

КЛАССИФИКАЦИЯ ВРАЧЕБНЫХ УЧАСТКОВ ГОРОДСКОГО АДМИНИСТРАТИВНОГО РАЙОНА ПО УРОВНЮ ЗАБОЛЕВАЕМОСТИ ДЕТЕЙ НА ОСНОВЕ МЕТОДОВ МНОГОМЕРНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА

В.Н. Коровин, О.В. Родионов

В статье представлены результаты классификации врачебных участков детской поликлиники № 5 Железнодорожного района города Воронежа на основе кластерного и дискриминантного анализа, в результате чего получены классификационные функции, которые используются для оценки риска заболеваемости детей по врачебным участкам в городском административном районе

Ключевые слова: детская заболеваемость, городской район, кластерный и дискриминантный анализ

Для классификации врачебных участков детской поликлиники № 5 Железнодорожного района города Воронежа по набору медицинских показателей, характеризующие состояние здоровья детей использовался кластерный анализ. Кластерный анализ представляет собой специфическую методологию проведения классификации неоднородных статистических совокупностей. Основная цель анализа – выделить в исходных многомерных данных такие однородные подмножества, чтобы объекты внутри групп были похожи в известном смысле друг на друга, а объекты из разных групп – не похожи. Под “похожестью” понимается близость объектов в многомерном пространстве признаков, и тогда задача сводится к выделению в этом пространстве естественных скоплений объектов, которые и считаются однородными группами [1, 2].

Проблема измерения близости объектов возникает при любых трактовках кластеров и различных методов классификации. Основные трудности, возникающие при этом: неоднозначность выбора способа нормировки и определения расстояния между объектами.

Расстоянием (метрикой) между объектами в пространстве параметров называется такая величина d_{ab} , которая удовлетворяет аксиомам:

$$A1. d_{ab} > 0, d_{aa} = 0;$$

$$A2. d_{ab} = d_{ba};$$

$$A3. d_{ab} + d_{bc} \geq d_{ac} \text{ (неравенство треугольника)}.$$

Мерой близости (сходства) обычно называется величина μ_{ab} имеющая предел и возрастающая с возрастанием близости объектов.

Существует возможность простого перехода от расстояний к мерам близости: достаточно применить, например, преобразование

$$\mu = 1/(1+d). \quad (1)$$

Не вполне строго будем считать, что адекватной является статистика, либо не меняющаяся при допустимых преобразованиях шкал, либо меняющаяся контролируемым способом.

Матрица расстояний или близости нередко задается непосредственно: либо как таблица экспертных оценок близости, либо как матрица прямых измерений сходства: межотраслевого баланса, степеней соседства географических регионов, взаимной цитируемости авторов и т. д. В таких случаях все поставленные выше проблемы адекватности расстояний и выбора мер сходства снимаются.

В настоящее время существует огромное количество алгоритмов кластер-анализа. Они отражают разнообразие не только вычислительных приемов, но и концепций, стоящих за ними.

Наиболее естественный путь нахождения образов заключается в том, что дается точное определение образа и отыскивается скопление точек, обладающее соответствующими свойствами. Например, образ (кластер) можно определить как такое скопление точек, в котором среднее межточечное расстояние меньше среднего расстояния от данных точек до остальных. Поэтому будем считать, что основой первого направления решения задачи структурной классификации является формулировка понятия кластера и разбиение совокупности на части, каждая из которых представляет собой кластер в данном смысле. Такой подход часто называется эвристическим. Однако многие свойства этих процедур изучены достаточно хорошо, а некоторые из алгоритмов находят локальный

Коровин Владимир Николаевич – ВГТУ, аспирант, тел. (4732) 46-76-99

Родионов Олег Валерьевич – ВГТУ, д-р техн. наук, профессор, тел. (4732) 46-76-99

экстремум определенному функционалу. Поэтому назовем группу алгоритмов, ориентированных на выделение кластеров с заранее заданными свойствами, процедурами прямой классификации [1, 2].

