Горбаченко В.И., Кузнецов Р.Н., Кузнецова О.Ю. Отбор информативных признаков для прогнозирования послеоперационных состояний при желчнокаменной болезни. // Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей XVI Междунар. научно-техн. конф. – Пенза: ПДЗ, 2016. – С. 91-97.

УДК 519.237.5

ОТБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОСЛЕОПЕРАЦИОННЫХ СОСТОЯНИЙ ПРИ ЖЕЛЧНОКАМЕННОЙ БОЛЕЗНИ

В.И. Горбаченко, Р.Н. Кузнецов, О.Ю. Кузнецова

FEATURE SELECTION FOR PREDICTING POSTOPERATIVE CONDITIONS IN CHOLELITHIASIS

V.I. Gorbachenko, R.N. Kuznetsov, O.Yu. Kuznetsova

Аннотация. В статье представлены результаты экспериментальной проверки возможности использования метода Forward Selection и метода Backward Elimination для отбора информативных признаков. Отбор важных признаков может помочь врачу выбрать наиболее важные признаки для диагностики заболевания.

Ключевые слова: метод Forward Selection и метод Backward Elimination, Feature Selection, желчнокаменная болезнь, отбор показателей, послеоперационные осложнения.

Abstract. The article presents the results of experimental verification of the possibility of using the method of Forward Selection and Backward Elimination method for feature selection. Selection of important features can help the physician to choose the most important features for the diagnosis of disease.

Keywords: forward Selection method and the method of Backward Elimination, Feature Selection, cholelithiasis, the selection of indicators, post-operative complications.

В медицинских исследованиях всегда имеется большое общее количество исследуемых связей. Однако многие из них будут статистически не значимыми, и при отсутствии клинической значимости их не надо принимать во внимание и интерпретировать. Сколько же будет ценных с точки зрения поставленных целей исследования взаимосвязей, заранее предположить невозможно. Как правило, статистически и клинически значимыми оказывается 5–25 % всех возможных связей.

Правильный выбор признаков может быть более значимой задачей, чем уменьшение времени обработки данных или улучшение точности классификации. Отбор важных признаков может помочь расшифровать механизмы, лежащие в основе проблемы, представляющей интерес для исследования.

В данной работе исследовано два основных подхода шаговой регрессии для отбора признаков. Данный подход обеспечивает получение коэффициентов модели для всех независимых переменных. Исходя из уровней значимости решается задача включения признаков в модель как значимых и достоверных. Для автоматического включения значимых признаков в модель и исключения незначимых рассматривался пошаговый регрессионный анализ в двух вариантах: метод Forward Selection и метод Backward Elimination [1–3]. Также приведен пример задачи, при решении которой один из методов показал хороший результат, рассмотрен пример работы некоторых из алгоритмов отбора признаков для задачи определения факторов, влияющих на послеоперационные осложнения при желчнокаменной болезни.

Материалом исследования послужили истории болезни 109 пациентов отделения торакальной хирургии областной клинической больницы им. Н.Н. Бурденко.

Из них 63 случая без осложнения, 46 осложненных случаев. Использовано 13 показателей на основе общего и биохимического анализа крови, а также 2 дополнительных показателя (длительность оперативных вмешательств, пол).

Алгоритм Forward Selection включает в себя следующие шаги [1, 3]:

1. Из списка всех возможных входных переменных в матрице значений независимых переменных X выбирается та, которая имеет наибольшую корреляцию с Y (зависимая переменная), после чего модель, содержащая лишь одну выбранную независимую переменную, проверяется на значимость при помощи частного F-критерия. Если значимость модели не подтверждается, то алгоритм на этом заканчивается за неимением существенных входных переменных. В противном случае эта переменная вводится в модель и осуществляется переход к следующему пункту алгоритма. Следует отметить, что в данном случае проверка на значимость всей модели в целом будет равносильна проверке на значимость выбранной независимой переменной, так как на данном этапе модель еще не содержит других входных переменных.

