MedVir: 3D Visual Interface Applied To Gene Profile Analysis

Gracia Berná, Antonio

DOCTORAL DISSERTATION COLLOQUIUM

EXTENDED ABSTRACT

Abstract— ***Background*: The use of data mining techniques for the gene profile discovery of diseases, such as cancer, is becoming usual in many researches. These techniques do not usually analyze the relationships between genes in depth, depending on the different variety of manifestations of the disease (related to patients). This kind of analysis takes a considerable amount of time and is not always the focus of the research. However, it is crucial in order to generate personalized treatments to fight the disease. Thus, this research focuses on finding a mechanism for gene profile analysis to be used by the medical and biologist experts. *Results*: In this research, the MedVir framework is proposed. It is an intuitive mechanism based on the visualization of medical data such as gene profiles, patients, clinical data, etc. MedVir, which is based on an Evolutionary Optimization technique, is a Dimensionality Reduction (DR) approach that presents the data in a three dimensional space. Furthermore, thanks to Virtual Reality technology, MedVir allows the expert to interact with the data in order to tailor it to the experience and knowledge of the expert.**

Keywords-component; dimensionality reduction, manifold learning, DNA microarray, visualization.

I. INTRODUCTION

Visualization techniques are being used by the scientific community to understand and obtain different conclusions about a particular dataset in an easy way. Nevertheless, these techniques are not frequently used to analyze very huge data volumes in life sciences field, particularly in genomics or biological domains, due to the high complexity of the data ('curse of dimensionality' [1]). There exist approaches that highlight the most important features of the data, and they make possible the construction of virtual reality spaces to visually understand the intrinsic nature of the data [2]. The benefits of representing n-dimensional data in tridimensional spaces are very well-known. Nowadays, these kinds of representations are carried out by means of dimensionality reduction and transformation of the data, and making use of a strong component of interaction methods.

At the same time, the use of data mining techniques for the gene profile discovery of diseases, such as cancer, is becoming usual in many researches. These techniques do not usually analyze the relationships between genes in depth, depending on the different variety of manifestations of the disease (related to patients). This kind of analysis takes a considerable amount of time and is not always the focus of the research. However, it is crucial in order to generate personalized treatments to fight the disease. Thus, this research focuses on finding a mechanism for gene profile analysis to be used by the medical and biologist experts.

In this research, the MedVir framework is proposed. It is an intuitive mechanism based on the visualization of medical data such as gene profiles, patients, clinical data, etc. MedVir, which is based on an Evolutionary Optimization technique, is a Dimensionality Reduction (DR) approach that presents the data in a three dimensional space. Furthermore, thanks to Virtual Reality technology, MedVir allows the expert to interact with the data in order to tailor it to the experience and knowledge of the expert.

II. MOTIVATIONS/RESEARCH

The origins for the proposed work arise in response to the increasing need for biologists to obtain tools for visual analysis of data obtained in a laboratory. When dealing with multidimensional data, such as biological data, the traditional data mining techniques can be a tedious, complex and limited task, even to some biologist experts. Therefore, it is necessary to develop useful visualization techniques that can complement the criterion of an expert and, at the same time, visually stimulate and make easier the process of obtaining knowledge from a dataset. Thus, the final process of interpretation and understanding of the data can be greatly enriched.

III. RELATED WORK

By analyzing the Dimensionality Reduction (DR) word in the literature, two general approaches for carrying out a DR process are presented [3]:

Feature Extraction: Transforming the existing features into a lower dimensional space.
Feature Selection: Selecting a subset of the existing features without a transformation [4, 11, 12].

MedVir algorithm is based on the first one, so this section focuses on Feature Extraction (FE). There are currently two canonical ways of dealing with the data when carrying out a DR process. The first one does so in a linear (Linear Dimensionality Reduction or LDR), whiles the second one is in a nonlinear way (Non Linear Dimensionality Reduction or NLDR). LDR handles datasets containing linear dependencies. However, they are not powerful enough to deal with complex datasets. The behavior of many datasets, such as a DNA Microarray, could not be explained by means of LDR because maybe it contains essential multiple nonlinear relationships between attributes that cannot simply be interpreted by using linear models [5]. This suggests the design of other techniques (NLDR methods) in order to highlight the true underlying structure of the data. These methods assume that data are generated according to a nonlinear model.

