

# Addressing class imbalance in Multilabel Prototype Generation for k-Nearest Neighbor classification

**Carlos Penarrubia<sup>1</sup>,**  
Jose J. Valero-Mas<sup>1,2</sup>, Antonio Javier Gallego<sup>1</sup>, and Jorge  
Calvo-Zaragoza<sup>1</sup>

<sup>1</sup>U.I. for Computer Research, University of Alicante

<sup>2</sup>Music Technology Group, Universitat Pompeu Fabra

11th Iberian Conference on Pattern Recognition and Image Analysis  
Alicante, June 2023

# Introduction

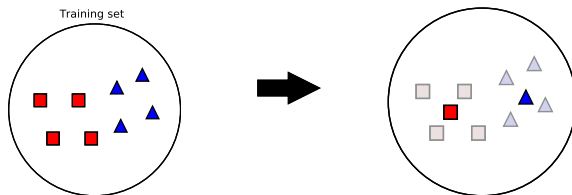
# The $k$ -Nearest Neighbor ( $k$ -NN) classifier

- ▶ Method for supervised classification
- ▶ Features
  - Compares each query to the whole dataset following a metric
  - Non-parametric method
- ▶ Drawbacks
  - Low efficiency
  - High memory usage

# The $k$ -Nearest Neighbor ( $k$ -NN) classifier

## Data Reduction

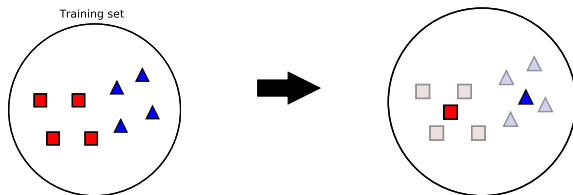
- ▶ Consists in reducing the size of the reference set
- ▶ Two main approaches:
  - \* Prototype selection (PS)
  - \* Prototype generation (PG)



# The $k$ -Nearest Neighbor ( $k$ -NN) classifier

## Data Reduction

- ▶ Consists in reducing the size of the reference set
- ▶ Two main approaches:
  - \* Prototype selection (PS)
  - \* Prototype generation (PG)

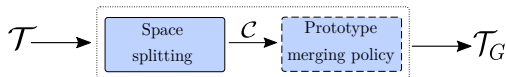


- ▶ However...
  - PG has been **scarcely** addressed in **multilabel** cases
  - Existing methods show **shortages** when addressing **imbalance** data
  - **Goal:** Tackle imbalance problems in multilabel PG

# Methodology

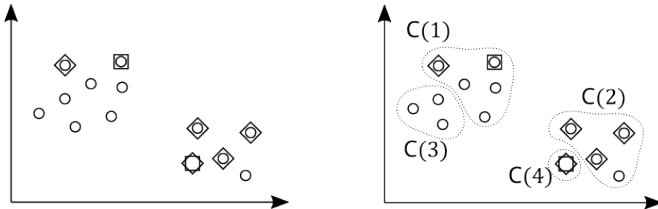
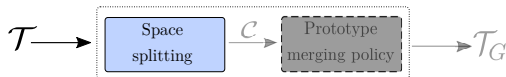
# Multilabel Prototype Generation

- ▶ Two stages
  1. Space splitting
  2. Prototype merging policy



# Multilabel Prototype Generation

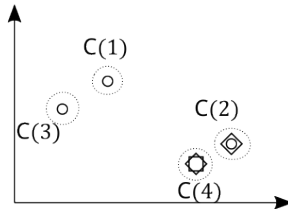
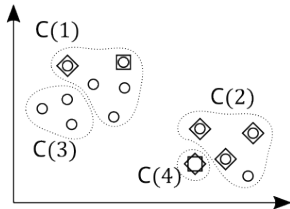
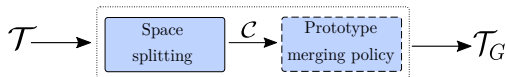
- ▶ Two stages
  1. Space splitting
  2. Prototype merging policy





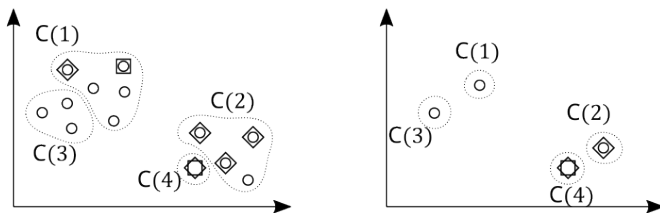
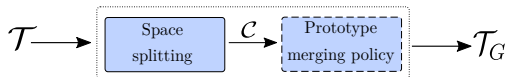
# Multilabel Prototype Generation

- ▶ Two stages
  1. Space splitting
  2. Prototype merging policy



# Multilabel Prototype Generation

- ▶ Two stages
  1. Space splitting
  2. Prototype merging policy



Is the square in C(1) **noise**? Or is it simply **underrepresented**?

