УДК 004.9

ОСТРОВНОЙ ГЕНЕТИЧЕСКИЙ АЛГОРИТМ В ЗАДАЧЕ КЛАССИФИКАЦИИ ДАННЫХ С УЧЕТОМ ИЗДЕРЖЕК

Демидова Л.А., Шыхыев А.А., Шаршатов М.А.

МИРЭА - Российский технологический университет», 119454, Россия, г. Москва, проспект Вернадского, 78, e-mail: alishykhyev@gmail.com, sharshatov99@mail.ru

В современном мире обработка и анализ больших объемов данных становятся все более важными для успешного функционирования информационных систем. Однако эффективная классификация данных может быть затруднена из-за различных издержек, таких как временные, вычислительные и обусловленные спецификой конкретной предметной области. В данной статье рассматриваются подходы к классификации данных с учетом этих издержек в информационных системах. Представляется обзор существующих методов учета издержек в классификации, а также описывается применение алгоритма классификации данных с учетом издержек в различных отраслях. Применяется островной генетический алгоритм для оптимизации параметров и подбора весов модели классификатора.

Ключевые слова: классификация, ошибки классификации, матрица издержек, сферы применения, учет издержек, информационная система, cost-sensitive, генетический алгоритм, островная модель.

ISLAND GENETIC ALGORITHM IN THE PROBLEM OF COST-SENSITIVE DATA CLASSIFICATION

Demidova L.A., Shykhyev A.A., Sharshatov M.A.

MIREA - Russian Technological University", 119454, Moscow, 78 Vernadskogo Avenue, Russia, e-mail: alishykhyev@gmail.com, sharshatov99@mail.ru

In the modern world, the processing and analysis of large volumes of data are becoming increasingly important for the successful functioning of information systems. However, effective data classification can be difficult due to various costs, such as time, computational, and those specific to a particular domain. This article considers approaches to data classification taking into account these costs in information systems. An overview of existing methods for cost-sensitive classification is presented, as well as the application of a data classification algorithm that takes into account costs in various industries. An island genetic algorithm is used to optimize the parameters and select the weights of the classifier model.

Keywords: classification, classification errors, cost matrix, application areas, cost consideration, information system, cost-sensitive, genetic algorithm, island model.

Введение

Одним из основных фундаментальных подходов в анализе данных является классификация данных, который позволяет автоматически разделять данные на

различные категории и группы в соответствии с заданными критериями [1]. Этот процесс может быть очень полезен в различных областях, включая медицину, бизнес, финансы, науку и технологии. В настоящее время объем данных, с которыми приходится работать, растет в геометрической прогрессии, что делает классификацию данных все более сложной задачей. Современные информационные системы обрабатывают огромные объемы структурированных и неструктурированных данных, которые требуют мощных алгоритмов и методов машинного обучения для обнаружения скрытых закономерностей и тенденций.

Развивающаяся отрасль машинного обучения, основанная на искусственных нейронных сетях, глубоком обучении и других техниках, предоставляет нам мощные инструменты для решения задач классификации данных. Однако, с ростом количества данных и сложности моделей, возникает необходимость в разработке более эффективных и точных методов классификации данных.

В этом контексте классификация с учетом издержек становится все более важной, так как она позволяет учитывать экономические или социальные издержки ошибок классификации и минимизировать общую стоимость ошибок, а не просто число ошибок. Разработка и применение эффективных методов классификации с учетом издержек в информационных системах имеет большое значение для успешного анализа и принятия решений в различных областях, и поэтому она является важной темой для исследования и разработки.

Применение популяционных алгоритмов в задачах классификации с учетом издержек может помочь найти баланс между точностью классификации и учетом различных видов ошибок. Эти алгоритмы могут эффективно исследовать пространство параметров модели и находить оптимальные настройки, которые минимизируют общую стоимость ошибок классификации, а не только количество ошибок

Классические методы классификации данных

На сегодняшний день существуют множество классических методов классификации данных, которые могут быть применены в зависимости от типа данных, задачи классификации и доступных ресурсов. Рассмотрим некоторые из них:

1. Линейная классификация — это один из самых простых методов классификации данных, который основан на использовании линейной разделяющей гиперплоскости для разбиения данных на две или более категории [2]. Этот метод широко применяется в областях, таких как анализ текста, распознавание образов, финансовый анализ и других.
2. Метод k-ближайших соседей — это метод классификации данных, который основан на нахождении k ближайших соседей к объекту, который нужно классифицировать, и присвоении этому объекту того же класса, что и класс большинства его ближайших соседей [3]. Этот метод часто используется в распознавании образов, рекомендательных системах и других приложениях.
3. Решающие деревья — это метод классификации данных, который описывает данные в виде дерева решений, где каждый узел представляет признак, а каждая ветвь — его значения [2]. Этот метод может быть использован для решения различных задач классификации, таких как распознавание образов, выявление мошенничества и других.
4. Наивный Байесовский классификатор — это метод классификации данных, который основан на вероятностных моделях, и который использует теорему Байеса для вычисления вероятности принадлежности объекта к тому или иному классу [2]. Этот метод часто используется в анализе текста, выявлении спама и других приложениях.

В целом, классические методы классификации данных помогают автоматически разделять данные на различные категории групп и в соответствии с заданными

критериями. Они широко применяются в различных областях, таких как бизнес, медицина, наука, технологии и многие другие.

Несмотря на то, что классические методы классификации данных являются важными инструментами анализа данных, они также имеют некоторые ограничения и недостатки такие как:

1. Недостаточная точность — классические методы классификации данных могут не обеспечивать достаточной точности при решении сложных задач классификации. Это связано с ограниченным числом признаков, которые используются для классификации, и неспособностью учитывать скрытые зависимости и взаимодействия между признаками.
2. Неумение обрабатывать большие объёмы данных — классические методы классификации данных могут столкнуться с проблемами обработки больших объемов данных, особенно если данные имеют сложную структуру или высокую размерность. Это может приводить к переобучению или недообучению моделей классификации и, как следствие, к низкой точности.
3. Неумение учитывать экономические или социальные издержки ошибок классификации — классические методы классификации данных не всегда учитывают экономические или социальные издержки ошибок классификации. Например, ошибка в диагностике может иметь серьезные последствия для здоровья пациента, и поэтому стоимость ложно отрицательного результата может быть гораздо выше, чем стоимость ложно положительного результата.
4. Сложность интерпретации — классические методы классификации данных могут быть сложными для интерпретации и объяснения результатов классификации. Например, решающие деревья и нейронные сети могут быть очень сложными моделями, которые не всегда легко интерпретировать.
5. Неумение работать с несбалансированными данными — классические методы классификации данных могут иметь проблемы при работе с несбалансированными данными, когда один класс представлен значительно меньшим количеством объектов, чем другой. Это может привести к низкой точности классификации для меньшего класса.

Современные методы классификации данных, такие как глубокое обучение и классификация с учетом издержек (cost-sensitive), разрабатываются с целью преодоления этих ограничений и улучшения точности и эффективности классификации.

Глубокое обучение, основанное на использовании искусственных нейронных сетей [4], может обеспечивать более высокую точность классификации и лучшую способность учитывать скрытые зависимости между признаками. Однако, такие модели могут быть очень сложными и требовательными к ресурсам, что может создавать проблемы в реальных приложениях.

Классификация данных с учетом издержек

Классификация с учетом издержек (cost-sensitive), в свою очередь, позволяет учитывать экономические и социальные издержки ошибок классификации, что делает её более привлекательное для решения задач, где ошибки могут иметь серьезные последствия.

Такая классификация в основном применяется в задачах, где ошибки одного типа могут иметь большие экономические или социальные издержки по сравнению с ошибками другого типа.

Однако, этим методы также могут сталкивать с проблемами несбалансированных данных и сложным выбором оптимальной функции стоимости.

Существуют несколько методов классификации данных с учетом издержек, которые могут быть применены в зависимости от типа данных и доступных ресурсов. Одним из основных методов является использование матрицы издержек, которая определяет стоимость ошибок в классификации каждого типа. Матрица издержек позволяет учитывать относительную важность различных типов ошибок и оптимизировать классификацию для минимизации суммарных издержек. Другой метод классификации данных с учетом издержек — это использование взвешенной логистической регрессии. В этом методе веса классов настраиваются таким образом, чтобы учитывать относительную важность каждого класса и минимизировать суммарные издержки. Это достигается путем умножения веса каждого класса на соответствующее значение функции потерь при обучении модели.

