Visual data mining of symbolic knowledge using rough sets, virtual reality and fuzzy techniques: an application in geophysical prospecting

Valdés, Julio

Publisher's version / Version de l'éditeur:

Proceedings of the International Conference on Conceptual Structures 2007 (ICCS 2007), 2007

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC·CNRC

*Visual Data Mining of Symbolic Knowledge using Rough Sets, Virtual Reality and Fuzzy Techniques: An Application in Geophysical Prospecting**

Valdés, J.
2007

* Proceedings: ICCS 2007. Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence Series. 2007. NRC 49300.

Canada

Visual Data Mining of Symbolic Knowledge using Rough Sets, Virtual Reality and Fuzzy Techniques: An Application in Geophysical Prospecting

Julio J. Valdés

National Research Council Canada
Institute for Information Technology
M50, 1200 Montreal Rd.,Ottawa, ON K1A 0R6
julio.valdes@nrc-cnrc.gc.ca,
WWW home page: http://iit-iti.nrc-cnrc.gc.ca

Abstract. Visual data mining using nonlinear virtual reality spaces (VR) is applied to symbolic knowledge in the form of production rules obtained by rough sets methods in a classification problem with partially defined and imprecise classes. In the context of a geophysical prospecting problem aiming at finding underground caves, a virtual reality nonlinear space for production rules is constructed. The distribution of the rough sets derived rules is characterized by a fuzzy model in both the original 5D space and in the 3D VR space. The membership function of the target class (the presence of a cave) is transferred from the rules to the data objects covered by the corresponding rules and mapped back to the original physical space. The fuzzy model built in the VR space predicted sites where new caves could be expected and one of them was confirmed.

1 Introduction

While applied frequently to databases, visualization techniques have not been applied often to the analysis of symbolic information. However, symbolic knowledge like for example, sets of production rules, are difficult to interpret for humans because of their more abstract nature and this is where visual methods become important aid. The purpose of this paper is to show that in addition to the understanding of symbolic knowledge provided by visual techniques, in particular virtual reality spaces [9], [11], mathematical models can be derived from the geometric properties of the symbolic objects in these spaces which can solve complex classification problems. In particular, situations where some of the classes are undefined because of lack of knowledge about class membership and where in addition, the classes themselves are *fuzzy*.

A general approach is proposed consisting of: *i)* use rough sets techniques for learning production rules from the original data using the imperfectly defined class labels *ii)* construct a nonlinear virtual reality space preserving the structure of the rules, *iii)* perform a data analysis in the new and the high dimensional space of the rules, *iv)* construct a fuzzy model based on the geometric properties of the rules in these spaces, *v)* induce the membership functions of the known classes to the database objects covered by the rules. This approach is applied to a real-world problem: the geophysical prospecting of

caves, where class membership can be defined only for a certain subset of the database objects and where the classes (presence/absence of a cave) are *fuzzy*.

2 Virtual Reality Representation of Information Systems

Several reasons make Virtual Reality (VR) a suitable paradigm: it is is *flexible*, it allows *immersion* and creates a *living* experience. Of no less importance is the fact that in order to interact with a virtual world, no mathematical knowledge is required. A virtual reality based visual data mining technique, extending the concept of 3D modeling to information systems and relational structures, was introduced in [9], [11]. It is oriented to the understanding of large heterogeneous, incomplete and imprecise data, which includes symbolic knowledge. The objects are considered as tuples from a heterogeneous space $\hat{\mathcal{H}}^n$ [10]. A *virtual reality space* is the tuple $\Upsilon =< \underline{O}, G, B, \Re^m, g_o, l, g_r, b, r >$, where $\underline{O}$ is a relational structure ($\underline{O} =< O, \Gamma^v >$, the O is a finite set of objects, and Γ^v is a set of relations), G is a non-empty set of *geometries* representing the different objects and relations. B is a non-empty set of *behaviors* of the objects in the virtual world. $\Re$ is the set of real numbers and $\Re^m \subset \mathbb{R}^m$ is a *metric space* of dimension m (Euclidean or not) which will be the actual virtual reality geometric space. The other elements are mappings: $g_o : O \rightarrow G$, $\varphi : O \rightarrow \Re^m$, $g_r : \Gamma^v \rightarrow G$, $b : O \rightarrow B$.