Основной чертой таких процедур является использование ими только одного понятия кластера. Скажем, в группе методов k -средних объекты попадают в тот класс, расстояния, до центра которого минимальны, т. е. реализуется одно из определений кластера. Это означает, что все классы разбиения будут удовлетворять именно этому определению. Если предположить, что некоторые исходные данные в самом деле имеют причудливый вид, то алгоритмы этого типа не смогут их разделить. Поэтому крайне интересно создать процедуры комбинированной прямой классификации, которые бы выделяли классы в смысле нескольких определений, т. е. подыскивали бы для каждого скопления свойственное ему определение кластера [1].

Требование к хорошей классификации предъявляют не только в терминах определений отдельных кластеров. Часто общее представление о качестве классификации формулируется в виде некоторого функционала, экстремальное значение которого соответствует наилучшей классификации.

Это второе оптимизационное направление в решении задачи кластер-анализа сформировалось позже первого и представляет богатую и разветвленную дисциплину. Остановимся на некоторых методических аспектах, касающихся связи двух подходов.

Оптимизационное направление пытается задачу кластер-анализа ввести в традиционное математическое русло, четко сформулировать критерий и добиваться его экстремизации. При этом, естественно, возникают чисто математические проблемы: определения свойств функционала, путем достижения оптимума, трудоемкости алгоритма [2].

Наконец можно выделить третье направление решения задачи кластеризации, наиболее позднее по срокам развития. Его можно называть аппроксимационным. Основная идея подхода заключается в следующем: отношения, заложенные в исходных данных, требуются наилучшим образом аппроксимировать отношением, отвечающим нашему представлению о классификации. Классификация обычно задает отношение эквивалентности, а исходные данные могут быть отражены по-разному.

В работе кластерный анализ применялся для классификации врачебных участков детской поликлиники № 5 Железнодорожного района города Воронежа по основным заболеваниям среди детей [3]. В качестве меры близости использовалось расстояние Евклида. Обработка проводилась при помощи пакета Statistica 5.5. Результат кластеризации врачебных участков детской поликлиники Железнодорожного района по основным заболеваниям детей приведен на рис. 1.

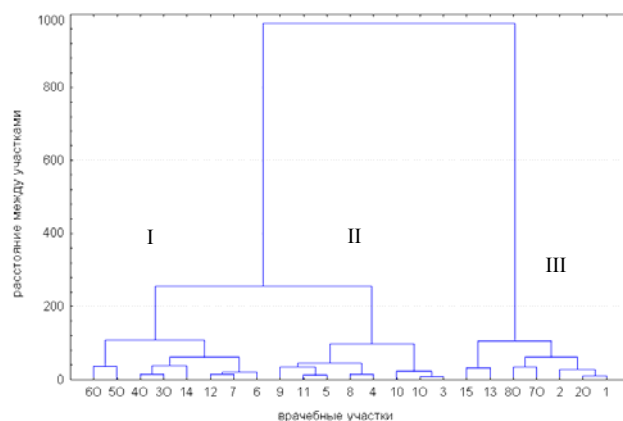


Рис. 1. Дендрограмма распределения врачебных участков Железнодорожного района на классы по основным заболеваниям у детей, где, I – класс участков с низким уровнем заболеваний; II – класс участков со средним уровнем заболеваний; III – класс участков с высоким уровнем участков

Для более точной классификации врачебных участков детской поликлиники № 5 и выделения трех классов в зависимости от уровня заболеваемости детей был применен метод k -средних, результаты которого представлены в табл. 1.

Таблица 1
Кластеры врачебных участков детской поликлиники № 5 Железнодорожного района по основным заболеваниям детей

№ класса	Участки
1	6, 7, 12, 14, 30, 40, 50
2	3, 4, 5, 8, 9, 10, 11, 10, 60
3	1, 2, 13, 15, 20, 70, 80

Средние количественные показатели уровней рассматриваемых заболеваний для каждого кластера приведены в табл. 2.

