A Comparison of PSO and GA Approaches for Gene Selection and Classification of Microarray Data

José García-Nieto,
Enrique Alba
Dept. de Leng. y Ciencias de la Computación
University of Málaga ETSI Informática,
Málaga - 29071, Spain
{jnieto,eat}@lcc.uma.es

Laetitia Jourdan,
El-Ghazali Talbi
LIFL-INRIA Futurs
Bât M3, Cité Scientifique
59655 Villeneuve d'Ascq, France
{jourdan,talbi}@lifl.fr

Feature selection for gene expression analysis in cancer prediction often uses wrapper classification methods to discriminate a type of tumor, to reduce the number of genes to investigate in case of a new patient. By creating clusters a big reduction of the number of considered genes and an improvement of the classification accuracy can be finally achieved. The definition of the feature selection problem is this: given a set of features $F = \{f_1, ..., f_i, ..., f_n\}$, find a subset $F' \subseteq F$ that maximizes a scoring function $\Theta : \Gamma \rightarrow G$ such that

$$F' = argmax_{G \subset \Gamma}\{\Theta(G)\}, \quad (1)$$

where Γ is the space of all possible feature subsets of F and G a subset of Γ. The optimal feature selection problem has been shown to be NP-hard. Therefore, only heuristics approaches are able to deal with large size problems.

In this work, we are interested in gene selection and classification of DNA Microarray data in order to distinguish tumor samples from normal ones. For this purpose, we propose two hybrid models that use metaheuristics and classification techniques. The first one consists of a Particle Swarm Optimization (PSO) combined with a SVM approach as wrapper method. The second model is based on the popular GA using a specialized SSOCF [1] crossover operator, that will be also combined with SVM in our approach. A second important contribution consists in the actual discovery of new and challenging results on six public datasets identifying significant in the development of a variety of cancers (leukemia, breast, colon, ovarian, prostate, and lung from the URL `http://sdmc.lit.org.sg/GEDatasets/Datasets.html`).

For our PSO_{SVM} approach, a binary version of PSO was implemented in C++ following the *skeleton* architecture of the MALLBA library [2]. For the GA_{SVM} approach the GA was implemented in C++ using the ParadisEO Framework. Since the position of a particle (chromosome in GA) x represents a gene subset, the evaluation is carried out by means of the SVM classifier to assess the quality of the represented gene subset. The fitness of a particle/chromosome x is calculated applying a Leave One Out Cross Validation (LOOCV) method to calculate the rate of correct classification (accuracy) of a SVM trained with this gene subset. The complete fitness function is described in Equation 2.

$$fitness(x) = \alpha \cdot (100/accuracy) + \beta \cdot \#features, \quad (2)$$

GECCO '07, July 7-11, 2007, London, England, United Kingdom.
ACM 978-1-59593-697-4/07/0007..

where α and β are weight values set to 0.75 and 0.25 respectively. The objective here consists of maximizing the accuracy and minimizing the number of genes ($\#features$). For convenience (only minimization of fitness) the first factor is presented as $(100/accuracy)$.

An special initialization method was adapted to gene selection as follows. The swarm/population was divided into four subsets of particles/chromosomes initialized in different ways depending on the number of features in each particle. That is, 10% of particles were initialized with N (prefixed value) selected genes (1s) located randomly. Another 20% of particles were initialized with $2N$ genes, 30% with $3N$ genes and finally, the rest of particles (40%) were initialized randomly and 50% of the genes were turned on.

Table 1: Subsets reported with 100% test accuracy

Dataset	Algorithm	Genes	
Leukemia	PSO_{SVM}	100(3)	*K01383_at, U03056_at, J04130_s_at*
Breast	PSO_{SVM}	100(4)	*Contig49744_RC, Contig26884_RC Contig25936_RC, Contig13846_RC*
Colon	PSO_{SVM}	100(3)	*H64398, H73758 U27699*
Lung	GA_{SVM}	100(3)	*33762_r_at, 34648_at 728_at, 829_s_at*
Ovarian	GA_{SVM}	100(2)	*MZ1154.6306 MZ2653.8464*
Prostate	GA_{SVM}	100(3)	*35935_at, 39801_at 40069_at*

In conclusion, both approaches were experimentally assessed on six well-known cancer datasets discovering new and challenging results, and identifying specific genes that our work suggests as significant ones. In this sense, comparisons with several state of art methods show competitive results according to standard evaluation. Results of 100% classification rate and few genes per subset (2, 3 and 4) are obtained in most of our executions (see Table 1). The use of an adapted initialization method has shown a great influence on the performance of proposed algorithms, since it introduces an early set of acceptable solutions in their evolution process.

1. REFERENCES

[1] L. Jourdan, C. Dhaenens, and E.-G. Talbi. A genetic algorithm for feature selection in data-mining for genetics. In *Proceedings of the 4th Metaheuristics International ConferencePorto (MIC'2001)*, pages 29–34, Porto, Portugal, 2001.

[2] E. Alba and M. group. Mallba: A Library of Skeletons for Combinatorial Optimisation. In B. Monien and R. Feldmann, editors, *Proceedings of the Euro-Par*, volume LNCS 2400, pages 927–932, 2002.