СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ ФАКТОРОВ, ВЛИЯЮЩИХ НА РАЗВИТИЕ СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ

Е.Ф. Гончарова [1], А.В. Гайдель [1,2], А.Г. Храмов [1]

[1] Самарский государственный аэрокосмический университет имени академика С.П. Королёва (национальный исследовательский университет) (СГАУ), Самара, Россия,
[2] Институт систем обработки изображений РАН, Самара, Россия

В работе исследуются различные признаки, влияющие на развитие фибрилляции предсердий у больных, проходящих лечение. Рассматривается метод отбора информационных признаков, основанный на дискриминантном анализе. В результате была выявлена степень влияния каждого отдельного признака на возможность классификации больных в зависимости от возможного (успешного или неуспешного) исхода лечения.

Ключевые слова: отбор признаков, дискриминантный анализ, критерий отбора.

Введение

В настоящее время наблюдается высокий уровень заболеваемости и смертности от сердечно-сосудистых заболеваний.

Для проведения диагностики, и впоследствии качественного лечения, необходимо рассматривать наиболее информативные признаки из всего набора данных, известных о пациенте.

В начале исследования имеется большой объем данных о каждом пациенте, поступающем на лечение. Для правильного проведения лечения врачу необходимо заранее знать потенциально возможный исход (смерть или выздоровление) для пациента, поступающего с определенными показателями здоровья. Однако необходимо учитывать не все многочисленные факторы, а лишь информативные для данной классификации[1], т.е. необходимо избавиться от несущественных и избыточных признаков. Таким образом, происходит снижение размерности пространства признаков. Можно выделить следующие причины, делающие задачу отбора признаков одним из важнейших этапов проведения медицинского исследования:

1) Для вычисления признаков часто расходуются временные, людские или материальные ресурсы. Чем больше признаков требуется оценить у объекта распознавания, тем сложнее это сделать. Снижение размерности пространства признаков позволяет существенно ускорить и удешевить процедуру распознавания.

2) При решении задачи классификации объектов на классы, классификатор не является идеальным, ошибка, которую он вносит в процедуру распознавания, сильно зависит от качества признакового пространства. Плохие шумовые признаки могут существенно затруднять построение качественного классификатора, поэтому для повышения эффективности распознавания нужно выделить набор признаков, в котором векторы признаков для объектов из различных классов сильно отличаются друг от друга и легко разделимы.

Различные методы отбора наиболее информативных признаков используются во многих работах по анализу биомедицинских изображений. Например, в работе [2] из 169 отдельных признаков, влияющих на диагностику хронической обструктивной болезни ХОБЛ, с помощью метода дискриминантного анализы, была выбрана группа из пяти признаков,

на которых достигается наилучшая в рамках выбранной процедуры ошибка классификации в 0,11.

В данной работе имеется информация о пациентах, страдающих фибрилляцией предсердий. Информация о каждом пациенте включает в себя множество факторов такие, как анкетные данные пациентов, результаты их медицинских анализов.

Целью данной работы является выявление группы наиболее информативных признаков методом дискриминантного анализа, влияющих на потенциальный исход лечения пациента для обеспечения возможности по показателям пациента заранее предсказывать опасность летального исхода.

1. Постановка задачи

Для решаемой нами задачи пусть имеется $\mathbf{\Omega}$ – множество объектов, подлежащих распознаванию. Это множество пациентов, которые могут быть равновероятно предъявлены нашей системе, или которые посещают врача в зависимости от конкретной цели исследования.

Множество было разбито на 2 класса с помощью разбиения $\mathbf{\Delta} = \left\{ \mathbf{\Omega}_j \right\}_{j=1}^{2}$. Классу $\mathbf{\Omega}_0$ принадлежат умершие пациенты, классу $\mathbf{\Omega}_1$ принадлежат выжившие пациенты.

Поскольку это разбиение, можно отметить два свойства:

1) $\bigcup_{j=1}^{2} \mathbf{\Omega}_j = \mathbf{\Omega}$,
2) $\forall i \neq j : \mathbf{\Omega}_i \cap \mathbf{\Omega}_j = \varnothing$.

В нашем случае имеется два класса: умершие пациенты и выжившие.

Было выбрано признаковое пространство $\mathbf{\Xi}(\mathbf{Q})$. Векторы признаков $x \in \mathbf{\Xi}(\mathbf{Q})$ – это случайные векторы, распределённые некоторым образом.

В качестве рассматриваемых признаков было выбрано 20 признаков:

1. Возраст
2. Форма ФП
3. Длительность ФП
4. Сопутствующие патологии
5. Давность КЭИ
6. Шкала HAS-BLED
7. Шкала NISSH
8. Длительность АГ
9. Уровень знаний
10. Изменение жизни
11. Риск инсульта
12. Прием аспирина
13. Важность приема антикоагулянтов
14. ПВ
15. Протромбин
16. АЧТВ

17. Фибриноген
18. САД
19. ДАД
20. Монр. шк.

