ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ ПРОГРАММЫ ОБУЧЕНИЯ ПАЦИЕНТОВ С ФИБРИЛЛЯЦИЕЙ ПРЕДСЕРДИЙ НА ОСНОВЕ МЕТОДОВ ОТБОРА ПРИЗНАКОВ

В.В. Кутикова [1], А.В. Гайдель [1,2], А.Г. Храмов [1,2]

[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 данной работе исследуется эффективность обучающей терапевтической программы школы «Стоп Инсульт», направленной на минимизацию факторов риска развития инсульта у пациентов с фибрилляцией предсердий. На основе двух методов отбора признаков, использующих критерий дискриминантного анализа для определения наилучшего подмножества признаков, был сделан вывод о том, что пациенты, принимавшие участие в обучающих занятиях школы, имеют более высокий уровень знаний о фибрилляции предсердий и длительное время принимают антикоагулянты, в отличие от пациентов, не проходивших обучение.

Ключевые слова: интеллектуальный анализ данных, отбор признаков, критерий дискриминантного анализа.

Введение

Снижение размерности признакового пространства является одной из центральных проблем интеллектуального анализа данных. Для решения большинства задач классификации и восстановления регрессии требуется выбрать наилучшее подмножество из заданного множества признаков. Это требование обусловлено тем, что использование большого количества признаков не только вычислительно трудоёмко, но и влияет на качество распознавания, так как могут быть использованы малоинформативные и избыточные признаки, усложняющие процесс принятия решений.

Методы отбора информативных признаков широко используются при анализе биомедицинских данных. Так, в работе [1] с помощью метода отбора, основанного на ковариационном анализе, из 22216 признаков, характеризующих уровни экспрессии различных генов, был найден биомаркер рака лёгких у курильщиков, включающий всего 80 признаков. Достоверность классификации с помощью полученного биомаркера составила 83 %. Для отбора из небольшого количества признаков допустимо использовать полный перебор [2]. В данной области также нашли своё применение последовательные алгоритмы поиска [3] и генетические алгоритмы [4], а методы отбора признаков, основанные на критерии дискриминантного анализа, показали свою эффективность в работах [5, 6] для анализа биомедицинских изображений.

Целью данной работы является исследование эффективности обучающей терапевтической программы школы «Стоп Инсульт», направленной на минимизацию факторов риска развития инсульта у пациентов с фибрилляцией предсердий (ФП). Исходные данные содержат наблюдения по 12 признакам, а в качестве классов выступают группы пациентов: основная группа и группа сравнения. В основную группу входят пациенты, принимавшие участие в обучающих занятиях школы, а в группу сравнения – те, кто наблюдался у врача, но не проходил обучение. Данные получены во время двух визитов пациентов к врачу: до и после проведения обучающих занятий.

Для определения эффективности обучающих занятий на основе данных, полученных во время второго визита, отбираются признаки, которые наилучшим образом разделяют исходные данные на две группы; проводится сравнение информативности полученных признаков для двух визитов, после чего делаются выводы об эффективности обучающих занятий.

В качестве основных инструментов исследования используются метод отбора признаков, основанный на оценивании информативности отдельных признаков с помощью критерия дискриминантного анализа, и метод, основанный на последовательном поиске лучшего подмножества признаков, также использующий критерий дискриминантного анализа.

1. Описание методов исследования

Упорядочение признаков в соответствии с критерием дискриминантного анализа

Согласно описанному в [4] методу, признаки упорядочиваются в соответствии с критерием дискриминантного анализа [7]:

$$J = \frac{\operatorname{tr} S_m}{\operatorname{tr} S_w}, \qquad (1)$$

где $\operatorname{tr} S_w$ – след матрицы рассеяния внутри классов, $\operatorname{tr} S_m$ – след матрицы рассеяния смеси распределений.

