# Pairwise Combination of Classifiers for Ensemble Learning on Data Streams

**By** Heitor Muril Gomes, Jean Paul Baraddal, Fabrício Enembreck
**From** Pontifical Catholic University of Paraná

# Introduction

- Despite attempts to create a diverse ensemble, there is always some amount of overlap between the component classifier models.

- Combining classifier pairs might alleviate incorrect predictions that would otherwise negatively impact the entire ensemble decision.

- Two voting strategies aimed at using these overlaps to support ensemble prediction.
  - Pairwise Accuracy (PA)
  - Pairwise Patterns (PP)

# Pairwise Accuracy

- Combines classifiers into pairs and weights predictions based on their shared estimated accuracy.

- Pairwise estimated accuracy.

- Shared estimated error rate.

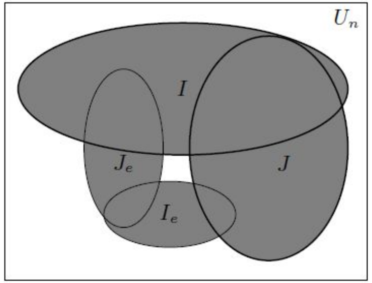


Figure 1: Venn diagram representation of window $n$ and classifiers $c_i$ and $c_j$ correctly and incorrectly classified subsets of instances

$$S_{acc}(c_i, c_j) = \frac{|I \cap J|}{|U_n^t|} \tag{1}$$

$$S_{err}(c_i, c_j) = \frac{|I_e \cap J_e|}{|U_n^t|} \tag{2}$$

$$acc(c_i) = \frac{|I|}{|U_n^t|} \tag{3}$$

# Pairwise Accuracy

$$\vec{v}(h_i(x)) := \vec{v}(h_i(x)) + S_{acc}(c_i, c_j) - S_{err}(c_i, c_j) \qquad (4)$$

$$\vec{v}(h_i(x)) := \vec{v}(h_i(x)) + acc(c_i) - S_{acc}(c_i, c_j) \qquad (5)$$

$$\vec{v}(h_j(x)) := \vec{v}(h_j(x)) + acc(c_j) - S_{acc}(c_i, c_j) \qquad (6)$$

- Employs a weighting function that prioritises equal pairwise prediction over individual prediction.
- Every position in vector $\vec{v}$ corresponds to a possible label, which is initialised to zero.
- Individual predictions for and instance $x$ are found and the vector position is updated with the new estimated accuracies.

# Pairwise Patterns

- Records prediction patterns during training and uses these patterns to weight decisions while predicting the label of an unknown instance $x$.
- Maintains a vector $\vec{p}$ with all possible prediction patterns, given two classifiers and $k$ classes. It also maintains a matrix with one column for each possible label and one row for each pattern.
- Classifiers $c_i$ and $c_j$ predict the label of an instance to form the pattern, and the correct label determines the position in the matrix to be incremented.

| $\vec{p}$ |
|---|
| (0,0) |
| (0,1) |
| $\vdots$ |
| (k-1,k-1) |

| Corr. 0 | Corr. 1 | ... | Corr. (k-1) |
|---|---|---|---|
| 12 | 4 | ... | 0 |
| 3 | 16 | ... | 1 |
| $\vdots$ | ... | ... | ... |
| 3 | 5 | ... | 18 |

Figure 2: Example of the data stored for a given pair of classifiers $c_i$ and $c_j$ for a classification problem with $k$ classes. Every entry in $\vec{p}$ has a one-to-one relation to a line in $M$.

# Ensemble Adaptations

- Generic Ensemble (GE) is an ensemble structure based on existing methods, used to test PA and PP.
  - Updating classifiers.
  - Windows.
- New classifiers can only replace existing ones.
- Creates a diverse set and can gradually adapt in case of concept drift.

# Experiments

- Comparing PA and PP.
- Prequential evaluation with a sample frequency 1/10 the total stream length.
- Maximum number of classifiers as 10.
- Window size at 1% the total stream length.

| ID | Data stream configuration | | |
|---|---|---|---|
| | Data generator | # drifts | Type of drift |
| RTG | RTG | - | - |
| AGR1 | AGRAWAL | 2 | A/A |
| AGR2 | AGRAWAL | 2 | G/G |
| SEA1 | SEA | 2 | A/A |
| SEA2 | SEA | 2 | G/G |
| HYPE | Hyperplane | - | I |

Table 1: Synthetic data streams configurations (A: Abrupt Drift, G: Gradual Drift, I: Incremental Drift)

# Experiments

- LevBag Vs. LevBag-PP

| *Dataset* | **LevBag-PP** | **LevBag** |
|---|---|---|
| AGR1 | 93.64 ± 0.15 | **93.69** ± 0.83 |
| AGR2 | 90.95 ± 1.17 | **91.07** ± 1.29 |
| AIRL | **63.7** ± 0.28 | 62.67 ± 0.25 |
| COVT | **92.95** ± 0.26 | 92.19 ± 0.28 |
| ELEC | 90.67 ± 0.24 | **90.82** ± 0.21 |
| RTS | 97.99 ± 0.1 | **98.21** ± 0.09 |
| SEA1 | **89.91** ± 0.04 | 89.64 ± 0.34 |
| SEA2 | **90.46** ± 0.03 | 90.45 ± 0.04 |
| SPAM | **93.89** ± 0.55 | 93.11 ± 0.35 |
| HYPE | **90.75** ± 0.11 | 90.29 ± 0.12 |

Table 2: Average accuracy for LevBag and LevBag-PP. The best accuracies per data stream are indicated in boldface.

