Learning class-imbalanced problems from the perspective of data intrinsic characteristics
Kong, J.

Citation
Kong, J. (2023, September 27). Learning class-imbalanced problems from the perspective of data intrinsic characteristics. Retrieved from https://hdl.handle.net/1887/3642254

Version: Publisher's Version
License: Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from: https://hdl.handle.net/1887/3642254

Note: To cite this publication please use the final published version (if applicable).
Learning Class-Imbalanced Problems from the Perspective of Data Intrinsic Characteristics

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
on gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 27 september 2023
klokke 13:45 uur

door

Jiawen Kong
geboren te Harbin, China
in 1995
Promotores:
Prof.dr. T.H.W. Bäck
Prof.dr. B. Sendhoff (TU Darmstadt, Germany)

Co-promotor:
Dr. W.J. Kowalczyk

Promotiecommissie: (Dutch)
Prof.dr. A. Plaat.
Prof.dr. M.M. Bonsangue
Dr. A.V. Kononova
Prof.dr. S. Mostaghim (Otto von Guericke University of Magdeburg, Germany)
Dr. M. López-Ibáñez (The University of Manchester, UK)

Copyright ©Jiawen Kong

This research has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement number 766186 (ECOLE).
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Research Questions</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Outline of the Thesis</td>
<td>6</td>
</tr>
<tr>
<td>1.4 Publications</td>
<td>7</td>
</tr>
<tr>
<td><strong>2 Preliminaries</strong></td>
<td>9</td>
</tr>
<tr>
<td>2.1 Binary Class Imbalance Learning</td>
<td>9</td>
</tr>
<tr>
<td>2.1.1 Existing Approaches</td>
<td>10</td>
</tr>
<tr>
<td>2.1.2 Performance Metrics</td>
<td>13</td>
</tr>
<tr>
<td>2.2 Multi-Class Imbalance Learning</td>
<td>17</td>
</tr>
<tr>
<td>2.2.1 Existing Approaches</td>
<td>18</td>
</tr>
<tr>
<td>2.2.2 Performance Metrics</td>
<td>21</td>
</tr>
<tr>
<td>2.3 Data Complexity for Imbalanced Datasets</td>
<td>22</td>
</tr>
<tr>
<td>2.3.1 Overlapping and Class Separability</td>
<td>23</td>
</tr>
<tr>
<td>2.3.2 Types of Sample in Imbalanced Domain</td>
<td>26</td>
</tr>
<tr>
<td>2.4 Imbalanced Benchmark Datasets and Applications</td>
<td>28</td>
</tr>
<tr>
<td>2.4.1 KEEL-Dataset Repository</td>
<td>28</td>
</tr>
<tr>
<td>2.4.2 Imbalanced Applications</td>
<td>29</td>
</tr>
<tr>
<td><strong>3 An Empirical Investigation Comparing Several Oversampling Techniques</strong></td>
<td>31</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>31</td>
</tr>
<tr>
<td>3.2 Related Work</td>
<td>33</td>
</tr>
</tbody>
</table>
# Improved Sample Type Identification for Multi-Class Imbalanced Classification

## 6.1 Introduction

## 6.2 Related Works

### 6.2.1 Studies on Types of Samples in Binary Scenarios

### 6.2.2 Problems When Extending to Multi-class Scenarios

## 6.3 New Identification Rule for Multi-class Scenarios

### 6.3.1 Adjusting $k$ according to Imbalance Ratio

### 6.3.2 Considering neighbourhood Information of the neighbours

## 6.4 Experiments

### 6.4.1 Information on the Datasets

### 6.4.2 Experimental Setup

### 6.4.3 Experimental Results and Discussion

## 6.5 Applications on the Detection of Surface Defects

### 6.5.1 Information on Surface Defects Dataset

### 6.5.2 Visualisation and Preprocessing

### 6.5.3 Experiments on Surface Defects Dataset

## 6.6 Conclusions and Future Work

## 7 Conclusions

### 7.1 Summary

### 7.2 Future Work

## Appendices

### A Additional Experimental Results

## Bibliography

## Samenvatting

## Summary

## Curriculum Vitae