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

Barton, Alan; Valdés, Julio

Publisher's version / Version de l'éditeur:

The 17th International Symposium on Methodologies for Intelligent Systems (ISMIS 2008) [Proceedings], 2008

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 Gastric and Liver Cancer Databases: An Evolutionary Computation Approach **

Barton, A., Valdés, J.
May 2008

* published at The 17th International Symposium on Methodologies for Intelligent Systems (ISMIS 2008). Toronto, Canada. May 20-23, 2008. Lecture Notes in Artificial Intelligence (LNAI 4994). Aijun An, Stan Matwin, Zbigniew W. Raś, Dominik Ślęzak (Eds.). pp. 256-261. NRC 49896.

Canada

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

Alan J. Barton and Julio J. Valdés

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

Abstract. This paper expands 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 two examples from cancer genomics data (e.g. gastric and liver 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

The World Health Organization (WHO) states that cancer is one of the leading causes of death in the world (http://www.who.int/cancer/en/) and that there are more than 100 types of cancers in which any part of the body may be affected. In particular, among men, the 5 most common types of cancer that kill are (in order of frequency): lung, stomach, liver, colorectal and oesophagus. As such, a previous study investigated lung cancer [14] and this new study investigates stomach and liver cancers. The presented approach provides 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. A VR technique for visual data mining on heterogeneous, imprecise and incomplete information systems was introduced in [12, 13] (see also http://www.hybridstrategies.com).

2 The multi-objective approach: A hybrid perspective

An enhancement to the traditional evolutionary algorithm [1], is to allow an individual to have more than one measure of fitness within a population (e.g. a weighted sum [2]). Multi-objective optimization, however, relies on the concept of a Pareto Front [10] of best current solutions, rather than a single best solution. One particular algorithm for multi-objective optimization is the elitist non-dominated sorting genetic algorithm (NSGA-II) [5], [4], [3], [2]. 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. Clearly, more requirements can be imposed on the solution by adding the corresponding objective functions. 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:

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

Examples of error measures frequently used for structure preservation are:

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

$$\text{Sammon error [11]} = \frac{1}{\sum_{i<j}\delta_{ij}}\frac{\sum_{i<j}(\delta_{ij} - \zeta_{ij})^2}{\delta_{ij}} \tag{3}$$

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

For heterogeneous data involving mixtures of nominal and ratio variables, the Gower similarity measure [6] has proven to be suitable. This measure can be easily extended for ordinal, interval, and other kind of variables.

2.1 Public Data

Each sample in this study is a vector in a high dimensional space. Direct inspection of the data structure and of the relationships between the descriptor variables (the genes) and the type of sample (normal/gastric cancer or control/liver tumor), is impossible. Moreover, within the collection of genes there is a mixture of potentially relevant genes with others which are irrelevant, noisy, etc.

Gastric Cancer: Gene expressions were compared in [7] to gain molecular understanding of gastric cancer. The public data contains 30 patient samples with 2 classes (8 samples of noncancerous gastric tissues and, 22 samples of primary human advanced gastric cancer tissues), with $7,129$ attributes (of which 34 values were missing) and was obtained from `http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=1210`.

Liver Cancer: Gene expressions were compared in [8] to gain molecular understanding of similarities between livers from zebrafish (Danio rerio) and 4 human tumor types (liver, gastric, prostate and lung). The public data contains 20 zebrafish samples with 2 classes (10 control samples and, 10 samples of zebrafish liver cancer), with $16,512$ attributes and was obtained from `http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=2220`.

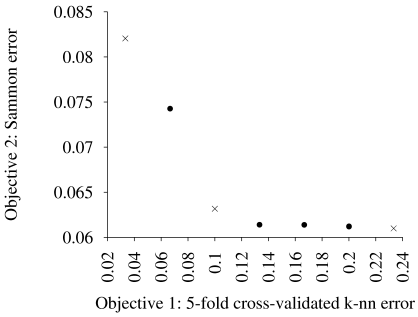
2.2 Results

Two sets of 100 non-dominated solutions were found (Fig-1(a) and Fig-2(a)) using the experimental settings in Table-1 for which the true location of the Pareto front is unknown. From these, three solutions were selected to investigate the *i)* best supervised solutions (Fig-1(b) and Fig-2(b)), resolving the respective classes at the cost of possible space distortions, *ii)* best unsupervised solutions (Fig-1(d) and Fig-2(d)), and *iii)* compromised solutions (Fig-1(c) and Fig-2(c)), of both class separation and internal data structure preservation.

