Learning class-imbalanced problems from the perspective of data intrinsic characteristics

Kong, J.

Learning Class-Imbalanced Problems from the Perspective of Data Intrinsic Characteristics

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op 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)

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

Contents

Contents iii

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Research Questions . . . 4
- 1.3 Outline of the Thesis . . . 6
- 1.4 Publications . . . 7

2 Preliminaries 9
- 2.1 Binary Class Imbalance Learning . . . 9
  - 2.1.1 Existing Approaches . . . 10
  - 2.1.2 Performance Metrics . . . 13
- 2.2 Multi-Class Imbalance Learning . . . 17
  - 2.2.1 Existing Approaches . . . 18
  - 2.2.2 Performance Metrics . . . 21
- 2.3 Data Complexity for Imbalanced Datasets . . . 22
  - 2.3.1 Overlapping and Class Separability . . . 23
  - 2.3.2 Types of Sample in Imbalanced Domain . . . 26
- 2.4 Imbalanced Benchmark Datasets and Applications . . . 28
  - 2.4.1 KEEL-Dataset Repository . . . 28
  - 2.4.2 Imbalanced Applications . . . 29

3 An Empirical Investigation Comparing Several Oversampling Techniques 31
- 3.1 Introduction . . . 31
- 3.2 Related Work . . . 33

3.2.1 Oversampling Techniques 34
3.2.2 Data Complexity 37
3.3 Experiments 38
3.3.1 Information on the Datasets 38
3.3.2 Cross-Validation in Imbalanced Learning 39
3.3.3 Experimental Setup 40
3.3.4 Experimental Results and Discussion 42
3.4 Efficient Oversampling for Engineering Vehicle Mesh Dataset . . . 48
3.4.1 Generation of a Synthetic Data Set 50
3.4.2 Experimental Results and Discussion 51
3.5 Conclusions 52

4 Hyperparameter Optimisation on Class-Imbalance Problems 55
4.1 Introduction 55
4.2 Related Works 57
4.2.1 Resampling Techniques 58
4.2.2 Hyperparameter Optimisation 59
4.3 Experiments 61
4.3.1 Information on the Datasets 61
4.3.2 Experimental Setup 62
4.3.3 Experimental Results and Discussions 63
4.4 Conclusions and Future Work 63

5 Improving Imbalanced Classification via Adding Additional Attributes 67
5.1 Introduction 67
5.2 Related Works 69
5.2.1 Resampling Techniques 69
5.2.2 Anomaly Detection 70
5.2.3 Four Types of Samples in Imbalanced Datasets 71
5.3 Experiments 73
5.3.1 Information on the Datasets 73
5.3.2 Experimental Setup 73
5.3.3 Experimental Results and Discussion 75
5.4 Conclusions and Future Work 76

6 Improved Sample Type Identification for Multi-Class Imbalanced Classification **85**
6.1 Introduction 85
6.2 Related Works 87
6.2.1 Studies on Types of Samples in Binary Scenarios 87
6.2.2 Problems When Extending to Multi-class Scenarios 88
6.3 New Identification Rule for Multi-class Scenarios 90
6.3.1 Adjusting k according to Imbalance Ratio 90
6.3.2 Considering neighbourhood Information of the neighbours 91
6.4 Experiments 92
6.4.1 Information on the Datasets 93
6.4.2 Experimental Setup 93
6.4.3 Experimental Results and Discussion 94
6.5 Applications on the Detection of Surface Defects 97
6.5.1 Information on Surface Defects Dataset 98
6.5.2 Visualisation and Preprocessing 99
6.5.3 Experiments on Surface Defects Dataset 101
6.6 Conclusions and Future Work 102

7 Conclusions **105**
7.1 Summary 105
7.2 Future Work 108

Appendices **111**

A Additional Experimental Results **113**

Bibliography **117**

Samenvatting **131**

Summary **133**

Curriculum Vitae **135**