# Classifier Pool Generation based on a Two-level Diversity Approach

Marcos Monteiro[1], Alceu de S. Britto Jr.[1,2], Jean P. Barddal[1], Luiz S. Oliveira[3] and Robert Sabourin[4]

*[1] Pontifícia Universidade Católica do Paraná*
*[2] Universidade Estadual de Ponta Grossa*
*[3] Universidade Federal do Paraná*
*[4] École de Technologie Supérieure*

# Introduction

- **Problem:** Diversity is essential in the process of pool generation. Training classifiers on different data subsets is usually the strategy applied to create homogeneous pools.

- **Challenge**: Create data subsets to promote pool diversity and accuracy.

# Objective

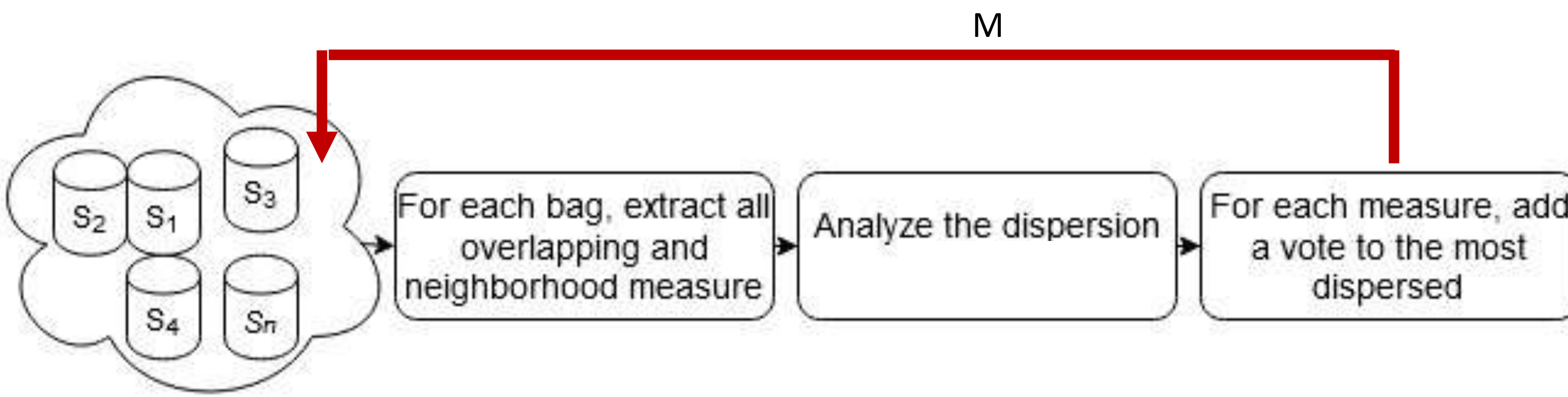
- Create a classifier pool generation method guided by diversity estimated on the data complexity and classifier decisions.
  - Select the best complexity measures for each classification problem.
  - Apply the selected measures and classifier decisions to generate a pool of diverse classifiers.

# Method

Pool generation based on diversity and complexity spaces (PGDCS).

- PGDCS is divided into **two steps:**
  - First step:
    - We select the most suitable complexity measures for each classification problem from 2 families of complexity measures.
  - Second step:
    - We generate a pool of classifiers using an optimization process to create data subsets that better cover the problem complexity space.

