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

Kong, J.

Samenvatting

Het ongebalanceerde klasse probleem is een uitdagende classificatie probleem en komt vaak voor in de praktijk in dagelijkse toepassingen. Er zijn verschillende technieken ontwikkeld om de onevenwichtige classificatieprestaties theoretisch en praktisch te verbeteren. Naast het ontwikkelen van nieuwe methodes, richten onderzoekers zich ook op het belang van het begrijpen van de data zelf, wat meer inzicht zal geven in wat de ongebalanceerde klasse prestaties daadwerkelijk belemmert.

In dit proefschrift is onderzoek gedaan naar het leren van ongebalanceerde klasse problemen vanuit het perspectief van data-intrinsieke kenmerken. Het empirische onderzoek waarbij verschillende algoritmen op data niveau werden vergeleken, toont aan dat over-sampling-benaderingen, rekening houdend met de minderheids-klasse-verdeling, in de meeste gevallen betere ongebalanceerde classificatieprestaties kunnen opleveren. Hoewel data complexe metingen geen richtlijn kunnen geven over de keuze van re-sampling technieken, vinden we dat de potentieel beste AUC-waarde kan worden voorspeld door de F1v-meting (de Directional-vector Maximum Fisher's Discriminant Ratio). Beide conclusies worden ook geverifieerd op een op de praktijk geïnspireerde voertuig mesh dataset van het Honda Research Institute.

Optimalisatie van hyperparameters is zeer effectief gebleken voor veel classificatiealgoritmen voor machine learning. De maximale diepte van de boom en het minimale aantal samples dat nodig is intern knooppunt te splitsen, zijn bijvoorbeeld cruciaal voor het afstemmen van de beslissingsboom om de beste prestaties te bereiken. we benadrukken daarom het belang van afstemming van hyperparameters voor benaderingen op data niveau.

Het afwijkingsdetectieprobleem is een ongebalanceerd-klasse-probleem met een

extreem onevenwichtige verhouding. Technieken voor afwijkingsdetectieprobleem kunnen worden toegepast op ongebalanceerde-klasse problemen met nauwkeurige afstelling. In dit proefschrift stellen we voor om de Local Outlier Score te introduceren, wat een belangrijke indicator om te evalueren of een steekproef een outlier is, als een extra toepassing van de originele ongebalanceerde dataset. Dit voorstel is meer dan het lenen van kennis uit afwijkingsdetectie onderzoeksveld, maar geeft onderzoekers ook de mogelijkheid om inzicht te krijgen in de data in plaats van te over- of onder-samplen.

In het laatste deel van het proefschrift wordt een verbeterde sampling type identificatie voorgesteld voor het omgaan met ongebalanceerde classificatie met meerdere klassen en toegepast op een dataset uit de praktijk voor oppervlaktedefecten van TATA Steel. Ondertussen gaan we in op het belang van het begrijpen van de verschillende intrinsieke gegevenskenmerken voor binaire scenario's en scenario's met meerdere klassen.