Как видно из представленных результатов, третий кластер характеризуется высоким уровнем заболеваний, практически по всем нозологическим формам, за исключением заболеваний системы кровообращения и болезней костно-мышечной системы и соединительной ткани, а первый кластер, наоборот, характеризуется

Таблица 2

Численные показатели основных нозологических форм для кластерных групп

Нозология	Уровень заболеваний (на 1000 детей)		
	1 кластер	2 кластер	3 кластер
Инфекционные и паразитарные болезни	47,22±16,08	37,75±17,96	51,13±14,72
Болезни крови	6,65±0,47	7,07±0,33	8,02±0,72
Болезни эндокринной системы	15,41±0,49	16,61±0,39	19,34±0,88
Болезни нервной системы	25,63±0,83	27,38±0,69	32,01±1,39
Болезни глаза и его придаточного аппарата	58,67±1,71	63,33±1,73	74,21±2,77
Болезни уха и сосцевидного отростка	46,37±1,31	50,25±1,38	58,76±2,27
Болезни системы кровообращения	1,12±0,22	1,94±0,70	1,61±0,28
Болезни органов дыхания	621,40±17,44	673,76±19,64	789,05±28,75
Болезни органов пищеварения	39,18±4,02	44,06±1,34	48,58±5,31
Болезни органов пищеварения	39,18±4,02	44,06±1,34	48,58±5,31
Болезни кожи и подкожной клетчатки	19,01±0,54	20,57±0,56	24,07±0,92
Болезни костно-мышечной системы и соединительной ткани	21,25±4,03	36,46±13,43	30,08±4,98
Болезни мочеполовой системы	14,80±0,42	16,04±0,43	18,75±0,73
Отдельные состояния возникшие в перинатальном периоде	32,75±0,88	35,79±1,06	41,74±1,67
Врожденные аномалии	21,16±7,19	17,05±8,07	23,01±6,61
Прочие заболевания	48,98±4,93	55,06±1,61	60,68±6,61
Отравления и травмы	41,88±1,48	56,17±1,49	65,61±2,63

низким уровнем заболеваний практически по всем рассматриваемым нозологическим формам, за исключением инфекционных заболеваний и врожденных аномалий.

Таким образом, результаты кластерного анализа показали, что на 1, 2, 13, 15 врачебных участках, а также на 2, 7 и 8 врачебных участках Отрожки отмечается наиболее высокий уровень заболеваемости детей практически по всем нозологическим формам.

Для уточнения и более детального исследования распространенности заболеваемости у детей по врачебным участкам детской поликлиники № 5 Железнодорожного района города Воронежа был применен дискриминантный анализ, позволяющий при помощи соответствующих дискриминантных классификационных функций распределять объекты, в данном случае врачебные участки по классам в зависимости от риска заболеваемости детей. В данном случае, имея показатели по рассматриваемым нозологическим формам для определенного врачебного участка, например прогнозные значения, и соответствующие дискриминантные классификационные функции можно определить врачебный участок в тот или иной класс, по уровню риска заболеваемости, не зная общей ситуации в целом по району, т.е. достаточно сведений только по конкретному врачебному участку.

Дискриминантный анализ является разделом многомерного статистического анализа,

который позволяет изучать различия между двумя и более группами объектов по нескольким переменным одновременно [4].

Для получения дискриминантных функций предложен вид в виде линейной комбинации для каждого класса (классифицирующая функция). Она имеет следующий вид:

$$h_k = b_{k0} + b_{k1}X_1 + b_{k2}X_2 + \dots + b_{kp}X_p \quad (2)$$

где h_k – значение функции для класса k ; X_i – значение i -ой дискриминантной переменной; p – количество дискриминантных переменных; b_{ki} – коэффициент, который необходимо определить.