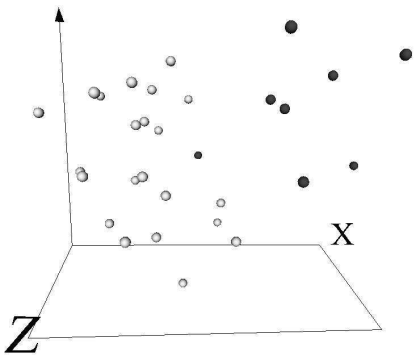
Table 1: Experimental settings for computing the pareto-optimal solution approximations by the multi-objective genetic algorithm (PGAPack [9] extended by embedding NSGA-II).

NumObjects	30 for Gastric Data	20 for Liver	
population size	100	number of generations	500
chromosome length	$= 3\cdot$ NumObjects	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

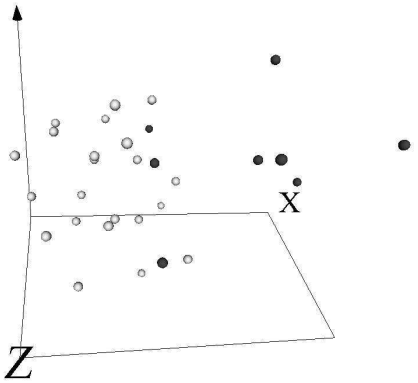
In general, different mappings lead to similar 3D visual representations; indicating good solution reproducibility even under the condition of potentially large amounts of attribute noise, redundancy, and irrelevancy within the sets of $7,129$ and $16,512$ original attributes. The major differences lie with local discrepancies with respect to the placement of some objects, which would need to be investigated further. For example, the object near the origin of Fig-2(b-d) is located differently in the three spaces.



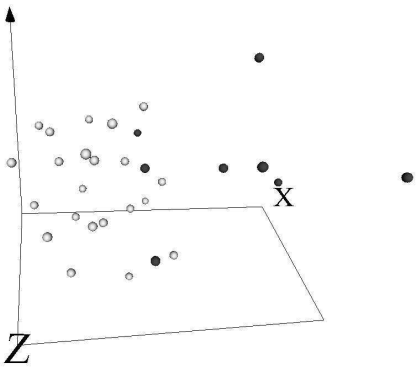
(a) Front obtained by multi-objective optimization (NSGA-II) that approximates the true Pareto Front for Gastric data.



(b) Computed minimum solution for objective 1 is chromosome 0 (5-fold CV k-nn Error: 0.0333333, Sammon Error: 0.0820461)

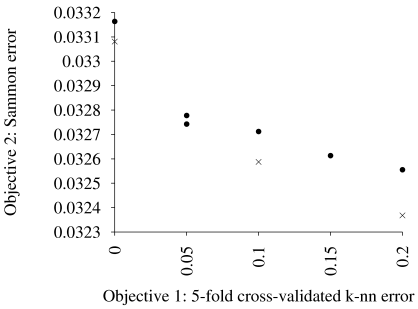


(c) Computed tradeoff solution for both objectives is chromosome 2 (5-fold CV k-nn Error: 0.100000, Sammon Error: 0.0631712)

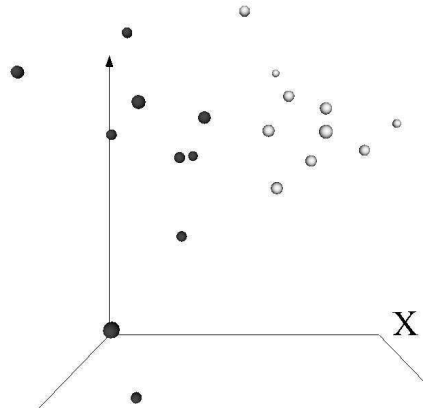


(d) Computed minimum solution for objective 2 is chromosome 1 (5-fold CV k-nn Error: 0.233333, Sammon Error: 0.0609954)

