Interpreting fuzzy clustering results with virtual reality-based visual data mining: application to microarray gene expression data

Valdes, Julio

Publisher's version / Version de l'éditeur:
https://doi.org/10.1109/NAFIPS.2004.1336296
IEEE Annual Meeting of the Fuzzy Information, 2004. Processing NAFIPS '04, pp. 302-307, 2004-09-27

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Interpreting Fuzzy Clustering Results With Virtual Reality-based Visual Data Mining: Application to Microarray Gene Expression Data *

Valdes, J.
March 2004

* published in North American Fuzzy Information Processing Society (NAFIPS) 2004. June 27-30, 2004. NRC 46560.

Canada

Interpreting Fuzzy Clustering Results With Virtual Reality-based Visual Data Mining: Application to Microarray Gene Expression Data*

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 paper combines fuzzy clustering with a virtual reality based technique for visual data mining. The purpose is to construct virtual reality spaces preserving as much structural information from the original data as possible, where the results of fuzzy clustering procedures can be displayed and analyzed. The construction of such spaces involves non-linear transformations of the original feature space, which can be either the space of the original attributes or the space of the fuzzy memberships with respect to the constructed fuzzy classes. In particular, the representation involves the centroids of the different classes, the individual memberships of all of the studied objects with respect to all of the fuzzy classes, and eventually their comparison with additional crisp partitions or partitions induced by a decision attribute. This approach is applied to different data sets from the fields of biology and medicine, including microarray gene expression data related to Alzheimer's disease and Leukemia. The visual inspection and the navigation in the virtual reality spaces, provided useful insights about *i)* the quality of the obtained classifications, *ii)* the overlapping of different classes, and *iii)* their relationships.**

I. Introduction

Fuzzy clustering [1], [2], [3], [4], [5] has been a very successful tool in data analysis, as demonstrated by many successful applications in different domains. In bioinformatics, fuzzy clustering can be an important tool in the understanding of microarray gene expression data. It is known that genes can have different functions, and due to the complex relationships between them, overlapping clusters can be expected when classifying either patient samples described by the expression behavior of sets of genes, or when the genes themselves are classified. Despite being a very effective tool, difficulties arise when interpreting fuzzy clustering results. In the case of large samples, the large number of membership values with respect to the constructed clusters makes it almost impossible to effectively compare the fuzzy properties of the objects. In the case of more than three or four classes, the mutual relationships between specific classes of interest can be masked. In addition, the relationships between data structure and fuzzy clustering results are difficult to understand when the dimensionality of the data set is large. In gene expression experiments, thousands of genes are normally used as attributes for characterizing samples. Even when the genes themselves are investigated (either their behavior in time, or in relation to patient or diseases), tens or hundreds of attributes are common.

The purpose of this paper is to use virtual reality representations of heterogeneous relational structures, as introduced in [6], [7], to visualize fuzzy clustering results. This approach allows the simultaneous analysis of data structure, crisp classifications defined on the data, and also fuzzy partitions. The advantages of a virtual reality environment from the point of view of navigation, data interaction, etc, creates an intuitively simple and at the same time powerful way to understand and interpret complex data.

II. Fuzzy Clustering

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 [1], and later on investigated by many others [2], [3], [4], [5]. In a fuzzy partition of n objects into K clusters, the state of clustering is by a n x K matrix $U=(u_{ik})$ where $u_{ik} \in [0,1]$, $i=1,\cdots,n$; $k=1,\cdots,K$, and the requirement that $\sum_{k=1}^{K} u_{ik}=1$. 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

* This research was performed in the framework of the Biomine Project. The author would like to thank Alan Barton, Robert Orchard and Fazel Famili from the Institute of Information Technology of the National Research Council of Canada.

