УДК 519.237

Феофанов В. А.

Пошаговая дискриминация, кросс-валидация и бутстрап в задаче классификации пострадавших с сочетанной травмой груди

Рекомендовано к публикации профессором Буре В. М.

Введение. Математические методы находят широкое применение в различных областях науки, и медицина не является исключением [1]. Так, важным исследованием является задача предсказания летальности исхода пострадавшего от определенной болезни. В таких исследованиях применяется теория классификации с учителем [2]. В данной работе проводится построение классификационного правила для пострадавших с сочетанной травмой груди. Данные представляют особый интерес, так как количество описывающих признаков значимо больше, чем наблюдений. В связи с этим, в работе рассматривается пошаговая процедура отбора признаков, а также несколько различных метод оценки точности классификатора.

Методика. Одним из самых популярных методов классификации является линейный дискриминантный анализ (ЛДА) [3]. Лежащий в его основе принцип разделения популяций гиперплоскостями до сих пор находит широкое применение [2, 4]. Однако, в случае, когда число переменных превышает количество наблюдений, непосредственное его применение не даст желаемого результата. ЛДА — параметрический метод, для реализации которого приходится оценивать большое количество параметров, поэтому само по себе большое число признаков приводит, как правило, к ухудшению точности [5]. Следовательно, возникает необходимость отбора признаков. Одним из подходов, решающих эту проблему является пошаговый дискриминантный анализ или еще известный, как пошаговый MANOVA [5]. На выходе пошагового анализа — набор признаков, который не гарантировано является оптимальным [5]. Для оценки точности классификатора, а также для сравнения нескольких наборов признаков,

Феофанов Василий Алексеевич – студент, Санкт-Петербургский государственный университет; e-mail: vasya3000-95@mail.ru, тел.: +7(911)900-80-39

здесь будут использоваться такие подходы, как cross-validation leave-one-out, bootstrap leave-one-out, bootstrap 0.632 и 0.632+, которые изложены в [6].

Классификация пострадавших. В исследовании рассматриваются 52 пациента с сочетанной травмой грудной клетки, госпитализированные в экстренном порядке. База содержит общую информацию о пострадавших, результаты лабораторных и инструментальных исследований, проведенных в течение первых 12 часов с момента поступления в больницу пострадавшего, а также информация о том, удалось ли спасти пострадавшего, или нет. После того, как была проведена предварительная чистка данных, осталось 160 признаков, описывающих каждого пациента. Как уже отмечалось выше, такое соотношение числа наблюдений и признаков влечет за собой большую ошибку при непосредственной классификации. Поэтому, целесообразно провести процедуру отбора признаков, разделив переменные на несколько групп и в каждой найти «хорошо дискриминирующие» признаки. Было решено разбить базу на одиннадцать групп. Для удобства, группы были выбраны тематически:

Группа 1. Общие данные.
Группа 2. Объективный статус при поступлении и шкалы.
Группа 3. Шкалы.
Группа 4. Структура повреждений внутренних органов.
Группа 5. Параметры искусственной вентиляции легких (ИВЛ).
Группа 6. Вариабельность сердечного ритма (ВСР), вариабельность систолического артериального давления (ВСАД) и вариабельность диастолического артериального давления (ВДАД).
Группа 7. Вариабельность дыхания и общие данные спироартериокардиоритмографа (САКР).
Группа 8. Электрокардиография (ЭКГ).
Группа 9. Общий анализ крови (ОАК).
Группа 10. Биохимия.
Группа 11. Маркеры повреждения сердца и газы крови.

После проведения пошагового дискриминантного анализа в каждой группе была получена новая подгруппа признаков. Результаты отбора изображены на рис. 1.

Группа	Признак 1	Признак 2	Признак 3	Признак 4	Признак 5	Λ_{total}
Группа 1	Возраст	Сроки госпитализации				0,6
Группа 2	ВПХ - П (МТ)	Речевой контакт	Величина кровопотери	САД при поступлении	Частота пульса При поступлении	0,4
Группа 3	ВПХ - голова	ВПХ - грудь	ВПХ - таз	AIS - грудь	AIS - таз	0,41
Группа 4	Повреждение ЦНС САК					0,85
Группа 5	ИВЛ ДАД					0,92
Группа 6	ВСР LF (n. u.)	Вар. САД TP	Вар. САД HF	Вар. САД LF (n. u.)	Вар. САД VLF	0,43
Группа 7	САКР PQ	САКР ЧССср	САКР САДмакс	САКР ДАДср	САКР Вар. Дых. HF n.	0,36
Группа 8	ЭКГ RR					0,92
Группа 9	ОАК Гемоглобин					0,77
Группа 10	Биохимия Натрий					0,63
Группа 11	Газы крови FiO2	Газы крови RI				0,67

Рис. 1. Результат отбора признаков с итоговыми статистиками лямбда Уилкса

Затем по полученным подгруппам переменных была оценена вероятность ошибочной классификации с помощью методов кросс-валидации и бутстрап, упомянутых выше (рис. 2). Получилось, что наилучшую точность имеет классификатор, построенный по отобранным признакам из группы 6. Теперь попробуем улучшить точность, объединяя группы. Все отобранные признаки из групп 1–3 образуют группу A, из групп 4–6 — группу B, из групп 7–11 — группу C. После проведения пошаговой процедуры и оценки ошибки для этих групп, вновь отобранные переменные образуют финальную группу F. Итоговые результаты представлены на рис. 3.

