# Simultaneous Gene Selection and Cancer Classification using a Hybrid Group Search Optimizer

Dattatraya Magatrao
Dept. of Computer Sc.
University of Pune, Pune, India
magatrao@gmail.com

Shameek Ghosh,
Jayaraman Valadi
C-DAC, Pune, India
Shiv Nadar University, Dadri, UP, India
shameekg@cdac.in, valadi@gmail.com

Patrick Siarry
LiSSi,
Université Paris-Est Créteil
Val-de-Marne, Créteil, France
siarry@u-pec.fr

## ABSTRACT
Constructing classifier models for gene expression datasets using informative features enhances prediction performance of concerned models. Here, we propose a hybrid Group Search based feature selection (GSO-FS) algorithm which can select relevant gene subsets that can optimally predict cancerous tissue samples. Our experimental results show that the GSO-FS algorithm in combination with SVM classifier performs quite well.

## Categories and Subject Descriptors
I.5.2 [**Pattern Recognition**]: Design Methodology – *Classifier design and evaluation, Feature evaluation and selection, Pattern analysis.*

## General Terms
Algorithms

## Keywords
Gene Selection, Cancer classification, Group Search Optimization, Information Gain.

## 1. GROUP SEARCH OPTIMIZATION
The challenging task of analyzing massive datasets to overcome dimensionality problems by selecting a relevant subset of genes is also known as Gene Selection or Feature Selection. In this work, we make use of the Group Search Optimization (GSO) [1] based hybrid filter-wrapper methodology to search for informative gene subsets and tend to improve the search by feeding an information gain gene ranking as a prior information heuristic, to the GSO based feature selection algorithm. GSO is a swarm intelligent mechanism proposed by S. He [1], based on animal search and foraging phenomenon. In the GSO based Producer-Scrounger (PS) model [1], a population may consist of producers, scroungers and rangers and together they form a group. Each individual of this group is known as a member. A producer is thus a member that uses its scanning mechanism to search for optimal solutions or resources in its nearby region. In contrast, a scrounger follows a policy of joining the producer in its search. In the artificial GSO optimization model, rangers are also introduced to perform a random walk for further improvements in the performance of the algorithm. Generally a group member may be any solution represented as an n-dimensional point in space and an associated head angle. Typically the producer has a search direction given by Cartesian co-ordinates transformations as in [1].

For simplification, a group normally consists of a single producer, which is selected as the one with the best fitness value. It then scans the environment to search for optimal resources (or points with better fitness values).The producer scanning field is thus considered within an n-dimensional space, with properties like maximum pursuit angle $\theta_{max} \epsilon R^1$ and maximum pursuit distance $l_{max} \epsilon R^1$. According to He [1], the maximum pursuit angle and distance are characteristic properties for a conical scanning field vision. Typically a producer may sample three points at zero degree, the right hand side hypercube and the left hand side hypercube. If the producer finds a better fitness value at any of the three points, then it moves to the same. Otherwise, it moves the head angle as given in [1]. After a number of iterations, if the producer does not find better resources, it turns its head back to zero degree [1].

At every iteration, a certain percentage of members selected as scroungers perform a random walk towards the producer based on the scrounger movement expression in [1]. Additionally, GSO also employs a specific number of rangers to perform a random walk which tends to avoid problems associated with local minima. This process of generating new sample points is repeated over till a termination criterion is reached.

## 2. GSO BASED FEATURE SELECTION (GSO-FS)
A solution or point in GSO may be represented as a set of feature indices. Thus one can create an initial population of random solutions, where each solution is represented by an n – dimensional point (n being the selected gene subset size). The n-dimensions are synonymous with n features of a subset. Thus depending on the input subset size, a random feature subset of cardinality n is represented as a point in an n-dimensional space. With the selected features of each such member, we apply the SVM classification function on the same to obtain a 10 fold cross validation classification accuracy, which is assigned as the fitness value. In our GSO-FS model, the individual having the highest 10 fold cross validation accuracy (CVA) is assigned the role of a producer in the group. Each member in the group, next performs its respective operations depending on whether it's a producer, scrounger or ranger as per equations as explained before in [1]. Owing to the GSO operations, it is possible that redundant indices may appear in the members of a group, thus creating a problem. To overcome this limitation, we introduce prior information while selecting genes for replacement. For this, we employed the information gain ranking for the probabilistic selection of genes. After generating a new group, the above process of GSO based gene selection, may be repeated for certain predefined number of iterations. Thus the member having the best resource in the final iteration is reported to have the most optimal gene subset.

## 3. DISCUSSION AND RESULTS
We obtained three datasets from the Kent Ridge Biomedical datasets repository and the libSVM repository (made available from various other original sources). Additionally, we also employed GSO-FS on the cervical cancer dataset extracted from


*GECCO'13 Companion*, July 6–10, 2013, Amsterdam, The Netherlands.
ACM 978-1-4503-1964-5/13/07.

[5]. The dimensions of the datasets are tabulated in Table 1. We performed extensive simulations by increasing the subset size incrementally for each dataset. In this context, 30 experiments were carried out and it was found (based on numerous simulations) that keeping maximum number of iterations to 80 was sufficient for convergence of GSO.