Матрица издержек определяет стоимость ошибок классификации каждого типа. Матрица издержек имеет размерность N x N, где N — число классов. Элементы матрицы издержек определяют стоимость ошибок классификации каждого типа. Обычно матрица издержек заполняется значениями 0 и 1, где 0 означает отсутствие издержек, а 1 — наличие издержек [5].

Например, в задаче медицинской диагностики [6], где одна из категорий является диагнозом болезни, а другая отсутствием болезни, матрица издержек может иметь вид, приведенный в таблице 1.

Таблица 1. Пример матрицы издержек

	Диагноз болезни	Нет диагноза болезни
Диагноз болезни	0 (TN)	10 (FP)
Нет диагноза болезни	100 (FN)	0 (TP)

TN, FN, FP и TP — это обозначения для четырех возможных комбинаций результатов классификации и фактического класса объекта. Каждая из этих комбинаций характеризует результат классификации для конкретного класса.

Опишем каждое обозначение.

1. TN (True Negative) — число верно классифицированных объектов отрицательного класса.
2. FN (False Negative) — число неверно классифицированных объектов отрицательного класса.
3. FP (False Positive) — число неверно классифицированных объектов положительного класса.
4. TP (True Positive) — число верно классифицированных объектов положительного класса.

Если в задаче бинарной классификации отрицательный класс обозначает «здоровый» и положительный класс обозначает «больной», то TN — означает количество здоровых, которые правильно были отнесены к здоровым, FN — количество здоровых, которые были ошибочно отнесены к больным, FP — количество больных, которые были

ошибочно отнесены к здоровым, а TP — количество больных, которые были правильно отнесены к больным.

Обозначения TN, FN, FP и TP используются для вычисления различных метрик классификации, таких как точность (accuracy), полнота (recall), точность (precision) и F_1-мера, которые используются для оценки качества работы алгоритмов классификации [7].

В таблице 1 представлен пример матрицы издержек в классификаторе, в котором стоимость ошибки, заключающейся в неверной диагностике факта болезни (false positive, FP), равна 10, а стоимость ошибки, заключающейся в неверной диагностике отсутствия болезни (false negative, FN), равна 100.

На основе построенной матрицы издержек можно определить общую стоимость ошибок классификации для конкретной модели. Эта стоимость может быть оптимизирована путем настройки параметров модели или выбора наилучшего алгоритма классификации.

Определение матрицы издержек — это важный шаг при использовании метода классификации с учетом издержек. Чтобы определить матрицу издержек, необходимо провести анализ задачи и определить стоимость различных типов ошибок классификации. Способ определения матрицы издержек зависит от конкретной задачи и может быть выполнен путем проведения экспертных оценок или анализа реальных данных.

Для проведения экспертной оценки матрицы издержек можно обратиться к специалистам в соответствующей области или провести опрос экспертов, чтобы определить стоимость ошибок классификации.

Важно отметить, что матрица издержек может меняться в зависимости от целей задачи и приоритетов. Поэтому необходимо проводить регулярный анализ и обновление матрицы издержек для оптимальной настройки модели классификации.

На рис. 1 представлена общая схема алгоритма классификации с учетом издержек.

В представленной схеме блоки представляют различные этапы работы алгоритма, а стрелки показывают направление потока данных и действий между блоками. Рассмотрим подробнее каждый шаг алгоритма.

1. Определение задачи и метрик оценки качества: определяется целевая переменная и выбираются метрики, используемые для оценки производительности модели классификации.
2. Анализ данных: производится анализ данных, включающий в себя их очистку, преобразование и визуализацию.
3. Определение матрицы издержек: проводится анализ задачи и определяются стоимость различных типов ошибок классификации, чтобы определить матрицу издержек. Этот шаг позволяет оптимизировать принимаемые решения, учитывая стоимость ошибок.
4. Разделение выборки на обучающую и тестовую: выборка данных разбивается на две части — обучающую и тестовую. Обучающая выборка используется для оценки качества работы модели.
5. Обучение модели классификации на обучающей выборке: модель классификации обучается, используя обучающую выборку данных и матрицу издержек. Можно использовать различные методы классификации, например, такие, как логистическая регрессия, машина опорных векторов (SVM), наивный байесовский классификатор

(Naive Bayes), случайный лес (Random Forest) и т.д.

