Advanced Science and Technology Letters
Vol.29 (CA 2013), pp.234-236
http://dx.doi.org/10.14257/astl.2013.29.49

Development of MATLAB based Data Visualization Tool for Early Cancer Detection

Ki-Seok Cheong[1,3],Yu-Seop Kim[2,3*], Chan-Young Park[2,3],Hye-Jung Song[2,3],Jong-Dae Kim[2,3]

[1]Dept. of Computer Enginerring, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, 200-702 Korea
[2]Dept. of Ubiquitous Computing, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, 200-702 Korea
[3]Bio-IT Research Center, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, 200-702 Korea
{vseominjungv, yskim01,cypark, hjsong, kimjd}@hallym.ac.kr

Abstract. In this paper, we developed an integrated system for bio-data analysis and a visualization tool by applying data mining technology. This system was written in MATLAB. This platform provides a high performance and facilitated flow of information among the appropriate analyses. This system included marker selection, data visualization and marker evaluation, which have been developed on the basis of the MATLAB. This system is tailor-made to the early diagnosis of overran cancer.

Keywords: Biomarker, Ovarian Cancer,Logistic Regression, MATLAB

1 Introduction

Medical diagnosis and biological data analysis requires a data mining tool appropriate for statistical evaluations and data visualization. The R project, SPSS, and GraphPad PRISM are commonly used, general data mining and data analysis system[1-3]. However, the statistical analysis systems require clinicians to have an extensive knowledge of statistics as well as the ability to use these tools. Data-driven reasoning processes ensure diagnostic accuracy by applying appropriate algorithms, and should be applicable to a variety of cases. Clinicians have their data analysis performed by statistical professionals. This can ensure diagnostic accuracy, but variously empirical knowledge cannot be taken into a account in the analysis. Thus, in order for clinicians to use their own empirical knowledge while conducting data analysis, it is necessary to provide them with technology for the visualization and analysis of the data, leading to the generation of useful knowledge[4].In this paper presents a statistical, data mining analysis system for identifying multi-marker data required for disease diagnosis in the field of diagnostic medical testing. The system includes visualization techniques that measure biomarker data needed to diagnose

*Corresponding author

ISSN: 2287-1233 ASTL

disease with the use of the Luminex equipment, converts the data into an easy-to-analyze format, and includes data mining techniques that find patterns inherent to the data and extract useful knowledge. The main modules include marker selection, data visualization and marker evaluation modules, include marker selection, data visualization and marker evaluation modules, developed using MATLAB that can enable uni-or multi-dimensional graphing and include various bioinformatics and statistical packages.

2 Data Visualization-

The concentrations of multiple analytic from the Luminex are in the form of numerical data that makes it difficult for analysts to assess the concentration. Therefore the values should be visualized on graphs so that the can be analyzed as easily and conveniently as possible. The proposed multi-marker analysis system provides various graphs to facilitate examination of combined multi-marker data.

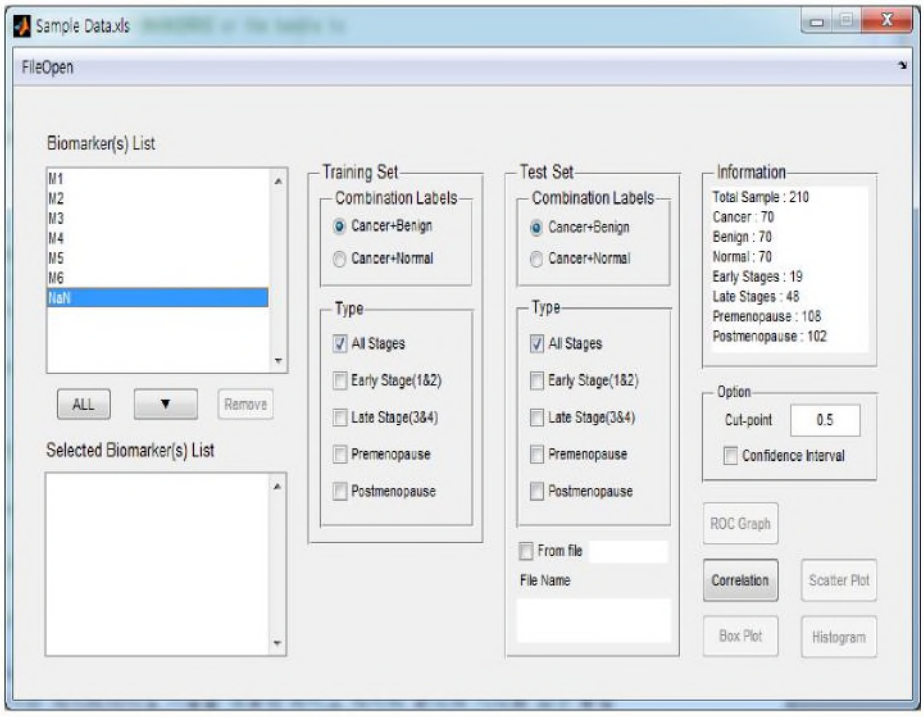

Fig. 1. Main user interface

Figure 1 shows the main screen that reads the sample datasets. The "Training Set" and the "Test Set" consist of all options for classes targeted by the respective classification models and the information related to other diseases. Cancer-benign and

cancer-normal groups are classification options in this system. In addition, the information related to the disease includes stages of cancer progression and menopausal status. Unlike the Training Set window, the Test Set window can load the additional data that can be used as the Test Set separately.

Once you use the Option window, you can set the cut-off point and choose whether you want to include the confidence interval (CI) into the performance result in the ROC graph. It is time consuming to calculate the CI, which makes it an option that is chosen only when absolutely necessary. All five buttons below the Option are for data visualization. In this system, the ROC Graph, Correlation, Scatter Plot, Box Plot and Histogram are provided as visualization options

3 Conclusion and discussion

This paper describes an analysis system with an intuitive user interface for interpreting multi-marker data. We developed an integrated system for bio-data analysis and visualization tool by applying data mining technology. . In this system, data extracted by multiple biomarkers tailored to the diagnosis of specific diseases are measured using Luminex equipment and converted into easily analyzable formats to be used for visualization, which facilitates the identification of data patterns, thus enabling simple automatic or visual discernment between normal and diseased conditions. Clinicians will be able to make more accurate judgments in the diagnosis of disease by using various simple tools utilizing the visualization, marker selection, and performance confirmation functions embedded in the system.

References

1. R Project, http://www.r-project.org
2. GraphPad Prism, http://graphpad.com
3. SPSS, httl://www.spss.com
4.Hendriks, Bart S., and Christopher W. Espelin. DataPflex: a MATLAB-based tool for the manipulation and visualization of multidimensional datasets. Bioinformatics**(2010)** 26.3 432-433.