Two main different taxonomies or classification models in DR-FE techniques have been proposed. Laurens van der Maaten et al. [6] carried out a thorough comparative review of the most important linear DR technique (PCA), and twelve front-ranked nonlinear DR techniques (Isomap, Kernel PCA/MVU, Diffusion maps, LLE, Laplacian Eigenmaps, Hessian LLE/LTSA, Sammon Mapping, LLC/Manifold Charting, Autoencoder).

On the other hand, John A. Lee et al. proposed a different taxonomy of DR techniques [5] in accordance with procedures that reduce the features or dimensionality of the data by preserving the global shape of the geometry, or by preserving the local properties and neighborhood information of the data [8]. That is distance and topology preservation, respectively. *Distance preservation algorithms*: MDS, Sammon Mapping, Curvilinear Component Analysis, Isomap, Geodesic NLM, Curvilinear distance analysis, Kernel PCA, Semidefinite embedding. *Topology preservation* algorithms: SOM's, Generative Topographic Mapping, LLE, Laplacian Eigenmaps, Isotop.

IV. APPROACH

The MedVir framework is based on a previously developed framework [9], and its aim is to provide hidden knowledge in the form of underlying patterns, relationships or trends in multidimensional biological data. In other words, the objective of MedVir is to make a tool available to the expert to visualize a biological dataset, apart from interacting with them by means of virtual reality techniques.

The proposed framework is divided into two main modules. Module I carries out an optimization process (OP). Here, a search algorithm attempts to find a tridimensional embedding of the original n-dimensional data which best preserves its intrinsic geometry. After the OP, it is time to represent the previous embedding visually into a tridimensional environment. Thus, the Module II deals with the data visualization. At the end of the pipeline, the expert would be able to obtain an interpretation of the model and formulate his own conclusions, through his knowledge and previous experience in the dominion of the data. This process can be strongly reinforced by means of the interaction between the expert and the elements of the representation.

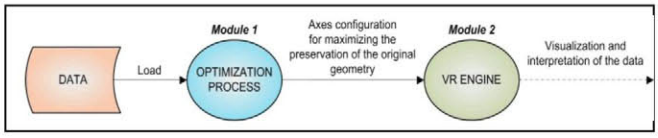

Figure 1. Pipeline of the framework (First, data are loaded into Module 1. After running the optimization process, a set of axes is obtained. Next, data are visually represented by means of the Module 2(VR engine). Finally, the expert interprets the visualization).

The OP is carried out by an evolutionary algorithm, particularly the Differential Evolution (DE) algorithm [10], and the VR engine is done by means of the game engine Unity 3D [7].

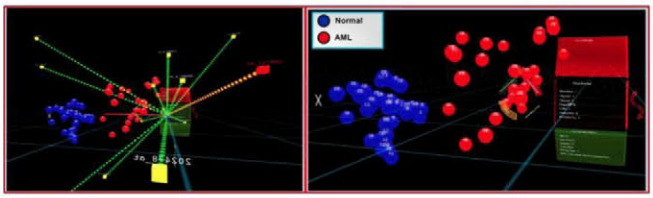

Figure 2. Visualization of the Stirewalt's dataset (MedVir applied to a DNA microarray dataset, particularly a leukemia dataset. The blue spheres represent normal tissue samples. Nevertheless, red spheres represent AML samples. The dotted green lines represent coordinate axes (genes). A separation of classes is

clearly visible. The interface shows the different clinical information of the sample, if it exists)

V. RESEARCH STATUS

The work schedule, here presented, has the following structure:

First, the application of MedVir framework to supervised, non-supervised or semi-supervised classification studies. Two different procedures are considered when analyzing datasets:

The first one is when the attributes or features of a particular dataset have been previously selected and validated by different authors in the literature. Here, the input set of attributes is supposed to be the correct one, and it is expected to obtain knowledge about this dataset in any of the three methods previously mentioned (supervised, non-supervised or semi-supervised).