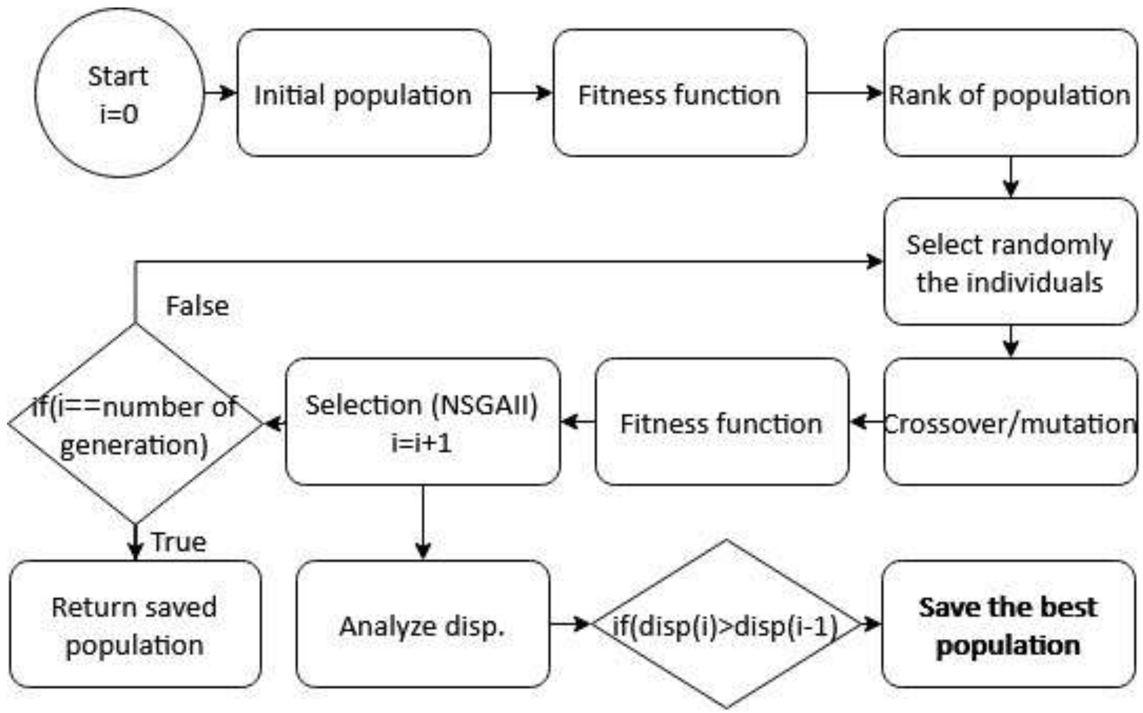
# Method - First Step




- Given the training data of a classification problem, two measures are selected:
  - A voting schema is used to select one complexity measure from each of two families: *neighborhood* and *overlapping*.
  - Subsets of data ($S_i$) with *N* samples are created randomly from the training set and analyzed concerning their dispersions in the complexity space.
  - The complexity measure presenting the greatest dispersion at each iteration received one vote.
  - The algorithm repeats the two previous steps *M* times.

# Method - Second Step

- Multi-Objective Genetic Algorithm (MOGA) is used to generate data subsets disperse in the complexity space.
- Each subset of data representing sub-problems with different levels of difficulty.
- A pool is obtained at each MOGA generation. However, the best pool in terms of diversity is selected.

