Литература:
1. *Безуглова Е.В.* Оползневый риск транспортных природно-технических систем: монография / Е. В. Безуглова, С. И. Маций, В. В. Подтелков. – Краснодар: КубГАУ, 2015. – 239 с.
2. *Маций С.И.* Противооползневая защита: монография. – Краснодар: АлВи-дизайн, 2010. – 288 с.
3. *Маций С.И.* Оценка оползневого риска транспортных сооружений: монография / С. И. Маций, Е. В. Безуглова, Д. В. Плешаков. – Краснодар: КубГАУ, 2015. – 120 с.
4. *Маций С.И.* Принятие решений при формировании при формировании природно-технических систем в условиях неопределенности и риска / С.И. Маций, Д.И. Кацко В сборнике: Системный анализ в проектировании и управлении. Сборник трудов XXII международной научно-практической конференции. – СПБ: СПбГПУ, 2018. С.268-273.
5. *Маций В.С.* Вариативные подходы к оценке и управлению оползневым риском транспортных систем / В.С. Маций, Д.И. Кацко. В сборнике трудов: IV Международной научно-практической молодежной конференции по геотехнике Тюмень: ТИУ, 2018. С.47-51.

Сидоренко В.Г., Кулагин М.А.

Прогнозирование совершения нарушения безопасности движения по вине локомотивной бригады с использованием современных методов машинного обучения

Аннотация: Статья посвящена вопросам определения и сбора показателей, влияющих на безопасность движения, а также расчету вероятности совершения нарушения локомотивной бригадой. В работе разработаны и проанализированы разнообразные алгоритмы машинного обучения, а именно: нейронные сети, градиентный бустинг над решающими деревьями и случайные леса.

Ключевые слова: локомотивная бригада, расчет вероятности, оценка влияния человеческого фактора, машинное обучение, искусственный интеллект

На данный момент в компании ОАО «РЖД» подавляющий объем перевозок выполняется людьми, от которых зависит высокое качество и безопасность работы железной дороги. Для того чтобы управлять целым составом требуются специалисты с высоким уровнем профессионализма. В компании ОАО «РЖД» насчитывает порядка 100 тысяч работников, управляющих локомотивом. Большинство нарушений, совершаемых

локомотивными бригадами, влечет за собой серьезные экономические потери для компании. Поэтому задача прогнозирования совершения нарушения по вине локомотивной бригады является актуальной на данный момент.

В рамках данной статьи представлен способ расчета вероятности совершения нарушения машинистом локомотива с использованием современных методов машинного обучения. Процедуру расчета вероятности можно разделить на следующие этапы:

1. Постановка задачи и определения допустимых критериев качества работы алгоритма.
2. Получение и исследование данных о машинисте.
3. Подготовка данных для алгоритмов машинного обучения.
4. Обучение и анализ качества работы разнообразных алгоритмов.
5. Проверка качества работы наилучшего с точки выбранного критерия алгоритма или композиции алгоритмов на тестовой выборке.

Схема разрабатываемой процедуры приведена на рис. 1.

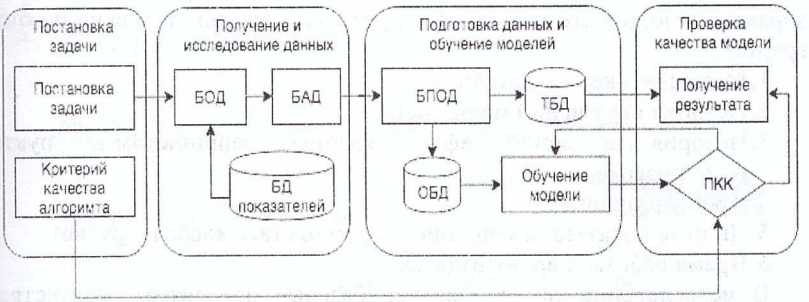

Рис. 1 – Блок-схема получения математической модели

(БОД – блок обработки данных; БАД – блок анализа данных; БД – база данных; БПОД – блок предварительной обработки данных; ОБД/ТБД – обучающая/тестовая база данных; ПКК – проверка критерия качества алгоритма)

Любое исследование начинается с постановки цели и задачи. В данной статье представлена классическая задача обучения с учителем [1]. В качестве признаков выступают показатели машиниста, собранные из различных автоматизированных систем. Меткой каждого маршрута выступает факт наличия вины машиниста в совершении нарушения (0 – вины машиниста нет; 1 – вина машиниста есть). В терминах машинного обучения можно данную задачу можно отнести к сфере задач бинарной классификация.