6. Оценка качества модели классификации на тестовой выборке: на тестовой выборке проводится оценка качества модели с использованием метрик, учитывающих матрицу издержек. Например, могут быть использованы такие метрики, как взвешенная точность (weighted precision), взвешенная полнота (weighted recall), взвешенная F_1-мера (weighted F_1-score) и т.д. Цель этого шага — оценить, насколько точно модель классификации способна классифицировать новые данные.

Рис.1 - Общая схема алгоритма классификации с учетом издержек.

7. Настройка параметров модели классификации для улучшения качества: на основе результатов оценки качества модели могут быть проведены дополнительные настройки параметров, чтобы улучшить качество классификации.

8. Проверка работы модели на новых данных: после проведения оценки качества модели и ее настройки, она может быть использована для классификации новых данных. Результаты работы модели могут быть оценены по метрикам качества, учитывающим матрицу издержек. Этот шаг может быть повторен периодически, чтобы обеспечить работу модели классификации на высоком уровне и соответствует бизнес-требованиям.

Каждый шаг алгоритма играет важную роль в обеспечении качественной классификации данных с учетом матрицы издержек. Описанный алгоритм помогает улучшить качество классификации и принимаемых решений. Он может быть применен в различных областях, где точность классификации имеет большое значение.

Сферы применения классификации данных с учетом издержек (cost-sensitive)

Сфера применения классификации с учетом издержек (cost-sensitive) имеет широкий спектр применений, где стоимость разных типов ошибок может быть очень разной. Перечислим основные сферы применения.

1. Медицинские исследования: в медицинских исследованиях классификация с учетом издержек может быть использована для правильного определения наличия определенного заболевания у пациента [8]. В этом случае стоимость ложноотрицательных (FN) ошибок может быть очень высокой, поскольку это может привести к пропуску лечения и ухудшению состояния пациента. С другой стороны, ложноположительные (FP) ошибки могут привести к назначению лишнего лечения и лишним расходам. В таком случае классификация с учетом издержек может помочь выбрать модель классификации, которая минимизирует общую стоимость ошибок.

2. Финансовые системы: в финансовых системах классификация с учетом издержек может быть использована для предсказания рисков и определения подходящих инвестиционных стратегий. Например, если стоимость ложноположительных ошибок высока, то модель должна быть настроена на уменьшение таких ошибок, даже если это приведет к увеличению количества ложноотрицательных ошибок.

3. Обработка естественного языка: в задачах обработки естественного языка классификация с учетом издержек может быть использована для классификации текстов по различным категориям, например, по тональности или теме. В таких задачах матрица издержек может быть определена на основе стоимости ложноотрицательных или ложноположительных ошибок.

4. Системы безопасности: в системах безопасности классификация с учетом издержек может быть использована для определения уровня риска различных действий или событий. Например, в системе контроля доступа к зоне с высокой конфиденциальностью стоимость ложноположительных ошибок может быть очень высокой, поэтому модель классификации должна быть настроена на минимизацию таких ошибок.

Применение классификации с учетом издержек (cost-sensitive) и результаты в сравнении с классическим алгоритмом классификации

Рассмотрим еще один подход к обучению классификации с учетом издержек (cost-sensitive) и сравним с классическим алгоритмом на примере набора данных “Обнаружение мошенничества с кредитными картами”. Набор данных содержит информацию о транзакциях на кредитных картах и имеет сильный дисбаланс классов: только 0,17% транзакций являются мошенническими.

Сравним два метода классификации: обычную логистическую регрессию и

взвешенную логистическую регрессию (cost-sensitive). Обычная логистическая регрессия не учитывает дисбаланс классов в данных, что может привести к ошибкам при классификации мошеннических транзакций, которые являются меньшинством. В свою очередь, взвешенная логистическая регрессия (cost-sensitive) учитывает и позволяет более точно классифицировать мошеннические транзакции. Для взвешенной классификации мы присвоим классу 0 (не мошеннических транзакций, что составляет 227451 элементов) вес равным 1, а классу 1 (мошеннических транзакций, что составляет 394 элементов) – присвоим вес равным 10, увеличивая штраф за ошибку второго рода на этапе обучения (False Negative). Обучив обычную логистическую регрессию и взвешенную логистическую регрессию на обучающей выборке, оценим их производительность на тестовой выборке, используя различные метрики качества, включая точность, полноту, F_1-меру и Precision Recall-curve.

Значения метрик качества для классической логистической регрессии приведены в таблице 2.

