A Survey on Methods for Solving Data Imbalance Problem for Classification

Volume 127
Number 15

Year of Publication: 2015

Authors:
Arpit Singh, Anuradha Purohit

Abstract

The term “data imbalance” in classification is a well established phenomenon in which data set contains unbalanced class distributions. Dataset is called unbalanced if it contains at least one class which is presented by very few examples. A range of solutions have been proposed for the problem of data imbalance including data sampling, cost evaluation of model, bagging, boosting, Genetic Programming (GP) based methods etc. This paper presents a survey of various methods introduced by researchers to handle data imbalance problem in order to improve classification performance and further the comparison between the methods on the basis of their advantages and disadvantages is done.

References

2. Urvesh Bhowan, Mark Johnston and Mengjie Zhang “Developing New Fitness Functions


Index Terms

Computer Science Artificial Intelligence

Keywords

Classification, data imbalance, genetic programming, boosting, bagging, sampling.