Матрица рассеяния внутри классов показывает разброс объектов относительно векторов математических ожиданий классов:

$$S_w = \sum_{i=1}^{2} p_i M\left\{\left(X^{(i)} - M_i\right)\left(X^{(i)} - M_i\right)^T\right\},$$

где p_i – априорные вероятности классов, $X^{(i)}$ – наблюдения, принадлежащие *i*-му классу, M_i – вектор математического ожидания смеси распределений.

Матрица рассеяния смеси характеризует разброс объектов относительно векторов математических ожиданий смеси

$$S_m = M\left\{\left(X - M_0\right)\left(X - M_0\right)^T\right\},$$

где $M_0 = p_1 M_1 + p_2 M_2$ – математическое ожидание смеси.

Чем больше значение критерия (1), тем лучше признак разделяет объекты, принадлежащие разным классам.

Последовательный поиск наилучшего подмножества признаков

Описанный выше подход к отбору признаков позволяет оценить качество каждого признака в отдельности, но не учитывает зависимости между признаками. Некоторые из них могут быть бесполезными сами по себе, но информативными при совместном использовании с другими признаками.

Последовательный алгоритм позволяет проводить поиск по пространству подмножеств признаков. Основная идея состоит в следующем: начиная с некоторого начального подмножества, на каждом шаге осуществляется переход в соседнее состояние, в котором либо из текущего подмножества удаляется один элемент, либо в него добавляется один элемент.

В данной работе переход осуществляется в соседнее подмножество, которому соответствует наибольшее значение критерия (1) среди подмножеств такой же размерности. Кроме того, используется методика «два шага вперёд, один назад», то есть через каждые два шага, на которых проводится добавление нового признака в текущее подмножество, необходимо исключить один признак.

Оценка эффективности программы обучения

Пусть J_0 – значение критерия (1) для некоторого подмножества признаков, полученное на основе данных первого визита, а J_1 – значение критерия, рассчитанное по данным второго визита. В качестве оценки эффективности программы обучения для рассматриваемого подмножества признаков используется отношение

$$E = \frac{J_1}{J_0}. \qquad (2)$$

На признаки, для которых $E > 1$, школа оказала «положительное» влияние, то есть по окончании обучающих занятий данные признаки стали более информативными. В случае, когда $E < 1$, обратная ситуация – признаки стали менее информативными, а на признаки, для которых $E = 1$, школа не оказала никакого влияния.

На основе полученных оценок (2) и значений внутригрупповых средних в данной работе делается вывод об эффективности обучающих занятий школы «Стоп Инсульт» для рассматриваемого подмножества признаков.

1. Экспериментальное исследование эффективности обучающих терапевтических программ школы «Стоп Инсульт»

В ходе исследования использовались 12 признаков: ответы на вопросы анкеты, которую заполняли участники обеих групп до и после проведения обучающих занятий, артериальное давление (систолическое или САД, диастолическое или ДАД), параметры гемостаза (протромбиновое время или ПВ, протромбин, фибриноген, активированное частичное тромбопластиновое время или АЧТВ). Набор данных включал 69 наблюдений (36 наблюдений, полученных от пациентов из основной группы, и 33 наблюдения – от пациентов из группы сравнения) для каждого визита пациентов к врачу.

Табл. 1. Эффективность терапевтических обучающих занятий для отдельных признаков

№	Признак	J_1	J_0	E
1	Антикоагулянтная терапия (0 – не принимает, 1 – принимает менее года, 2 – от 1 до 5 лет, 3 – более 5 лет)	9,35	0,97	9,64
2	Как вы оцениваете свой уровень знаний о ФП? (1 – низкий, 5 – высокий)	5,21	0,98	5,33
3	Насколько важно регулярно принимать препарат для профилактики инсульта в соответствии с назначениями? (1 – совершенно неважно, 5 – очень важно)	3,07	3,52	0,87
4	Как вы оцениваете свои знания о риске инсульта как об основном осложнении ФП? (1 – низкие, 5 – высокие)	2,58	1,03	2,50
5	Протромбин	1,19	1,08	1,10