Таблица 2. Значения метрик качества для классической логистической регрессии

	Precision	Recall	F_1-score
0	1.00	1.00	1.00
1	0.86	0.60	0.71
Confusion Matrix			
85294 (TN)		13 (FP)	
54 (FN)		82 (TP)	

Опишем подробнее значения таблицы:

1. Precision (точность) — показывает, насколько точно модель определяет положительный класс (1) из всех предсказанных положительных классов.
2. Recall (полнота) — показывает, какую долю объектов положительного класса модель определяет корректно из всех объектов этого класса.
3. F_1-score (F_1-мера) — является гармоническим средним между точностью и полнотой.
4. Accuracy (точность) — это метрика, которая показывает, какую долю объектов модель предсказывает правильно относительно всех объектов в выборке.
5. Confusion Matrix (матрица ошибок) — это таблица, которая показывает количество верно и неверно предсказанных объектов для каждого класса. В этом случае, матрица ошибок имеет размерность 2x2, где строки соответствуют реальным классам, а столбцы - предсказанным классам.

Значения метрик качества для взвешенной логистической регрессии приведены в таблице 3.

Таблица 3. Значения метрик качества для взвешенной логистической регрессии

	Precision	Recall	F_1-score
0	1.00	1.00	1.00
1	0.78	0.84	0.81
Confusion Matrix			
85274 (TN)		33 (FP)	
22 (FN)		114 (TP)	

Результаты показывают, что взвешенная логистическая регрессия с учетом издержек (cost-sensitive) и обычная логистическая регрессия имеют разные значения метрик качества. Взвешенная логистическая регрессия имеет гораздо более высокий показатель полноты (recall) для класса 1 (мошеннические операции), но ниже точность (precision) по сравнению с обычной логистической регрессией. Таким образом, взвешенная логистическая регрессия лучше обнаруживает мошеннические операции (больше истинно-положительных результатов), но также больше склонна к ложным срабатываниям (ложно-положительные результаты). В контексте задачи обнаружения мошеннических транзакций, использование обычной логистической регрессии может обеспечить более высокую общую точность в определении транзакций, но может не быть наилучшим выбором, если приложение должно эффективно выявлять мошеннические операции. Вместо этого, модель с более высокой полнотой и f1-мерой для класса мошеннических транзакций, даже если это сопряжено с более высокой ошибкой ложно-положительных предсказаний, может быть более эффективной в обнаружении мошеннических транзакций и более соответствовать бизнес-задаче.

В данном случае, на основе F_1-меры для класса 1 (мошеннические операции), можно сделать вывод, что взвешенная логистическая регрессия с учетом издержек (cost-sensitive) показывает лучше результаты (F_1-меры = 0.81) чем обычная логистическая регрессия (F_1-меры = 0.71). Метрика Recall (полнота) взвешенной логической регрессии имеет значительно более высокую полноту (Recall = 0.84), по сравнению с обычной логистической регрессией (Recall = 0.60). Исходя из этого, взвешенная логистическая регрессия лучше определяет мошеннические транзакции чем классическая логистическая регрессия.

Для визуальной оценки воспользуемся RR-curve кривой (Precision-Recall curve).

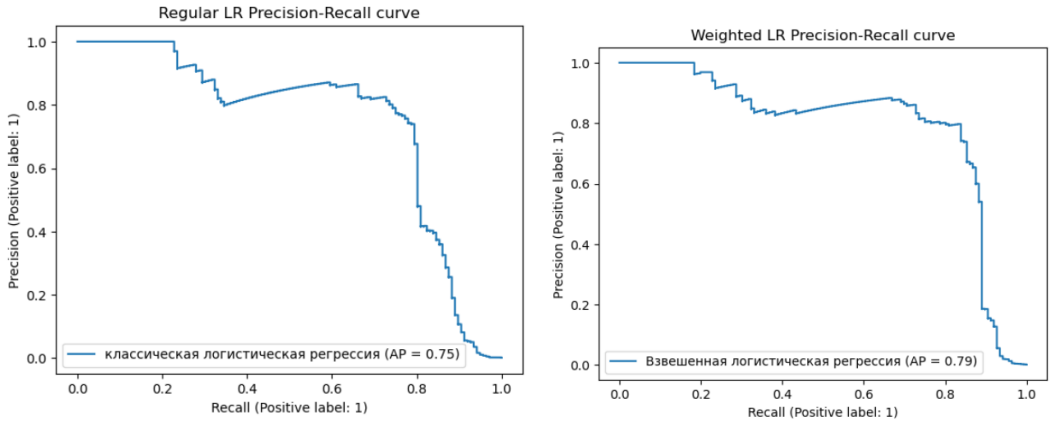

Рис.2. - PR-curve кривая логистической регрессии и взвешенной логистической регрессии

