УДК 519.24

ПРОГНОЗИРОВАНИЕ ПЕРФОРАЦИЙ КОРОНАРНЫХ АРТЕРИЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

М.С. Абрамович[1], В.И. Стельмашок[2], Г.С. Дорофеев[1]

[1]Научно-исследовательский институт прикладных проблем математики и информатики, Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь, abramovichms@bsu.by
[2]Республиканский научно-практический центр «Кардиология», ул. Розы Люксембург, 110Б, 220036, г. Минск, Беларусь, stelval@yandex.by

Рассматривается проблема прогнозирования перфорации коронарных артерий при рентгенэндоваскулярных операциях. На обучающей выборке, состоящей из групп пациентов успешно прооперированных и прооперированных с перфорациями коронарных артерий, с использованием точного критерия Фишера отобраны два набора информативных признаков. Предложена процедура устранения дисбаланса классов обучающей выборки. В качестве классификаторов использовались алгоритмы машинного обучения и ансамбли алгоритмов.

Ключевые слова: Коронарная артерия; перфорация; критерий Фишера; дисбаланс классов; машинное обучение; ROC-анализ.

PREDICTION OF CORONARY ARTERY PERFORATIONS USING MACHINE LEARNING METHODS

M.S. Abramovich[1],V.I. Stelmashok[2], G.S. Dorofeev[1]

[1]Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University, 4 Niezalieznasci Avenue, Minsk, 220030, Belarus
[2]Republican Scientific and Practical Center «Cardiology», st. Rozy Luksemburg, 110B, Minsk, 220036, Belarus.
Corresponding author: abramovichms@bsu.by

In this work we study the problem of predicting whether coronary artery perforation will take place as a result of rentgen-endovascular surgery. The training set consists of the patients who were successfully operated on and the patients who ended up having coronary artery perforation. We used Fisher's exact test to select two sets of informative features. We proposed a procedure for handling class imbalance problem. Machine learning algorithms and ensemble methods were used as classifiers.

Keywords: Coronary artery; perforation; Fisher's exact test; class imbalance; machine learning; ROC-analysis.

Введение

Успешное восстановление кровотока в хронически окклюзированной коронарной артерии рентгенэндоваскулярными способами позволяет улучшить качество жизни и прогноз у оперируемых пациентов [1]. Вместе с тем проведение подобных вмешательств нередко ассоциируется с развитием осложнений, наиболее тяжелым из которых является перфорация стенки коронарного сосуда. В связи с этим представляется целесообразным разработать систему прогнозирования риска развития перфораций в ходе выполнения оперативного вмешательства. Это позволит предотвратить данные осложнения.

Отметим что при этом количество пациентов с перфорацией коронарных артерий существенно меньше, чем без перфорации. Поэтому проводилось устранения дисбаланса классов обучающей выборки с использованием методов увеличения класса меньшего объема (oversampling) и уменьшения класса большего объема (undersampling).

Прогнозирование (классификация) перфораций коронарных артерий пациентов проводилась с использованием следующих методов машинного обучения: логистической регрессии, метода опорных векторов, метода *k* - ближайших соседей, случайного леса и градиентного бустинга. Проводился также подбор гиперпараметров алгоритмов классификации, обеспечивающих максимальную точность классификации пациентов.

1. Методология исследования / теоретические основы

В ходе детального ретроспективного изучения ангиограмм включенных в исследование пациентов выполнялась оценка развития перфорации стенки коронарной артерии в соответствии с критериями, предложенными в работе [2]. В зависимости от наличия, либо отсутствия возникновения отмеченного осложнения в ходе антеградного проведения коронарного проводника через зону хронического окклюзионного поражения, все пациенты были разделены на 2 группы: лица у которых была зарегистрирована перфорация – 28 пациентов (миноритарный класс), у которых не было отмечено данного осложнения – 369 пациентов (мажоритарный класс).

Последующий этап анализа включал определение перечня потенциальных предикторов, влияющих на развитие перфорации коронарной артерии в ходе процедуры реканализации хронических тотальных окклюзий антеградным доступом. К таким были отнесены клинико-демографические показатели, рентгенанатомические и рентгенморфологические показатели, характеризующие пораженную артерию, а также процедурные аспекты выполняемой операции.

Отбор информативных признаков проводился с использованием фильтрационного теста – точного критерия Фишера [3]. Выбор данного теста бы обусловлен фактом наличия ожидаемой частоты значений хотя бы в одной из ячеек таблицы 2х2 менее 10% (что нередко встречается в малых выборках). Таким образом, для всех бинарных признаков сравнивались частоты групп с возникшей и не возникшей перфорацией в ходе реканализации хронических тотальных окклюзий. Было сформировано два набора информативных признаков. В первый набор включались те показатели, которые имели существенное различие на уровне значимости не менее чем 0.05, а во второй набор – на уровне значимости не менее чем 0.2. В первом наборе оказалось 10 показателей, во втором – 37 .

