Федеральное государственное бюджетное образовательное учреждение высшего образования

Московский авиационный институт
(национальный исследовательский университет)

Кафедра 804 «Теория вероятностей и компьютерное моделирование»

О т ч е т
по научно-исследовательской работе

направление подготовки 01.04.04 «Прикладная математика»

Студента (ки) Осокиной Анастасии Сергеевны

Курс 1 группа 8О-104М-19

Научный руководитель Платонов Е.Н.

_____

Дата оценка подпись

Москва 2019

ЗАДАНИЕ

Тема: Решение задачи классификации для несбалансированных данных

Руководитель практики от института _____

ФИО, подпись

ОТЧЕТ

Решение задачи классификации для несбалансированных данных.

Вероятностная постановка задачи обучения

В задачах обучения по прецедентам элементы множества X — это доступные данные о реальных объектах них. Данные могут быть неточными, поскольку измерения значений признаков f_j (x) и целевой зависимости $y^*(x)$ обычно выполняются с погрешностями. Данные могут быть неполными, поскольку измеряются не все мыслимые признаки, а лишь физически доступные для измерения. В результате одному и тому же описанию x могут соответствовать различные объекты и различные ответы. В таком случае $y^*(x)$, строго говоря, не является функцией. Устранить эту некорректность позволяет вероятностная постановка задачи.
Вместо существования неизвестной целевой зависимости $y^*(x)$ предположим существование неизвестного вероятностного распределения на множестве $X \times Y$ с плотностью $p(x, y)$, из которого случайно и независимо выбираются ℓ наблюдений $X^l = (xi, yi)_{i=1}^{l}$. Такие выборки называются простыми или случайными одинаково распределёнными.

Вероятностная постановка задачи считается более общей, так как функциональную зависимость $y^*(x)$ можно представить в виде вероятностного распределениях (x, y) = $p(x)p(y|x)$, положив $p(y|x)$ = δ (y – $y^*(x)$), где δ(z) — дельта-функция.

Предметом исследования являются методы классификации несбалансированных выборок.

Достаточно распространенным явлением является ситуация, когда в выборке экземпляров одного класса значительно больше (мажоритарный класс) чем экземпляров другого класса (миноритарный класс).

В таких условиях большинство методов машинного обучения приводят к получению моделей, которые неправильно определяют редкие экземпляры миноритарного класса из-за подавления экземплярами мажоритарного класса экземпляров миноритарного класса при обучении модели.

Например, для решения задачи бинарной классификации, где класса всего два, несбалансированной выборкой считается ситуация, где объектов одного класса $\leq 10\%$ от общего числа объектов выборки.

Однако именно миноритарный класс может иметь первостепенную важность в таких прикладных задачах, как медицинская диагностика, кредитный скоринг, выявление мошенничества с кредитными картами, защита компьютерных сетей.

Пример задачи предсказания и классификации:

В задачах медицинской диагностики в роли объектов выступают пациенты. Признаки характеризуют результаты обследований, симптомы заболевания

и применявшиеся методы лечения. Примеры бинарных признаков — пол, наличие головной боли, слабости, тошноты, и т. д. Порядковый признак — тяжесть состояния (удовлетворительное, средней тяжести, тяжёлое, крайне тяжёлое). Количественные признаки — возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т. д. Признаковое описание пациента является, по сути дела, формализованной историей болезни. Накопив достаточное количество прецедентов, можно решать различные задачи: классифицировать вид заболевания (дифференциальная диагностика);

определять наиболее целесообразный способ лечения; предсказывать длительность и исход заболевания; оценивать риск осложнений; находить синдромы — наиболее характерные для данного заболевания совокупности симптомов. Ценность такого рода систем в том, что они способны мгновенно анализировать и обобщать огромное количество прецедентов — возможность, недоступная человеку.

Существуют два подхода к работе с несбалансированными выборками:

-undersampling (удаляют экземпляры мажоритарного класса)

- oversampling (добавляют (синтезируют) экземпляры миноритарного класса).