# Multilabel Prototype Generation

Imbalance metrics: IRLbl and MeanIR

- Imbalance ratio per label ( $\lambda$ ):

$$\text{IRLbl}(\lambda) = \frac{\max_{\forall \lambda' \in \mathcal{Y}} \left( \sum_{i=1}^{|\mathcal{T}|} \lambda' \in \mathbf{y}_i \right)}{\sum_{i=1}^{|\mathcal{T}|} \lambda \in \mathbf{y}_i} \quad (1)$$

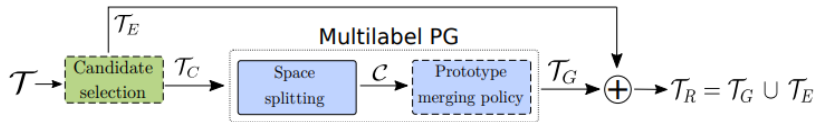
- Example  $\lambda = \square$ :

$$\text{IRLbl}(\square) = \frac{\max(12, 4, 2)}{2} = 6 \quad (2)$$

- Mean imbalance ratio:

$$\text{MeanIR} = \frac{1}{|\mathcal{Y}|} \sum_{\lambda \in \mathcal{Y}} \text{IRLbl}(\lambda) = \frac{1 + 2.4 + 6}{3} = 3.13 \quad (3)$$

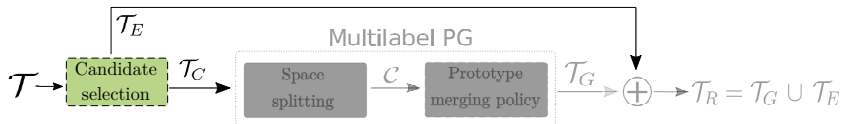
# Proposal



- ▶ Two additional **imbalance-aware** mechanisms:
  - Candidate selection
  - Prototype merging policies

# Proposal

## Candidate selection

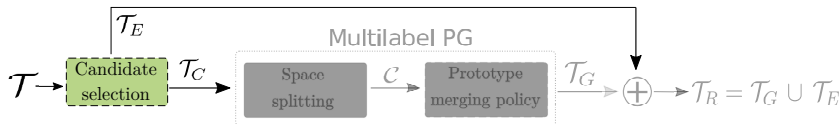


► Initial set  $\mathcal{T}$  is **split**:

- Set  $\mathcal{T}_E$  of samples with **imbalanced** samples
- Set  $\mathcal{T}_C$  of samples with **non-imbalanced** samples

# Proposal

## Candidate selection

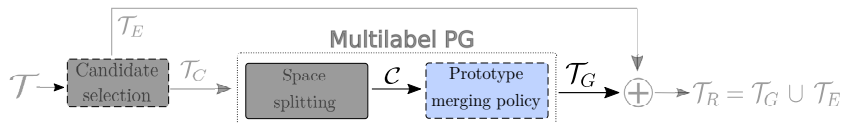


- ▶ Initial set  $\mathcal{T}$  is **split**:
  - Set  $\mathcal{T}_E$  of samples with **imbalanced** samples
  - Set  $\mathcal{T}_C$  of samples with **non-imbalanced** samples
- ▶ Imbalanced samples:

$$\mathcal{T}_E = (\mathbf{x}_i, \mathbf{y}_i) : \text{IRLbl}(\lambda) > \text{MeanIR} \quad \forall \lambda \in \mathbf{y}_i$$

# Proposal

## Prototype merging policy

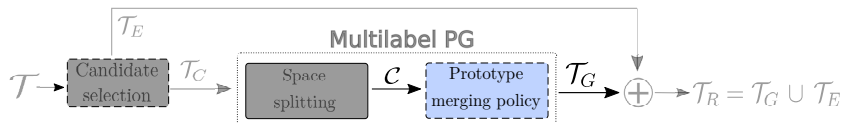


► Policies for **merging** prototypes in  $\mathcal{C}$ :

- Features ( $\mathbf{x}_i$ ): Feature-wise mean
- Label space ( $\mathbf{y}_i$ ):
  - Base case:  $|C(m)|_{\lambda} \geq \frac{|C(m)|}{2}$