2. По всем оставшимся переменным рассчитывается значение статистики γ, которая представляет собой отношение прироста суммы квадратов регрессии, достигаемой за счет ввода в модель соответствующей дополнительной переменной, к величине суммы квадратов ошибок.

3. Из всех переменных-претендентов на включение в модель выбирается та, которая имеет наибольшее значение критерия, рассчитанного в пункте 2.

4. Проводится проверка на значимость выбранной в пункте 3 независимой переменной. Если ее значимость подтверждается, то она включается в модель, и осуществляется переход к пункту 2 (но уже с новой независимой переменной в составе модели). В противном случае алгоритм останавливается.

Метод обратного исключения (Backward Elimination) похож на предыдущий метод, но с тем отличием, что все переменные изначально включены в модель и постепенно осуществляется «отсеивание» тех из них, которые не проходят проверку на значимость [1, 3].

1. В модель включаются все имеющиеся независимые переменные.

2. По переменным, включенным к данному моменту в модель, рассчитывается величина, представляющая собой разность между суммой квадратов регрессии, построенной по всем текущим переменным модели, и аналогичным показателем, рассчитанным теперь уже без учета одной переменной, для которой вычисляется данный показатель. По каждой найденной величине рассчитывается статистика γ.

3. Выбирается переменная с минимальным значением γ.

4. Решается вопрос о целесообразности присутствия в модели выбранной в пункте 3 переменной. Если она не проходит проверку на значимость, то производится ее исключение из модели, после чего осуществляется переход к пункту 2 алгоритма, но уже из расчета, что указанная переменная в модели не присутствует. В противном случае, когда переменная оказывается значимой, алгоритм останавливается.

Для проведения исследования использовалась функция последовательного выбора признаков, на ее основе разработан сценарий, использующий функцию `sequentialfs` пакета Statistics and Machine Learning Toolbox системы MATLAB.

Обращение к функции имеет вид (для простоты необязательные параметры опущены)

```
inmodel=sequentialfs(@fun,X,y).
```

Здесь `@fun` – дескриптор функции `fun`, которая определяет критерий, используемый для выбора переменных. Входными данными для функции `sequentialfs` являются матрица независимых переменных `X` размерностью 15*109 и вектор зависимой переменной y размерностью 1*109. Строка матрицы `X` соответствует наблюдениям; столбцы матрицы соответствуют показателям. Элементы вектора *y* соответствуют целевому классу для каждого наблюдения в матрице *X*. Выходом `inmodel` является логическим вектор, указывающий, какие признаки окончательно выбраны. Начиная с пустого набора переменных функция `sequentialfs` создает кандидата в подмножество переменных путем последовательного добавления каждой из переменных. Для каждого кандидата в подмножество `sequentialfs` выполняет кросс-проверку, неоднократно вызывая функцию `fun` с различными обучающими и тестовыми подмножествами, следующим образом:

```
criterion = fun(XTRAIN,ytrain,XTEST,ytest).
```

В функцию `fun` передавались обучающие и тестовые выборки, полученные случайным выбором из входных данных: 50% для обучающей и 50% для тестовой выборки. Каждый раз при вызове функция возвращает скалярное значение критерия. Функция fun использует `XTRAIN` и `ytrain` для обучения модели и XTEST и `ytest` для проверки качества модели. В данной работе функция `fun` производит классификацию с использованием функции дискриминантного анализа `classify` и возвращает сумму ошибок классификации на каждом наборе переменных. В перекрестной проверке для каждого набора кандидатов `sequentialfs` суммирует значения, возвращаемые `fun`, и делит эту сумму на общее число тестовых наблюдений. Затем это среднее значение используется для оценки каждого кандидата в подмножество. После вычисления средних значений критерия для каждого кандидата в подмножество `sequentialfs` выбирает кандидата в подмножество, который минимизирует среднее значение критерия. Этот процесс продолжается до тех пор, пока добавление новых кандидатов не перестанет снижать критерий.

