Abstract

Outlier detection and class imbalance modeling process play significant roles to enable effective and efficient algorithms for statistic analysis, data mining, machine learning, and knowledge discovery frameworks working on imbalanced datasets. Although there has been vast literature on imbalanced datasets, the shortcomings of distance-based functions in response to a varied density of data points have not been solving yet.

The primary aim of this dissertation was to exploit a new alternative approach for local outlier detection tasks by fundamentally changing the way to measure the outlier degree of each data point. To achieve this goal, we developed a mass-based approach to measure the dissimilarity between data points. Then, we introduced a new outlier scoring method by employing mass-based dissimilarity and probability modeling to detect the local outliers in a given dataset. The experimental study tested on artificial datasets and real application datasets show that our proposed MLOS approach is competitive with the state-of-the-art approaches.

In the same manner, to exploit the mass-based measurement for learning from the imbalanced datasets, we introduce the other two new methods for the class imbalance task. The first model is a simple application of weighted sum. The second model is an integration of the mass estimation and the Dempster-Shafer theory of evidence. These proposed models were assessed by using significant evaluation metrics such as F1 score, Brier score, ROCAUC, and PR-AUC score testing on a wide range of benchmark datasets. In addition, all experimental results were validated using the non-parametric statistical Wilcoxon signed ranks test.

This dissertation was the first study, regarding to our knowledge, to investigate the local outlier detection problem using mass-based dissimilarity measurement; the key finding was that the proposed MLOS approach presents an alternative way to score the outlierness of each data point in a given dataset. Secondly, the simulation results showed that our proposed new models for the class imbalance task outperformed the other 11 competitive methods. The experiments were conducted on a wide varying application domains, a varied imbalance ratio, and the number of instances.

Keywords: Imbalanced data, outlier detection, outlier modeling, mass-based dissimilarity, weighted sum, Dempster-Shafer theory.