	(в процентах)			
6	САД (в мм рт. ст.)	1,13	0,99	1,14
7	ПВ (в секундах)	1,13	0,99	1,14
8	АЧТВ (в секундах)	1,07	1,03	1,03
9	Фибриноген (г/л)	1,02	1,00	1,02
10	Аспирин (1 – принимает, 0 – не принимает)	1,02	1,02	1,00
11	ДАД (в мм рт. ст.)	0,99	0,99	1,00
12	Насколько диагноз ФП изменил вашу повседневную жизнь? (1 – совсем не изменил, 5 – изменил в значительной степени)	0,99	0,99	1,00

Эффективность терапевтических обучающих занятий для отдельных признаков

В табл. 1 представлены результаты оценки эффективности программы обучения школы «Стоп Инсульт» для отдельных признаков. Для каждого признака указано значение критерия (1), рассчитанного по данным первого J_0 и второго J_1 визитов, а также значение эффективности *E*, рассчитанное по формуле (2). Признаки упорядочены по убыванию значений J_1. В табл. 2 приведены средние значения для признаков, указанных в табл.1.

Табл. 2. Средние значения признаков внутри групп

№	Основная группа		Группа сравнения	
	Визит 1	Визит 2	Визит 1	Визит 2
1	0,06	2,06	0,12	0,12
2	1,47	4,31	1,55	1,48
3	4,08	4,15	1,39	1,61
4	1,49	4,28	1,89	2,08
5	88,58	84,02	97,34	95,18
6	166,94	140,78	168,48	144,3
7	13,13	13,78	13,06	12,98
8	31,70	31,90	30,19	30,01
9	4,66	4,26	4,58	4,37
10	0,94	0,94	0,79	0,79
11	98,17	87,47	97,70	88,24
12	3,02	3,56	3,06	3,30

Согласно табл. 1, после проведения обучающих занятий школы признаки 1, 2, 4 стали хорошо разделять группы пациентов, что не наблюдается при первом визите, то есть значение эффективности E достаточно большое для данных признаков. Из табл. 2 видно, что средние значения внутри основной группы для признаков 1, 2, 4 увеличились от первого визита ко второму, а в группе сравнения изменились незначительно. Следовательно, обучающие терапевтические занятия оказались эффективными для признаков 1, 2, 4.

Видно также, что для признака под номером 3 значение J_0 больше значения J_1, то есть данный признак до начала обучающих занятий разделял группы лучше, чем после окончания занятий. Такой эффект объясняется тем, что средние значения внутри обеих групп от первого визита ко второму немного изменились в лучшую сторону (пациенты обеих групп стали сознавать важность приёма препарата) и вместе с тем увеличилась внутригрупповая дисперсия признака.

Эффективность терапевтических обучающих занятий для подмножеств признаков

В табл. 3 приведены результаты оценки эффективности программы обучения школы «Стоп Инсульт» для подмножеств, полученных на основе последовательного метода отбора признаков. В табл. 3 для 10 лучших подмножеств признаков из табл. 1 указаны значения J_0 и J_1 критерия дискриминантного анализа, а также значение эффективности E.

Видно, что все десять подмножеств после завершения обучающих занятий школы разделяют группы лучше, чем до начала занятий. Однако первые 3 подмножества, состоящие из признаков 1, 2, 10 имеют большее значение эффективности E. Учитывая, что средние значения признаков 1 и 2 для основной группы от первого ко второму визиту увеличились, а для группы сравнения практически не изменились, можно сделать вывод о том, что обучающие занятия оказались эффективными для данных признаков. То есть, пациенты, принимавшие участие в обучающих занятиях школы, имеют более высокий уровень знаний о ФП и длительное время принимают антикоагулянты, в отличие от пациентов, не проходивших обучение.