Several desiderata can be considered for building a VR-space [11]. From an unsupervised perspective, the role of φ could be to maximize some metric/non-metric structure preservation criteria (e.g. similarity) [2]. 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). An error measure frequently used is [7]:

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

3 The Data Mining Process

The original data is processed with Rough Sets techniques and rules are obtained relating the prediction attributes with the classes (this result is considered partial, if not all of the class are known). Then a virtual reality space for the obtained rules is built (using the method of Fletcher-Reeves [5]) and an analysis of the rules in the original and in the new spaces is made. Finally, the results of the analysis are mapped back into the original physical space for interpretation. Fig.-1.

3.1 Application to a Geophysical Prospecting Problem

Cave detection is a very important problem in civil and geological engineering. Typically caves are not opened to the surface and geophysical methods are required for their detection, which is a complex task. In a pilot investigation, geophysical methods and a topographic survey were used with the goal of deriving criteria for predicting the presence of underground caves [8]. In the studied area, a cave was known to exist, but the

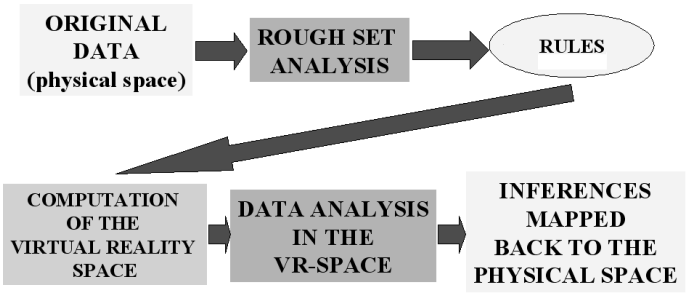

Fig. 1: The data analysis process.

presence of others was suspected. This is a problem with *partially* defined classes: the existence of a cave beneath a measurement station is either known for sure or *unknown* (i.e. only one class membership is really defined). Moreover, the classes themselves are also imprecise or *fuzzy*, as there are no sharp boundaries between the classes. The problem is not the typical two-classes presence/absence one because only one class is known with certainty: a combination of unsupervised and supervised approaches is required.

Five geophysical methods were measured on a regular grid containing 1225 measurement stations (the objects) [8]. The measured fields were: *i)* the spontaneous electric potential (SP_{dry}) at the earth's surface measured during the dry season, *ii)* the vertical component of the electro- magnetic field in the VLF region of the spectrum (frequency range $[3-30]$ kHz), *iii)* the spontaneous electric potential measured during the rainy season (SP_{dry}), *iv)* the gamma ray intensity (Rad) and *v)* the topography (Alt). A data preprocessing process was performed consisting of: *i)* conversion of each physical field to standard scores (zero mean and unit variance), *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)* fitting a least squares two-dimensional linear trend $\hat{t}(x,y) = c_0 + c_1 x + c_2 y$ and computation of the residual: $\hat{r}(x,y) = f(x,y) - \hat{t}(x,y)$, *iv)* Convolution of the residual with a low-pass zero-phase shift two-dimensional digital filter [3] to attenuate the noise component, and *v)* Re-computation of the standard scores and addition of a class attribute indicating whether a cave is known to exist below the corresponding measurement station or if it is unknown. The pre-processed data set will be called *prp-data*.

Rough set analysis was performed using the Rosetta system [6]. The *prp-data* was discretized using the Boolean Reasoning algorithm and reducts were computed. Only one reduct was found containing all of the five attributes, indicating that none of the observed geophysical fields can be discarded without loosing discernibility. A set of 345 rules were obtained from the reduct and the following are two examples:

SPdry([*, -1.50209)) AND VLF([*, -1.14882)) AND
SPrain([*, -0.46789)) AND Rad([*, -1.54413)) AND
Alt([*, -1.22398)) => (CAVE is present) (6 objects)

SPdry([-0.16981, *)) AND VLF([-0.75462, *)) AND
SPrain([0.48744, *)) AND Rad([-0.21015, *)) AND
Alt([0.00346, *)) => (CAVE is unknown) (123 objects)

