Universität Rostock

Traditio et Innovatio

DISSERTATION

zur

zur Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)

Improved imbalanced classification through convex space learning

Promotionsgebiet Systembiologie und Bioinformatik

Fakultät für Informatik und Elektrotechnik

UNIVERSITÄT ROSTOCK

vorgelegt von

SAPTARSHI BEJ

geboren am 24. Juli, 1991, Bardhaman, West Bengal, Indien

Supervisor: Professor Olaf Wolkenhauer
Universität Rostock

Rostock, August 2021

Abstract

Imbalanced datasets are abundant in several real-life classification problems where Machine Learning (ML) finds its application. Such problems are characterised by classes with an unequal distribution of samples over several classes. For imbalanced datasets during the training process, the ML models encounter a larger number of majority class samples and, thus, tends to become biased towards the majority class.

One of the principal disadvantages of the SMOTE algorithm is its tendency to over-generalise the minority class. This causes the associated classifiers to misclassify the majority class data points. Moreover, evidence from recent comparative studies reveals that the performance of the extensions of the SMOTE algorithm vary depending on the classifiers they are implemented with.

In this thesis, these limitations of SMOTE-based oversampling algorithms are addressed through the novel idea of *convex space learning*. In an analytical explanation behind the idea, I show that SMOTE-based oversampling algorithms generate synthetic samples with high variance in a minority class data neighbourhood. I developed the LoRAS algorithm that can model the convex space of the minority class using multiple convex combinations of shadowsamples in a minority class neighbourhood.

To address the issue of classifier dependence of SMOTE-based oversampling algorithms, I proposed the ProWRAS algorithm. By controlling the variance of the synthetic samples, as well as a proximity-weighted clustering system of the minority class data, the ProWRAS algorithm improves the performance, compared to algorithms that generate synthetic samples through modelling high dimensional convex spaces of the minority class. Most importantly, the performance of ProWRAS with proper choice of oversampling schemes, is independent of the classifier used.

The success of the proposed algorithms has been demonstrated using rigorous benchmarking studies for over thirty publicly available datasets. The most challenging datasets are chosen with several well-defined criteria such as high imbalance, high dimensionality, and high absolute imbalance, ensuring impartiality in the benchmarking studies. From the benchmarking studies, where the proposed algorithms are compared with over ten SMOTE-based algorithms including some models with best known performance, for diverse performance measures, it is comprehensively established that the proposed algorithms can out-perform state-of-the-art algorithms.

Zusammenfassung

Unausgewogene Datensätze sind in mehreren realen Klassifizierungsproblemen, in denen Maschinelles Lernen (ML) seine Anwendung findet, häufig anzutreffen. Solche Probleme sind durch Klassen mit einer ungleichen Verteilung der Proben auf mehrere Klassen gekennzeichnet. Bei unausgewogenen Datensätzen treffen die ML-Modelle während des Trainingsprozesses auf eine gröSSere Anzahl von Stichproben der Mehrheitsklasse und neigen daher dazu, in Richtung der Mehrheitsklasse verzerrt zu werden.

Einer der Hauptnachteile des SMOTE-Algorithmus ist seine Tendenz zur Übergeneralisierung der Minderheitenklasse. Dies führt dazu, dass die zugehörigen Klassifikatoren Datenpunkte der Mehrheitsklasse falsch klassifizieren. Darüber hinaus zeigen aktuelle Vergleichsstudien, dass die Leistung der Erweiterungen des SMOTE-Algorithmus in Abhängigkeit von den Klassifikatoren, mit denen sie implementiert werden, variiert.

In dieser Arbeit werden diese Einschränkungen der SMOTE-basierten Oversampling-Algorithmen durch die neuartige Idee des konvexen Raumlernens angegangen. In einer analytischen Erklärung hinter der Idee zeigen wir, dass SMOTE-basierte Oversampling-Algorithmen synthetische Stichproben mit hoher Varianz in einer Minderheitenklassen-Datenumgebung erzeugen. Ich habe den LoRAS-Algorithmus entwickelt, der den konvexen Raum der Minderheitenklasse mit mehreren konvexen Kombinationen von Schattensamples in einer Minderheitenklassen-Nachbarschaft modellieren kann.

Um das Problem der Klassifikatorabhängigkeit von SMOTE-basierten Oversampling-Algorithmen zu lösen, habe ich den ProWRAS-Algorithmus vorgeschlagen. Durch die Steuerung der Varianz der synthetischen Stichproben sowie eines näherungsgewichteten Clustersystems der Daten der Minderheitenklasse verbessert der ProWRAS-Algorithmus die Leistung im Vergleich zu Algorithmen, die synthetische Stichproben durch Modellierung hochdimensionaler konvexer Räume der Minderheitenklasse erzeugen. Am wichtigsten ist jedoch, dass die Leistung von ProWRAS bei richtiger Wahl der Oversampling-Schemata unabhängig vom verwendeten Klassifikator ist.

Der Erfolg der vorgeschlagenen Algorithmen wurde durch strenge Benchmarking-Studien für über dreiSSig öffentlich verfügbare Datensätze nachgewiesen. Die anspruchsvollsten Datensätze werden nach mehreren wohldefinierten Kriterien ausgewählt, wie z. B. hohe Ungleichheit, hohe Dimensionalität und hohe absolute Ungleichheit, um die Unparteilichkeit in den Benchmarking-Studien zu gewährleisten. Aus den Benchmarking-Studien, in denen die vorgeschlagenen Algorithmen mit mehr als zehn SMOTE-basierten Algorithmen, einschlieSSlich einiger Modelle mit der besten

bekannten Leistung, für verschiedene LeistungsmaSSe verglichen werden, wird umfassend festgestellt, dass die vorgeschlagenen Algorithmen die State-of-the-Art-Algorithmen übertreffen können.