The second one occurs when dealing with a dataset which has not a previous feature subset selection. Thus, a supervised and non-supervised feature subset selection package will be implemented. This package will provide MedVir the dataset with the features that have demonstrated to be the most discriminatory or important for this dataset.

Another possible line of study is based on bioinformatics research field, i.e, all related to the collection of clinical data, biomarkers, and to try to integrate it with what is already done. The important part is the study of techniques for validating our methods. Finally, it would be advisable to study the integration of virtual reality techniques, such as Kinect device, and thus interact with data through a more intuitive interface.

VI. EXPECTED CONTRIBUTIONS

First, a biological dataset is considered. Here, the instances represent the different samples and the attributes are intrinsic features that are considered for each of the instances and define them uniquely. The initial hypothesis of this work is based on the assertion that it is intended to obtain an understanding in a visual way of a set of biological data.

The integration of data mining techniques with visualization techniques provides richness to KDD process and underlying implicit in a biological database. In addition, the application of visualization techniques to the process of knowledge discovery brings a great agility to the process of KDD (Knowledge Discovery in Databases).

Moreover, the initial hypothesis leads to the assertion that MedVir could be applied to supervised, semi-supervised and unsupervised classification.

Regarding the expected outputs, it is supposed to be able to achieve the formulation of a thesis which allows concluding a successful process of obtaining knowledge fully and uniquely from a dataset. This process will be carried out in a visual and interactive way, and it should also allow verifying or rejecting claims about the initial data. For example, the application of MedVir to a case of supervised classification for a particular subset of attributes should allow concluding whether these selected attributes are the best ones that discriminate between classes in the dataset.

VII. CONCLUSIONS

MedVir can visualize multidimensional data in 3D, so this allows conclusions to be obtained in a more simple and intuitive way. MedVir has been designed to provide a framework that makes easier the interaction of the expert with the data representation, for example, by asking for additional information about a patient or modifying the importance of a gene expression from one specific study. In addition, because the data used in this research are about the classification of patients based on their gene profiles, MedVir allows these gene profiles to be analyzed using the 3D representation of data colored according to their class.

Thus, promising results have been obtained so far, which could allow in the near future achieving the considered expectations. For example, the results have been published in a conference and two research articles are being formalized to send to international journals with impact factor about the application of this contribution in the biological world.

REFERENCES

[1] R. Bellman. *Adaptative Control Processes: A guided Tour*. Princeton University Press, Princeton, NJ, 1961.

[2] J. Valdés, Alan J. Barton, R. Orchard: Genetic Programming for Exploring Medical Data using Visual Spaces. Genetic and Evolutionary Computation: Medical Applications 2011.

[3] D. Addison, Intelligent Computing Techniques: A Review. Telos Pr 2004.

[4] T. Jirapech-Umpai, S. Aitken, "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes". BMC Bioinformatics 2005, 6:148.

[5] J.A. Lee, M. Verleysen: Nonlinear dimensionality reduction. New York; London: Springer 2007.

[6] L.J.P Van der Maaten, E.O. Postma, H.J. van den Herik, "Dimensionality Reduction: A Comparative Review", 2007.

[7] Unity Technologies: Unity3D 2009, [http://www.unity3d.com/].

[8] V. De Silva, J.B. Tenenbaum, "Global Versus Local Methods in Nonlinear Dimensionality Reduction". In Advances in Neural Information Processing Systems 15 2003:705{712.

[9] A. Gracia, S. González , J. Veiga, V. Robles, "VR BioViewer - A new interactive-visual model to represent medical information". In MSV '11: Proceedings of the 2011 International Conference on Modeling, Simulation and Visualization Methods., Las Vegas, NV, USA 2011:40{46.

[10] R. Storn, K. Price, "Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces". Journal of Global Optimization 1997, (4):341-359.

[11] Y. Su, T.M Murali, V. Pavlovic, M. Shcaffer, S. Kasif, "RankGene: identification of diagnostic genes based on expression data". Bioinformatics 2003,(12):1578-1579.

[12] T. Li, C. Zhang, M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression". Bioinformatics (Oxford,England) 2004, (15):2429-2437.