Visualizing Cancer Databases Using Hybrid Spaces

Julio J. Valdés
National Research Council Canada, Canada

Alan J. Barton
National Research Council Canada, Canada

INTRODUCTION

According to the World Health Organization(WHO), the directing and coordinating authority for health within the United Nations system *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.

BACKGROUND

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 (Valdés, J.J., 2002) (Valdés, J.J., 2003) (see also *http://www.hybridstrategies.com*).

The purpose of this article is to explore the construction of high quality VR spaces for visual data mining (in opposition to classical data mining (Fayyad, U., Piatesky-Shapiro, G., & Smyth, P., 1996)) 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 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.

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 (Sammon, J.W., 1969) 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:

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

Structure Preservation: An Unsupervised Perspective

Examples of error measures frequently used for structure preservation (Kruskal, J., 1964) (Sammon, J.W., 1969) (Borg, I., & Lingoes, J., 1987) 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 (Gower, J.C., 1973) 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.

Multi-Objective Optimization Using Genetic Algorithms

An enhancement to the traditional evolutionary algorithm (Bäck T., Fogel, D.B., & Michalewicz, Z, 1997), 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 (Burke, E.K., & Kendall, G., 2005). 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 (Pareto, V., 1896) of best current individuals, rather than a single best individual. Most (Burke, E.K., & Kendall, G., 2005) multi-objective algorithms use the concept of dominance to address this issue.

A solution $x_{(1)}$ is said to dominate (Burke, E.K., & Kendall, G., 2005) a solution $x_{(2)}$ for a set of m objective functions $<f_1(x), f_2(x), ..., f_m(x)>$ if

- $x_{(1)}$ is not worse than $x_{(2)}$ over all objectives. For example, $f_3(x_{(1)}) \leq f_3(x_{(2)})$ if $f_3(x)$ is a minimization objective.