DOI 10.37882/2223-2966.2023.8-2.29

ПРИМЕНЕНИЕ МНОГОМЕРНОГО МЕТОДА ТОЧЕЧНЫХ РАСПРЕДЕЛЕНИЙ В ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ С НЕСБАЛАНСИРОВАННЫМИ ДАННЫМИ

APPLICATION OF THE MULTIDIMENSIONAL KERNEL DENSITY ESTIMATION METHOD IN MACHINE LEARNING TASKS WITH IMBALANCED DATA

V. Popukaylo
A. Shmelyova

Summary. The article addresses the problem of using imbalanced data in multi-class classification tasks. It briefly examines the main existing approaches and proposes the application of the multidimensional kernel density estimation method to balance classes. The algorithm for applying this method is described, and an experiment is conducted using synthetic data. The results are compared with existing algorithms such as random oversampling of the small class, ADASYN, SMOTE, ASMO, SVMSMOTE. The article shows the possibility of using the multidimensional kernel density estimation method in principle to improve the quality of machine learning algorithms in conditions of imbalanced data.

Keywords: machine Learning, classification task, tabular data processing, imbalanced data, multidimensional kernel density estimation method.

Попукайло Владимир Сергеевич
Кандидат технических наук, доцент, Приднестровский государственный университет им. Т.Г. Шевченко
vsp.science@gmail.com
Шмелёва Анастасия Владимировна
Аспирант, Приднестровский государственный университет им. Т.Г. Шевченко
avshmlva@gmail.com

Аннотация. В статье описана проблема использования несбалансированных данных при решении задач многоклассовой классификации, кратко рассмотрены основные существующие подходы, предложено применение многомерного метода точечных распределений для балансировки классов, описан алгоритм применения данного метода, проведён эксперимент на синтетических данных, представлены результаты сравнения с существующими алгоритмами, такими как: случайное увеличение числа наблюдений малого класса, ADASYN, SMOTE, ASMO, SVMSMOTE, показана принципиальная возможность использования многомерного метода точечных распределений для решения задачи улучшения качества алгоритмов машинного обучения в условиях несбалансированных данных.

Ключевые слова: машинное обучение, задача классификации, обработка табличных данных, несбалансированные данные, многомерный метод точечных распределения.

Введение

Классификация — это метод контролируемого машинного обучения, позволяющий прогнозировать распределение данных по заранее определенному и четкому количеству классов. В реальном мире большинство этих наборов данных несбалансированы, классическими примерами таких задач могут быть задачи: оттока клиентов компаний, фрода при финансовых операциях или выявления редких заболеваний. Если один из классов содержит значительно меньше наблюдений, чем другие классы, этот класс называется миноритарным, а этот набор данных называется несбалансированным набором данных. Свойство несбалансированности набора данных сильно повлияло на эффективность традиционных методов классификации, и классификаторы стали смещаться в сторону мажоритарного классах [1]. Для классификации несбалансированного набора данных исследователями предложены различные методы машинного обучения. В этой статье предпринята попытка применить многомерный метод точечных распределения для увеличения числа примеров миноритарного класса, что может привести к улучшению качества работы алгоритмов машинного обучения.

Обзор состояния проблемы

Рассмотрим несколько основных подходов для решения проблемы дисбаланса классов на уровне данных [2]:

1. Случайное увеличение числа наблюдений малого класса (случайный оверсэмплинг).
2. Случайное уменьшение числа наблюдений преобладающего класса (случайный андерсэмплинг).
3. Информативное увеличение числа наблюдений малого класса (при котором не создаются новые объекты, а выбор наблюдений для ресэмплинга является целенаправленным, а не случайным).
4. Информативное уменьшение числа наблюдений преобладающего класса (выбор наблюдений для удаления является целенаправленным).
5. Увеличение числа наблюдений малого класса путем генерации новых синтетических данных.
6. Комбинации вышеуказанных техник.

Ресэмплинг является часто используемым методом для решения проблемы несбалансированных классов. При использовании данного подхода необходимо решить, как определить оптимальное распределение классов для данного набора данных, а также как провести