# Proposal

## Prototype merging policy



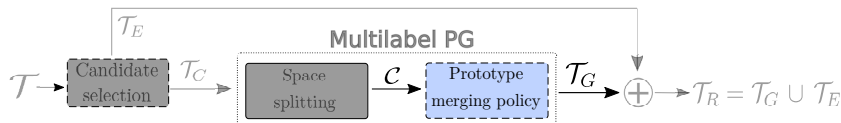
### ► Policies for **merging** prototypes in $\mathcal{C}$ :

- Features ( $\mathbf{x}_i$ ): Feature-wise mean
- Label space ( $\mathbf{y}_i$ ):
  - Base case:  $|C(m)|_\lambda \geq \frac{|C(m)|}{2}$
  - Proposal I:  $|C(m)|_\lambda \geq \left\lfloor \frac{|C(m)|}{2 \cdot \text{IRLb}(\lambda)} \right\rfloor$



# Proposal

## Prototype merging policy



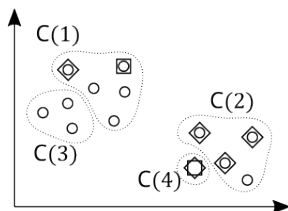
### ► Policies for **merging** prototypes in $\mathcal{C}$ :

- Features ( $\mathbf{x}_i$ ): Feature-wise mean
- Label space ( $\mathbf{y}_i$ ):
  - Base case:  $|C(m)|_\lambda \geq \frac{|C(m)|}{2}$
  - Proposal I:  $|C(m)|_\lambda \geq \left\lfloor \frac{|C(m)|}{2 \cdot \text{IRLbl}(\lambda)} \right\rfloor$
  - Proposal II: (Base case)  $\vee$  ( $\text{IRLbl}(\lambda) > \text{MeanIR}$ )

# Proposal

## Prototype merging policy

### ► Example:

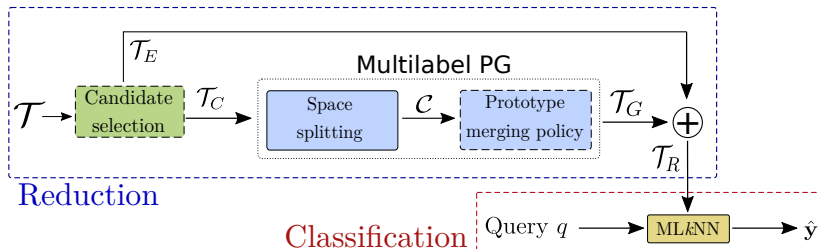


- Base policy:  $C(1) = \{\circ\}$
- Policy 1:  $C(1) = \{\circ, \square\}$
- Policy 2:  $C(1) = \{\circ, \square\}$

## Experimental set-up and results

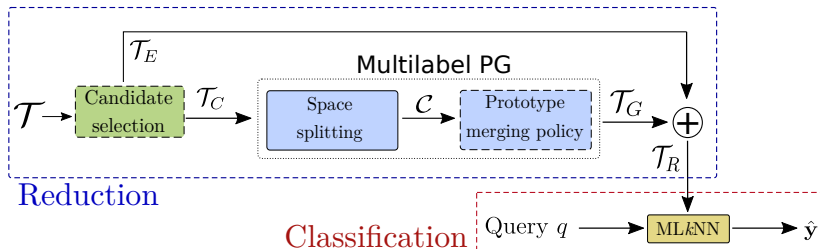
# Experimental set-up

## Scheme and algorithms



# Experimental set-up

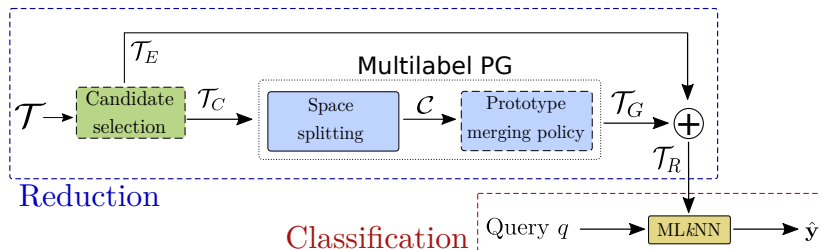
## Scheme and algorithms



- Multilabel  $k$ -NN algorithm:
  - ML $k$ NN ( $k = 1$ )

# Experimental set-up

## Scheme and algorithms



- ▶ Multilabel  $k$ -NN algorithm:
  - ML $k$ NN ( $k = 1$ )
- ▶ Multilabel PG methods:
  - Multilabel Reduction through Homogeneous Clustering (MRHC)
  - Multilabel Chen (MChen)
  - Multilabel Reduction through Space Partitioning (MRSP3)