**Table 1**: *Dataset Specifications*

| *Cancer Dataset* | *No. of genes* | *No. of classes* | *No. of Samples* |
|---|---|---|---|
| ***Colon*** | 2000 | 2 | 62 |
| ***Breast*** | 7129 | 2 | 44 |
| ***Leukemia*** | 7129 | 2 | 72 |
| ***Cervical*** | 14023 | 3 | 38 |

Here, we report the mean 10 fold cross validation classification accuracies obtained using the final gene subset, after carrying out 30 simulations, for each dataset. For colon cancer, GSO-FS reported 98.9% (10 fold CVA) which has performed well against ACO-RF (95.47%), ACO-SVM (96.77%), BBO-SVM (98.39%) reported earlier [2, 6]. For the duke breast cancer data, with a 10 fold CVA of 99.1% GSO-FS has performed well in contrast to some of the more powerful models based on Bagging (92%), BBO-SVM (99.56%) and Ensemble (94%) techniques [3, 6]. GSO-FS with leukemia reported a 10 fold CVA of 99.2%, which was compared with a baseline SVM model (97.06%), BBO-SVM (99.60%) and ACO-AM (96%), reported earlier [4, 6]. For the cervical cancer dataset, we obtained a 10 fold CVA of 97% using a feature subset size of 12. Earlier literature on this dataset mostly takes p-values while assessing the goodness of feature subsets.

**Table 2**: *GSO-SVM Parameters*

| ***GSO Algorithm Parameters*** | ***Values*** |
|---|---|
| No. of members | 40 |
| No. of Generations | 80 |
| $a$ | $\sqrt{(feat_{max}+1)}$ |
| Maximum pursuit angle ($\theta_{max}$) | $\pi/(2a^2)$ |
| Maximum turning angle($\alpha_{max}$) | $\theta max/2$ |
| Maximum Pursuit Distance($l_{max}$) | $\sqrt{\sum(feat_{\max}-1)^2}$ |
| cost,gamma(for RBF as SVM kernel),Folds | 50,0.02,10 |

Additionally to assess the statistical significance of the feature subsets obtained we performed 30 permutation test simulations for each dataset. In each simulation, the number of generated permutations of the class labels was 3000. Based on a significance threshold of 0.05, the obtained p-values indicate that most of our final subsets (duke = 0.027, leukemia = 0.015, colon = 0.0054, cervical=0.0055) were statistically significant thus helping to reject the null hypothesis, which states that the classification model is not reliable enough to predict the labels from the extracted set of features. The alternative hypothesis states that a classifier with a high 10 fold CVA can be trained from the selected features. For permutation tests, the p-value was simply the relative frequency of executions that resulted in similar or better 10 fold CV performances than the CV performance of the model trained on the non-permuted ( i.e. actual ) labels. Extensive tuning of parameters had been carried out to obtain the best performances as described in Table 2.

The comparisons for the first three data sets were carried out using the popular ant colony based SVM approach, which has been known to report very good classification accuracies for the concerned datasets. Although the comparisons are not rigorous due to different simulation conditions, we can say in general that GSO-SVM has performed quite well. Additional tests were also carried on a Cervical Cancer dataset having 38 samples and 14023 features, where GSO-SVM reported a 97% 10 fold CVA for subset sizes of 12 (p-value=0.0055). The GSO based gene subset sizes selected were 15 for Colon, 15 for Breast, 15 for Leukemia and 12 for cervical cancer. GSO based feature selection may thus further be explored by introducing newer scanning/foraging mechanisms and analogous animal behaviour operations.

## 4. ACKNOWLEDGMENTS

Jayaraman Valadi gratefully acknowledges the Department of Science and Technology, New Delhi, India for financial support.

## 5. REFERENCES

[1] He, S., Q. H. Wu, and J. R. Saunders. 2009. *Group search optimizer: an optimization algorithm inspired by animal searching behavior*. Evolutionary Computation, IEEE Transactions on 13, no. 5, 973-990.

[2] S. Sharma et al. 2012. *Simultaneous informative gene extraction and cancer classification using aco-antminer and aco-random forests*. In Proceedings of the International Conference on Information Systems Design and Intelligent Applications, AISC Springer, vol. 132, pp. 755–761.

[3] M. Martn-Merino, A. Blanco and J. De Las Rivas. 2008. *Combining dissimilarity based classifiers for cancer prediction using gene expression profiles*. BMC Bioinformatics, vol. 8.

[4] G. Cong et al. 2005. *Mining top-k covering rule groups for gene expression data*. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, ser. SIGMOD '05. ACM, pp. 670–681.

[5] Zhai Y, Kuick R, Nan B, Ota I et al. *Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies HOXC10 as a key mediator of invasion*. Cancer Res 2007 Nov 1;67(21):10163-72.

[6] S Nikumbh, S Ghosh, and V. K. Jayaraman. 2012. *Biogeography-based informative gene selection and cancer classification using SVM and Random Forests*. In Evolutionary Computation (CEC), IEEE Congress on, pp. 1-6. IEEE, 2012.