Критерием оценки качества алгоритма будет выступать:

– Accuracy – доля верного срабатывания алгоритма;

– AUC-ROC – площадь по ROC кривой (соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущих признак, и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущих признак) [2];
– F-мера – гармоническое среднее между Precision (доля действительно верного определения нарушения алгоритмом к общему количеству выявленных нарушений) и Recall (доля найденных классификатором нарушений относительно всех нарушений в тестовой выборке) [3].

В качестве объекта исследования был выбран машинист и все его поездки за 2017-2018 год. Машинист - это работник, осуществляющий обслуживание и управление локомотивом, ведение поезда с точным соблюдением графика движения поездов, обеспечивающий требования безопасности, безусловное выполнение установленного регламента переговоров, сохранность грузов и подвижного состава, а также рациональный режим ведения поезда при минимальном расходе топлива и электроэнергии. У машиниста было выделено порядка 50 признаков, характеризующих его работу. Все признаки машиниста можно разбить на группы:

1. Медицинские показатели.
2. История нарушений машиниста.
3. История нарушений, зафиксированных машинистом-инструктором при проверке.
4. Уровень знаний.
5. Личные показатели машинистов (депо, стаж, класс и другие).
6. Время работы и время отдыха.

В исследовании использовались базовые алгоритмы: искусственная нейронная сеть (ИНС), градиентный бустинг (ГБ), случайные леса (СЛ).

На момент написания статьи алгоритм проходил обучение на выборке размером около 500 тысяч поездок и тестировался на выборке 100 тысяч поездок. Ключевой проблемой решаемой задачи является несбалансированность обучающей выборки (рис. 2).

Из рисунка 2 можно заключить, что количество случаев с нарушениями по вине машиниста занимают около 7,5 % места в выборке (597247 – поездок без вины; 45109 – поездок с виной). Данная проблема является ключевой трудностью на пути к получению рационального математического алгоритма. К основным приемам, которые используются учеными и специалистами по анализу данных, относятся:

Поиск возможностей получения дополнительных данных о минорном классе [3].

Размножение малочисленного целевого (минорного) класса путем копирования. Количество элементов минорного класса за счет

множественного дублирования приравнивается к количеству доминирующего [3,4].

Балансирование обучающей выборки до уровня целевого класса. Количество элементов доминирующего класса приравнивается количеству объектов минорного [3,4].

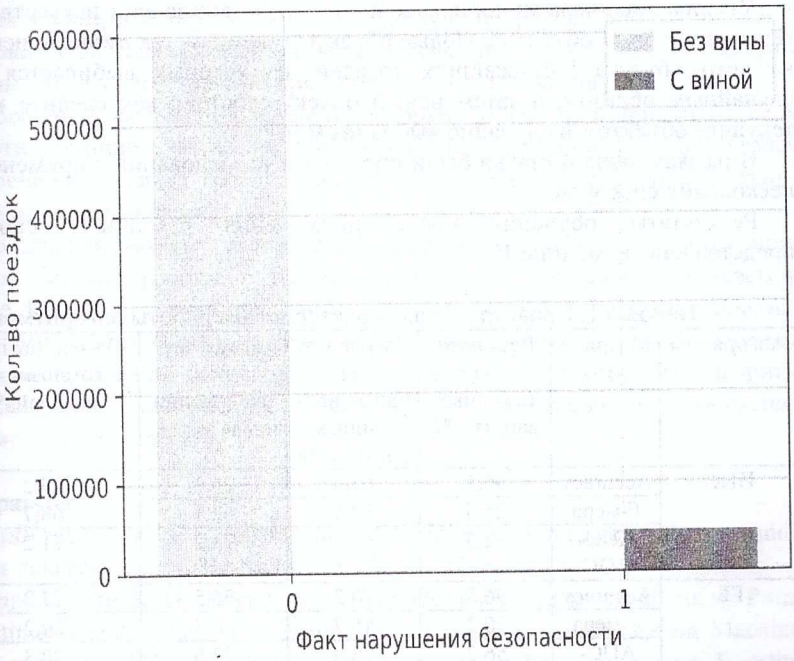

Рис. 2 – Гистограмма распределения меток в базе данных

Создание синтетических образцов. Данный способ подходит к задачам, связанным с компьютерным зрением, так как суть метода заключается в увеличении минорного класса за счет небольших изменений значений признаков объектов. Например, поворот или зеркальное отображение. В нашей задаче данный подход не имеет большого смысла [4].