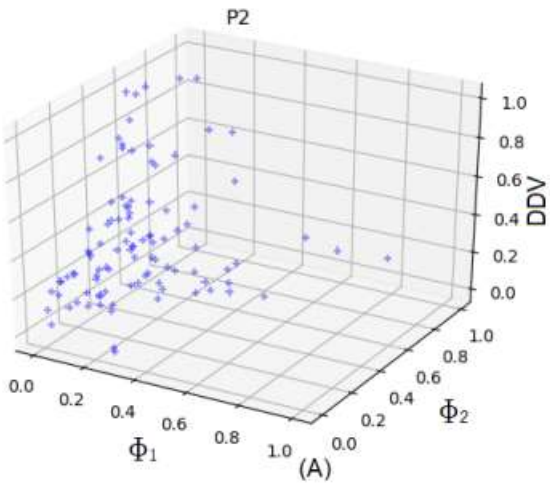


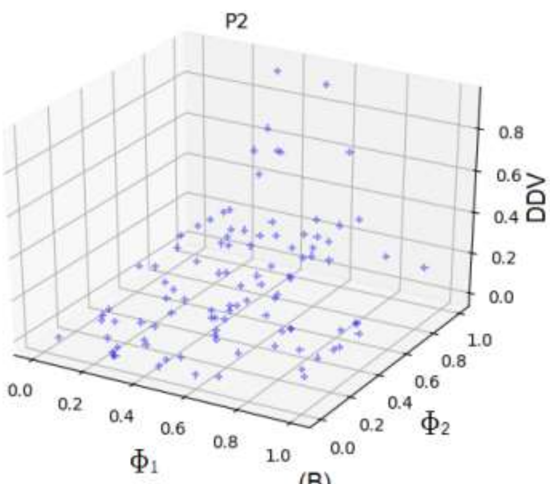
# Results – First Step

| Dataset | F1 | F1v | F2 | F3 | F4 | N1 | N2 | N3 | N4 | T1 | LSC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Australian | **4** | 3 | 1 | 2 | 0 | 0 | 0 | 2 | **5** | 1 | 2 |
| Banana | 0 | 2 | 3 | **4** | 1 | 1 | 0 | 3 | 0 | **4** | 2 |
| Blood | 2 | 0 | **5** | 3 | 0 | **3** | 1 | 1 | 2 | 1 | 2 |
| CTG | **4** | 3 | 2 | 1 | 0 | 1 | **3** | 2 | 1 | 2 | 1 |
| Diabetes | 1 | 1 | 2 | **5** | 1 | 1 | 3 | 1 | 0 | 1 | **4** |
| Faults | 0 | **10** | 0 | 0 | 0 | 1 | 1 | **4** | 1 | 1 | 2 |
| German | 0 | 0 | **9** | 1 | 0 | 3 | 1 | 0 | **4** | 1 | 1 |
| Haberman | 2 | 2 | **4** | 0 | 2 | 0 | 1 | 3 | **4** | 1 | 1 |
| Heart | 0 | 0 | 0 | 0 | **10** | 0 | 0 | 0 | **6** | 3 | 1 |
| ILPD | 1 | 3 | 0 | **5** | 1 | **4** | 2 | 2 | 1 | 0 | 1 |
| Ionosphere | **7** | 2 | 0 | 1 | 0 | 0 | 1 | 3 | **5** | 1 | 0 |
| Laryngeal1 | 1 | 0 | 0 | 1 | **8** | 0 | 2 | 2 | **4** | 0 | 2 |
| Laryngeal3 | 0 | 3 | 0 | 2 | **5** | 2 | **4** | 2 | 1 | 0 | 1 |
| Lithuanian | 0 | 0 | 2 | **6** | 2 | 2 | 2 | 1 | 1 | 1 | **3** |
| Liver | 1 | 5 | 2 | 0 | 2 | 2 | 1 | **3** | 2 | 2 | 0 |
| Mammo | 1 | 2 | **7** | 0 | 0 | 2 | 1 | 1 | **3** | 2 | 1 |
| Monk | 0 | 0 | 4 | 0 | **6** | 1 | **4** | 1 | 2 | 2 | 0 |
| P2 | 2 | **5** | 3 | 0 | 0 | 1 | 2 | **4** | 2 | 0 | 1 |
| Phoneme | 1 | **4** | 1 | 2 | 2 | 1 | 3 | **4** | 1 | 0 | 1 |
| Segmentation | 0 | **10** | 0 | 0 | 0 | 0 | 2 | **4** | 1 | 2 | 1 |
| Sonar | **5** | 2 | 0 | 1 | 2 | 2 | 0 | 2 | **5** | 1 | 0 |
| Thyroid | 2 | 0 | 0 | **5** | 3 | 0 | 2 | 2 | 1 | 0 | **5** |
| Vehicle | 0 | 0 | 0 | 1 | **9** | 1 | 0 | 3 | 0 | **4** | 2 |
| Vertebral | 1 | 2 | 1 | 0 | **6** | 2 | **4** | 2 | 2 | 0 | 0 |
| WBC | 3 | 2 | 0 | **5** | 0 | 1 | 2 | 0 | **3** | 2 | 2 |
| WDVG | **5** | 0 | 0 | 2 | 3 | 3 | 1 | 2 | 0 | **4** | 0 |
| Weaning | 3 | 3 | 0 | **4** | 0 | 0 | 2 | 0 | **6** | 0 | 2 |
| Wine | 1 | 1 | 0 | **8** | 0 | 4 | 0 | **5** | 0 | 1 | 0 |
| Average | 1.7 | 2.3 | 1.6 | 2.1 | 2.3 | 1.4 | 1.6 | 2.1 | 2.3 | 1.3 | 1.4 |

- Result of the first step for different classification problems considering measures of overlapping (F1, F1v, F2, F3 and F4) and neighborhood (N1, N2, N3, N4, T1, LSC).
- We can see the total of votes each complexity measure received.
- For instance, for the Australian dataset the following measures were selected: F1 and N4.