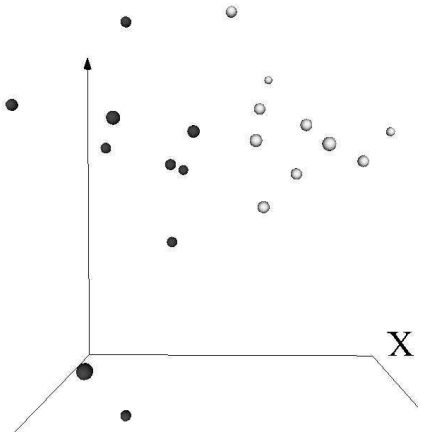
Fig. 1: Set of computed 100 multi-objective solutions for gastric cancer dataset. Those along the Pareto front approximation progressively span the extremes between minimum classification error and minimum dissimilarity loss. 3 solutions were selected and snapshots of computed VR spaces taken. Geometries: "light grey spheres" = normal samples, "dark grey spheres" = cancer samples. Behavior = static. The axis in the 3D views are highly non-linear maps from the original space ($7,129$ dimensions) to the respective 3D spaces.



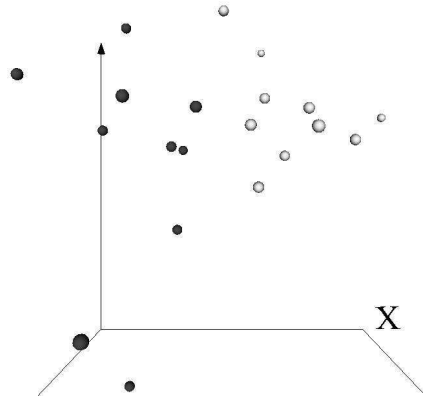
(a) Front obtained by multi-objective optimization (NSGA-II) that approximates the true Pareto Front for Liver data.



(b) Computed minimum solution for objective 1 is chromosome 9 (5-fold CV k-nn Error: 0.00000, Sammon Error: 0.0330806)



(c) Computed tradeoff solution for both objectives is chromosome 8 (5-fold CV k-nn Error: 0.100000, Sammon Error: 0.0325873)



(d) Computed minimum solution for objective 2 is chromosome 0 (5-fold CV k-nn Error: 0.200000, Sammon Error: 0.0323679)

Fig. 2: Set of computed 100 multi-objective solutions for liver cancer dataset. 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 computed VR spaces taken. Geometries: "light grey spheres" = control samples, "dark grey spheres" = liver tumor samples. Behavior = static. The axis in the 3D views are highly non-linear maps from the original space $(16,512$ dimensions) to the respective 3D spaces.

3 Conclusions

Analysis of high dimensional genomic data collected in the framework of Gastric and Liver cancer research was performed within the context of visual data mining and knowledge discovery research. Sequences of visualizations showing progression from spaces with minimum class separation and poor similarity preservation to spaces with reversed characteristics were reported. Solutions with reasonable compromises between the two criteria were identified. These preliminary research results expand the set of previously investigated real world cancer data sets. They also show the large potential for such an approach. Further investigations are required.

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. 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.
3. 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.
4. 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.
5. 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.
6. J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 1(27):857–871, 1973.
7. Y. Hippo, H. Taniguchi, S. Tsutsumi, N. Machida, J. Chong, M. Fukayama, T. Kidama, H. Aburatani. Global Gene Expression Analysis of Gastric Cancer by Oligonucleotide Microarrays. Cancer Research, **62** (2002) 233–240. PMID: 11782383.
8. S. H. Lam, Y. L. Wu, V. B. Vega, L. D. Miller, J. Spitsbergen, Y. Tong, H. Zhan, K. R. Govindarajan, S. Lee, S. Mathavan, K. R. Krishna Murthy, D. R. Buhler, E. T. Liu, Z. Gong. Conservation of gene expression signatures between zebrafish and human liver tumors and tumor progression. Nature Biotechnology, **24** (2006) 73–75. PMID: 16327811
9. D. Levine. *Users Guide to the PGAPack Parallel Genetic Algorithm Library*. Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, January 1996.
10. V. Pareto. *Cours D'Economie Politique*, volume I and II. F. Rouge, Lausanne, 1896.
11. J. W. Sammon. A non-linear mapping for data structure analysis. *IEEE Trans. Computers*, C18:401–408, 1969.
12. 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.
13. J. J. Valdés. VR representation of information systems and decision rules. In *Lecture Notes in Artificial Intelligence*, volume 2639 of *LNAI*, pages 615–618. Springer-Verlag, 2003.
14. J. J. Valdés. and A. J. Barton. Hybrid Unsupervised/Supervised Virtual Reality Spaces for Visualizing Cancer Databases: An Evolutionary Computation Approach. Proceedings: IWANN 2007. Lecture Notes in Computer Science. Vol. 4507. 2007. NRC 49295.