Объект относится к классу с наибольшим значением классифицирующей функции. Коэффициент для классифицирующей функции определяется по формуле:

$$b_{ki} = (n - g) \sum_{j=1}^p a_{ij} X_{jk} \quad (3)$$

где n – общее число объектов; g – число классов; a_{ij} – коэффициенты матрицы, обратной к внутригрупповой матрице сумм попарных произведений W_{ij} :

$$W_{ij} = \sum_{k=1}^g \sum_{m=1}^p (X_{ikm} - X_{ik})(X_{jkm} - X_{jk}) \quad (4)$$

где X_{ikm} – значение i -ой переменной для m -го наблюдения в k -м классе; X_{ik} – среднее значение i -ой переменной в k -м классе; X_i – значение

i -ой переменной по всем классам (общее среднее).

Постоянный член определяется по формуле:

$$b_{k0} = -0,5 \sum_{j=1}^p b_{kj} X_{jk} \quad (5)$$

Для проверки статистической значимости чаще всего используют понятие остаточной дискриминантной способности (способность переменных различать классы до определения функции, если исключить информацию, полученную с помощью ранее вычисленных функций). Для этого используется Λ -статистика Уилкса [4]:

$$\Lambda = \prod_{i=k+1}^g \frac{1}{1 + \lambda_i}, \quad (6)$$

где k – число функций; λ_i – собственное значение матрицы коэффициентов функций.

Λ -статистика Уилкса может принимать значения в диапазоне $[0; 1]$, причем при $\Lambda = 0$ имеем высокую дискриминацию, то есть классы хорошо различаются, соответственно $\Lambda = 1$ говорит о низкой дискриминации.

В результате проведенного дискриминантного анализа по основным, наиболее распространенным заболеваниям среди детей по данным детской поликлиники № 5 Железнодорожного района: X_1 – инфекционные и паразитарные болезни; X_2 – болезни глаза и его придаточного аппарата; X_3 – болезни уха и сосцевидного отростка; X_4 – болезни органов дыхания; X_5 – болезни органов пищеварения; X_6 – прочие заболевания; X_7 – травмы и отравления были получены следующие виды дискриминантных классификационных функций для каждого класса (H_1 – низкий уровень заболеваемости у детей; H_2 – средний уровень заболеваемости; H_3 – высокий уровень заболеваемости):

$$H_1 = -447,645 - 1,144 * X_1 + 35,405 * X_2 - 87,004 * X_3 + 1,746 * X_4 + 63,484 * X_5 - 50,610 * X_6 + 34,845 * X_7,$$

$$H_2 = -539,884 - 1,395 * X_1 + 31,322 * X_2 - 53,826 * X_3 + 1,577 * X_4 + 41,436 * X_5 - 32,405 * X_6 + 13,155 * X_7,$$

$$H_3 = -722,051 - 1,507 * X_1 + 41,158 * X_2 - 80,489 * X_3 + 1,821 * X_4 + 78,453 * X_5 - 62,531 * X_6 + 26,524 * X_7.$$

Значение критерия Уилкса: $\Lambda = 0,003153$ стремиться к нулю, что говорит о хорошем различии между классами. Значение F – статистики, связанной с критерием Уилкса, F

$(14,28) = 9,2633$, уровень значимости $P < 0,00005$ [5].

Распределение врачебных участков по классам на основе дискриминантного анализа представлено на рис. 2. Значения вероятностей попадания каждого врачебного участка в тот или иной класс на основе дискриминантных функций представлены в табл. 3.

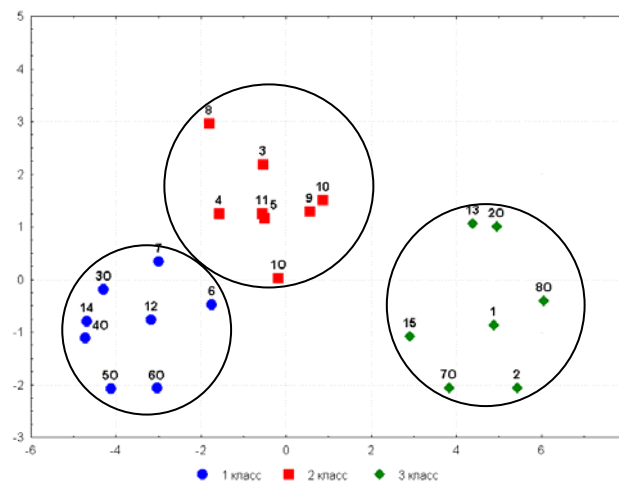


Рис. 2. Диаграмма рассеивания врачебных участков по плоскости дискриминантных функций по основным заболеваниям