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:

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

where $\vec{x}_i$ is a vector representing data object i, $\vec{v}_k$ 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 (1). The classical algorithm proceeds by obtaining successive approximations by first estimating the centroids

$$\vec{v}_{ka} = \frac{\sum_{i=1}^{n} (u_{ik})^m \; \vec{x}_{ia}}{\sum_{i=1}^{n} (u_{ik})^m}$$, where $a = 1, \cdots, p$ (p is the number of attributes of the data objects). Then, the memberships are approximated according to

$$u_{ik} = \left[\sum_{j=1}^{K} \left(\frac{d(\vec{x}_i, \vec{v}_k)}{d(\vec{x}_j, \vec{v}_k)} \right)^{1/m-1} \right]^{-1}$$

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 (1). Among them are the partition coefficient F_c, and the entropy H_c of the partition *U*, given by

$$F_c(U) = \sum_{k=1}^{K} \sum_{i=1}^{n} \frac{(u_{ik})^2}{n} \qquad (3)$$

$$H_c(U) = -\sum_{k=1}^{K} \sum_{i=1}^{n} \frac{u_{ik} \ln(u_{ik})}{n} \qquad (4)$$

III VIRTUAL REALITY AS A DATA MINING TOOL

This is a technique for visual data mining of heterogeneous relational structures (like databases or knowledge bases), based on virtual reality (http://www.hybridstrategies.com), [6] and [7]. 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* [8], given by a Cartesian product of different *source* sets like: nominal, ordinal, real-valued, fuzzy-valued, image-valued, time-series-valued, graph-valued, etc. A set of relations of different arities may be defined over these objects. The construction of a VR-space requires the specification of several sets and a collection of extra mappings, which may be defined in infinitely many ways. A desideratum for the VR-space is to keep as many properties from the original space as possible, in particular, the similarity structure of the data [9]. The method is based on parameterized mappings between the heterogeneous space $\hat{H}$ representing the original data and the virtual reality space. The former can also be constructed for unions of information systems (e.g. heterogeneous and incomplete data sets together with knowledge bases composed by decision rules), simplifying the process of discovery of interesting patterns as well as relationships between the original data and the symbolic expressions representing the structured knowledge.

A virtual reality space is composed by different sets and functions: $\Omega = < \underline{O}, G, B, \Re^m, g_0, l, g_r, b, r >$, where $\underline{O}$ is a relational structure (a set of objects *O*, and attributes, endowed with a set Γ^v relations defined over the objects), G is a non-empty set of *geometries* representing the different objects and their relationship in the visual space (an *empty* or *invisible* geometry is a possibility), 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.), $\Re^m$ is a *metric space* of dimension m which will be the actual virtual reality geometric space (usually $m = 3$). The rest of the elements are mappings: $g_0 : O \to G$, $l : O \to \Re^m$, $g_r : \Gamma^v \to G$, and r is a collection of characteristic functions for Γ^v.

The representation of an extended information system (i.e. database) $\hat{S}$ implies the construction of another one $\hat{S}^v$ in the virtual world. It requires the specification of several sets and a collection of extra mappings. There are many ways in which it can be done. A desideratum for the virtual reality heterogeneous space $\hat{H}^v$ is to keep as many properties from $\hat{S}$ as possible, in particular, the similarity structure of the original data. In this sense, the idea is to optimize some metric/non-metric structure preservation criteria as in multidimensional scaling [10] and [11]. If δ_{ij} is a dissimilarity measure between any two objects i, j, and ξ_{ij} is another dissimilarity measure defined on objects i^v, j^v in the virtual reality space (the images of the original objects i, j), an error measure frequently used is the *Sammon error:*

$$E = \frac{1}{\sum_{i^v < j^v} \delta_{i^v j^v}} \frac{\sum_{i^v < j^v} \left(\delta_{i^v j^v} - \xi_{i^v j^v}\right)^2}{\delta_{i^v j^v}} \qquad (5)$$

The transformation l obtained by solving (5) is implicit, as no analytic representations are found.

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. This provides a rich framework where appropriate measures capable of detecting interrelationships hidden in the data can be found, more suited to both its internal structure and to external criteria.

IV. Virtual Reality Spaces for representing fuzzy clustering results

A. *Gene Expression Data from Alzheimer's disease*

Alzheimer's disease (AD) is a chronic, progressive, debilitating condition which, along with other neuro-degenerative diseases, represents the largest area of unmet need in modern medicine [12], and 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. 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 comprehensive study of gene expression Alzheimer's data from a data mining perspective is presented in [12]. Among other techniques, visual data mining using VR spaces was used with very good results.

