Сравнение возможностей логистической регрессии и искусственных нейронных сетей в прогнозировании результатов исследования на малой выборке

Базылев В.В.*, Карнахин В.А.*
*Федеральный центр сердечно-сосудистой хирургии
Министерства здравоохранения России

В работе доказаны теоремы, позволяющие находить асимптотический дефект выборочной медианы, основанной на выборках случайного объема. Это делает возможным сравнивать в терминах добавочного числа наблюдений качество выборочной медианы, основанной на выборках случайного и неслучайного объемов. Рассмотрены случаи биномиального и распределения Пуассона.

Ключевые слова: искусственные нейронные сети, логистическая регрессия, статистика, выборка, математическая модель, регрессия.

Введение

Прогностические модели искусственных нейронных сетей (ИНС) и логистической регрессии используются в различных областях медицины для решения различных задач (Han, 2018; García-Reiriz, 2007; Bhatikar, 2005; Zhang, 1998; Waisman, 2019; Cireşan, 2012; Parisi, 2019). ИНС, создание которых было вдохновлено нейробиологией и архитектурой человеческого мозга являются непараметрическими методами распознавания образов, которые выявляют скрытые связи между зависимыми и независимыми переменными (Haykin, 1999; Lobo, 2018; Bazrafkan, 2018; Plis, 2018; Soltoggio, 2018; Parisi, 2019; Vellappally, 2019; Hyvärinen, 2000). В последние годы нейронные сети получили широкое распространение во многих дисциплинах науки и медицины. Модели нейронных сетей могут учиться на примерах, включать большое количество переменных и предоставлять адекватный и быстрый ответ на новые входящие данные (Tavanaei, 2019; Kulkarni, 2018; Graves, 2005; Boutin, 2018; Zhang, 2018; Coninck, 2018; Wideman, 2018). В настоящее время все чаще предпринимаются попытки сравнивать различные количественные модели для решения конкретных задач классификации данных (Sargent, 2001; Dreiseitl, 2002; Zurada, 1994). D.J. Sargent (Sargent, 2001) представил метаанализ, сравнив ИНС с регрессионными моделями в 28 исследованиях и обнаружил, что в 36% случаев ИНС оказались более эффективны, чем регрессионные модели. Логистическая регрессия обладала преимуществом в 14% работ, в оставшихся 50% исследований производительность моделей была одинаковой. S. Dreiseitl (Dreiseitl, 2002) проанализировал 72 статьи, сравнивая ИНС с логистической регрессией и обнаружил преимущество нейронных сетей в 18% случаев, а логистической регрессии только в 1%, в 42% случаев не было никакой разницы между двумя моделями. При этом в литературе отсутствуют данные о сравнении математических моделей в условиях малых выборок и сложных клинических ситуаций. Для ответа на вопрос о преимуществе в прогнозировании результатов работы при малых выборках выполнено симуляционное исследование.

Цель работы: сравнить производительность моделей ИНС и логистической регрессии в прогнозировании результатов исследования в условиях малой выборки.

Материалы и методы

Логистическая регрессия

Логистическая регрессия — это статистическая модель, используемая для предсказания вероятности возникновения некоторого события путём подгонки данных к логистической кривой. Методика используется для определения предсказания вероятности возникновения некоторого события по значениям множества признаков. Логарифм правдоподобия логистической регрессии имеет вид:

$$L(\beta) = \sum_{i=1}^{n} (l_i(\beta)) \tag{1}$$

где, β - вектор параметров.

Обучение логистической рагрессии на языке Python и библиотеки scikit-learn:

```
# 1. Обучение логистической регрессии
logreg = LogisticRegression(max_iter=1000)
logreg.fit(x_train, y_train)

# 2. Вывод весовых коэффициентов
print("Весовые коэффициенты модели:")
print(pd.DataFrame({'feature': x_train.columns, 'coef': logreg.coef_[0]}))

# 3. Вывод значений, возвращаемых моделью (степень уверенности)
# для первых 10 примеров
probabilities = logreg.predict_proba(x_train[:10])
print("\nСтепень уверенности модели для первых 10 примеров:")
print(probabilities)
print("\nВероятность выжить для первых 10 примеров:")
print(probabilities[:, 1])

# 4. Вывод предсказаний модели и истинных ответов для первых 10 примеров
predictions = logreg.predict(x_train[:10])
print("\nПредсказания модели для первых 10 примеров:")
print(predictions)
print("\nИстинные ответы для первых 10 примеров:")
print(y_train[:10].values)

# Совместный вывод
print("\nПредсказания и истинные ответы для первых 10 примеров:")
for i in range(10):
        print(f"Пример {i+1}: Предсказание = {predictions[i]},
                Истинный ответ = {y_train.values[i]}")
```

