УДК 519.254

© 2010 г. **Л.В. Горбов,** канд. мед. наук,
Л.А. Лазарева, канд. мед. наук
(Кубанский государственный медицинский университет, Краснодар),
Д.В. Лесик, канд. мед. наук,
А.А. Сухинин, канд. мед. наук
(Кубанский медицинский институт, Краснодар)

СТАТИСТИЧЕСКИЕ ПОДХОДЫ К АНАЛИЗУ ДАННЫХ НА ПРИМЕРЕ ДИСКРИМИНАНТНОЙ МОДЕЛИ У БОЛЬНЫХ С ОСТРОЙ НЕЙРОСЕНСОРНОЙ ТУГОУХОСТЬЮ

Методы многомерного статистического исследования во многих случаях позиционируется как «анализ данных». Тем не менее решение во многом зависит от конкретной реализации выборки, на которой оно получено. В работе показан алгоритм получения устойчивого решения дискриминантного анализа путем многократного моделирования обучающих выборок с последующим отбором устойчивых предикторов.

Ключевые слова: дискриминантный анализ, алгоритм получения устойчивого решения.

Введение

В последние десятилетия в медицинской информатике получили широкое распространение методы прикладной статистики, не аппелирующие к вероятностной природе обрабатываемых данных, нацеленные на геометрическую интерпретацию природы получаемых решений. К таким методам относят кластерный и факторный анализ, многомерное шкалирование, целенаправленное проецирование многомерных данных, анализ главных компонент и пр. Другие методы – такие как многомерные регрессионный и дискриминантный анализ – базируются на основе параметрической модели данных [1].

Вместе с тем, зачастую многие авторы рекомендуют [2] и используют эти методы анализа без учета возможного влияния конкретной реализации обучающей выборки на полученные результаты и их устойчивость. Чаще всего исследователи даже не делают попыток при проведении дискриминантного анализа разделить выборку на обучающую и контрольную группы и проверить результат применения решающего правила к независимым данным [3] или не считают необходимым сообщить об этом в разделе «Материалы и методы».

Целью настоящей работы явилась разработка алгоритма получения устойчивого результата дискриминантного анализа путем многократного случайного разделения выборки больных на обучающую и контрольную группы с проведением выбора предикатов, входящих в итоговую модель, и последующее определение соответствующих коэффициентов и свободного члена модели.

Материалы и методы. Объектом исследования явились результаты иммунограммы у больных с острой нейросенсорной тугоухостью (ОНСТ), полученные в первый день поступления в клинику. Как свидетельствуют наши наблюдения, часть больных с острой нейросенсорной тугоухостью, поступивших на лечение в стационар, к моменту выписки полностью восстанавливает слух. Однако также велика доля тех больных, кто выписывается из больницы с остаточным дефектом слухового анализатора. Принцип данного методического подхода был предложен В. Эфроном в 1979 году [4], однако широкого распространения не получил. Расчеты проводились с использованием программ «Excel 2003 MS» и «Statistica 6.0».

Выбор предикаторных переменных

Так как дискриминантный анализ является статистическим методом исследования, то предикаторные переменные и сила их влияния на прогноз являются случайными величинами. Для выбора переменных, которые должны быть включены в окончательную предсказывающую модель, была построена 21 промежуточная дискриминантная модель на обучающих выборках больных. Больные, не включенные в обучающие выборки, отбирались при помощи генератора случайных чисел (функция MS Excel «=RND()»). Из всей популяции обследованных лиц (больных и здоровых) в обучающую выборку случайным образом отбиралось 75% всех обследованных лиц. Так как показатели здоровых людей не участвовали в построении дискриминантной модели, то доля лиц, включенных в численный эксперимент, колебалась от 60,7 до 78,6%, а медиана и мода доли составили 71,4 и 75,0% всех больных соответственно. Количество обучающих выборок, равное 21, было взято произвольно, исходя из баланса между количеством вычислительной работы и желанием получить репрезентативные результаты.