In that study, the data set was composed by 23 samples taken from Alzheimer and non-Alzheimer cases. These samples are described in terms of 9600 genes. 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, and a virtual reality representation for the data was computed.

With this data set, a collection of fuzzy clustering experiments was performed with the following parameters: number of clusters={2, 3, 4}, fuzzy exponent *m*={1.25, 1.5, 1.75, 2, 3, 4, 5, 6, 7, 8, 9, 10}, dissimilarity functions given by distance metrics (Euclidean, diagonal and Mahalanobis). Collections of 10 random approximations were tried for each configuration when computing the virtual reality space (see below) for a total of 1080 experiments. Equation (1) was the objective function to minimize. The partition coefficient (2) and the entropy (3) were computed for each solution. In 179 solutions the partition coefficient was at least 0.9, with an entropy range of [0.00540, 0.35132]. Within this set, 42.5% (76) of the solutions were obtained with Euclidean distance, 40.8% (73) with diagonal metrics and the remaining 16.7% (30), with the Mahalanobis distance. An idea of the different degrees of fuzziness found is given by the distribution of the values of the exponent *m*. (50.28% with *m*= 1.25, 33.52% with *m*= 1.5, and 16.2% to *m*= 1.75). Subsequent lower thresholds in the partition coefficient lead to different distributions of the mentioned parameters.

The measure used as structure link function between the space of the original data and a 3D space suitable for visualization was the Sammon error (5). Gower's dissimilarity coefficient [13] was used as δ_{ij}, whereas Euclidean distance in the target space was used as ξ_{ij}. An implicit representation was computed via deterministic optimization with a Newton-type gradient descent technique. In this classical algorithm, a random approximation is used as initial solution, and it is refined in successive iterations by multiplying each of the coordinates of the data objects in the target space by a correction factor given by $\eta\left(E^1 / E^{11}\right)$, where η is the step size, and E^1, E^{11} are the first and second partial derivatives of (5) w.r.t the coordinates of the new space. In this study the step size was kept fixed at a value equal to 0.15.

For obvious reasons it is not possible to present a virtual reality environment on hardcopy media. Navigation, interaction and many other features inherent to the functionalities required by this approach to visual data mining are completely lost. Thus, only snapshots of specific regions can be shown. Moreover, the color information has to be transformed into gray level tones. 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-1.

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*.

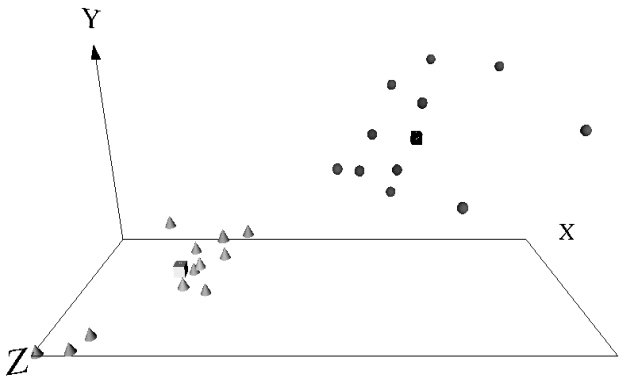

Figure. 1. Snapshot of part of the virtual reality representation of the Alzheimer's data (with four selected genes). The cones represents the samples from the Alzheimer class, and the spheres the samples from the non-Alzheimer class. The cubes are the centroids of the corresponding classes (pure white 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.

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.

B. Gene Expression Data from Leukemia

The dataset used is that of [14], and consists of 7129 genes where patients are separated into *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. Note that, the test set contains a much broader range of biological samples, including those from peripheral blood rather than bone marrow, from childhood AML patients, and from different reference laboratories that used different sample preparation protocols. Further, the dataset is known to have two types of ALL, namely B-cell and T-cell. For the purposes of investigation, only the AML and ALL distinction was made. The dataset distributed by [14] contains preprocessed intensity values, which were obtained by re-scaling such that overall intensities for each chip are equivalent.