- GE Vs. GE-PA Vs. GA-PP

| *Dataset* | **GE-PA** | **GE-PP** | **GE** |
|---|---|---|---|
| AGR1 | **94.2** ± 0.27 | 87.43 | 92.04 ± 0.08 |
| AGR2 | **92.42** ± 0.41 | 82.7 | 90.87 ± 0.01 |
| AIRL | **66.38** ± 0.15 | 62.67 | 66.21 ± 0.02 |
| COVT | 87.72 ± 0.43 | **89.67** | 88.31 ± 0.16 |
| ELEC | **86.03** ± 0.17 | 84.76 | 85.97 ± 0.15 |
| RTS | 95.13 ± 0.08 | **96.57** | 95.19 ± 0.02 |
| SEA1 | 89.32 ± 0.22 | 86.54 | **89.33** ± 0 |
| SEA2 | 89.41 ± 0.07 | 86.51 | **89.43** ± 0 |
| SPAM | 87.19 ± 0.06 | **88.76** | 87.1 ± 0.02 |
| HYPE | **91.16** ± 0.06 | 86.42 | 91.15 ± 0.06 |

Table 3: Average accuracy for GE, GE-PA and GE-PP. The best accuracies per data stream are indicated in boldface. GE-PP standard deviation was below $10^{-9}$ for all experiments.

# Experiments

- GE-PA & LevBag-PP Vs. Other ensemble methods.

| *Dataset* | **LevBag-PP** | **GE-PA** | **ADWBag** | **DWM** | **OAUE** | **SFNC** | **SAE2** |
|---|---|---|---|---|---|---|---|
| AGR1 | 93.64 ± 0.15 | 94.2 ± 0.27 | 94.36 ± 0.2 | 86.5 | 93.77 | 93.33 | **94.68** ± 0.15 |
| AGR2 | 90.95 ± 1.17 | 92.42 ± 0.41 | 90.69 ± 1.32 | 82.41 | **93.24** | 92.31 | 89.79 ± 2.09 |
| AIRL | 63.7 ± 0.28 | 66.38 ± 0.15 | 66.05 ± 0.32 | 61.46 | 64.48 | **66.42** | 60.8 ± 0.58 |
| COVT | 92.95 ± 0.26 | 87.72 ± 0.43 | 85.67 ± 0.25 | 91.28 | **93.55** | 85.85 | 86.59 ± 0.53 |
| ELEC | **90.67** ± 0.24 | 86.03 ± 0.17 | 85.05 ± 0.33 | 84.69 | 89.38 | 85.38 | 85.8 ± 0.48 |
| RTS | **97.99** ± 0.1 | 95.13 ± 0.08 | 95.6 ± 0.09 | 93.64 | 97.35 | 95.09 | 95.06 ± 0.12 |
| SEA1 | 89.91 ± 0.04 | 89.32 ± 0.22 | 88.63 ± 0.48 | 88.6 | **90.02** | 89.54 | 89.86 ± 0.17 |
| SEA2 | **90.46** ± 0.03 | 89.41 ± 0.07 | 90.15 ± 0.08 | 88.63 | 90.25 | 89.16 | 90.12 ± 0.1 |
| SPAM | **93.89** ± 0.55 | 87.19 ± 0.06 | 88.34 ± 0.9 | 88.21 | 67.23 | 86.5 | 87.76 ± 0.42 |
| HYPE | 90.75 ± 0.11 | **91.16** ± 0.06 | 90.5 ± 0.12 | 88.2 | 90.41 | 90.91 | 90.88 ± 0.12 |
| *Avg. Rank* | 2.6 | 3.4 | 4.2 | 6.1 | 3 | 4.4 | 4.3 |

Table 4: Comparison of average accuracy. The best accuracies per data stream are indicated in boldface. SFNC standard deviation were below $10^{-14}$ for all experiments.

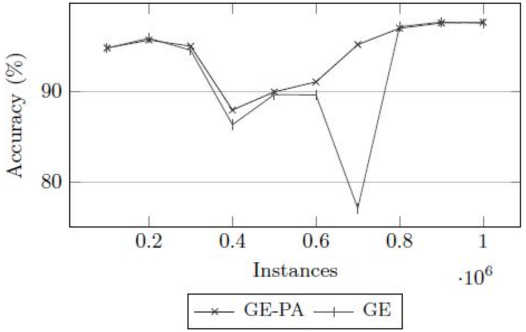


Figure 3: Accuracy on the AGR1 experiment (2 abrupt concept drifts around instances $3.33\times10^5$ and $6.66\times10^5$)

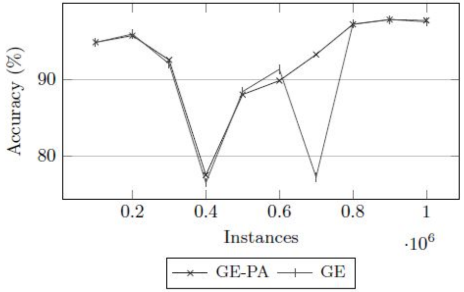


Figure 4: Accuracy on the AGR2 experiment (2 gradual concept drifts around instances $3.33\times10^5$ and $6.66\times10^5$)

# Conclusion

- Presented two voting strategies.
- The experiments performed have provided moderate accuracy improvements.
- Combining classifiers into pairs may allow more sophisticated weighting mechanisms.
- Future work?

# Questions?

Feel free to ask!