Hybrid Unsupervised/Supervised Virtual Reality Spaces for Visualizing Cancer Databases: An Evolutionary Computation Approach

Valdés, Julio; Barton, Alan

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*Hybrid Unsupervised/Supervised Virtual Reality Spaces for Visualizing Cancer Databases: An Evolutionary Computation Approach **

Valdés, J. and Barton, A.
2007

* proceedings of IWANN 2007. Lecture Notes in Computer Science. Vol. 4507. 2007. NRC 49295.

Hybrid Unsupervised/Supervised Virtual Reality Spaces for Visualizing Cancer Databases: An Evolutionary Computation Approach

Julio J. Valdés and Alan J. Barton

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

Abstract. This paper introduces a multi-objective optimization approach to the problem of computing virtual reality spaces for the visual representation of relational structures (e.g. databases), symbolic knowledge and others, in the context of visual data mining and knowledge discovery. Procedures based on evolutionary computation are discussed. In particular, the NSGA-II algorithm is used as a framework for an instance of this methodology; simultaneously minimizing Sammon's error for dissimilarity measures, and mean cross-validation error on a k-nn pattern classifier. The proposed approach is illustrated with an example from cancer genomics data (e.g. lung cancer) by constructing virtual reality spaces resulting from multi-objective optimization. Selected solutions along the Pareto front approximation are used as nonlinearly transformed features for new spaces that compromise similarity structure preservation (from an unsupervised perspective) and class separability (from a supervised pattern recognition perspective), simultaneously. The possibility of spanning a range of solutions between these two important goals, is a benefit for the knowledge discovery and data understanding process. The quality of the set of discovered solutions is superior to the ones obtained separately, from the point of view of visual data mining.

1 Introduction

According to the World Health Organization (WHO) 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 worldwide. This places cancer as one of the leading causes of death in the world, with lung cancer (the main cancer leading to mortality) accounting for 1.3 million deaths per year. Thus the importance of understanding the mechanisms of lung cancer is clear. One approach is through the rapid quantification of the gene expression levels of samples of healthy and diseased lung tissue. This new field blending the knowledge from biologists, computer scientists and mathematicians is known as Bioinformatics and is yielding large quantities of data of a very high dimensional nature that needs to be understood.

The increasing complexity of the data analysis procedures makes it more difficult for the user (not necessarily a mathematician or data mining expert), to extract useful information out of the results generated by the various techniques. This makes

graphical representation directly appealing; for which Virtual Reality (VR) is a suitable paradigm. Virtual Reality is *flexible*, it allows the construction of different virtual worlds representing *the same* underlying information, but with a different look and feel. VR allows *immersion*, that is, the user can navigate inside the data, interact with the objects in the world. VR creates a *living* experience. The user is not merely a passive observer but an actor in the world. VR is *broad and deep*. The user may see the VR world as a whole, and/or concentrate the focus of attention on specific details of the world. Of no less importance is the fact that in order to interact with a Virtual World, no mathematical knowledge is required, and the user only needs minimal computer skills. A virtual reality technique for visual data mining on heterogeneous, imprecise and incomplete information systems was introduced in [24, 25] (see also `http://www.hybridstrategies.com`).

The purpose of this paper is to explore the construction of high quality VR spaces for visual data mining using a multi-objective optimization technique applied to the understanding of a publicly available lung cancer gene expression data set. This approach provides both a solution for the previously discussed problem, and the possibility of obtaining a set of spaces in which the different objectives are expressed in different degrees, with the proviso that no other spaces could improve any of the considered criteria individually (if spaces are constructed using the solutions along the Pareto front). This strategy clearly represents a conceptual improvement in comparison with spaces computed from the solutions obtained by single-objective optimization algorithms in which the objective function is a weighted composition involving different criteria.

2 The multi-objective approach: A hybrid perspective

In order to establish a formulation of the problem based on multi-objective optimization, a set of objective functions has to be specified, representing the corresponding criteria that must be simultaneously satisfied by the solution. The minimization of a measure of similarity information loss between the original and the transformed spaces and a classification error measure over the objects in the new space can be used in a first approximation. Clearly, more requirements can be imposed on the solution by adding the corresponding objective functions. Following a principle of parsimony this paper will consider the use of only two criteria, namely, Sammon's error (Eq-3) for the unsupervised case and mean cross-validated classification error with a k-nearest neighbour pattern recognizer for the supervised case.

