universe is endowed with a set of relations of different arities. Let $t = \langle t_1, \dots, t_p \rangle$ be a sequence of p natural integers, called *type*, and $\underline{Y} = \langle Y, \gamma_1, \dots, \gamma_p \rangle$ the extended information system will be $\hat{\mathcal{S}} = \langle U, A, \Gamma \rangle$, endowed with the relational system $\underline{U} = \langle U, \Gamma \rangle$.

A *virtual reality space* is a structure composed by different sets and functions defined as $\mathcal{Y} = \langle \underline{Q}, G, B, \mathfrak{R}^m, g_o, l, g_r, b, r \rangle$. \underline{Q} is a relational structure defined as above ($\underline{Q} = \langle O, \Gamma^v \rangle$, $\Gamma^v = \langle \gamma_1^v, \dots, \gamma_q^v \rangle$, $q \in \mathbb{N}^+$ and the $o \in O$ are objects), G is a non-empty set of *geometries* representing the different objects and relations (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.). $\mathfrak{R}^m \subset \mathbb{R}^m$ is a *metric space* of dimension m (the actual virtual reality geometric space). The other elements are mappings: $g_o : O \rightarrow G$, $l : O \rightarrow \mathfrak{R}^m$, $g_r : \Gamma^v \rightarrow G$, $b : O \rightarrow B$, r is a collection of characteristic functions for Γ^v , (r_1, \dots, r_q) s.t. $r_i : \gamma_i^{v t_i} \rightarrow \{0, 1\}$, according to the type t associated with Γ^v . The representation of an extended

information system \hat{S} in a virtual world requires the specification of several sets and a collection of extra mappings: $\hat{S}^v = \langle O, A^v, T^v \rangle, \underline{Q}$ in \mathcal{T} , which can be done in many ways. A desideratum for \hat{S}^v is to keep as many properties from \hat{S} as possible. Thus, a requirement is that U and O are in one-to-one correspondence (with a mapping $\xi : U \rightarrow O$). The structural link is given by a mapping $f : \hat{\mathcal{H}}^n \rightarrow \mathfrak{R}^m$. If $u = \langle f_{a_1}(u), \dots, f_{a_n}(u) \rangle$ and $\xi(u) = o$, then $l(o) = f(\xi(\langle f_{a_1}(u), \dots, f_{a_n}(u) \rangle)) = \langle f_{a_1^v}(o), \dots, f_{a_m^v}(o) \rangle$ ($f_{a_i^v}$ are the evaluation functions of A^v). It is natural to require that $T^v \subseteq T$, thus having a virtual world portraying selected relations from the information system. Function f can be constructed as to maximize some metric/non-metric structure preservation criteria as is typical in multidimensional scaling [1], or minimize some error measure of information loss [8], [4].

3 Examples

Clearly, a VR environment can not be shown on paper, and only simplified, grey level screen snapshots from two examples are shown just to give an idea. The VR spaces were kept simple in terms of the geometries used. The f transform used was Sammon error, with ζ_{ij} given by the Euclidean distance in \mathcal{T} and $\delta_{ij} = (1 - \hat{s}_{ij})/\hat{s}_{ij}$, where \hat{s}_{ij} is Gower's similarity [3]. For genomic research in neurology, time-varying expression data for 2611 genes in 8 time attributes were measured. Fig-1(a) shows the representation in \mathcal{T} of the information system and the result of a previous rough k-means clustering [5]. Besides showing that there is no differentiated class structure in this data, the undecidable region between the two enforced classes is perfectly clear. 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 discovered when examining the VR space and was considered interesting by the domain experts. This pattern remained

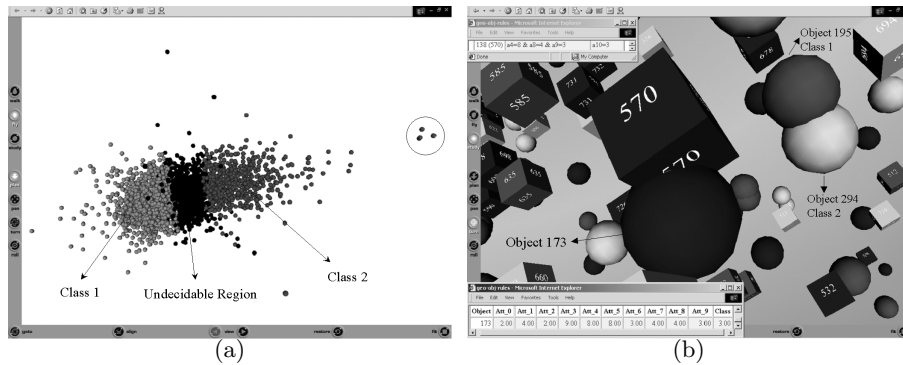


Fig. 1. VR spaces of (a) a genomic data base (with rough clusters), and (b) a geologic data base with decision rules build with rough set methods.

unnoticed since it was masked by the clustering procedure (its objects were assigned to the nearby bigger cluster).

When data sets and decision rules are combined, the information systems are of the form $S = \langle U, A \cup \{d\} \rangle$, $S_r = \langle R, A \cup \{d\} \rangle$ (for the rules), where $\{d\}$ is the decision attribute. Decision rules are of the form $\bigwedge_{i=1}^p (A_{\tau_i} = v_{\eta_i}^{\tau_i}) \rightarrow (d = v_j^d)$, where the $A_{\tau_i} \subseteq A$, the $v_{\eta_i}^{\tau_i} \in V_{\tau_i}$ and $v_j^d \in V_d$. The \hat{s}_{ij} used for δ_{ij} in A , was given by: $\hat{s}_{ij} = \frac{1}{\sum_{a \in \check{A}} \omega_{ij}} \sum_{a \in \check{A}} (\omega_{ij} \cdot s_{ij})$, where: $\check{A} = A^u$ if $i, j \in U$, A^r if $i, j \in R$ and $A^u \cap A^r$ 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 [2]. The last attribute was considered the decision attribute and the rules correspond to the *very fast* strategy, giving 99% accuracy. The join VR space is shown in Fig-1(b) where objects are spheres and rules cubes, respectively. According to RSL results, Rule 570 is supported by object 173, and they appear very close in \mathcal{Y} . Also, data objects 195 and 294 are very similar and they appear very close in \mathcal{Y} .

References

1. Borg, I., Lingoes, J.: Multidimensional Similarity Structure Analysis. Springer-Verlag 1987.
2. Gawrys, M., Sienkiewicz, J. : Rough Set Library User's Manual (version 2.0). Inst. of Computer Science. Warsaw Univ. of Technology (1993)
3. Gower, J.C.: A General Coefficient of Similarity and Some of its Properties. *Biometrics* Vol.1 No. 27 (1973) pp. 857-871
4. 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
5. 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
6. Pawlak, Z. : Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrecht, Netherlands. (1991)
7. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Series in Machine Learning (1992)
8. Sammon, J. W. A non-linear mapping for data structure analysis. *IEEE Trans. Computers*, C-18, 401-408, (1969)
9. 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)
10. 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