A data mining procedure combining different clustering methods, rough set, and other techniques, was applied to this dataset [15]. The procedure consists of a series of staged experiments were each stage feeds its results to the next stage. After each clustering solution, training and test subsets of the original raw data are constructed using cluster-derived leaders (data objects selected as class representatives). The training set is discretized with a boolean reasoning algorithm, and then reducts and decision rules are computed. The test set is discretized according to the training cuts, and classified using the training decision rules. The process is illustrated in Fig-2.

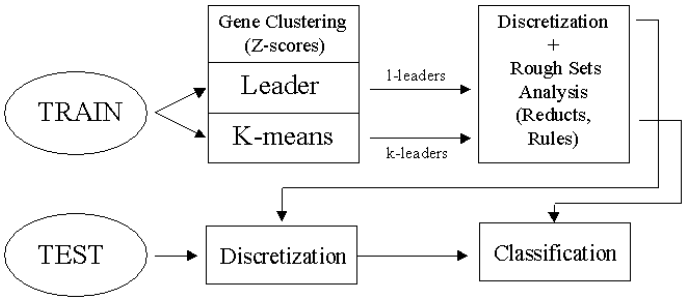

Figure. 2. Data processing strategy combining clustering with Rough Set analysis applied to Leukemia data.

The procedure leads to the identification of a subset of four very relevant genes. Some of them were found in other studies using the same data, whereas others were not previously reported. Then, a new dataset was constructed by taking all of the original objects, but described only in terms of the four relevant genes found.

With this data set, a collection of 1080 fuzzy clustering experiments was performed using the same settings as with the Alzheimer's data. In 134 solutions the partition coefficient was at least 0.9, with an entropy range of [0.0094, 0.1975]. Within this set, 61.94% (83) of the solutions were obtained with Euclidean distance, 30.6% (41) with diagonal metrics and the remaining 7.5% (10), with the Mahalanobis distance. An indication of the different degrees of fuzziness found is given by the distribution of the values of the exponent *m*. (52.24% with *m*= 1.25, 31.34% with *m*= 1.5, 8.96% to *m*= 1.75, and 7.46 to *m*= 2). Clearly, the amount of fuzziness in this data is larger than in the Alzheimer's case. From the point of view of the partition coefficient, the best solution was obtained with Euclidean distance and *m*= 1.25. The coefficient was equal to 0.99512, also with the lowest entropy.

The criteria for the computation of the virtual reality representation of the Leukemia data described by the selected genes were also the same as those used with Alzheimer's data.

Therefore, similarly, the new space can be used for displaying the results of fuzzy clustering experiments, and the relationships between the crisp partition given by the ALL and AML classes of Leukemia, with the fuzzy classifications can be visualized (Fig-3). The snapshot corresponds to a two-cluster solution with m= 4, and partition coefficient equal to 0.695258.

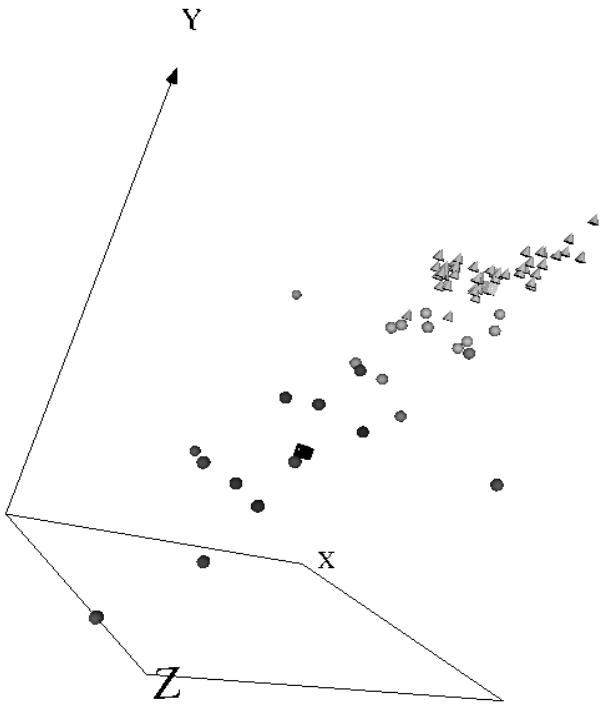

Figure. 3. Snapshot of part of the virtual reality representation of the Leukemia data (with selected genes {X95735_at, D26308_at, D21063_at, M27891_at}). The cones represents the samples from the ALL class, and the spheres the samples from the AML class. The cubes are the centroids of the corresponding classes (pure white for the ALL class, and pure black for the AML). 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 0.034.

