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 , R. Imaculada Conceição, 1155 - Curitiba (PR) 80215-901 - Brazil
[2] Universidade Estadual de Ponta Grossa, Pr. Santos Andrade s/n, Ponta Grossa (PR) 84100-000 - Brazil
[3] Universidade Federal do Paraná, R. XV de Novembro, 1299, Curitiba (PR) 80060-000 - Brazil
[4] École de Technologie Supérieure, 1100, Rue Notre Dame Ouest, Montreal (QC) H3C1K3 - Canada

ICPR 2020 — 25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION — Milan, Italy 10 | 15 January 2021

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.

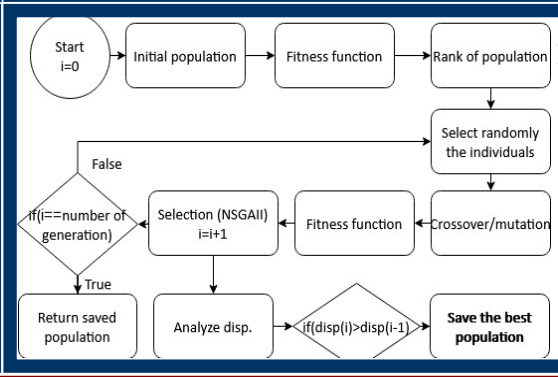
➢ **Objective:** A classifier pool generation method guided by diversity estimated on the data complexity and classifier decisions.

POOL GENERATION BASED ON DIVERSITY AND COMPLEXITY SPACES (PGDCS) METHOD

Selection of Complexity Measures

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

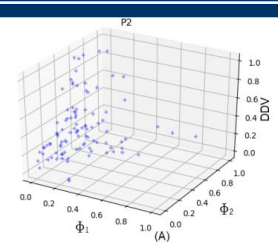
Overview of the Pool Generation Process

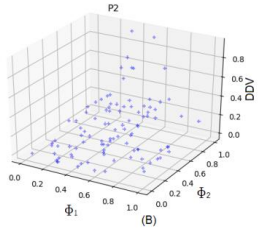


➢ Generation of data subsets disperse in the complexity space. Each subset representing sub-problems with different levels of difficulty.

ANALYSIS OF THE PGDCS METHOD

Second Step (Evolution - Pool Adaptation)





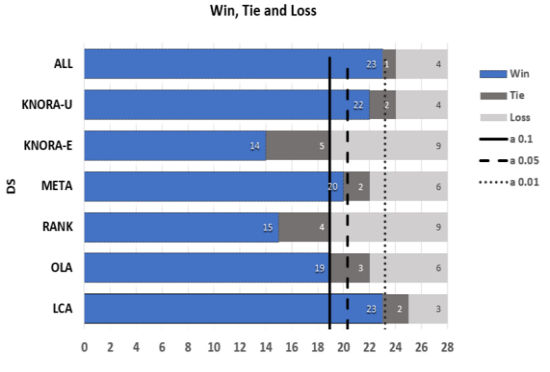
- In Figures A and B, the blue dots represent data subsets of a classification problem.
- Figure A presents the subsets' dispersion in the first generation, where each ϕ is a complexity measures and DDV is the diversity in the complexity space.
- Figure B shows the subsets after executing PGDCS. We can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

Results - General Overview

Method \ Result	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)



An important impact on Dynamic Selection Methods

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 cases.

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