Табл. 3. Эффективность терапевтических обучающих занятий для подмножеств признаков

Признаки	J_1	J_0	***E***
1	9,35	0,97	9.64
1, 2	6,01	0,98	6,14
1, 2, 10	5,20	0,98	5,29
1, 2, 3, 10	4,08	2,16	1,89
1, 2, 3, 9, 10	3,59	1,95	1,84
1, 2, 3, 4, 9, 10	3,29	1,72	1,91
1, 2, 3, 4, 7, 9, 10	2,63	1,46	1,81
1, 2, 3, 4, 7, 9, 10, 12	2,18	1,32	1,66
1, 2, 3, 4, 7, 8, 9, 10, 12	1,42	1,10	1,29
1, 2, 3, 4, 7, 8, 9, 10, 11, 12	1,24	1,05	1,18

Заключение

В работе на основе методов отбора признаков исследовалась эффективность обучающей терапевтической программы школы «Стоп Инсульт». На основе данных о пациентах, полученных после проведения обучающих занятий школы, отбирались признаки, которые наилучшим образом разделяют пациентов основной группы и группы сравнения. На основе значений эффективности и внутригрупповых средних для отобранных признаков были сделаны выводы об эффективности обучающей программы в целом.

Исследование эффективности обучающих занятий для отдельных признаков показало, что обучение пациентов в школе «Стоп Инсульт» эффективно для признаков «Антикоагулянтная терапия» (*E = 9,62*), «Как вы оцениваете свой уровень знаний о ФП?» (*E = 5,33*) и «Как вы оцениваете свои знания о риске инсульта как об основном осложнении ФП?» (*E = 2,50*). Похожие результаты были получены при оценивании эффективности занятий для подмножеств признаков. В этом случае обучающая программа школы оказалась эффективна для пары признаков «Антикоагулянтная терапия» и «Как вы оцениваете свой уровень знаний о ФП?» (*E = 6,14*). Другими словами, пациенты, принимавшие участие в обучающих занятиях школы, имеют более высокий уровень знаний о ФП и дли-

тельное время принимают антикоагулянты, в отличие от пациентов, не проходивших обучение.

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. Spira, A. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer / A. Spira, J.E Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y.-M. Dumas, P. Calner, P. Sebastiani, S. Sridhar, J. Beamis, C. Lamb, T. Anderson, N. Gerry, J. Keane, M. E Lenburg, J. S Brody // Nature Medicine. – 2007. – Vol. 13, № 3. – P. 361-366.
2. Ильясова, Н.Ю. Формирование признаков для повышения качества медицинской диагностики на основе методов дискриминантного анализа / Н.Ю. Ильясова, А.В. Куприянов, Р.А. Парингер // Компьютерная оптика. – 2014. – Т. 38, № 4. – С. 851-855.
3. Peng, Y. A novel feature selection approach for biomedical data classification / Y. Peng, Z. Wu, J. Jiang // Journal of Biomedical Informatics. – 2010. – Vol. 43. – P. 15-23.
4. Tsai, C.-F. Genetic algorithms in feature and instance selection / C.-F. Tsai, W. Eberle, C.Y. Chu // Knowledge-Based Systems. – 2013. – Vol. 39. – P. 240-247.
5. Гайдель, А.В. Исследование текстурных признаков для диагностики заболеваний костной ткани по рентгеновским изображениям / А.В. Гайдель, С.С. Первушкин // Компьютерная оптика. – 2013. – Т. 37, № 1. – С. 113-119.
6. Кутикова, В.В. Исследование методов отбора информативных признаков для задачи распознавания текстурных изображений с помощью масок Лавса / В.В. Кутикова, А.В. Гайдель // Компьютерная оптика. – 2015. – Т. 39, № 5. – С. 744-750.
7. Fukunaga, K. Introduction to statistical pattern recognition / K. Fukunaga. – San Diego: Academic Press, 1990. – 592 p.