# Scatter Search for high-dimensional feature selection using feature grouping

Miguel García-Torres
Universidad Pablo de Olavide
Seville, Spain
Universidad Americana
Asunción, Paraguay
mgarciat@upo.es

Francisco Gómez-Vela
Federico Divina
Universidad Pablo de Olavide
Seville, Spain

Diego P. Pinto-Roa
José Luis Vázquez Noguera
Julio C. Mello Román
Universidad Americana
Asunción, Paraguay

## ABSTRACT

In feature selection tasks, finding the optimal subset of features is unfeasible due to the increase of the search space with the dimensionality. In order to reduce the complexity of the space, feature grouping approach aims to generate subsets of correlated features. In this context, evolutionary algorithms have proven to achieve competitive solutions. In this work we propose a novel Scatter Search (SS) strategy that uses feature grouping to generate a population of diverse and high quality solutions. Solutions are evolved by applying random mechanisms in combination with the feature group structure to maintain the diversity and the quality of the solutions during the search. We test the proposed strategy on high dimensional data from biomedical domains and compare the performance against the first adaptation of the SS to the feature selection problem. Results show that our proposal is able to find smaller subsets of features while keeping a similar predictive power of the classifier models. Finally, a case of study regarding melanoma skin cancer is analysed using the proposed strategy.

## CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; • **Computing methodologies** → **Feature selection**; **Randomized search**;

## KEYWORDS

Feature selection, feature grouping, high dimensionality, metaheuristic

**ACM Reference Format:**
Miguel García-Torres, Francisco Gómez-Vela, Federico Divina, Diego P. Pinto-Roa, José Luis Vázquez Noguera, and Julio C. Mello Román. 2021. Scatter Search for high-dimensional feature selection using feature grouping. In *Proceedings of the Genetic and Evolutionary Computation Conference 2021 (GECCO '21 Companion).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3459481

## 1 BACKGROUND

The proposed SS, called Predominant Group based Scatter Search (PGSS), uses the Greedy Predominant Groups Generator [1], called GreedyPGG, to reduce the search space by generating subsets of correlated features and reduce the search space.

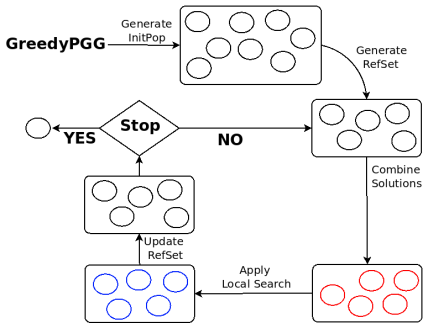


**Figure 1: Scatter Search workflow.**

The workflow of the SS is presented in Figure 1. First, it applies the GreedyPGG strategy. Then, the initial population (InitPop) is generated. For each new solution, the number of features is randomly set to a value ranging from 1 to $|\mathcal{G}|$, where $\mathcal{G}$ is the subset of feature groups. Features are also randomly selected from $\mathcal{G}$.

In the next step a subset of solutions, called Reference Set (RefSet), is selected from InitPop. Such subset is usually small and consist of the best and most diverse solutions found. The diversity between solutions is measured using the symmetric difference metric.

The algorithm continues by combining all pairwise solutions from RefSet. Given two solutions $S_1$ and $S_2$, it first creates the subset $C = S_1 \cup S_2$. Then, the GreedyPGG is applied to $C$. The solution $S_1^{'}$ is comprised by the predominant features found in $C$ while $S_2^{'}$ is generated by randomly selecting from all the features from $C$. The size of $S_2^{'}$ is fixed to the number of predominant groups found in $C$.

Each new solution is improved with the Sequential Forward Selection (SFS). The features of the solutions are used as the starting subset. Finally, the RefSet is updated with the new improved solutions. This process is repeated until a maximum number of iterations is reached.

SS and PGSS parameters were set following the recommendations found in [2]. $|InitPop| = 100$, $|RefSet| = 10$, $|RefSetSize1| = |RefSetSize2| = 5$ and $NumMaxIter = 10$. The quality of feature subsets $S$ was measured using *Correlation Feature Selection* [3].


*GECCO '21 Companion, July 10–14, 2021, Lille, France*

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3459481

## 2 RESULTS ON BIOLOGICAL DATA

The experiments were conducted using a 5-fold cross validation. For the predictive modes, the accuracy averaged over the folds is reported with its corresponding standard deviation. For feature selection algorithms, the average number of features with the standard deviation is presented. Finally, to support the conclusions, we applied the Wilcoxon signed-rank test. A summary of the datasets is presented in Table 1.

**Table 1: Characteristics of the high-dimensional datasets.**

| Dataset | Id | #Inst. | #Feat. | Labels | #Inst./label |
|---|---|---|---|---|---|
| colon | cln | 62 | 2000 | normal/tumor | 22/40 |
| lymphoma | lym | 77 | 2647 | diffuse/follicular | 58/19 |
| breast | bcg | 168 | 2905 | good/poor | 111/57 |
| lung | lng | 181 | 12533 | MPM/ADCA | 31/150 |
| breast | brc | 118 | 22215 | positive/negative | 75/43 |
| breast/colon | bco | 104 | 22283 | breast/colon | 62/42 |
| crohn | cro | 127 | 22283 | normal/colitis/crohn | 42/26/59 |

The performance of the classifier and the number of features selected are presented in Table 2. On average, feature selection improves the predictive power in most datasets. Comparing the results achieved by SS and PGSS, we can observe that results of SS and PGSS are similar and, therefore, the differences found are not statistically significant. In contrast, PGSS is the strategy that finds the smallest feature subsets in all cases. In this case, the differences found are statistically significant with a confidence level of 95% ($\alpha = 0.05$).