The proximity (or similarity) of an object to another object may be defined by a distance (or similarity) calculated over the independent variables and can be defined by using a variety of measures. In the present case a normalized Euclidean distance is chosen:

$$d_{\overleftarrow{x}\,\overleftarrow{t}} = \sqrt{(1/p)\sum_{j=1}^{p}(x_{ij} - t_{kj})^2} \tag{1}$$

2.1 Structure Preservation: An unsupervised perspective

Examples of error measures frequently used for structure preservation are:

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

$$\text{Sammon error} = \frac{1}{\sum_{i<j} \delta_{ij}} \frac{\sum_{i<j} (\delta_{ij} - \zeta_{ij})^2}{\delta_{ij}} \tag{3}$$

$$\text{Quadratic Loss} = \sum_{i<j} (\delta_{ij} - \zeta_{ij})^2 \tag{4}$$

For heterogeneous data involving mixtures of nominal and ratio variables, the Gower similarity measure [11] has proven to be suitable. The similarity between objects i and j is given by

$$S_{ij} = \sum_{k=1}^{p} s_{ijk} / \sum_{k=1}^{p} w_{ijk} \tag{5}$$

where the weight of the attribute (w_{ijk}) is set equal to 0 or 1 depending on whether the comparison is considered valid for attribute k. If $v_k(i), v_k(j)$ are the values of attribute k for objects i and j respectively, an invalid comparison occurs when at least one them is missing. In this situation w_{ijk} is set to 0.

For quantitative attributes (like the ones of the datasets used in the paper), the scores s_{ijk} are assigned as

$$s_{ijk} = 1 - |v_k(i) - v_k(j)|/R_k$$

where R_k is the range of attribute k. For nominal attributes

$$s_{ijk} = \begin{cases} 1 \text{ if } v_k(i) = v_k(j) \\ 0 \text{ otherwise} \end{cases}$$

This measure can be easily extended for ordinal, interval, and other kind of variables. Also, weighting schemes can be incorporated for considering differential importance of the descriptor variables.

2.2 Multi-objective Optimization Using Genetic Algorithms

An enhancement to the traditional evolutionary algorithm[1], is to allow an individual to have more than one measure of fitness within a population. One way in which such an enhancement may be applied, is through the use of, for example, a weighted sum of more than one fitness value [3]. Multi-objective optimization, however, offers another possible way for enabling such an enhancement. In the latter case, the problem arises for the evolutionary algorithm to select individuals for inclusion in the next population, because a set of individuals contained in one population exhibits a Pareto Front[19] of best current individuals, rather than a single best individual. Most [3] multi-objective algorithms use the concept of dominance to address this issue.

A solution $\overleftarrow{x}_{(1)}$ is said to dominate [3] a solution $\overleftarrow{x}_{(2)}$ for a set of m objective functions $< f_1(\overleftarrow{x}), f_2(\overleftarrow{x}), ..., f_m(\overleftarrow{x}) >$ if

1. $\overleftarrow{x}_{(1)}$ is not worse than $\overleftarrow{x}_{(2)}$ over all objectives.
   For example, $f_3(\overleftarrow{x}_{(1)}) \leq f_3(\overleftarrow{x}_{(2)})$ if $f_3(\overleftarrow{x})$ is a minimization objective.
2. $\overleftarrow{x}_{(1)}$ is strictly better than $\overleftarrow{x}_{(2)}$ in at least one objective. For example, $f_6(\overleftarrow{x}_{(1)}) > f_6(\overleftarrow{x}_{(2)})$ if $f_6(\overleftarrow{x})$ is a maximization objective.

One particular algorithm for multi-objective optimization is the elitist non-dominated sorting genetic algorithm (NSGA-II) [7], [6], [5], [3]. It has the features that it *i)* uses elitism, *ii)* uses an explicit diversity preserving mechanism, and *iii)* emphasizes the non-dominated solutions.

2.3 Original Study

Gene expressions were compared in [21] for 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). 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 $9,336$ 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 was obtained from `http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=737`.