Группы 1, 2 и 6 продемонстривовали хорошие результаты по отдельности, объединение первой и второй дало наилучший результат, тогда как признаки из шестой группы дополнительного улучшения не принесли. Факт того, что на группе F была получена меньшая точность, чем на группе A, подтверждает, что пошаговый дискриминантный анализ не выдает гарантированно оптимальный набор признаков.

Заключение. Исследовалась база данных пострадавших с сочетанной травмой груди. Для непосредственной классификации использовался линейный дискриминантный анализ, для отбора признаков — пошаговая дискриминация, для оценки ошибки — кросс-

Группа	Ошибка на обучении	Cross-validation Leave-one-out	Bootstrap Leave-one-out	Bootstrap 0.632	Bootstrap 0.632+
Группа 1	0,289	0,333	0,399	0,358	0,367
Группа 2	0,118	0,157	0,177	0,155	0,157
Группа 3	0,178	0,222	0,242	0,219	0,221
Группа 4	0,280	0,280	0,305	0,296	0,296
Группа 5	0,306	0,347	0,345	0,331	0,332
Группа 6	0,114	**0,114**	**0,164**	**0,146**	**0,147**
Группа 7	**0,074**	0,185	0,222	0,168	0,179
Группа 8	0,304	0,326	0,375	0,349	0,353
Группа 9	0,362	0,426	0,385	0,376	0,377
Группа 10	0,267	0,267	0,291	0,282	0,282
Группа 11	0,250	0,296	0,309	0,287	0,289

Рис. 2. Оценки вероятности ошибочной классификации для каждой группы

Группа	Признак 1	Признак 2	Признак 3	Признак 4	Признак 5	Признак 6	Признак 7	Λ_{total}
Группа A	ВПХ - П (МТ)	Речевой контакт	Возраст	Величина кровопотери	САД при поступлении	Сроки госпи-тализации		0,3
Группа B	ВСР LF (n. u.)	Вар. САД VLF	ИВЛ АДД					0,38
Группа C	Газы крови FiO2	Газы крови RI						0,77
Группа F	ВПХ - П (МТ)	Возраст	Вар. САД LF (n. u.)	Речевой контакт	Величина кровопотери	Сроки госпи-тализации	САД при поступлении	0,26

Группа	Ошибка на обучении	Cross-validation Leave-one-out	Bootstrap Leave-one-out	Bootstrap 0.632	Bootstrap 0.632+
Группа A	**0,044**	**0,044**	**0,099**	**0,079**	**0,08**
Группа B	0,116	0,163	0,161	0,144	0,145
Группа C	0,25	0,296	0,312	0,289	0,291
Группа F	0,075	0,1	0,124	0,106	0,107

Рис. 3. Финальный результат отбора и оценки ошибки

валидация и три бутстрап метода. Благодаря такому подходу, удалось добиться довольно высокой точности (90,1–95,6%) в ситуации, когда число признаков значительно превышает количество наблюдений. Более того, данный результат оказался лучше в сравнении с прошлогодним исследованием этой базы данных [7]. Методы оценки вероятности ошибочной классификации показали примерно одинаковые результаты. Вывод о том, какой из подходов дает наилучшую оценку в данном исследовании, представляет собой отдельную задачу, требующую тщательного анализа. Все рассматриваемые методы оценки вероятности ошибки не показывают надежных результатов на данных с малым числом объектов и большим количеством признаков. Некоторые исследования на этот счет были проделаны в работах [8, 9].

Литература

1. Буре В. М., Парилина Е. М., Рубша А. И., Свиркина Л. А. Анализ выживаемости по медицинской базе данных больных раком предстательной железы // Вестник Санкт-Петербургского университета. Серия 10: Прикладная математика. Информатика. Процессы управления. 2014. № 2. С. 27–35.

2. Буре В. М., Щербакова А. А. Применение дискриминантного анализа и метода деревьев принятия решений для диагностики офтальмологических заболеваний // Вестник Санкт-Петербургского университета. Серия 10: Прикладная математика. Информатика. Процессы управления. 2013. № 1. С. 70–76.

3. Рао С. Р. Линейные статистические методы и их применения / науч. ред. Линник Ю. В. / пер. с англ. Калинина В. М. и др. М.: Наука, 1968. 548 с.

4. Zavorka S., Perrett J. J. Minimum sample size considerations for two-group linear and quadratic discriminant analysis with rare populations // Communications in Statistics – Simulation and Computation. 2014. Vol. 43 (7). P. 1726–1739.

5. Rencher A. C. Methods of Multivariate Analysis. 2nd Ed. New York: John Wiley & Sons, Inc., 2002. 738 p.

6. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Ed. New York: Springer-Verlag, 2009. 745 p.

7. Семенчиков Д. Н. Классификация больных с тяжёлой сочетанной травмой грудной клетки // Процессы управления и устойчивость. 2015. Т. 2. № 1. С. 317–321.

8. Fu W. J., Carroll R. J., Wang S. Estimating misclassification error with small samples via bootstrap cross-validation // Bioinformatics. 2005. Vol. 21 (9). P. 1979–1986.

9. Molinaro A. M., Simon R., Pfeiffer R. M. Prediction error estimation: a comparison of resampling methods // Bioinformatics. 2005. Vol. 21 (15). P. 3301–3307.