**Table 2: Performance of Naive Bayes and number of features selected by SS and PGSS.**

| | Naive Bayes | | | #Features | |
|---|---|---|---|---|---|
| Id | Baseline | SS | PGSS | SS | PGSS |
| cln | 54.87 ± 25.00 | 82.56 ± 16.50 | 84.10 ± 13.15 | 25.80 ± 5.90 | 13.20 ± 3.90 |
| lym | 81.83 ± 10.39 | 92.25 ± 6.80 | 90.83 ± 5.83 | 49.80 ± 3.56 | 32.20 ± 5.17 |
| bcg | 70.21 ± 7.61 | 74.92 ± 11.79 | 71.96 ± 8.00 | 75.00 ± 6.04 | 32.80 ± 2.17 |
| lng | 97.78 ± 3.62 | 96.67 ± 3.04 | 95.02 ± 2.34 | 19.00 ± 11.58 | 6.2 ± 2.40 |
| brc | 85.65 ± 8.32 | 82.25 ± 10.93 | 82.25 ± 6.85 | 104.60 ± 7.10 | 71.60 ± 11.01 |
| bco | 69.24 ± 7.91 | 94.19 ± 5.31 | 93.24 ± 4.34 | 18.20 ± 15.16 | 7.80 ± 2.17 |
| cro | 74.06 ± 8.06 | 87.45 ± 3.06 | 85.11 ± 6.30 | 137.40 ± 12.28 | 69.80 ± 6.30 |
| avg | 76.23 | 87.18 | 86.07 | 61.40 | 33.37 |
| pval | 0.156 | 0.142 | | 0.017 | |

## 3 CASE STUDY: MELANOMA DATASET

The features of this dataset are reported in Table 3. A complete description of the dataset can be found in [4].

**Table 3: Features of the melanoma dataset.**

| Criterion | Feature | Id | Description |
|---|---|---|---|
| asymmetry | index of asymmetry | $as$ | #pixels into irregular disjoint areas |
| borders | segment 1-8 | $b1-8$ | variation of colors from center pixel to border pixels |
| colors | white | $wh$ | #pixels with tis color |
| | light brown | $lb$ | |
| | dark brown | $db$ | |
| | black | $bk$ | |
| dermatoscopic structures | linear branches | $lr$ | variation of distance from center to border |
| | irregular pigment network | $ip$ | number of unconnected pixels |
| | | $ne$ | micro-regions |
| | structureless areas | $sa$ | number of pixel into the area |
| | dots and globules | $dg$ | number of dots and globules |

The performance of the predictive model is reported in Table 4. Results show that feature selection improves the predictive model while reducing, on average, the number of features to 6.60 and 4.20 with SS and PGS respectively.

**Table 4: Performance comparison of baseline classifier, SS and PGSS.**

| Performance measure | Baseline | SS | PGSS |
|---|---|---|---|
| Accuracy | 78.86 ± 7.92 | 82.81 ± 7.88 | 81.81 ± 6.05 |
| $\|S\|$ | - | 4.60 ± 0.55 | 4.20 ± 0.84 |

Finally, the structure of the feature grouping is shown in Figure 2. Each predominant group is represented with a different colour. The largest size of the circle correspond to the predominant features found $b6$, $lr$, $bk$ and $db$. The top 4 most selected features are $b6$, $lr$, $bk$ and $lb$.

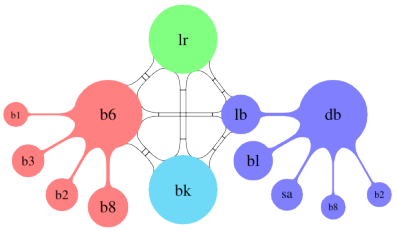


**Figure 2: Feature grouping structure and the top 4 most selected features by PGSS.**

## 4 CONCLUSIONS

In this work we propose to address the feature selection problem on high dimensional data using a feature grouping based Scatter Search strategy, called PGSS. For feature grouping the GeedyPGG algorithm is considered. PGSS is tested on high dimensional data from biomedical domain. Results show that PGSS is a competitive strategy that is able to find a reduced subset of features while keeping the predictive power of the classifier. Finally PGSS was also tested in a case study regarding melanoma. The information that PGSS provides allows to understand how features are related.

## ACKNOWLEDGMENTS

This work was supported by the CONACYT, Paraguay, under Grant PINV18-1129.

## REFERENCES

[1] M. García-Torres, F. Gómez-Vela, B. Melián-Batista, and J. M. Moreno-Vega. 2016. High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach. *Information Sciences* 326 (2016), 102–118.
[2] F. C. García-López, M. García-Torres, B. Melián-Batista, J. A. Moreno-Pérez, and J. M. Moreno-Vega. 2006. Solving the Feature Selection Problem by a Parallel Scatter Search. *European Journal of Operations Research* 169, 2 (2006), 477–489.
[3] M. A. Hall. 1999. *Correlation-based Feature Subset Selection for Machine Learning*. Ph.D. Dissertation. University of Waikato, Hamilton, New Zealand.
[4] D. N. Leguizamon Correa, L. R. Bareiro Paniagua, J. L. Vazquez Noguera, D. P. Pinto-Roa, and L. A. Salgueiro Toledo. 2015. Computerized Diagnosis of Melanocytic Lesions Based on the ABCD Method. In *2015 Latin American Computing Conference (CLEI)*. 1–12.