Промежуточные модели были построены на основе всех изученных признаков – абсолютного количества (Ч10^6/л) лейкоцитов, нейтрофилов, лимфоцитов, моноцитов, эозинофилов, CD3, CD4, CD8, CD19, CD(16+56), отношения CD4/CD8 (ед.), а также концентрации иммуноглобулинов (г/л) классов A, M и G. Под CD… подразумеваются клетки крови, несущие на своей поверхности определенные комбинации макромолекул – кластеры дифференциации (**C**laster **D**ifferentiation), определяющие ее функциональную принадлежность – Т-лимфоциты, В-лимфоциты, натуральные киллеры (NK-клетки) и другие субпопуляции лимфоцитов. Построение промежуточных моделей производилось методом пошагового включения переменных (*Forward stepwise*).

В табл. 1 указаны результаты построения промежуточных моделей дискриминантного анализа (внизу таблицы указан уровень значимости различий, полученных в дисперсионном анализе при соответствующей кластеризации). При анализе полученных результатов видно, что ни один из признаков не входит в дис-

криминантную модель при анализе различных обучающих выборок каждый раз. Вероятность для признака участвовать в построении модели колеблется от 14 (CD3) до почти 91% (абсолютное содержание эозинофилов и CD(16+56)-клеток).

Таблица 1

Признак	Сколько раз признак включен в модель		Значение коэффициента		Отношение меньшего к большему	Кластер	Включение в итоговую модель
	число	кластер	>0	<0			
Лейкоциты$^{абс.}$	10	**1**	3	7	0,429	2	
Нейтр$^{абс.}$	11	**1**	9	2	0,222	**1**	**да**
Лимф$^{абс.}$	6	2	20	4	0,500	2	
Мон$^{абс.}$	10	**1**	1	9	0,111	**1**	**да**
Эоз$^{абс.}$	19	**1**	19	0	0,000	**1**	**да**
CD3$^{абс.}$	3	2	1	2	0,500	2	
CD4$^{абс.}$	9	2	1	8	0,125	**1**	
CD8$^{абс.}$	7	2	7	0	0,000	**1**	
CD4/CD8	5	2	0	5	0,000	**1**	
CD19$^{абс.}$	10	**1**	9	1	0,111	**1**	**да**
CD(16+56)$^{абс.}$	19	**1**	0	19	0,000	**1**	**да**
IgA	15	**1**	1	14	0,071	**1**	**да**
IgM	5	2	3	2	0,667	2	
IgG	6	2	6	0	0,000	**1**	
	p=0,000926				p=0,000001		

Понятно, что для улучшения качества разделения в качестве предикаторных переменных необходимо использовать те из них, которые наиболее часто включаются в дискриминантную модель. При этом не совсем очевидно, на каком уровне вероятности включения признака в модель следует остановиться – 90, 70, 50 или 40%. Для уменьшения волюнтаризма при выборе переменных нами использован кластерный анализ по методу k-средних по переменной «Сколько раз признак включен в модель» с разделением на две группы. В первый кластер вошли признаки, вероятность которых участвовать в построении модели составила более 47%. Очевидно, что выбор признаков для построения окончательной модели должен производиться именно из первого кластера.

Существует еще одно обстоятельство, накладывающее ограничение на использование переменной в итоговой моделей дискриминантного анализа. При реализации анализа на различных обучающих выборках коэффициенты дискриминантной функции могут менять знак. Очевидно, что наилучший результат дискриминации может быть получен в том случае, когда смена знака или не происходит, или происходит достаточно редко. Для поиска переменных коэффициентов дискриминантной функции, при которых не меняют знак, определено их (коэффициентов) количество меньшее и большее нуля и рассчитано отношение между наименьшим и наибольшим из этих чисел. Ясно, что чем ближе к нулю это отношение, тем реже коэффициент меняет знак (табл. 1).

Для того чтобы определить, какой уровень данного отношения представляется еще допустимым для включения переменной в итоговую дискриминантную

модель, нами повторно был использован кластерный анализ по методу k-средних по переменной «Отношение меньшего к большему» с разделением на две группы (табл. 1). В этом случае кластер, включающий оптимальные для успешной дискриминации переменные, также был обозначен номером один.