# Results – Second step



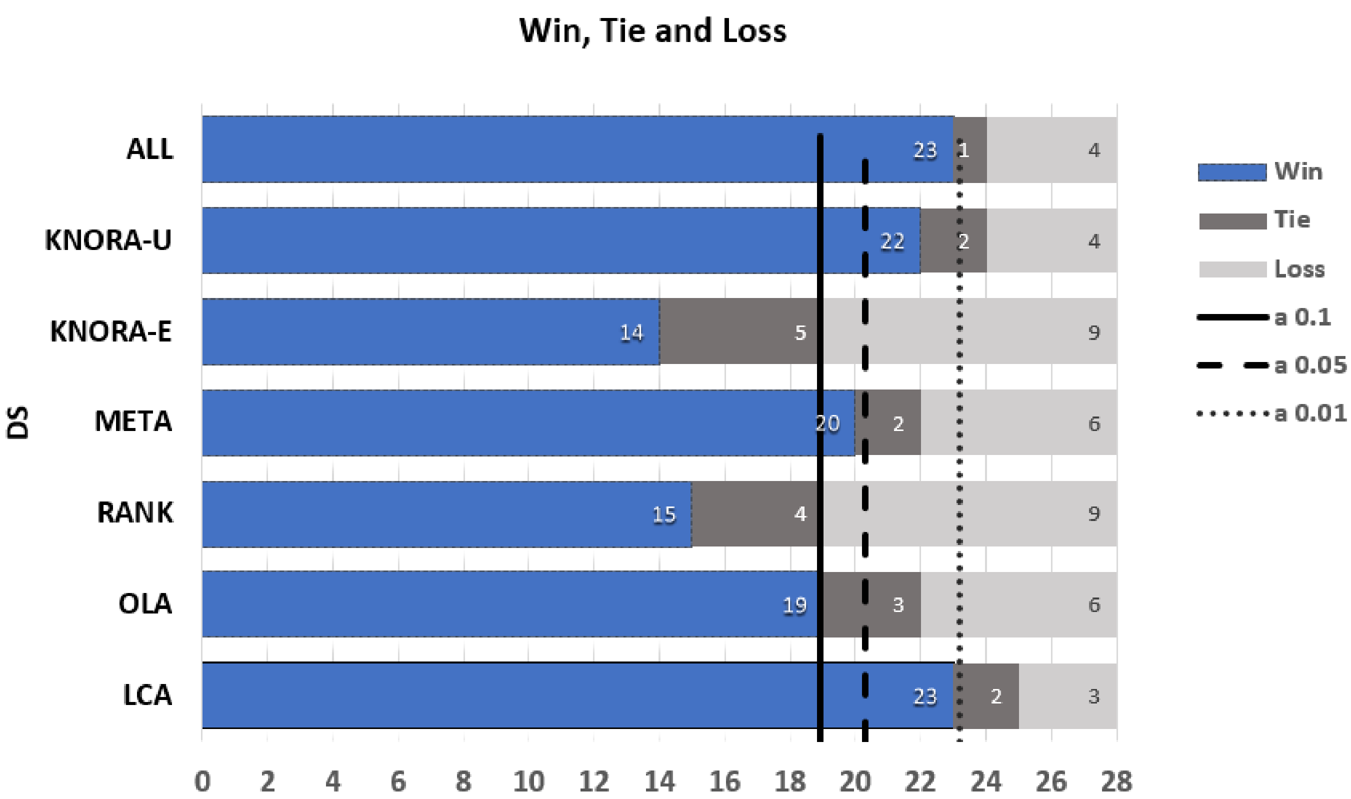

- In Figures A and B, the blue dots represent data subsets of a classification problem.
- Figure A presents the subsets' dispersion in the first generation, where each ϕ is a complexity measures and DDV is the diversity in the complexity space.
- Figure B shows the subsets after executing PGDCS. We can see them better representing the whole complexity space.

# Results

| Method | PGDCS vs Bagging | | | |
|---|---|---|---|---|
| | Win | Tie | Loss | Total experiments |
| Majority Vote | 23 | 1 | 4 | 28 |
| Dynamic Classifier Selection | 57 | 9 | 18 | 84 |
| Dynamic Ensemble Selection | 56 | 9 | 19 | 84 |
| **Overall result** | **136** | **19** | **41** | **196** |

- 20 Replications
- 196 Experiments
- 69.4% Win
- 9.6% Tie
- 20.9% Loss

# Results - Impact on dynamic selections (DS) and majority vote (ALL)




- We can see an important impact on Dynamic Selection Methods since the PGDCS generated pools composed of classifiers trained on data subsets with different levels of difficulty.

# Conclusion

- We proposed a new approach for creating a pool of diverse classifiers.
- PGDCS uses diversity in both complexity and decision spaces to generate a homogeneous pool of classifier.
- As a result, we observed that our proposal outperforms existing approaches in 69.4% of the experiments.

# Future Works

- Future works will consider different strategies to select the best pool generation.

- Compare PGDCS with another methods of pool generation.

# Acknowledgment

PUCPR

CAPES