PR-кривая (Precision-Recall curve) является графическим представлением производительности бинарного классификатора на основе метрик точности (precision) и полноты (recall). На PR-кривой ось X представляет собой полноту (Recall), а ось Y – точность (Precision). Каждая точка на PR-кривой представляет собой различное значение порога (threshold) для классификации объектов, при котором точность и полнота различны. PR-кривая позволяет определить оптимальное значение порога,

обеспечивающее баланс между точностью и полнотой классификатора. Наглядная визуализация помогает увидеть, что взвешенная логистическая регрессия повышает факт обнаружения мошеннических транзакций, но при этом повышается ложно-положительная классификация, что является допустимым для обнаружения мошеннических транзакций. Average Precision (AP = 0.75) для классической логистической регрессии означает, что средняя точность на интервале полноты составляет 0.75, а Average Precision (AP = 0.79) для взвешенной логистической регрессии означает, что средняя точность на интервале полноты составляет 0.79. Это означает, что взвешенная логистическая регрессия имеет более высокую точность при выявлении объектов класса 1 (мошеннических транзакций) в сравнении с классической логистической регрессией.

Применение островного эволюционного алгоритма оптимизации в разработке классификатора с учетом издержек

Эволюционные алгоритмы оптимизации являются методами глобальной оптимизации, которые основаны на принципах естественного отбора и генетической мутации. Эти методы используются для поиска оптимальных решений в задачах оптимизации, когда требуется найти наилучшее решение среди множества возможных вариантов.

Основная идея эволюционных алгоритмов заключается в создании популяции потенциальных решений, которые затем эволюционируют в соответствии с определенными правилами, которые определяют, какие решения лучше подходят для решения задачи. В ходе эволюции создаются новые потенциальные решения, которые могут быть более оптимальными, чем предыдущие, а менее эффективные решения постепенно исключаются из популяции.

Сам процесс эволюции включает в себя несколько шагов, включая инициализацию начальной популяции, выбор родительских особей, мутацию и скрещивание, а также выбор наилучших решений для создания следующего поколения.

Преимуществом эволюционных алгоритмов является возможность нахождения глобального оптимума, а не локального, что обеспечивает максимальную точность и эффективность в решении задач оптимизации.

Существуют различные типы эволюционных алгоритмов, такие как генетические алгоритмы [9], стратегии эволюции, рой частиц, дифференциальная эволюция и другие.

Для оптимизации параметров классификатора с учетом издержек и определения веса для обучения логистической регрессии взвешенным методом. Рассмотрим улучшенную вариацию генетического алгоритма.

Островной генетический алгоритм (Island Genetic Algorithm) представляет собой вариант генетического алгоритма, в котором популяция разделена на несколько подгрупп, называемых островами [10]. Каждый остров имеет свою собственную популяцию, которая эволюционирует независимо от других островов. Острова периодические обмениваются информацией и особями, что способствует разнообразию и глобальному поиску оптимального решения.

Рассмотрим основные шаги, которые выполняются в островном алгоритме.

1. Инициализация — создаются несколько независимых островов, каждый из которых имеет свою начальную популяцию индивидов. Начальные индивиды обычно генерируются случайным образом.

2. Оценка — каждый остров оценивает индивидов своей популяции с использованием функции оценки, которая определяет качество решения, представляемого каждым индивидом. Оценка может основываться на достижении целевой функции или других критериев, зависящих от задачи.

3. Миграция — периодически происходит обмен индивидами между островами. Лучшие индивиды или группа индивидов мигрируют с одного острова на другой. Миграция позволяет обменяться информацией о лучших решениях и разнообразии между островами.

4. Эволюция — на каждом острове выполняются операции эволюции, такие как селекция, скрещивание и мутация, чтобы создавать новые поколения индивидов. Операции эволюции направлены на поиск более оптимальных решений в рамках каждого острова.

5. Повторение — шаги 2-4 повторяются до достижения заданного условия остановки, например, определенного количества поколений или достижения достаточно хорошего решения.