Очевидно, что наилучшими разделяющими свойствами будет обладать модель, включающая переменные, относящиеся к первому кластеру как при первой, так и при второй классификациях. Как видно из таблицы, это абсолютное количество нейтрофилов, эозинофилов, моноцитов, CD19 и CD(16+56)-клеток, а также концентрация иммуноглобулина А. Таким образом, включение этих переменных в модель дискриминантного анализа представляется наиболее вероятным.

Как мы уже убедились, использование различных обучающих выборок приводит к вариабельности предикатов и только выбор наиболее часто встречающихся из них с учетом наименьшей вероятности смены знака перед коэффициентом может привести к устойчивости предлагаемого решения дискриминантного анализа.

Определение коэффициентов дискриминантного уравнения

Еще одной задачей, без разрешения которой невозможно получить устойчивое решение, является поиск коэффициентов и свободного члена дискриминантного уравнения. Интуитивно понятно, что не только предикаты, но и значения коэффициентов перед ними в дискриминантном уравнении также зависят от конкретной реализации обучающей выборки. Как видно из табл. 2, во многих случаях имеет место даже смена знака коэффициентов.

Таблица 2

Признак	Коэффициент		Значение коэффициента		Отличие коэффициента от нуля
	значение	стандартное отклонение	>0	<0	
Нейтр$^{абс.}$	-0,16	2,72	15	6	0,7888
Мон$^{абс.}$	-12,00	15,06	2	19	**0,0016**
Эоз$^{абс.}$	83,44	47,66	21	0	**0,0000**
CD19$^{абс.}$	13,29	12,15	19	2	**0,0001**
CD(16+56)$^{абс.}$	-49,30	21,34	0	21	**0,0000**
IgA	-3,78	1,44	0	21	**0,0000**
Константа	24,45	25,22	21	0	**0,0003**

Представляется оправданным провести усреднение полученных коэффициентов с расчетом соответствующих ошибок и сравнение их с нулем. При отсутствии значимых отличий коэффициента в дискриминантном уравнении от нуля соответствующий предикат должен быть элиминирован.

С этой целью проведено повторное моделирование 21 обучающей выборки по вышеуказанному методу. При этом в модели дискриминантного анализа были использованы определенные выше предикаты. Изучали величину коэффициентов и значения константы дискриминантной функции (табл. 2).

При изучении представленных результатов обращает на себя внимание коэффициент, стоящий перед признаком «абсолютное количество нейтрофилов», абсолютная величина которого практически в 20 раз меньше его стандартного отклонения. Очевидной причиной этого является то, что почти в каждом 4 случае коэффициент меняет знак, в результате – среднее значение коэффициента оказывается близким к нулю.

Проверка значимости отличий коэффициентов и константы уравнения от нуля, проведенная с использованием t-критерия Стьюдента, показала достоверные отличия ($p<0,002$) для всех, кроме коэффициента признака «абсолютное количество нейтрофилов».

Как свидетельствуют основные теоретические предпосылки прикладного статистического анализа, включение в уравнение признаков с «нулевым» коэффициентом не улучшает качества решения, а вносит лишний «шум» в полученный статистический вывод [1]. Исходя из этого, признак «абсолютное количество нейтрофилов» необходимо исключить из итоговой дискриминантной модели.

Для проверки устойчивости определения коэффициентов дискриминантной модели, руководствуясь вышеописанными принципами, была сгенерирована еще 21 обучающая выборка. На этих выборках был проведен дискриминантный анализ с участием тех же признаков, кроме абсолютного количества нейтрофилов. Результаты приведены в табл. 3.

Как можно видеть из представленных данных, средние значения коэффициентов с учетом их стандартных отклонений в исходной модели и после исключения признака «абсолютное количество нейтрофилов» достоверно не отличаются друг от друга. Значения коэффициентов сравнивали при помощи t-критерия Стьюдента с учетом двусторонних различий. Коэффициент при исключаемом признаке сравнивали с константой, равной нулю.