Увеличение штрафа за ошибки на целевом классе. При обучении алгоритма за каждую ошибку на минорном классе штрафовать алгоритм сильнее, чем за ошибки на доминирующем [3].

Построение композиции алгоритмов и усреднение результатов обучения. Суть метода заключается в балансировании выборки, но только не один раз, как это было описано в пункте 4, а несколько, при этом каждый раз с разным набором доминирующего класса. Затем каждая выборка используется для обучения, а результаты усредняются [5].

Использование алгоритмов обучения без учителя (поиск аномалий). Зачастую, в зависимости от задачи, необязательно строить алгоритм классификации по классическому принципу. Достаточно отыскать признаки с аномальными значениями, которые и являются представителями минорного класса [5].

Увеличение числа минорного класса происходит за счет искусственной генерации новых объектов. Новые объекты генерируется путем поиска для каждого объекта k-ближайших соседей, из которых выбирается один случайным образом, а затем новый объект строится как среднее между текущим объектом и случайно выбранным [6].

В рамках данной статьи были проведены исследования с применением нескольких способов.

Результаты обучения после применения различных способов представлены в таблице 1.

Таблица 1. Сравнительная характеристика работы алгоритмов

Алгоритмы	Метрики	Результат на исходных данных, %	Результат на балансиро-ванных данных, %	Результат при увеличении числа меток, %	Результат при композиции моделей, %
ИНС	Accuracy	95,1	67,4	76,1	79,2
	F-мера	25,5	32.1	62,1	68,7
	AUC-ROC	52,5	68,6	73,2	71,2
ГБ	Accuracy	96,3	70,2	76,5	77,2
	F-мера	30,2	45.7	66,3	69,2
	AUC-ROC	56,2	76,9	72,6	70,3
СЛ	Accuracy	93,6	62,7	68,1	70,2
	F-мера	22.4	40.8	56,5	54,9
	AUC-ROC	51,0	66,2	70,8	69,6

При решении задачи бинарной классификации требуется использовать только тот способ оценки качества работы модели, который отражает объективную ситуацию. Например, в данной задаче корректнее использовать AUC-ROC и F-меру. Согласно полученным результатам, следует, что ни один способ работы с несбалансированными выборками не показывает хороший результат для данного набора данных. Отсюда следует, что проблема несбалансированной выборки остается актуальной, как минимум, для решаемой задачи.

Метод уменьшения 0-го класса до уровня 1-го приводит к тому, что мы строим вероятностную модель только на выборочной совокупности данных, которая значительно меньше всей, а это значит, что модель не видела всех данных в выборке, и в будущем появится много ложных срабатываний алгоритма.

Два последних метода показали наилучшие результаты на тестовой выборке. Причины успеха, возможно, связаны с тем, что алгоритмы используют всю выборку, но при этом сохраняется баланс классов.

Проблема несбалансированной выборки до сих пор присутствует в области машинного обучения, так как большинство данных, используемых для решения разного рода задач, имеют смещение в сторону одного доминирующего класса.

На данный момент стоит задача разработки нового или адаптации старого метода работы с несбалансированной выборкой к решаемой задаче. Кроме того, очевидно, требуется увеличение обучающей выборки.

В рамках данной статьи предложен способ обучения, поднята проблема несбалансированной выборки, а также представлен возможный алгоритм предсказания совершения нарушения безопасности движения локомотивной бригадой.

Литература:

1. *Muller* A.C., *Guido* S. Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc., — 2016.
2. *Boyd* K., *Eng* K. H., *Page* C. D. Area under the precision-recall curve: Point estimates and condence intervals //Joint European Conference on Machine Learning and K nowledge D iscovery in D atabases. S pringer, B erlin, Heidelberg — 2013. — C. 451-466.
3. *Powers D.M.* Evaluation: from precision, recall and F -measure to R OC, informedness, markedness and correlation — 2011.
4. *Lemaitre, Guillaume, Fernando Nogueira*, and *Christos* K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research 18.1— 2017. C. 559-563.
5. *He; Haibo*, and *Yunqian* Ma, eds. Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons, — 2013.
6. *Chawla* N., *Bowyer* K., *Hall* L., Kegelmeyer W . SMOTE: Synthetic Minority Over-sampling Technique. // Journal of Artificial Intelligence Research, — 2002. — C. 321–357.