# Experimental set-up

## Datasets and metrics

### ► Datasets:

Name	Set size		MeanIR
	Train	Test	
Low imbalance			
Scene	1,211	1,196	1.33
Emotions	391	202	1.49
Birds	322	323	6.10
Yeast	1,500	917	7.27
Bibtex	4,880	2,515	12.78
High imbalance			
Genbase	463	199	31.60
Medical	333	645	48.59
rcv1subset4	3,000	3,000	170.84
rcv1subset2	3,000	3,000	177.89
Corel5k	4,500	500	183.29
rcv1subset1	3,000	3,000	191.42
rcv1subset3	3,000	3,000	192.48

# Experimental set-up

## Datasets and metrics

### ► Datasets:

Name	Set size		MeanIR
	Train	Test	
Low imbalance			
Scene	1,211	1,196	1.33
Emotions	391	202	1.49
Birds	322	323	6.10
Yeast	1,500	917	7.27
Bibtex	4,880	2,515	12.78
High imbalance			
Genbase	463	199	31.60
Medical	333	645	48.59
rcv1subset4	3,000	3,000	170.84
rcv1subset2	3,000	3,000	177.89
Corel5k	4,500	500	183.29
rcv1subset1	3,000	3,000	191.42
rcv1subset3	3,000	3,000	192.48

### ► Metrics:

- Macro F1 score
- Reduction rate:  $|\mathcal{T}_R|/|\mathcal{T}|$



# Results

	No candidate selection ( $\mathcal{T}_C = \mathcal{T}$ )				Using candidate selection ( $\mathcal{T}_C \subseteq \mathcal{T}$ )			
	Size (%)	Merging policy			Size (%)	Merging policy		
		Base	Policy 1	Policy 2		Base	Policy 1	Policy 2
<b>Low imbalance</b>								
MRHC	57.83	42.44	<b>43.70</b>	42.87	70.65	42.34	43.36	42.34
MChen <sub>10</sub>	9.98	30.11	36.87	27.81	40.98	36.36	<b>40.33</b>	36.36
MChen <sub>50</sub>	49.97	37.15	41.64	38.26	67.17	40.46	<b>41.66</b>	40.46
MChen <sub>90</sub>	89.89	42.20	42.29	42.26	93.23	42.99	<b>43.00</b>	42.99
MRSP3	66.84	40.73	<b>43.58</b>	41.43	78.12	41.76	43.04	41.76

- General improvement with proposed policies
- Best results with Policy 1
- Worse efficiency when using Candidate Selection

# Results

	No candidate selection ( $\mathcal{T}_C = \mathcal{T}$ )				Using candidate selection ( $\mathcal{T}_C \subseteq \mathcal{T}$ )			
	Size (%)	Merging policy			Size (%)	Merging policy		
		Base	Policy 1	Policy 2		Base	Policy 1	Policy 2
<b>High imbalance</b>								
MRHC	47.55	12.03	<b>12.48</b>	12.03	46.89	11.92	<b>12.48</b>	11.92
MChen <sub>10</sub>	9.98	7.23	9.80	7.36	12.64	7.48	<b>9.94</b>	7.48
MChen <sub>50</sub>	49.96	9.96	<b>11.93</b>	10.04	51.45	9.97	11.78	9.97
MChen <sub>90</sub>	89.58	11.91	<b>12.08</b>	11.91	89.74	11.83	12.04	11.83
MRSP3	60.94	11.65	<b>13.96</b>	12.11	61.08	8.80	10.59	8.80

- General improvement with proposed policies
- Best results Policy 1
- Similar efficiency when using Candidate Selection

# Conclusions

# Conclusions

- ▶ Two novel policy methods for imbalance aware
- ▶ Mechanism to prevent severely imbalanced samples from undergoing a reduction process
- ▶ Experimental validation in low imbalance and high imbalance datasets
- ▶ Promising results with different levels of imbalance ratio

# Future works

- ▶ Develop similar policies for other stages of Multilabel PG
- ▶ Full pipeline
- ▶ Use other measurement for imbalance

# Acknowledgments

This work was supported by the I+D+i project TED2021-132103A-I00 (DOREMI), funded by MCIN/AEI/10.13039/501100011033.



**Financiado por  
la Unión Europea**

NextGenerationEU

# Addressing class imbalance in Multilabel Prototype Generation for k-Nearest Neighbor classification

**Carlos Penarrubia<sup>1</sup>,**  
Jose J. Valero-Mas<sup>1,2</sup>, Antonio Javier Gallego<sup>1</sup>, and Jorge  
Calvo-Zaragoza<sup>1</sup>

<sup>1</sup>U.I. for Computer Research, University of Alicante

<sup>2</sup>Music Technology Group, Universitat Pompeu Fabra

11th Iberian Conference on Pattern Recognition and Image Analysis  
Alicante, June 2023