Таблица 3

Признак	До		После		Отличие коэффициентов
	значение	стандартное отклонение	значение	стандартное отклонение	
Нейтр$^{абс.}$	-0,16	2,72	0	-	0,7903
Мон$^{абс.}$	-12,00	15,06	-7,62	7,28	0,2372
Эоз$^{абс.}$	83,44	47,66	73,95	20,33	0,4063
CD19$^{абс.}$	13,29	12,15	11,99	11,96	0,7291
CD(16+56)^{абс}	-49,30	21,34	-45,10	15,40	0,469
IgA	-3,78	1,44	-3,28	0,94	0,1899
Константа	24,45	25,22	18,95	6,06	0,3376

Таким образом, нами получен окончательный вид дискриминантной функции: **ДФ** = – 7,62ЧМон + 20,33ЧЭоз + 11,99ЧCD19 – 45,10ЧCD(16+56) – 3,28Ч IgA + +18,95, подставив в которую значения соответствующих признаков, можно прогнозировать благополучное течение заболевания в случае получения положительного результата (>0). Если значение функции окажется отрицательным, то с большой долей вероятности можно ожидать развития осложнений в виде снижения слуха.

При расчете значений дискриминантной функции с помощью табличного редактора MS Excel (рис. 1) по полной исследуемой выборке больных видно, что неправильно классифицированными оказывается около 3,5% всех наблюдений. Как видно из рисунка, небольшая часть больных, выписанных из стационара с дефектом слуха, решающим правилом отнесена к группе больных с благоприятным течением заболевания. В то же время в группе с прогнозируемым благоприятным исходом нет больных, выписанных с дефектом слухового анализатора.