For each pair of rules, a similarity measure was computed using the condition attributes. In this case the measure used was Gower's coefficient (s) [4], converted into a dissimilarity measure δ using the transformation $\delta = (1/s) - 1$. A VR-space minimizing Eq.1 was computed as described in [9], [11] and a snapshot is shown in Fig.2 as a static picture. Each sphere is a rough set rule from the knowledge base: dark objects represent rules leading to the Cave class and lighter objects represent rules leading to the Unknown class. The wrapping surface is the convex hull of the Cave class wrapping all of its rules (computed according to [1]) and the star indicates its centroid. There are rules leading to the unknown class which are within the hull of the cave class, indicating that they are similar to those concluding about the presence of a cave. Another subset of the rules concluding about the unknown class is located outside of the surface enclosing the set of rules of the cave class. They are more representative of the *no-cave* situation. In Fig.2, the distance d between any rule in the space and the centroid of

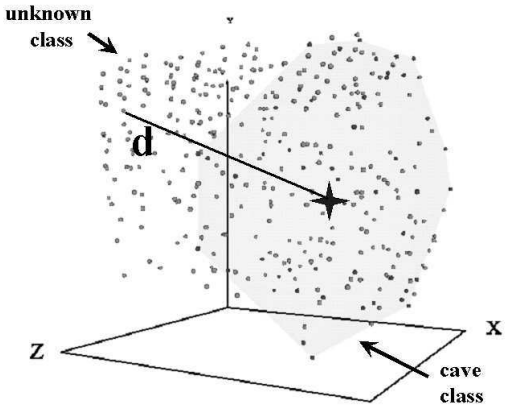

Fig. 2: Snapshot of the VR space containing the rules obtained via Rough Sets. Dark objects: cave class. Light objects: unknown class. Many objects of the unknown class are within the cave class.

the Cave class is shown. The distance between any rule in the space and this centroid gives an indication about how similar the corresponding rule is of being a descriptor of the cave properties as an abstract concept. This notion can be formalized as a fuzzy

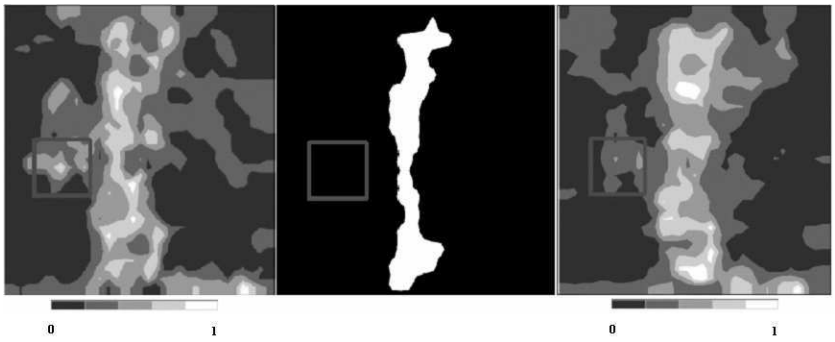

Fig. 3: Spatial distribution of the μ_{cave} function derived from the rough sets rules. Left: μ_{cave} from the original 5-D space. Center: The area with the known cave. Right: μ_{cave} from the VR space. In all images, a square shows the area where a borehole hit a previously unknown cave.

property with a membership function constructed, among many others, as:

$$\mu_{cave}(r_i) = 1 - \frac{d(r_i, c)}{d_{max}} \tag{2}$$

where $\mu_{cave}(r_i)$ is the membership of rule r_i to the cave class, $d(r_i, c)$ is the distance between the i-th rule and the centroid c of the cave class and d_{max} is the maximal distance between the centroid and the farthest rule.

