



Contribution ID: 102

Type: not specified

Evaluating classifier performance in hard classification tasks

Tuesday, 17 September 2024 10:25 (20 minutes)

Different data difficulty factors (e.g., class imbalance, class overlapping, presence of outliers and noisy observations and difficult border decisions) make classification tasks challenging in many practical applications and are hot topics in the domain of pattern recognition, machine learning and deep learning. Data complexity factors have been widely discussed in specialized literature from a model-based or a data-based perspective, conversely less research efforts have been devoted to investigating their effect on the behavior of classifier predictive performance measures. Our study tries to address this issue by investigating the impact of data complexity on the behavior of several measures of classifier predictive performance. The investigation has been conducted via an extensive study based on numerical experiments using artificial data sets. The data generation process has been controlled through a set of parameters (e.g., number of features; class frequency distributions; frequency distributions of safe and unsafe instances) defining the characteristics of generated data. The artificial data sets have been classified using several algorithms whose predictive performances have been evaluated through the measures under study. Study results highlight that, although the investigated performance measures quite agree for easy classification tasks (i.e., with balanced datasets containing only safe instances), their behavior significantly differs when dealing with difficult classification tasks (i.e., increasing data complexity) which is a rule in many real-word classification problems.

Acknowledgements

This study was carried out within the MICS (Made in Italy–Circular and Sustainable) Extended Partnership and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) –MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 –D.D. 1551.11-10-2022, PE00000004). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Type of presentation

Talk

Classification

Both methodology and application

Keywords

Classifier predictive performance, Data Complexity, Artificial data

Primary author: VANACORE, Amalia (Department of Industrial Engineering University of Naples Federico II)

Co-author: Dr CIARDIELLO, Armando (Dept. of Industrial Engineering, University of Naples Federico II)

Presenters: VANACORE, Amalia (Department of Industrial Engineering University of Naples Federico II); Dr
CIARDIELLO, Armando (Dept. of Industrial Engineering, University of Naples Federico II)

Session Classification: Data Science

Track Classification: Data Science