Генетический островной алгоритм применяется для создания и оптимизации индивидуальных наборов гиперпараметров модели и веса класса в данных.

Укажем оптимизационные параметры, которые подаются на вход:

– параметр регуляризации C;

– алгоритм (ядро), используемый для решения оптимизационной задачи логистической регрессии;

– веса классов, используемые для учета дисбаланса классов в данных с диапазоном от 1 до 10 для мошеннических транзакций.

Хромосома в островном генетическом алгоритме кодируется в виде словаря, где каждый ключ представляет ген или генотип, а значение ключа определяет конкретное значение гена.

Функция приспособленности принимает индивидуум (хромосому) в качестве входных данных и оценивает его качество. Функция приспособленности создает модель взвешенной логистической регрессии с параметрами, заданными в индивидууме, и вычисляет среднее значение метрики F_1 на кросс-валидации, с использованием обучающих данных и меток. Значение метрики F_1 является оценкой качества классификации модели.

После оптимизации был отобран лучший индивид со всех островов с лучшими параметрами и весом равным 7 для мошеннических транзакций, найденный островным генетическим алгоритмом.

Значения метрик качества для оптимизированной взвешенной логистической регрессии приведены в таблице 4.

Таблица 4. Значения метрик качества для оптимизированной взвешенной логистической регрессии.

	Precision	Recall	F1-score
0	1.00	1.00	1.00
1	0.86	0.82	0.84
Confusion Matrix			
85289 (TN)		18 (FP)	
24 (FN)		112 (TP)	

Визуализируем PR-curve кривой для модели, оптимизированной взвешенной логистической регрессии.

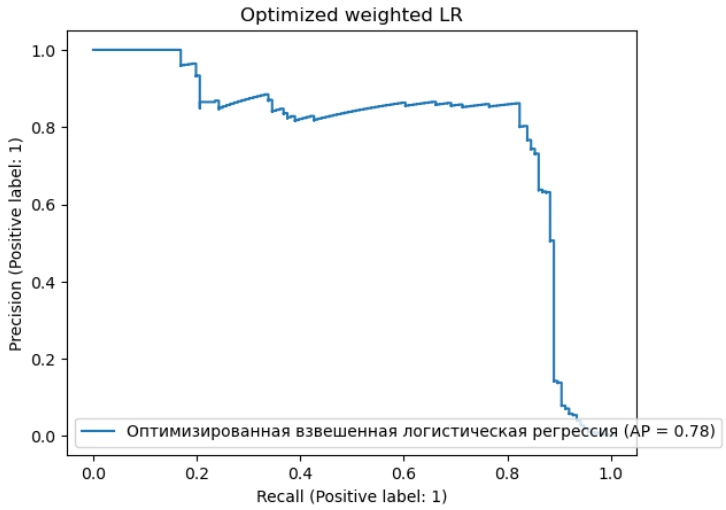

Рис.3 - PR-curve кривая оптимизированной взвешенной логистической регрессии

Оптимизированная взвешенная логистическая регрессия показывает более высокую точность (Precision = 0.86) для класса 1 (мошеннические операции), но немного ниже показатель полноты (Recall = 0.82) по сравнению с взвешенной логистической регрессией без оптимизации (Recall = 0.84). Это указывает на то, что оптимизированная модель менее склонна к ложно-положительным результатам.

В контексте обнаружения мошеннических транзакций, оптимизированная взвешенная логистическая регрессия, обеспечит более высокую точность в определении мошеннических транзакций.

Рассматривая F1-меру, оптимизированная взвешенная логистическая регрессия показывает немного лучшие результаты (F1-мера = 0.84) по сравнению с взвешенной логистической регрессией без оптимизации (F1-мера = 0.81). Это свидетельствует о том, что оптимизированная модель более сбалансирована и обеспечивает лучший компромисс между точностью и полнотой.

Заключение

Таким образом, классификация с учетом издержек может быть применена во многих различных сферах, где стоимость разных типов ошибок может быть очень разной. Правильное определение матрицы издержек и использование ее при классификации данных позволяет выбрать наиболее подходящую модель классификации и улучшить качество классификации данных.

