Málaga, Octubre de 2012

Informe Ejecutivo

TÍTULO: BIO-3.1-2012: Metaheurísticas para la selección de genes en pacientes con cáncer de mama.

RESUMEN: El problema de la selección de genes relevantes en la clasificación de enfermedades como la leucemia, el cancer de cólon o de pulmón, ya ha sido previamente estudiado en el Proyecto DIRICOM. Para afrontar estos problemas se usó, con éxito, un Algoritmo basado en Cúmulos de Partículas Geométrico (GPSO). En este informe técnico presentamos un nuevo problema: la clasificación de pacientes con cáncer de mama. Además, los buenos resultados obtenidos con GPSO nos llevó a incluirlo como nuevo paquete para la selección de atributos en la biblioteca WEKA. Por eso, mostraremos también el resultado de incluir GPSO en WEKA.

OBJETIVOS:

1. Describir un nuevo problema de clasificación: pacientes con cáncer de mama.
2. Presentar el algoritmo GPSO implementado en la biblioteca WEKA.

CONCLUSIONES:

1. GPSO es una técnica competitiva para la selección de genes en la clasificación de pacientes con cáncer de mama.
2. Publicar GPSO como paquete WEKA permite gran difusión del trabajo realizado en el Proyecto DIRICOM.

RELACIÓN CON ENTREGABLES:

PRE: BIO-2.0-2010 (lectura obligatoria)

PRE: PSO-1.0-2008 (lectura obligatoria)

Málaga, October, 2012

Executive Summary

TITLE: BIO-3.1-2012 Metaheuristics for Gene Selection in Breast Cancer Patient Classification

ABSTRACT: The problem of gene selection in the classification of diseases like leukemia, colon and lung cancer, has been already studied in the DIRICOM Project. The Geometric Particle Swarm Optimization algorithm (GPSO) was successfully used to face these problems. In this technical report we present a new problem: the classification of breast cancer patients. Moreover, the excellent results obtained with GPSO was a motivation for us to include this algorithm as a new package for feature selection in the WEKA library. Therefore, we also present the result of including GPSO in WEKA.

GOALS:

1. Describing a new classification problem: breast cancer patients.
2. Presenting the GPSO algorithm implemented in the WEKA library.

CONCLUSIONS:

1. The GPSO algorithm is a competitive technique for gene selection in the classification of breast cancer patients.
2. Publishing the GPSO algorithm as a new WEKA package allows the broadcasting of the work carried out in the DIRICOM Project.

RELATION WITH DELIVERABLES:

PRE: BIO-2.0-2010 (mandatory reading)

PRE: PSO-1.0-2008 (mandatory reading)

Metaheuristics for Gene Selection in Breast Cancer Patient Classification

DIRICOM

October 2012

1. Introduction

The problem of feature (gene) selection and supervised classification of gene expression data obtained with DNA microarrays has already been presented in a technical report of the DIRICOM Project [3]. In that technical report the following classification problems were considered: ALL-AML leukemia, colon tumour, diffuse large B-cell lymphoma and lung cancer. Herein, however, we will consider a new one: breast cancer.

Feature selection is crucial to improve the classification of gene expression data. *Geometric Particle Swarm Optimization* GPSO algorithm [5] has been successfully employed in the DIRICOM Project [2] as a new feature selection technique that finds small subset of relevant genes related to different cancer diseases. Due to its success, we decided to include GPSO as a new software package in the WEKA library [4], in order to share its benefits with this popular data mining tool.

The rest of this document is structured as follows: Section 2 describes the new classification problem (breast cancer) addressed in the DIRICOM Project. Section 3 presents how to use the GPSO algorithm already included in the WEKA library and, finally, Section 4 discusses our conclusions.