эффективный ресэмплинг обучающих данных. Самые простые методы сэмплирования — это удаление объектов мажоритарного класса или дублирование примеров миноритарного класса. В зависимости от того, какое соотношение классов необходимо, выбирается количество случайных записей для данной операции. Случайная выборка проста, но недостаточна во многих случаях. Если проблема несбалансированности классов набора данных представлена внутриклассовыми различиями, случайный оверсэмплинг может привести к повторному дублированию наблюдений в некоторых частях и меньшему количеству в других, в тоже время случайный андерсэмплинг может ухудшить вариативность данных, что также может сказаться на качестве работы некоторых классификаторов. Более предпочтительным процессом ресэмплинга будет определение подвыборок, составляющих класс, а затем увеличение числа наблюдений каждой концепции в отдельности для сбалансированного общего распределения. Информативное сэмплирование с целью сделать выборочные наблюдения более представительными повышает стоимость анализа данных, так как добавляет необходимость определения критерия выбора наблюдений. Например, если наблюдения измеряются с помощью некоторых измерений расстояния, те наблюдения мажоритарного класса, которые находятся относительно далеко от наблюдений миноритарного класса, могут лучше представлять признаки большинства, в то время как те, которые находятся относительно близко к наблюдениям меньшинства, могут быть важными для принятия решения о границе класса некоторыми алгоритмами обучения классификатора. Существуют различные техники для информационного сэмплирования, каждая из которых может быть эффективной при применении в определенном контексте. Для увеличения числа наблюдений малого класса путем генерации новых синтетических данных используются различные алгоритмы, такие как:

— SMOTE (англ. Synthetic Minority Oversampling Technique). Этот алгоритм основан на идее генерации некоторого количества искусственных наблюдений, которые были бы похожи на имеющиеся в миноритарном классе, но при этом не дублировали их. Для создания новой записи используя алгоритм ближайшего соседа KNN. Алгоритм *SMOTE* позволяет задавать количество записей, которое необходимо искусственно сгенерировать. Степень сходства примеров можно регулировать путем изменения числа ближайших соседей *k*.
— ASMO (англ. Adaptive Synthetic Minority Oversampling). Данный алгоритм является модификацией SMOTE, который использует предварительную кластеризацию (например алгоритмом k-means) для улучшения качества сэмплирования если миноритарные наблюдения равномерно распределены среди мажоритарных и имеют низкую плотность.
— Алгоритм Метрополис-Гастингса. Алгоритм позволяет сэмплировать любую функцию распределения. Он основан на создании цепи Маркова, то есть на каждом шаге алгоритма новое выбранное значение зависит только от предыдущего.

Многомерный метод точечных распределений

В работе [3] был предложен метод точечных распределений, который позволяет обработать выборку малого объема и получить так называемую виртуальную или эквивалентную выборку большого объема. В развитие этого метода был предложен многомерный метод точечных распределений, который может быть использован для улучшения качества моделирования данных малого объёма [4]. В работе [5] также было показано, что применение данного подхода позволяет сохранить знание о виде закона распределения случайной величины и о величине линейной коррелияционной связи между исследуемыми факторами.

Рассмотрим возможности применения метода точечных распределений в задаче увеличения числа наблюдений малого класса путем генерации новых синтетических данных. Рассмотрим пример, предложенный в Python библиотеке для решения проблемы несбалансированных данных [6] для сравнения различных алгоритмов оверсэмплинга. Для анализа воспользуемся библиотекой sklearn [7], которая позволяет создать кластеры точек, нормально распределенных вокруг вершин n-информативно-мерного гиперкуба со сторонами длиной 2*L и сгенерируем набор данных со следующими параметрами:

— Количество наблюдений: 1000.
— Количество классов: 3.
— Веса классов: 0.01, 0.01, 0.98.
— Размер гиперкуба (L): 0.8.
— Количество кластеров: 1.

В качестве алгоритма классификации будем использовать логистическую регрессию. На рисунке 1 визуализированы исходные данные, а также показана функция принятия решения для алгоритма.

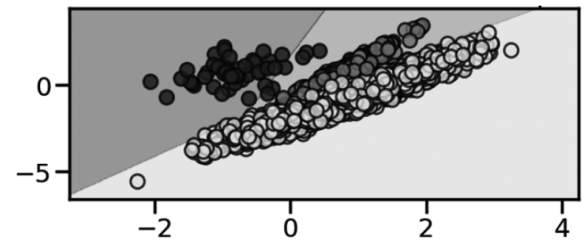

Рис. 1. Логистическая регрессия на исходных данных

На рисунке 2 продемонстрируем как на вид функции принятия решения влияют различные алгоритмы оверсэмплинга, такие как: случайное увеличение чис-