In the virtual reality space, the ALL and AML classes are almost linearly separable. The AML class is more spread-out (i.e. less homogeneous) than the ALL class, and the effect of the fuzziness in the data can be appreciated by observing that the AML objects (the spheres) are progressively lighter in colour as they are approaching the vicinity of the ALL class. Those are cases of objects exhibiting hybrid behaviour, and it would have been more difficult to discover them by direct inspection of the fuzzy membership matrix.

Clearly, in the presence of much larger data sets, with fuzzy partitions targeting more than three classes, the difficulties would be much greater when traditional methods for interpretation are used, as compared to a virtual reality space (provided that the space accurately preserves the internal structure of the data, as given by the value of the chosen mapping function). On the other hand, the increase in complexity of the datasets does not imply a proportional increase when virtual reality spaces are used.

C. Virtual Reality Representation of the Space of Fuzzy Memberships

In the previous two examples the virtual reality spaces represented mappings of relational systems describing data sets resulting from direct observation (or measurements). However, a fuzzy partition can be formally described as a relational system in which the attributes are the fuzzy memberships u_{ik} w.r.t each of the k-fuzzy classes whose centroids were computed. Then, it is natural to apply the same principle to understand the classification structure of a fuzzy partition U. In the context of the Neurobiology program of the Institute for Biological Sciences (National Research Council of Canada), a microarray experiment produced a dataset composed by 2611 genes whose intensities were observed at 8 different times (Dr. R. Walker personal communication). As part of the data mining process, crisp and fuzzy partitions were computed, in particular, targeting different numbers of clusters. Figure 4 (left) shows a snapshot of the virtual reality space corresponding to the representation of the original data matrix (2611 genes observed at 8 different times). The grey levels indicate a crisp 5-cluster partition, which does not correspond to the natural similarity structure of the data, evidenced in the virtual reality space. When another space is computed for the fuzzy 5-cluster partition representing the memberships w.r.t the five fuzzy classes (Fig-4 right), it is clearly seen that the membership structure is that of a single, quasi-isometric cloud, thus indicating that the original data has little or no group structure. This example illustrates how the same visual data mining approach can be used for representing conceptually different entities, but also how these representations can complement each other.

V. CONCLUSIONS

The virtual reality approach for representing, in the same space, databases, crisp relations and fuzzy partitions simultaneously, is an intuitive and simple tool for visual data mining. It effectively highlights structural properties of the data from the point of view of the distribution of the natural classes. The relation between decision attributes and the existence of hybrid objects, as represented by fuzzy membership values can be more clearly distinguished. Further experiences of the use of this technique are necessary, specially when studying large data sets on which fuzzy partitions with respect to many classes have been computed.

REFERENCES

[1] E. Ruspini, "A New Approach to Clustering," *Inform. Control*, vol. 15, no. 1 pp. 22-32, April 1969.

[2] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1989.

[3] D.E. Gustafson and W.C.Kessel, "*Fuzzy clustering with a covariance matrix*," in *IEEE Conference on Decision and Control*, 1979, pp. 761-766.

[4] I. Gath and A. Deva, "*Unsupervised Optimal Fuzzy Clustering*," IEEE Trans. Of Pat Anal. and Mach Intel, 1989, pp 773-781.

[5] M. Sato, Y. Sato and L.C. Jain "*Fuzzy Clustering Models and Applications*," *Physica Verlag, Heidelberg, New York 1997*.

[6] J.J. Valdés "*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, 2002, November 14-16.

[7] J.J. Valdés, "*Virtual reality representation of information systems and decision rules: An exploratory tool for understanding data and knowledge*". Proc. 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, Springer-Verlag, 2003, pp.615-618.

[8] J.J. Valdés, "*Similarity-based heterogeneous neurons in the context of general observational models*". Neural Network World. Vol 12, No. 5, 2002, pp 499-508.