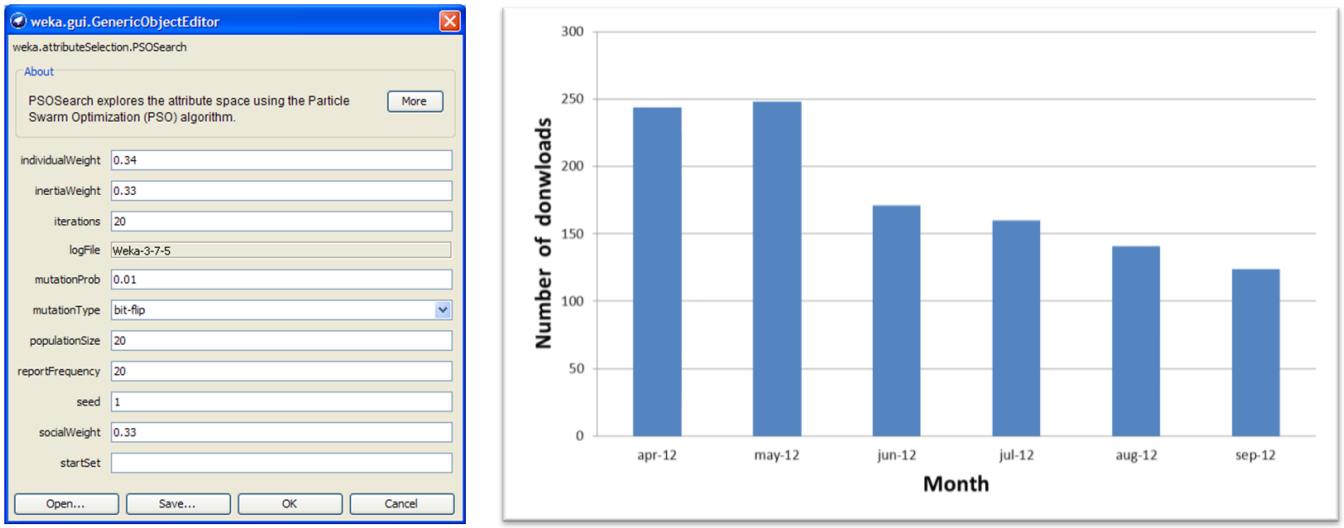
2. Breast Cancer Classification

Gene expression data obtained with DNA microarrays is used to characterize the behaviour of a cell inside the human body under concrete circumstances. Two publications in important scientific journals [1] [6] have previously analyzed the gene expression data of early-stage breast cancer women in order to predict distant metastasis recurrence in 5 years after surgery to remove their tumour. In the DIRICOM Project, we use the gene expression data published in these previous works in order to test the performance of GPSO in breast cancer classification.

The works of Wang et al. [6] and Desmedt et al. [1] extract the gene expression data from breast cancer patients after removing the tumor with surgery. Concretely, this works study the relationship between gene expression levels and breast cancer relapse (*distant metastasis*) within 5 years after surgery to remove the tomour. This way, they define two class of patients: the class 'relapse' for those patients who develop distant metastasis within 5 years after surgery; and the class 'no relapse' for those patients remaining metastasis-free after 5 years from surgery. This is the *class definition* that we also use in our work to build the classifiers. The ultimate goal is to develop a reliable prognostic test, based on gene expression data, to classify patients into two groups: the *good* prognosis group (class 'no relapse') in which the patients are not going to relapse, and the *poor* prognosis group (class 'relapse') where patients are going to present distant metastasis after surgery. Therefore, the patients classified into the good prognosis group will safely omit the adjuvant chemotherapy, improving their living quality since they will avoid the corresponding side effects of such treatment.

3. GPSO in WEKA

The GPSO algorithm [5] has been previously explained in a technical report for the DIRICOM Project [2]. In this Section we describe how this algorithm can be used after its implementation in the WEKA library.



a) GUI in WEKA

b) Download statistics

Figura 1: GPSO in WEKA.

Figure 1a shows the user interface for the GPSO algorithm in WEKA. This software package is freely available to download and install from the WEKA repository with name: **PSOSearch**. Once the WEKA Explorer is started, go to Select attributes → Search Method → Choose → PSOSearch and the window presented in Figure 1a is opened. In this window, each of the input parameters for the GPSO algorithm (explained in [2]) can be configured: number of particles in the swarm (populationSize), number of times to move the swarm (iterations), the *3PMBCX* crossover probabilities (individualWeight, inertiaWeight and socialWeight) along with the mutation probability and type (mutationProb and mutationType). Additionally, you can select the seed for this stochastic technique, one specific particle to be added in the initial swarm (startSet) and a file to log the evolution of the best fitness and the final result (logFile).

Figure 1b presents how many times per month has been installed the PSOSearch package from the official repository (ranging from April to September 2012). As can be seen in Figure 1b, the number of downloads per moth has been decreasing as the time goes by. However, this number is always above 100 downloads per month. This is a good opportunity to disseminate part of the work carried out inside the DIRICOM Project.

4. Conclusions

In this technical report we have presented a new problem for gene selection: the classification of gene expression data coming from breast cancer patients. The GPSO algorithm [2] was already used to face similar problems successfully in previous works of the DIRICOM Project [3]. That is why we also used the GPSO algorithm to perform gene selection in breast cancer classification. More importantly, we have included the GPSO algorithm as a new software package in the WEKA data mining tool in order to share its benefits with everyone. This software package is freely available to download from the WEKA official repository. As the download statistics suggest (Figure 1b in Section 3), this is a good approach to disseminate the DIRICOM Project results.

Referencias

[1] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. D'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. L. Harris, J. G. M. Klijn, J. A. Foekens, F. Cardoso, M. J. Piccart, M. Buyse, and C. Sotiriou. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical Cancer Research*, 13(11):3207–3214, June 2007.

[2] DIRICOM. Geometric Particle Swarm Optimization Implementation. Technical Report PSO-1.0-2008, University of Malaga, 2008.

[3] DIRICOM. Gene Selection and Classification in DNA Microarrays. Technical Report BIO-2.0-2010, University of Malaga, 2010.

[4] E. Frank, M. Hall, L. Trigg, G. Holmes, and I.H. Witten. Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–2481, 2004.

[5] A. Moraglio, C. Di Chio, and R. Poli. Geometric particle swarm optimisation. In *Proceedings of the 10th European conference on Genetic programming*, EuroGP'07, pages 125–136, Berlin, Heidelberg, 2007. Springer-Verlag.

[6] Y. Wang, J. G. M. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-Van Gelder, J. Yu, T. Jatkoe, E. M. J. J. Berns, D. Atkins, and J. A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–679, February 2005.