Искусственные нейронные сети. Многослойный перцептрон

Перцептрон представляет собой сеть формальных нейронов МакКаллока и Питтса, состоящую из нескольких последовательно соединенных слоев (Wang, 2017). Входной слой нейронов состоит из сенсорных элементов, выполняет функцию приема и распространения по сети входной информации. $W^{k+1} = W^k - \eta((a(X) - Y) \cdot X)$ – в векторно-матричном виде. Затем идет один или несколько скрытых слоев. Все нейроны в скрытом слое имеют несколько входов, сообщающихся с выходами нейронов предыдущего слоя и один выход. Задача нейрона состоит в вычислении взвешенной суммы его входов с дальнейшим преобразованием ее в выходной сигнал. Нейроны суммируют поступающие к ним сигналы от нейронов предыдущего уровня иерархии с весами, определяемыми состояниями синапсов, и формирует ответ, если полученная сумма выше порогового значения. Сеть переводит входной образ, определяющий степени возбуждения нейронов самого нижнего уровня иерархии, в выходной образ, определяемый нейронами самого верхнего уровня. Возбуждение нейрона на верхнем уровне говорит о принадлежности входного образа к той или иной категории. Процедура многослойного перцептрона создает прогностическую модель для одной или нескольких зависимых переменных на основании значений переменных предикторов (Esfandiari, 2017).

Проведено симуляционное исследование для оценки эффективности выявления предикторов с помощью логистической регрессии и ИНС. С целью усложнения работы прогностических моделей взята малая выборка больных. В симуляцию включена группа больных из 50 человек, которым была выполнена пластическая операция на митральном клапане. Для симуляции выбраны пять независимых переменных: пол, возраст, индекс массы тела (ИМТ), методика аппроксимации папиллярных мышц. Две переменных – пол и аппроксимация папиллярных мышц – категориальные, все остальные непрерывные. Зависимая переменная – регургитация на митральном клапане в отдаленном периоде. Число событий в зависимой переменной - 4. В симуляции получилась малая выборка с небольшим количеством событий. Все случаи регургитации произошли в группе больных, где не была выполнена аппроксимация папиллярных мышц. На лицо феномен сепарации данных, что значительно усложняет работу модели логистической регресии. Все случае рецидива митральной недостаточности произошли у больных старше 60 лет, хотя возраст не является предиктором появления повторной регургитации по литературным данным [1]. Таким образом, выявление предикторов в такой ситуации представляется трудной задачей. Клинико-демографические характеристики по изучаемым переменным представлены в таблице 1.

Таблица 1: Клинико-демографические характеристики больных

Фактор	**N=50, n; % (M±SD)**
Мужчины	43(86%)
Возраст	59,8±6,4
ИМТ	29,9±3,4
Аппроксимация	4(8%)

Поиск предикторов выполнен с помощью множественной логистической ре-

грессии и многослойного перцептрона. Алгоритм реализован с помощью SPSS версии 23 (SPSS, Chicago, IL, USA).

Результаты

Согласно результатам, пол, ИМТ - статистически незначимые факторы (p=1,1 и 0,6 соответственно). Возраст, несмотря на то, что только у больных старше 60 лет произошло событие, статистически незначимый фактор (p=0,2). Как и предполагалось, в случае аппроксимации возник феномен разделения данных и получена огромная среднеквадратичная ошибка. Несмотря на то, что фактор, возможно, значимый, невозможно интерпретировать данные логистической регрессии [2]. ROC-кривая в отношении предиктора возраст представлена на рисунке 1.

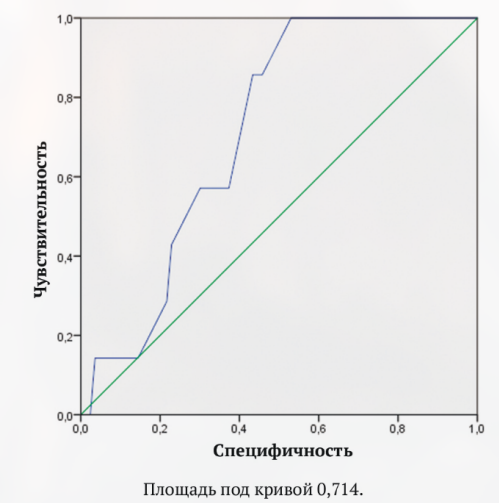
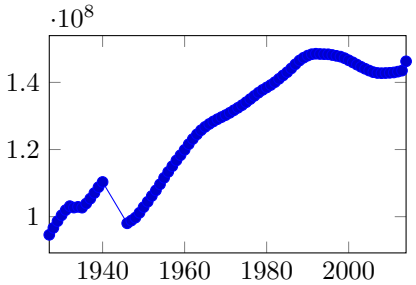


Рис. 1: ROC кривая. Логистическая регрессия. Предиктор возраст

По результатам анализа ROC-кривой выявлена зависимость между предиктором возраст и регургитацией на митральном клапане, площадь под кривой говорит о среднем уровне взаимосвязи.



Заключение

При малой выборке с небольшим количеством событий СИН имеют преимущество над другими методиками при определении предикторов влияния на зависимую переменную [3].

ИНС позволяет нивелировать феномен разделения данных.

Необходимо более широко использовать методику в медицинской статистике.

Список литературы

[1] Draper Peter. The title of the work // The title of the book / Ed. by The editor ; The organization. — Vol. 4 of 5. — The address of the publisher : The publisher, 1993. — P. 213.

[2] Сычёв М. С. История Астраханского казачьего войска: учебное пособие. — Астрахань : Волга, 2009. — 231 с.

[3] Носовский А. М. Статистика малых выборок в медицинских исследованиях // Российский медицинский журнал. — 2013. — № 6. — С. 19–34.