Эксперименты с методом Backward Elimination показали, что данный метод отобрал только один показатель: Общий билирубин (рис. 1).

```
Initial columns included:  none
Columns that can not be included:  none
Step 1, added column 10, criterion value 0.311927
Final columns included:  10

fs =

     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0


history =

      In: [0 0 0 0 0 0 0 0 0 1 0 0 0 0 0]
    Crit: 0.3119
```

Рис. 1. Результат работы алгоритма обратного исключения переменных

На рисунке показаны шаги уменьшения значения критерия, при котором новая переменная будет добавлена в модель. Конечный результат представляет собой уменьшенную модель только с одной из пятнадцати переменных: столбец 10. Эти переменные указаны в логическом векторе `fs`, возвращаемом `sequentialfs`.

Эксперименты с методом Forward Selection показали, что данный метод отобрал восемь показателей: Лейкоциты, Нейтрофилы палочкоядерные, Нейтрофилы сегментоядерные, Лимфоциты, Общий билирубин, Длительность оперативных вмешательств (рис. 2).

На следующем этапе исследования с помощью регрессионного анализа уточнялись выводы алгоритмов о выборе наиболее значимых независимых переменных. Расчеты по линейной регрессионной модели проводились с помощью статистического пакета StatSoft STATISTICA 10.0. В качестве независимых переменных использовали переменные, полученные отбором при помощи алгоритмов прямого отбора и обратного исключения. В качестве зависимой переменной выступал показатель отсутствия или наличия послеоперационных осложнений.

```
Initial columns included:  all
Columns that must be included:  none
Step 1, used initial columns, criterion value 0.33945
Step 2, removed column 3, criterion value 0.321101
Step 3, removed column 11, criterion value 0.293578
Step 4, removed column 2, criterion value 0.293578
Step 5, removed column 8, criterion value 0.293578
Step 6, removed column 15, criterion value 0.284404
Step 7, removed column 13, criterion value 0.284404
Step 8, removed column 1, criterion value 0.284404
Step 9, removed column 9, criterion value 0.284404
Step 10, removed column 12, criterion value 0.266055
Final columns included:  4 5 6 7 10 14
fs =

     0     0     0     1     1     1     1     0     0     1     0     0     0     1     0


history =

      In: [9x15 logical]
    Crit: [0.3394 0.3211 0.2936 0.2936 0.2936 0.2844 0.2844 0.2844 0.2844 0.2660]
```

Рис. 2. Результат работы алгоритма прямого отбора переменных

В ходе экспериментов моделью множественной линейной регрессии на данных, полученных от алгоритма прямого отбора, было распознано на обучающей выборке 54,02%, на тестовой – 45,45%. Моделью множественной линейной регрессии на данных, полученных от алгоритма обратного исключения, было распознано на обучающей выборке 18,39%, на тестовой – 04,55 %. Эксперименты показали, что метод прямого отбора лучше отслеживает взаимосвязи между переменными.

Библиографический список

1. Методы отбора переменных в регрессионные модели. URL: https: // basegroup.ru/community/articles/feature-selection

2. Kursa M. B., Rudnicki W. R. Feature Selection with the Boruta Package // Journal of Statistical Software. 2010, Vol. 36. Issue 11. P. 1–13.

3. Tuv E., Borisov A., Runger G., Torkkola K. Feature selection with ensembles, artificial variables, and redundancy elimination // The Journal of Machine Learning Research. 2009, Vol. 10. P. 1341–1366.

Горбаченко Владимир Иванович
Пензенский государственный университет, г. Пенза, Россия
E-mail: gorvi@mail.ru

Gorbachenko V.I.
Penza State University, Penza, Russia

Кузнецов Роман Николаевич
Пензенский государственный университет, г. Пенза, Россия
E-mail: nahab007@rambler.ru

Kuznetsov R.N.
Penza State University, Penza, Russia

Кузнецова Ольга Юрьевна
Пензенский государственный университет, г. Пенза, Россия
E-mail: ellekasandra@yandex.ru

Kuznetsova O.Yu.
Penza State University, Penza, Russia