В данной статье была рассмотрена важность классификации данных в динамически развивающейся отрасли анализа данных. Была рассмотрена матрица издержек и подходы к её определению. Также были описаны сферы применения классификации с учетом издержек, включая медицинскую и финансовую область. Был применен островной генетический алгоритм оптимизации для определения оптимальных параметров и подбора весов для мошеннических транзакций, что позволило нам улучшить результаты значения метрик.

Применение классификации с учетом издержек в информационных системах может быть критически важным, особенно в случаях, когда стоимость разных типов ошибок может быть очень разной. Это позволяет получать более точные результаты

классификации и улучшать качество принимаемых решений. Однако использование классификации с учетом издержек может быть нежелательным или нецелесообразным в целях некоторых случаях, особенно если стоимость разных типов ошибок одинакова или очень близка.

Классификация с учетом издержек является отличным инструментом для анализа данных, который может быть эффективно использован в различных областях. Однако, для того чтобы получить наибольшую пользу классификации, необходимо учитывать специфику задачи и особенности данных.

Список литературы

1. М.А. Поручиков. Анализ данных / - М.: Изд-во Самарского университета, 2016. - 37 с.

2. Т.В. Батура. Методы автоматической классификации текстов. // Программные продукты и системы / Software & Systems. — №1 — 89-99 с.

3. Демидова Л.А., Соколова Ю.С. Классификация данных на основе SVM-алгоритма и алгоритма k-ближайших соседей // Вестник Рязанского государственного радиотехнического университета. 2017. № 62. С. 119-132.

4. M.Manoj Krishna, M. Neelima, M. Harshali, M. Venu Gopala Rao. Image classification using Deep Learning. // International Journal of Engineering & Technology. 2018. 614-617 с.

5. P.Turney. Types of Cost in Inductive Concept Learning. 2 Apr 2000.

6. Demidova L.A. A Novel Approach to Decision-Making on Diagnosing Oncological Diseases Using Machine Learning Classifiers Based on Datasets Combining Known and/or New Generated Features of a Different Nature // Mathematics. 2023. Vol. 11(4). C. 792.

7. Aurelien Geron. Hands-On Machine Learning with Scikit-Learn and TensorFlow. 127-151 с.

8. Chaohong Song, Xinran Li. Cost-Sensitive KNN Algorithm for Cancer Prediction based on Entropy Analysis. Entropy MDPI. 2021.

9. Tegar Arifin Prasetyo, Roberd Saragih and Dewi Handayani. Genetic algorithm to optimization mobility-based dengue mathematical model. International Journal of Electrical and Computer Engineering

10. Darrell Whitley, Soraya B. Rana and Robert B.Heckendorn. Colorado State University. The Island Model Genetic Algorithm: On Separability, Population Size and Convergence. 27 December 2015.

References

U1. M.A. Poruchikov. Analiz dannyh / - M.: Izd-vo Samarskogo universiteta, 2016. - 37 s.

2. T.V. Batura. Metody avtomaticheskoj klassifikacii tekstov. // Programmnye produkty i sistemy / Software & Systems. — №1 — 89-99 с.

3. Demidova L.A., Sokolova YU.S. Klassifikaciya dannyh na osnove SVM-algoritma i algoritma k-blizhajshih sosedej // Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta. 2017. № 62. S. 119-132.

4. M.Manoj Krishna, M. Neelima, M. Harshali, M. Venu Gopala Rao. Image classification using Deep Learning. // International Journal of Engineering & Technology. 2018. 614-617 s.

5. P.Turney. Types of Cost in Inductive Concept Learning. 2 Apr 2000.

6. Demidova L.A. A Novel Approach to Decision-Making on Diagnosing Oncological Diseases Using Machine Learning Classifiers Based on Datasets Combining Known and/or New Generated Features of a Different Nature // Mathematics. 2023. Vol. 11(4). C. 792.

7. Aurelien Geron. Hands-On Machine Learning with Scikit-Learn and TensorFlow. 127-151 c.

8. Chaohong Song, Xinran Li. Cost-Sensitive KNN Algorithm for Cancer Prediction based on Entropy Analysis. Entropy MDPI. 2021.

9. Tegar Arifin Prasetyo, Roberd Saragih and Dewi Handayani. Genetic algorithm to optimization mobility-based dengue mathematical model. International Journal of Electrical and Computer Engineering

10. Darrell Whitley, Soraya B. Rana and Robert B. Heckendorn. Colorado State University. The Island Model Genetic Algorithm: On Separability, Population Size and Convergence. 27 December 2015.