Two μ_{cave} functions were computed: *i)* in the original 5-D space of the attributes appearing in the condition part of the rules and *ii)* in the VR 3-D space. Since rules are abstract symbolic entities (without any physical location), these results have to be mapped back to the physical space. This was done by transferring the membership to the cave class from each given rule to the objects covered by the rule, which correspond to the measurement points on the earth's surface (the physical space). Thus, each fuzzy membership function $\mu_{cave}(r_i)$ of the rules leads to a two dimensional fuzzy membership function of the objects with respect to the cave class ${\mu_{cave}}^{o_j}(x, y)$, where (x, y) are the coordinates corresponding to the j-th data object o_j, covered by rule r_i. The distribution of the fuzzy memberships computed in the original 5-D space and in the 3-D VR-spaces are shown in Fig.3.1(left and right respectively), as well as the map of the area, with the location of the known cave (Fig.3.1-center).

The fuzzy membership function in the original 5D space (Fig.3.1-left), has a central narrow band of high values which corresponds to the location of the known cave. In addition, there are other areas of high values located at the center-left and bottom-right, both beyond the outline of the surveyed cave. This suggests the presence of other caves, not opened to the surface. These areas are wrapped by a medium membership value enclosure emerging from the one enclosing the known cave which suggests that they might be a part of the same cave system.

A similar behavior is exhibited by the fuzzy membership function in the nonlinear VR space, shown in Fig.3.1 bottom-left. The patterns observed are the same in terms of the appearance of a central band of high values and the two additional areas of high

membership values. The results can be perceived as more clear because the function is smoother. This indicates that the information lost during the nonlinear mapping of the original 5D space to the 3D space actually increased the signal to noise ratio, which is a very important feature. Some time after the geophysical investigation was made, a borehole was drilled in the location corresponding to the center-left area of high membership indicated above. A a cave was hit, thus confirming the results suggested by the presented approach.

4 Conclusions

Visual data mining of symbolic knowledge obtained with rough sets proved to be effective in understanding complex problems with partially defined and imprecise classes. Fuzzy models derived from the original rough set rules and from a virtual reality space obtained from them by nonlinear mapping, revealed the essential properties of the target class. In the studied case of a geophysical prospecting problem, it allowed the identification of areas where the presence of new hidden caves could be expected and one was confirmed by drilling. The comparisson of the fuzzy membership function in the original and in the VR space turned out to be a very effective noise reduction filter, which also preserves most of the information associated with the target class. This approach is domain-independent and could be applied to similar problems in other areas.

References

1. Barber, C. B., Dobkin, D. P., Huhdanpaa, H. T.: The Quickhull algorithm for convex hulls. ACM Trans. on Mathematical Software, 22(4),(1996). pp469–483.
2. Borg, I., and Lingoes, J., Multidimensional similarity structure analysis: Springer-Verlag, (1987), 390 p.
3. D.E. Dudgeon, R.M. Mersereau. Multidimensional Signal Processing. Prentice Hall, 1984.
4. Gower, J.C., A general coefficient of similarity and some of its properties: Biometrics, v.**1**, no. 27, p. 857–871. (1973).
5. Press, W.H.,Teukolsky S.A., Vetterling W.T., Flannery B.P: Numerical Recipes in C. The Art of Scientific Computing. Cambridge Univ. Press, 994 p. (1992).
6. Øhrn A., Komorowski J.: Rosetta- A Rough Set Toolkit for the Analysis of Data. Proc. of Third Int. Join Conf. on Information Sciences (JCIS97), Durham, USA, (1997), 403–407.
7. Sammon, J.W. A non-linear mapping for data structure analysis. IEEE Trans. on Computers C18, p 401–409 (1969).
8. Valdés J.J, Gil J. L. Joint use of geophysical and geomathematical methods in the study of experimental karst areas. *XXVII International Geological Congress*, pp 214, Moscow, 1984.
9. Valdés, J.J.: Virtual Reality Representation of Relational Systems and Decision Rules: An exploratory Tool for understanding Data Structure. In Theory and Application of Relational Structures as Knowledge Instruments. Meeting of the COST Action 274 (P. Hajek. Ed). Prague, November 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, pp 499–508, (2002).
11. Valdés, J.J.: Virtual Reality Representation of Information Systems and Decision Rules: An Exploratory Tool for Understanding Data and Knowledge. Lecture Notes in Artificial Intelligence LNAI 2639, pp. 615-618. Springer-Verlag (2003).