Принципом undersampling является удаление примеров объектов из больших классов.

Этапы стратегии undersampling:

- множество экземпляров мажоритарного класса разбивается на число кластеров, равное числу экземпляров миноритарного класса

- выбирается по одному экземпляру из каждого кластера

- удаляются все остальные экземпляры мажоритарного класса

Случайное удаление экземпляров мажоритарного класса (random undersampling) – наиболее простая стратегия, в которой случайным образом удаляются экземпляры мажоритарного класса для достижения необходимого соотношения классов. Уровень соотношения классов подбирается эмпирическим путем. Достоинствами стратегии являются высокая скорость работы, уменьшение размера выборки и простота реализации, а недостатками – высокая вероятность потери значимых данных.

Дублирование экземпляров миноритарного класса (oversampling) – это стратегия, в которой для достижения необходимого соотношения классов, дублируются экземпляры миноритарного класса. Достоинствами стратегии являются высокая скорость работы и простота реализации, а недостатками – возможность переобучения модели и увеличение размера выборки.

На практике стратегия undersampling работает более эффективно, чем стратегия oversampling. Это связано с тем, что стратегия oversampling увеличивает размер выборки, что может повысить вероятность переобучения и время работы классификатора. Однако, при применении стратегии undersampling существует вероятность потери важной информации.

Стратегия искусственного увеличения экземпляров миноритарного класса SMOTE– одна из популярных стратегий сэмплинга, которая базируется на технологии oversampling. Данная стратегия предполагает синтез искусственных экземпляров путем создания одного или нескольких ближайших соседей для экземпляров миноритарного класса, в зависимости от необходимого соотношения классов. Достоинствами стратегии являются высокая скорость работы и простота, а недостатками – возможное переобучения построенной модели, увеличение размера формируемой выборки.

Существует ряд реализаций алгоритма SMOTE, например:

- В Python, посмотрите на «UnbalancedDatasetМодуль. Он предоставляет ряд реализаций SMOTE, а также различные другие методы повторной выборки, которые вы можете попробовать.
- В R Пакет DMwR обеспечивает реализацию SMOTE.

Библиотека Imbalanced Classification в языке R объединяет большую коллекцию решений к бинарной несбалансированной задаче классификации. А

также библиотека обеспечивает реализацию новых алгоритмов, которые не являются доступными в других пакетах языка R.

Пакет ROSE обеспечивает как стандартные, так и более совершенные инструменты для улучшения задачи бинарной классификации с несбалансированными данными.

Пакет включает функцию ROSE, которая генерирует синтетические сбалансированные образцы и, таким образом, позволяет усилить последующую оценку любого двоичного классификатора.

Библиотека ROSE помогает выполнить задачу двоичной классификации при наличии редких классов. Он создает искусственную сбалансированную выборку данных, смоделированных в соответствии с подходом сглаженной начальной загрузки.

Применение ROSE:

$ROSE(formula, data, N, p, hmult.majo, hmult.mino, subset, na.actio, seed)$

Название параметра	Описание	Обязательность заполнения
formula	Объект класса. Левая сторона должна быть вектором, определяющим метки класса. Правая часть должна быть последовательностью векторов с предикторами.	O

data	Фрейм данных, в котором предпочтительно интерпретировать «формулу».	H
N	Размер выборки результирующего набора данных, сгенерированного ROSE.	H
p	Вероятность примеров миноритарных классов в результирующем наборе данных, сгенерированном ROSE.	O
hmult.majo	Коэффициент сжатия для умножения на параметры сглаживания для оценки условной плотности ядра в мажоритарном классе.	H
hmult.mino	Коэффициент сжатия, который необходимо умножить на	H

	параметры сглаживания для оценки условной плотности ядра в миноритарном классе.	
subset	Вектор, определяющий подмножество наблюдений	H
na.actio	Функция, которая указывает, что должно происходить, когда данные содержат «NA»	H
seed	Значение для отслеживания сгенерированного образца	O

Дата подпись