[9] I. Borg and J. Lingoes, "*Multidimensional Similarity Structure Analysis*", Springer-Verlag, New York, NY, 1987, 390 p.

[10] I. Kruskal, "*Nonmetric multidimensional scaling: A numerical method*", Psychometrica, vol. 29, 1964, pp 115-129.

[11] J.W. Sammon, "*A non-linear mapping for data structure analysis*". IEEE Trans. on Computers C18, 1969, p 401-409.

[12] R. Walker et.al. "*Data mining of gene expression changes in alzheimer brain*". Int. Jour. Of Artificial Intelligence in Medicine, Elsevier Acad. Pub. 2004 (in press).

[13] J.C. Gower, "*A general coefficient of similarity and some of its properties*", Biometrics, v.1, no. 27, 1973, p. 857-871.

[14] T.R. Golub, et. al., "*Molecular classification of cancer: class discovery and class prediction by gene expression monitoring"*. Science, vol. 286, 1999, pp531-537.

[15] J.J. Valdés and A.J. Barton, "*Gene Discovery in Leukemia Revisited: A Computational Intelligence Perspective"*. Proc. Seventeenth International Conference on Industrial Engineering Applications of Artificial Intelligence & Expert Systems (IEA/AIE 2004). Ottawa, Canada, 2004 (in press).

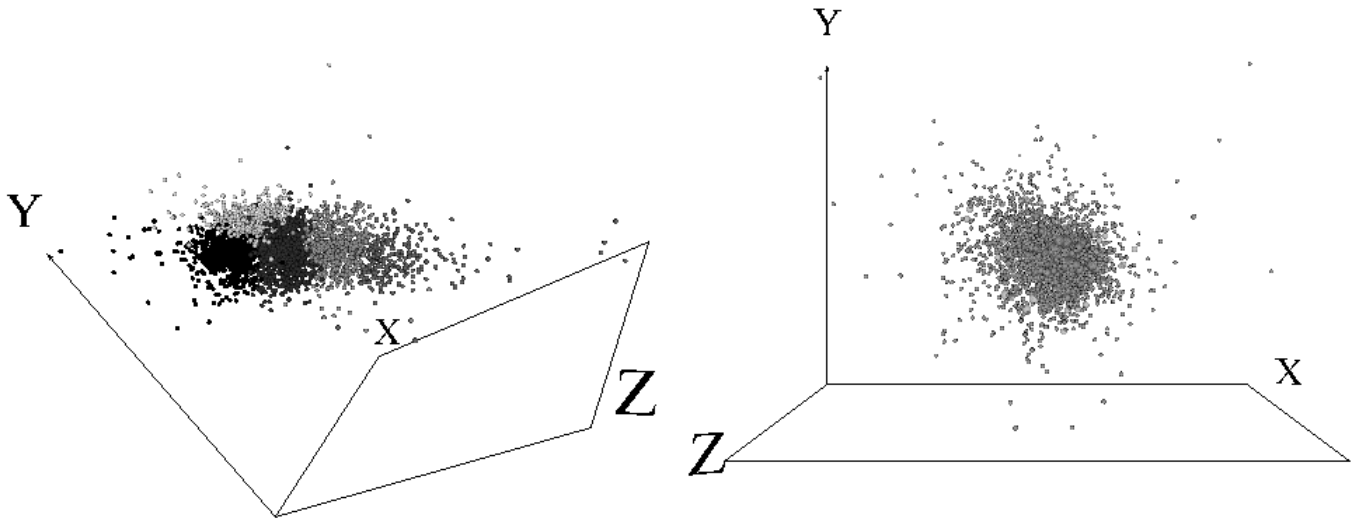

Figure. 4. Neurogenesis data : Snapshots of part of two virtual reality spaces . Left: Five grey levels represent a k-means crisp clustering.computed on the original attributes (gene intensities at 8 different times) (Sammon error = 0.06 after 100 iterations). Right: Five dimensional fuzzy membership matrix (2611 genes w.r.t. 5 fuzzy c-means classes) (Sammon error = 0.026 after 100 iterations).