Таблица 3
Вероятность распределения врачебных участков в классы по уровню заболеваемости детей

Врачебный участок	p_1	p_2	p_3	Номер класса	Уровень заболеваний
1	0	0	1	3	высокий
2	0	0	1	3	высокий
3	0,0001	0,9999	0	2	средний
4	0,0242	0,9758	0	2	средний
5	0,0011	0,9989	0	2	средний
6	0,7075	0,2925	0	1	низкий
7	0,9452	0,0548	0	1	низкий
8	0,0009	0,9991	0	2	средний
9	0,0001	0,9998	0,0001	2	средний
10	0	0,9999	0,0001	2	средний
11	0,0011	0,9989	0	2	средний
12	0,9976	0,0024	0	1	низкий
13	0	0,0001	0,9999	3	высокий
14	0,9999	0,0001	0	1	низкий
15	0	0,0008	0,9992	3	высокий
1O	0,0055	0,9944	0,0001	2	средний
2O	0	0,0001	0,9999	3	высокий
3O	0,9997	0,0003	0	1	низкий
4O	0,9999	0,0001	0	1	низкий
5O	0,9999	0,0001	0	1	низкий
6O	0,9999	0,0001	0	1	низкий
7O	0	0,0001	0,9999	3	высокий
8O	0	0	1	3	высокий

Результаты дискриминантного анализа подтвердили общую обстановку по основным заболеваниям среди детей Железнодорожного района города Воронежа, в некоторой степени скорректировав результаты кластерного анализа и распределив врачебные участки детской поликлиники № 5 в три класса, в зависимости от уровня риска заболеваемости детей.

Литература

1. Дюран Н., Оделл П. Кластерный анализ. М.: Статистика, 1987.
2. Мандель И.Д. Кластерный анализ. М.: ФиС, 1988.
3. Жабин В.А., Коровин Е.Н., Нехаенко Н.Е. Классификация врачебных участков детской поликлиники муниципального района городского округа по уровню заболеваемости детей на основе кластерного анализа // Актуальные вопросы организации,

оказания первичной, специализированной медицинской помощи в условиях многопрофильного стационара и на догоспитальном этапе: Материалы 3 научно-практической конференции, посвященной 20-летию ГКБСМП № 10. Воронеж, 2009.

4. Факторный, дискриминантный и кластерный анализ: Пер. с англ. Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка и др.; Под. ред. И.С. Енюкова. М.: Финансы и статистика, 1989.

5. Жабин В.А., Коровин Е.Н., Нехаенко Н.Е. Методы рационального управления системой медицинского обслуживания детей городского административного района на основе оценок комфортности проживания и риска заболеваемости // Системный анализ и управление в биомедицинских системах: журнал практической и теоретической биологии и медицины. М., 2009. Т.8. №. 1.

Воронежский государственный технический университет

CLASSIFICATION OF MEDICAL SITES CITY ADMINISTRATIVE AREA ON THE LEVEL OF DISEASE OF CHILDREN ON THE BASIS OF METHODS OF THE MULTIVARIATE STATISTICAL ANALYSIS

V.N. Korovin, O.V. Rodionov

In article results of classification of medical sites of a children's polyclinic 5 Zheleznodorozhniy areas of city of Voronezh on the basis of cluster and discriminant analyses therefore classification functions which are used for an estimation of risk of disease of children on medical sites in city administrative area are received are presented

Keywords: children's disease, city area, cluster and the discriminant analysis