2.4 Experimental Settings

Each sample in this study is a vector in a high dimensional space, and therefore, direct inspection of the structure of this data, and of the relationship between the descriptor variables (the genes) and the type of sample (normal or cancer), is impossible. Moreover, within the collection of genes there is a mixture of potentially relevant genes with others which are irrelevant, noisy, etc. The need of simultaneously finding a visual representation (3D) respecting (as much as possible) the set of object interrelationships as defined by the original attributes, and the construction of a new feature space effectively differentiating the two classes of objects present, makes this problem suitable for a multi-objective optimization approach.

The collection of parameters describing the application of the NSGA-II algorithm is shown in Table-1. A modest population size and number of generations were used, with a relatively high mutation probability in order to enable richer genetic diversity. Randomization of the set of data objects was applied in order to reduce the bias in the composition of the cross-validated folds by providing a more even class distribution between successive training and test subsets. The number of folds was set in consideration of the sample size.

Table 1: Experimental settings for computing the pareto-optimal solution approximations by the multi-objective genetic algorithm (PGAPack extended by NSGA-II).

population size	100	number of generations	500
chromosome length	90 ($= 3 \cdot 30$)	ga seed	101
No. new inds. in ($i + 1$st) pop.	20	objective functions should be	minimized
chromosome data representation	real	crossover probability	0.8
crossover type	uniform (prob. 0.6)	mutation probability	0.4
mutation type	gaussian	selection type	tournament
tournament probability	0.6	mutation and crossover	yes
population initialization	random, bounded	lower bound for initialization	-2
upper bound for initialization	2	fitness values	raw
stopping criteria	maximum iterations	restart ga during execution	no
parallel populations	no		
number of objectives	2	number of constraints	0
pre-computed diss. matrix	Gower dissimilarity		
evaluation functions	mean cross-validated k-nn error and Sammon error		
cross-validation (c.v.)	5 folds	randomize before c.v.	yes
knn seed	101	k nearest neighbors	3
non-linear mapping measure	Sammon	dimension of the new space	3

2.5 Results

The set of non-dominated solutions obtained by the NSGA-II algorithm is shown in the scatter plot of Fig-1(a), where the horizontal axis is the mean cross-validated knn error and the vertical axis the Sammon error. The approximate location of the Pareto front is defined by the convex polygon joining the solutions provided by chromosomes 2, 1, 10, etc. Chromosome 2 defines a space with a perfect resolution of the supervised problem in terms of the "*no or mild emphysema*" and "*severe emphysema*" classes (knn error $= 0$), but at the cost of a severe distortion of the space. Whereas, chromosome 1 approximates a pure unsupervised solution (with low Sammon error). Its classification error is large indicating that few non-linear features preserving the similarity structure lacks classification power. This may be due to the large amount of attribute noise, redundancy, and irrelevancy within the set of $22,283$ original genes.

Clearly, it is impossible to represent virtual reality spaces on a static medium. However, a composition of snapshots of the VR spaces using the solutions along the Pareto front approximation is shown in Fig-1(b-d). Different mappings (even with important differences from the point of view of the mapping error) lead to similar 3D visual representations, which indicates good reproducibility of the solutions. The similarities are associated to the main distributions of the clouds of points, which are preserved, while there might be local discrepancies with respect to the placement of some objects.

A solution satisfying classification error as much as possible (actually with 0-error) is shown in Fig-1(b) where both classes are separated into 2 main clouds of points and a distinct point, Object 6, positioned separately from the clouds. It can be seen that Object 6 is positioned relatively differently in the spaces that comprise the best Sammon error Fig-1(d) and tradeoff between the classification error objective and Sammon er-

ror objective Fig-1(c). This is why, visually, the latter space represents a compromised solution between the two goals and a tradeoff between the two objective functions. It should be remembered that the class information is not used at all for computing the spaces. Chromosome 10, according to Fig-1(a) and Fig-1(c), can be considered to be the best multi-objective compromised solution in which both error criteria are simultaneously as low as possible. It shows a reasonable class discrimination with a non-large similarity structure distortion, which is a very meaningful result.