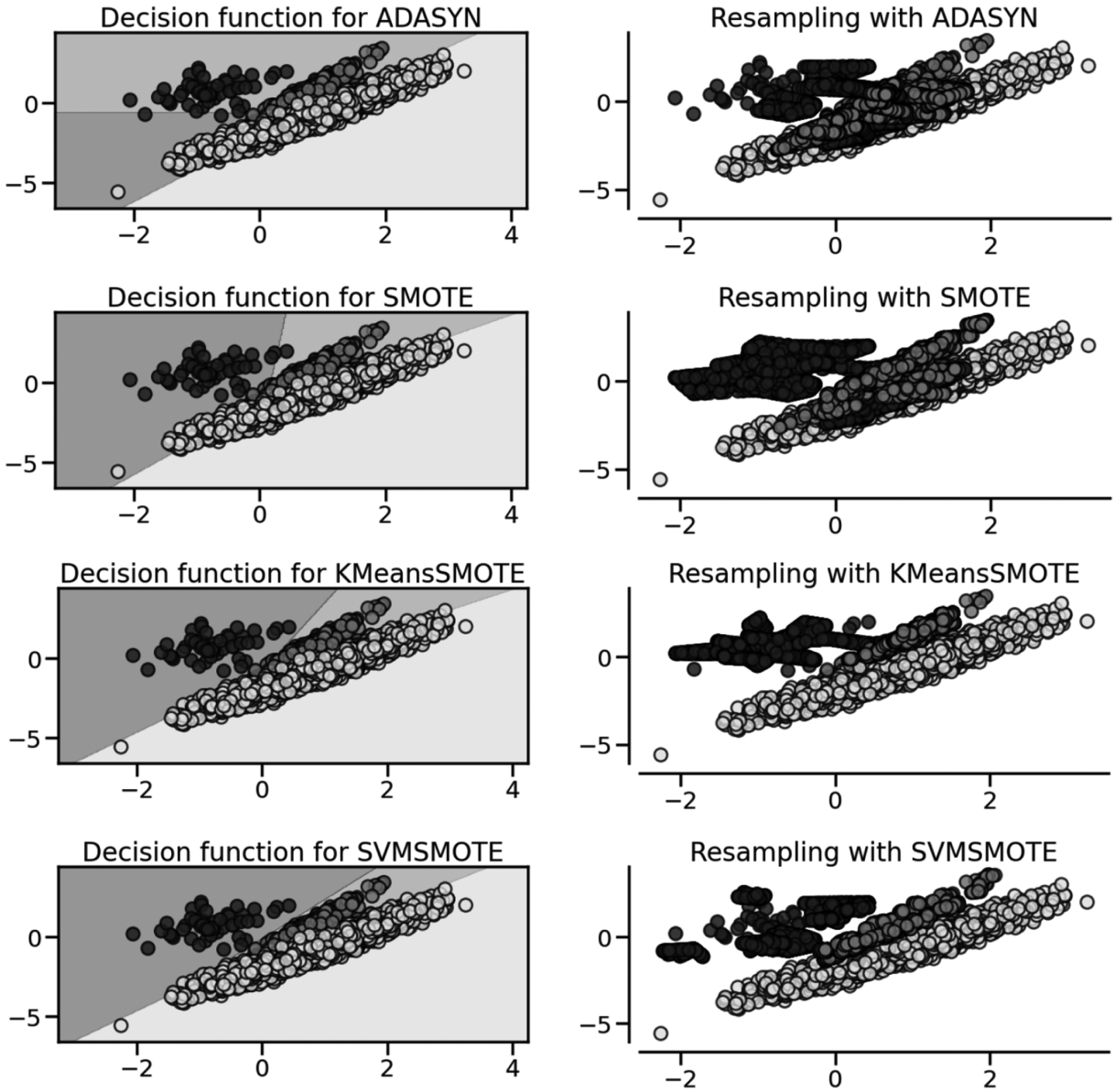

Рис. 2. Результаты работы различных алгоритмов сэмплирования

ла наблюдений малого класса, ADASYN, SMOTE, ASMO (K-meansSMOTE), SVMSMOTE [8].

Проведём оверэмплинг на основе многомерного метода точечных распределений, с параметрами по умолчанию (n=30, нормальный закон распределения) [9] для этого необходимо:

1. Выделить наборы данных, относящиеся к каждому из миноритарных классов.
2. Для каждого из миноритарных классов, с помощью метода точечных распределений для всех *Xi* построить таблицы расчёта ненормированных плотностей вероятности в виртуальной области, для каждой строки исходных экспериментальных данных построить таблицы данных методом точечных распределений, в которые вносить одновременно величины двух столбцов *Xij* из соответствующей таблицы ненормированных плотностей вероятностей и столбца *Xif*. Выравнивание (состыковка) столбцов *Xij* и *Xif* должно происходить по уровню максимальной ненормированной плотности вероятности.
3. Объединить синтетически полученные наборы данных с мажоритарным классом.

Следующим шагом проведём классификацию логистической регрессией и визуализируем результаты на рисунке 3.