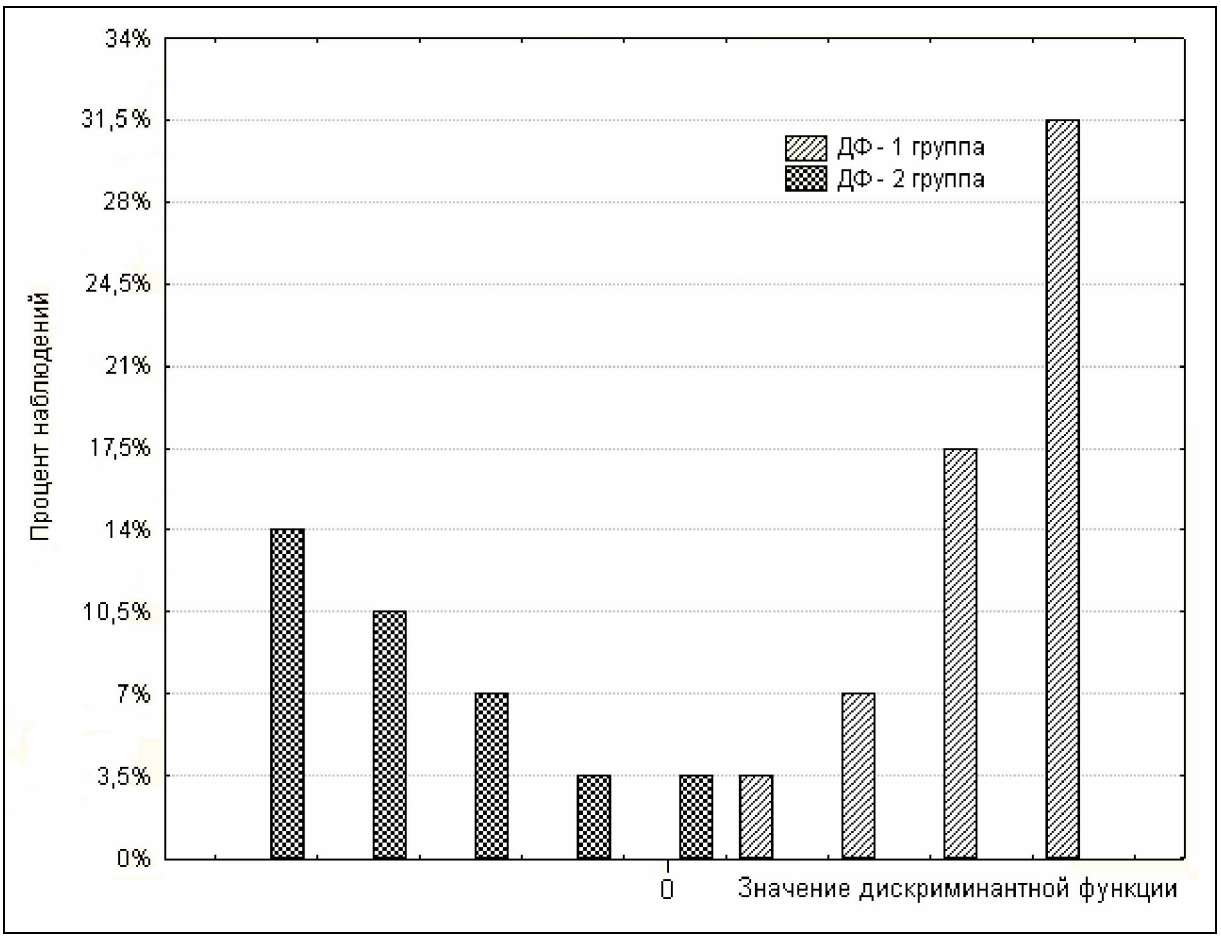

Рис. 1. Расчет значений дискриминантной функции
у всех обследованных больных с ОНСТ.

Такое решение может быть как свидетельством некоторой ошибки автоматизированного определения прогноза течения заболевания (вполне допустимой по величине), так и наличием ошибок в классификации врача-клинициста. Тем не менее, поскольку распределения больных обеих групп по величине дискриминантной функции не перекрываются (максимальное значение ДФ у неправильно классифицированного больного 2 группы составляет 1,96; тогда как минимальное значение ДФ у правильно классифицированных больных 1 группы превышает 5,04), простое уменьшение константы дискриминантной функции на 2,00 позволило бы полностью обеспечить 100% правильную дискриминацию в изученной выборке больных. Однако, так как по нашему глубокому убеждению, абсолютно правильное предсказание, с привлечением любых методов диагностики и анализа, представляется практически маловероятным, а искусственное занижение константы дискриминантного уравнения может в будущем привести к ошибке друго-

го рода – отнесении больного с положительным результатом лечения во 2 группу, мы решили не добиваться 100% точности дискриминации, допуская ошибку классификации менее 5%.

Кроме того, как можно видеть из рис. 1, в целом распределение значений дискриминантной функции носит U-образный характер, что способствует минимизации вероятности ошибочной классификации. Центры тяжести распределений стремятся отдалиться друг от друга по шкале дискриминантной функции, что обеспечивает успешность предлагаемого классификационного алгоритма.

Заключение

Несмотря на удовлетворительное развитие методов дискриминантного анализа, теоретически позволяющего получить диагностические и прогностические результаты, до настоящего времени его применение не обрело широкого распространения в практике. По-видимому, это связано с тем, что получаемое решение зависит от объема выборки, на которой получена дискриминантная функция. Кроме того, при ограниченном объеме выборки (менее нескольких сотен) дискриминантная функция также зависит от конкретной реализации выборки, т.е. от того, вошел или нет в нее тот или иной объект (предмет, человек и пр.).

Предложенный алгоритм построения дискриминантной модели, принимающий во внимание статистический характер анализируемых данных, уменьшает случайность в выборе предикаторных переменных, позволяя получить устойчивое решение удовлетворительного качества. Такой подход представляется особенно важным в случае использования дискриминантной модели в прикладных целях для задач диагностики и прогнозирования, поскольку позволяет избежать мало обоснованных решений, имеющих ограниченную применимость в случае использования на другой выборке больных.

ЛИТЕРАТУРА

1. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Классификация и снижение размерности.– М.: Финансы и статистика, 1989.
2. *Афифи А., Эйзен С.* Статистический анализ: Подход с использованием ЭВМ. – М.: Мир, 1982.
3. *Андреева Е.О, Корякина Л.Б., Курильская Т.Е. и др.* Дисфункции эндотелия у больных стенокардией напряжения II и III функционального класса.// Клиническая лабораторная диагностика. – 2008. – № 10. – С.15-17.
4. *Efron B.* Bootstrap methods another at the jackknife // Ann. Statist. – 1979. – Vol. 7, N 1.– P.1-26.

Статья представлена к публикации членом редколлегии Ю.М. Перельманом.

E-mail:
Горбов Л.В. – hamp2@rambler.ru