3 Conclusions

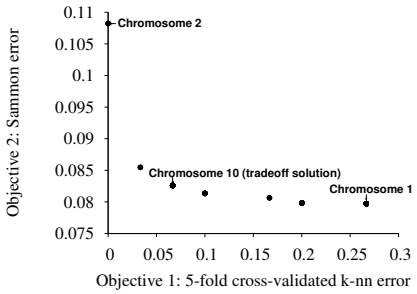
A multi-objective optimization approach was introduced for the problem of computing virtual reality spaces in the context of visual data mining and knowledge discovery applied to relational structures (e.g. databases). The multi-objective procedure was based on NSGA-II using two objective functions representative of unsupervised and supervised criteria (mean cross-validated knn error as a measure of miss-classification, and Sammon error as a measure of similarity structure loss). This methodology was applied to the analysis of high dimensional genomic data collected in the framework of Lung cancer research. A Pareto front approximation was recognizable from within the solutions provided by the final population. Selected solutions from along that approximation were used for the construction of a sequence of visualizations showing the progression from spaces with complete class separation and poor similarity preservation to spaces with reversed characteristics. A solution with a reasonable compromise between the two criteria was identified and clearly contained properties of both extreme solution spaces. These research results, although preliminary, showed large potential and further investigation is required.

4 Acknowledgments

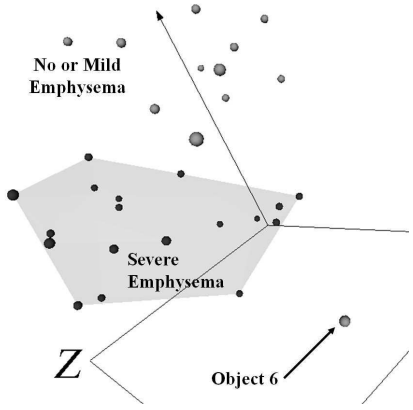
The authors would like to thank Robert Orchard from the Integrated Reasoning Group (National Research Council Canada, Institute for Information Technology) for his constructive criticism of the first draft of this paper.

References

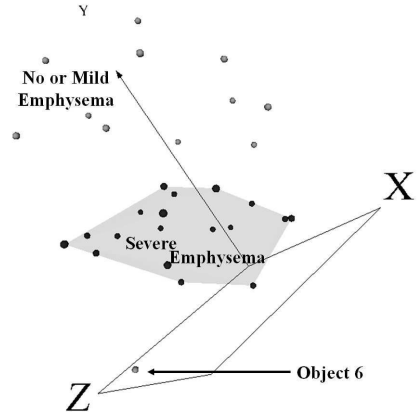
1. T. Bäck, D. B. Fogel, and Z. Michalewicz. *Handbook of Evolutionary Computation*. Institute of Physics Publishing and Oxford Univ. Press, New York, Oxford, 1997.
2. I. Borg and J. Lingoes. *Multidimensional similarity structure analysis*. Springer-Verlag, 1987.
3. E. K. Burke and G. Kendall. *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*. Number 0-387-23460-8. Springer Science and Business Media, Inc., 233 Spring Street, New York, NY 10013, USA, 2005.
4. J. L. Chandon and S. Pinson. *Analyse typologique. Théorie et applications*. Masson, Paris, 1981.
5. K. Deb, S. Agarwal, and T. Meyarivan. A fast and elitist multi-objective genetic algorithm: Nsga-ii. In *IEEE Transaction on Evolutionary Computation*, volume 6 (2), pages 181–197, 2002.



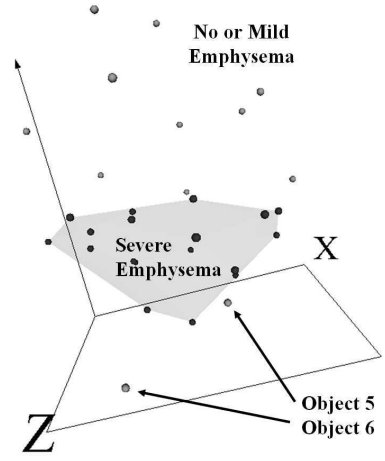
(a) Front obtained by multi-objective optimization (NSGA-II) that approximates the true Pareto Front.