Анализ результатов

На представленных выше изображениях видно, что применение многомерного метода точечных распре-

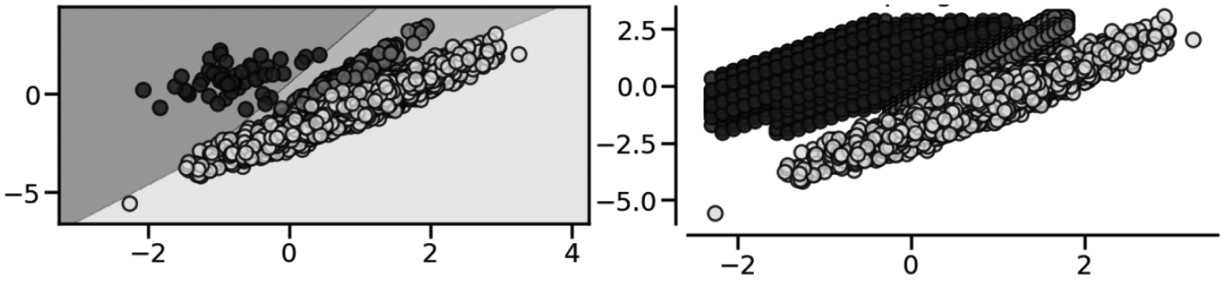

Рис. 3. Результаты применения многомерного метода точечных распределений для сэмплирования данных

делений позволило улучшить функцию принятия решений за счёт генерации синтетических данных, похожих по своей структуре на исходные объекты. При этом, так как использовалась реализация алгоритма по умолчанию, было сгенерировано избыточное количество данных для решения этой задачи.

В таблице 1 показаны метрики качества классификации для приведённых выше алгоритмов:

Таблица 1.

Метрики качества классификации

Алгоритм	Weighted F-1 score	Geometric mean score	Index balanced accuracy
Без сэмплирования	0.99	0.91	0.67
Случайное сэмплирование	0.94	0.92	0.85
ADASYN	0.81	0.82	0.67
SMOTE	0.96	0.92	0.85
KmeansSMOTE	0.95	0.93	0.87
SVMSMOTE	0.97	0.94	0.90
Многомерный метод точечных распределений	0.97	0.94	0.89

Анализ полученных метрик позволяет сделать выводы о возможном применении многомерного метода точечных распределений для улучшения качества алгоритмов машинного обучения в условиях несбалансированных данных. За рамками данного исследования остаётся вопрос о границах применения данного метода на данных различной природы, а также с различными методами построения математических моделей.

Выводы

В данной статье была показана принципиальная возможность использования многомерного метода точечных распределений для решения задачи улучшения качества алгоритмов машинного обучения в условиях несбалансированных данных. В связи с этим дальнейшим направлением исследований может быть автоматический подбор параметров метода точечных распределений для обеспечения наилучшего качества на различных наборах данных, а также определение типов задач, в которых данный подход может быть использован наиболее эффективно для улучшения качества классификации.

ЛИТЕРАТУРА

1. Kumar, A., Goel, S., Sinha, N., Bhardwaj, A. A Review on Unbalanced Data Classification. //Proceedings of International Joint Conference on Advances in Computational Intelligence. Algorithms for Intelligent Systems. Springer, Singapore. — 2022.
2. Yanminsun, & Wong, Andrew & Kamel, Mohamed S. Classification of imbalanced data: a review. //International Journal of Pattern Recognition and Artificial Intelligence. — 2009. — Vol. 23, №4. — pp. 687–719
3. Столяренко Ю.А. Метод точечных распределений //Радіоелектронні і комп'ютерні системи. — 2012. — №. 6. — С. 75–77
4. Popukaylo V. Small size sample mathematical modeling. //Meridian Ingineresc. — 2015. — № 4. — pp. 25–30.
5. Попукайло В.С. Исследование линейной корреляционной связи в многомерном методе точечных распределений //Информационно-управляющие системы. — 2016. — №6. — с. 96–98
6. Lemaître, G. Nogueira, F. Aridas, Ch.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning //Journal of Machine Learning Research. — 2017. — vol. 18, №17. — pp. 1–5.
7. Pedregosa et al., Scikit-learn: Machine Learning in Python //Journal of Machine Learning Research. — 2011. — №12. — pp. 2825–2830.
8. Compare over-sampling samples. [Электронный источник] // URL: https://imbalanced-learn.org/stable/auto_examples/over-sampling/plot_comparison_over_sampling.html
9. Попукайло В.С., Столяренко Ю.А. Программная реализация метода точечных распределений //Информационно-телекоммуникационные системы и технологии. — 2015. — С. 260–260.

© Попукайло Владимир Сергеевич (vsp.science@gmail.com); Шмелёва Анастасия Владимировна (avshmlva@gmail.com)
Журнал «Современная наука: актуальные проблемы теории и практики»