Для обучения и тестирования модели был использован метод скользящего экзамена (leave one out). Модель обучается на всех наблюдениях, кроме одного, и выдает предсказание для этого наблюдения. Учитывая наличие истинных меток пациентов, можно вычислять различные метрики машинного обучения, такие как точность, полнота и f1-мера [4]. При этом каждый раз обучение происходит почти на всей выборке, а оценка эффективности модели на тесте более надежна, чем в случае одинарного разбиения на обучающую и тестовую выборки. Если используется метод oversampling, то необходимо убедиться в том, что на кросс-валидации синтетические объекты ни разу не попадут в тестовую выборку. Иначе полученная оценка может быть завышенной. Таким образом, кросс-валидация проводится только на оригинальных объектах.

Далее для устранения дисбаланса классов применялся метод undersampling в комбинации с методом oversampling. В качестве метода oversampling был применен алгоритм SMOTE [5]. В качестве метода undersampling использовался алгоритм Edited Near Neighbors [6]. Его суть заключается в следующем: для каждого объекта алгоритм *k*-ближайших соседей (KNN) пытается определить метку класса. Если объект принадлежит мажоритарному классу, а алгоритм KNN ошибается и присваивает метку миноритарного, то объект удаляется из выборки. Если объект принадлежит миноритарному классу и алгоритм KNN ошибается, то удаляются наблюдения мажоритарного класса, являющиеся соседями объекта. После этой процедуры для первого набора информативных признаков из 369 объектов мажоритарного класса осталось 273. После применения алгоритма SMOTE количество объектов миноритарного класса тоже стало 273. Кросс-валидация по отдельным объектам проводилась только для оригинальных объектов (273 мажоритарного класса и 28 миноритарного).

Кроме этого проводилась кросс-валидация для подбора параметров модели, обеспечивающих наибольшую точность классификации.

Результаты и их обсуждение

В таблице 1 представлены метрики точности классификации на кросс-валидации по отдельным объектам для первого набора информативных признаков.

Таблица 1 – Метрики точности классификации для первого набора информативных признаков

Метод машинного обучения	Точность	Полнота	F1-мера	Показатель ROC-AUC
Случайный лес	0.34	0.79	0.48	0.92
Логистическая регрессия	0.30	0.75	0.43	0.87
Метод *k*-ближайших соседей	0.67	0.71	0.69	0.83
Метод опорных векторов	0.73	0.79	0.70	0.93
Градиентный бустинг	0.42	0.79	0.55	0.91

Как следует из таблицы 1, наибольшая точность классификации была достигнута методом опорных векторов.

При отборе признаков с помощью точного критерия Фишера, изменение уровня значимости влияет на количество признаков, которое будет отобрано для обучения и чем больше уровень значимости, тем больше признаков будет отобрано. Необходимо найти баланс между недообучением модели с одной стороны и переобучением с другой.

Так как набор данных небольшой, то необходимо, чтобы при применении метода undersampling, из выборки удалялось минимальное количество наблюдений. Если это требование удовлетворяется, а метрики качества становятся лучше, то предполагается, что удаленные наблюдения были шумовыми, и их удаление поспособствовало лучшей классификации. Для второго набора информативных признаков лучшей оказалась следующая конфигурация: алгоритм undersampling, количество ближайших соседей равно 2, алгоритм oversampling SMOTE, количество наблюдений мажоритарного класса равно 286.

В таблице 2 представлены метрики точности классификации на кросс-валидации по отдельным объектам для второго набора информативных признаков.

Таблица 2 – Метрики точности классификации для второго набора информативных признаков

Метод машинного обучения	Точность	Полнота	F1-мера	Показатель ROC-AUC
Случайный лес	0.57	0.86	0.69	0.94
Логистическая регрессия	0.34	0.46	0.39	0.86
Метод *k*-ближайших соседей	0.87	0.93	0.90	0.96
Метод опорных векторов	0.79	0.79	0.79	0.96
Градиентный бустинг	0.75	0.86	0.80	0.93

Как следует из таблицы 2, метод *k*-ближайших соседей показал наибольшую точность классификации.

Таким образом, полученные результаты показывают, что даже в условиях существенного дисбаланса классов обучающей выборки, удалось для некоторых алгоритмов добиться большой точности прогнозирования риска развития перфораций в ходе выполнения оперативного вмешательства на коронарных артериях.

Библиографические ссылки

1. Patel Y., Depta J., De Martini T. Complications of chronic total occlusion percutaneous coronary intervention // Intervention Cardiology. 2013. Vol. 5, № 5. P. 567–575.
2. Ellis S. G. Increased coronary perforation in new device era. Incidence, classification, management and outcome // Circulation. 1994. Vol. 90, № 6. P. 2725–2730.
3. Afifi A. A., Azen S. P. Statistical analysis: A Computer Oriented Approach, 2nded. New York: Academic Press, 1979. 442 p.
4. Harrington P. Machine Learning in Action. New York: Manning. 2012. 382 p.
5. Chawla N., Bowyer K., Hall L., Kegelmeyer W. SMOTE: Synthetic Minority Oversampling Technique // Journal of Artificial Intelligence Research. 2002. № 16. P. 341–378.
6. Wilson L. Asymptotic Properties of Nearest NeighborRules Using Edited Data // IEEE Transactions on Systems. 1992. Vol. SMC-2, № 3. P. 34–42.