В данной группе признаков «форма ФП» – форма фибрилляции предсердий (1 - постоянная, 2 - пароксизмальная), «длительность ФП» (1 - менее года, 2 - от 1-го года до 5-ти лет, 3 - более 5-ти лет), сопутствующая патология (1 - сахарный диабет, 2 - ишемическая болезнь сердца, 3 - хроническая болезнь почек, 4 - инфаркт миокарда), «Шкала HAS-BLED» (максимум – 9 баллов, сумма баллов ≥ 3 указывает на высокий риск кровотечения, и применение любого антитромботического препарата требует особой осторожности, результат, 2 бала и менее — нет значимого повышения риска кровотечений, но необходим тщательный контроль). Шкала NISSH (0 - состояние удовлетворительное, 3–8 - неврологические нарушения легкой степени, 9–12 – неврологические нарушения средней степени, 13–15 - тяжелые неврологические нарушения, 16–34 – неврологические нарушения крайней степени тяжести, 34 – кома), давность КЭИ –давность кардиоэмболического инсульта, длительность АГ – длительность получения антикоагулянтной терапии, ПВ - протромбиновое время (лабораторный показатель, определяемые для оценки внешнего пути свёртывания крови), АЧТВ – активированное частичное тромбопластиновое время, САД и ДАД соответственно нижнее и верхнее артериальное давление, монр. шк. – монреальская шкала когнитивной оценки (максимально возможное количество баллов – 30, 26 баллов и более считается нормальным).

2. Отбор признаков с помощью дискриминантного анализа

Пусть каждому объекту поставлен в соответствие вектор его признаков $x(k)$.

Определим на этом множестве оператор $\Phi(\omega):\mathbf{\Omega}\to\mathbf{\Delta}$, который переводит объект распознавания в его класс.

Также существует обучающая выборка $\mathbf{U}\subseteq\mathbf{\Omega}$, для объектов которой нам известен класс $|\mathbf{U}|$=40.

Сначала стандартизуем вектор признаков для каждого отдельного признака. Для этого необходимо оценить математическое ожидание:

$$M(k)=\frac{1}{|\mathbf{U}|}\sum_{x\in\mathbf{U}}x(k)$$

и коэффициент корреляции:

$$R(k,l)=\frac{1}{|\mathbf{U}|}\sum_{x\in U}\Big(x\big(k\big)-M(k)\Big)\Big(x(l)-M(l)\Big)$$

для каждого вектора признака.

Затем воспользуемся формулой:

$$x=\left(\frac{x(k)-M(k)}{\sqrt{R(k,k)}}\right).$$

Таким образом, векторы признаков будут иметь нулевое математическое ожидание и единичную дисперсию.

Для выявления наиболее информативных признаков оценим внутриклассовое математическое ожидание $\mathrm{E}\left\{x \mid \mathbf{\Omega}_j\right\}$:

$$\overline{x}_j = \frac{1}{\left|\mathbf{U}\right|} \sum_{\omega \in \mathbf{U} \cap \mathbf{\Omega}_j} \Psi\left(\omega\right),$$

и внутриклассовую корреляционную матрицу $\mathrm{E}\left\{\left(x-\overline{x}\right)^T\left(x-\overline{x}\right) \mid \mathbf{\Omega}_j\right\}$:

$R_j = \frac{1}{\left|\mathbf{U}\right|} \sum_{\omega \in \mathbf{U} \cap \mathbf{\Omega}_j} \left(\Psi\left(\omega\right)-\overline{x}_j\right)^T \left(\Psi\left(\omega\right)-\overline{x}_j\right)$ кроме того, предположим, что нам известны априорные вероятности $\mathrm{P}\left(\mathbf{\Omega}_j\right)$ появления объектов из каждого класса $\mathrm{P}\left(\mathbf{\Omega}_0\right) = \frac{7}{40}$ и $\mathrm{P}\left(\mathbf{\Omega}_1\right) = \frac{33}{40}$.

Далее можем определить математическое ожидание смеси распределений по следующей формуле:

$$\overline{x} = \sum_{j=1}^{2} \overline{x}_j \, \mathrm{P}\left(\Omega_j\right)$$

и корреляционную матрицу смеси распределений

$$R = R_0 \, \mathrm{P}\left(\Omega_0\right) + R_1 \, \mathrm{P}\left(\Omega_1\right).$$

Определим критерий, соответствующий одному из критериев дискриминантного анализа[3], рассеяния смеси распределений как след корреляционной матрицы $\mathrm{tr}\, R$, а рассеяние внутри *j*-го класса – как $\mathrm{tr}\, R_j$. Выбранный набор признаков тем лучше, чем сильнее рассеяние смеси распределений превышает среднее внутриклассовое рассеяние:

$$J\left(\mathbf{Q}\right) = \frac{\mathrm{tr}\, R}{\sum_{j=1}^{2} \mathrm{P}\left(\Omega_j\right) \mathrm{tr}\, R_j}.$$

Рассмотри жадный алгоритм выбора оптимального набора признаков.