(b) Chromosome 2 (5-fold CV k-nn Error: 0.0000, Sammon Error: 0.1082)



(c) Chromosome 10 (5-fold CV k-nn Error: 0.0667, Sammon Error: 0.0826)



(d) Chromosome 1 (5-fold CV k-nn Error: 0.2667, Sammon Error: 0.0797)

Fig. 1: Set of 100 multi-objective solutions. Those along the Pareto front approximation progressively span the extremes between minimum classification error and minimum dissimilarity loss. 3 solutions were selected and snapshots of VR spaces computed. Geometries: "light grey spheres" = no or mild emphysema samples, "dark grey spheres encased within a convex hull" = severe emphysema samples. Behavior = static.

6. K. Deb, S. Agarwal, A. Pratap, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In *Proceedings of the Parallel Problem Solving from Nature VI Conference*, pages 849–858, Paris, France, 16-20 September 2000.
7. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multi-objective genetic algorithm: Nsga-ii. Technical Report 2000001, Kanpur Genetic Algorithms Laboratory (KanGAL), Indian Institute of Technology Kanpur, 2000.
8. R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley New York, 1972.
9. U. Fayyad, G. Piatesky-Shapiro, and P. Smyth. From data mining to knowledge discovery. In U. F. et al., editor, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press, 1996.
10. K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1972.
11. J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 1(27):857–871, 1973.
12. A. K. Jain and J. Mao. Artificial neural networks for nonlinear projection of multivariate data. In *1992 IEEE joint Conf. on Neural Networks*, pages 335–340, Baltimore, MD, 1992.
13. M. Jianchang and A. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans. On Neural Networks*, 6(2):1–27, 1995.
14. J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psichometrika*, 29:1–27, 1964.
15. D. Levine. *Users Guide to the PGAPack Parallel Genetic Algorithm Library*. Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, January 1996.
16. J. Mao and A. K. Jain. Discriminant analysis neural networks. In *1993 IEEE International Conference on Neural Networks*, pages 300–305, San Francisco, California, March 1993.
17. J. Mao and A. K. Jain. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans. on Neural Networks*, 6:296–317, 1995.
18. T. Masters. *Advanced Algorithms for Neural Networks*. John Wiley & Sons, 1993.
19. V. Pareto. *Cours D'Economie Politique*, volume I and II. F. Rouge, Lausanne, 1896.
20. J. W. Sammon. A non-linear mapping for data structure analysis. *IEEE Trans. Computers*, C18:401–408, 1969.
21. Spira, A., Beane, J., Pinto-Plata, V., Kadar, A., Liu, G., Shah, V., Celli, B., Brody, J.S.: Gene Expression Profiling of Human Lung Tissue from Smokers with Severe Emphysema. American Journal of Respiratory Cell and Molecular Biology **31** (2004) 601–610
22. J. Valdés. Building virtual reality spaces for visual data mining with hybrid evolutionary-classical optimization: Application to microarray gene expression data. In *2004 IASTED International Joint Conference on Artificial Intelligence and Soft Computing, ASC'2004*, pages 161–166, Marbella, Spain, September 2004. IASTED, ACTA Press, Anaheim, USA.
23. J. J. Valdés. Similarity-based heterogeneous neurons in the context of general. *Neural Network World*, 12(5):499–508, 2002.
24. J. J. Valdés. Virtual reality representation of relational systems and decision rules:. In P. Hajek, editor, *Theory and Application of Relational Structures as Knowledge Instruments*, Prague, Nov 2002. Meeting of the COST Action 274.
25. J. J. Valdés. Virtual reality representation of information systems and decision rules:. In *Lecture Notes in Artificial Intelligence*, volume 2639 of *LNAI*, pages 615–618. Springer-Verlag, 2003.
26. P. Walker, B. Smith, Y. Qing, F. Famili, J. J. Valdés, L. Ziying, and L. Boleslaw. Data mining of gene expression changes in alzheimer brain. *Artificial Intelligence in Medicine*, 31:137–154, 2004.
27. A. R. Webb and D. Lowe. The optimized internal representation of a multilayer classifier. *Neural Networks*, 3:367–375, 1990.