Пусть изначально множество признаков $\mathbf{Q}_{(0)} = \varnothing$. Для каждого признака вычислим его индивидуальный критерий качества $J_j = J\left(\left\{j\right\}\right)$, который зависит от того как *j*-й признак разделяет векторы признаков в одномерном признаковом пространстве. Упорядочим признаки в порядке убывания критерия J_j и на очередном шаге будем жадно добавлять в очередное множество $\mathbf{Q}_{(i)} = \varnothing$ очередной признак с наибольшим значением критерия качества из ещё не добавленных:

$$\mathbf{Q}_{(i)} = \mathbf{Q}_{(i-1)} \cup \left\{ \underset{j \in [1;K] \cap \mathbf{Z} \setminus \mathbf{Q}_{(i-1)}}{\arg\max} J_j \right\}.$$

В соответствии с вышеизложенным алгоритмом были получены результаты, представленные в таблице 1. Записи в таблице расположены в порядке убывания индивидуального критерия.

Табл. 1. Критерий индивидуальных признаков

Номер	Признак	*J(Q)*
1	АЧТВ	34.980
2	Протромбин	32.472
3	ПВ	27.369
4	Важность приема антикоагулянтов	25.579
5	Прием аспирина	24.230
6	Шкала NISSH	19.712
7	САД	19.106
8	Сопутствующие патологии	18.910
9	Фибриноген	18.820
10	Шкала HAS-BLED	18.813
11	Монр. шк.	18.538
12	Уровень знаний	18.049
13	Длительность ФП	17.982
14	Давность КЭИ	17.811
15	Длительность АГ	17.712
16	Форма ФП	17.641
17	ДАД	17.609
18	Изменение жизни	17.607
19	Возраст	17.601
20	Риск инсульта	17.600

Для первых трех признаков рассчитаем среднее значение этих признаков в каждом классе и разброс значений. Результаты представлены в таблице 2.

Табл. 2. Сравнение признаков

Класс	Средн. значение	Разброс
Ω_0	38.071 66.814 15.129	1.515 7.128 0.607
Ω_1	30.191 97.336 13.064	3.304 13.760 1.147

Анализ данных из таблицы 2 показывает, что первые три признака действительно оказывают влияние на классификацию объектов на соответствующие классы выживших и умерших пациентов, наблюдается значительное различие средних значений данных признаков внутри каждого класса, с учетом разброса.

Заключение

В результате проведенного исследования был рассмотрен метод отбора признаков с помощью дискриминантного анализа. Жадный алгоритм отбора информативных признаков показал, что наиболее информативными в данной задаче являются первые признаки, указанные в таблице 1. Так, например, средние значения признаков «Активированное ча-

стичное тромбопластиновое время», «Протромбин» и «Протромбиновое время» в классе Ω_0 ($\overline{x^0_{АЧТВ}}$ = 38.071, $\overline{x^0_{Протр.}}$ = 66.814, $\overline{x^0_{ПВ}}$ = 15.129) имеют суще-ственное различие от средних значений соответствующих признаков из класса Ω_1 ($\overline{x^1_{АЧТВ}}$ = 30.191, $\overline{x^1_{Протр.}}$ = 97.336, $\overline{x^1_{ПВ}}$ = 13.064). В то же время, последние признаки не вносят большой вклад в изменение критерия J. Например, такие параметры как «Возраст» и «Нижнее артериальное давление» имеют небольшие различия в средних значениях внутри каждого класса ($\overline{x^0_{Возраст}}$ = 98.0, $\overline{x^0_{ДАД}}$ = 68.857 для класса Ω_0; $\overline{x^1_{Возраст}}$ = 97.697, $\overline{x^1_{ДАД}}$ = 69.152 для класса Ω_1), следовательно, они не играют решающей роли при проведении классификации объектов. Таким образом, проведенное исследование показало, что при построении классификатора, способного предсказать угрозу смерти пациента или же возможность его выздоровления необходимо учитывать именно наиболее информативные признаки, в то время как признаки, представленные нижними строками таблицы 1, не несут полезной информации для данной классификации, и их не следует рассматривать при построении классификатора.

Благодарности

Работа выполнена при поддержке гранта РФФИ 14-07-97040-р_поволжье_а и Министерства образования и науки РФ в рамках мероприятий Программы повышения конкурентоспособности СГАУ среди ведущих мировых научно-образовательных центров на 2013-2020 годы, а также Программы фундаментальных исследований ОНИТ РАН «Биоинформатика, современные информационные технологии и математические методы в медицине».

Литература

1. Дронов, С.В. Многомерный статистический анализ: Учебное пособие. – Барнаул: Издательство Алтайского государственного университета, 2003. – 213 с.
2. Гайдель, А.В. Возможности текстурного анализа компьютерных томограмм в диагностике хронической обструктивной болезни / А.В. Гайдель, П.М. Зельтер, А.В. Капишников, А.Г. Храмов // Компьютерная оптика. – 2014. – Т. 38, № 4. – С. 843-850.
3. Fukunaga, K. Introduction to statistical pattern recognition / K. Fukunaga. – San Diego: Academic Press, 1990. – 592 p.