ВАРИАНТЫ КЛАССИФИКАЦИИ ПО БАЙЕСУ ДЛЯ ПРОГНОЗА РАЗВИТИЯ ПНЕВМОНИИ ПРИ ОСТРЫХ ОТРАВЛЕНИЯХ ПСИХОТРОПНЫМИ СРЕДСТВАМИ

А.Н.Ельков, К.К.Ильяшенко

НИИ скорой помощи им. Н.В. Склифосовского,
ИПМ РАН им. М.В.Келдыша, Москва
Работа выполнена при поддержке гранта РФФИ; код проекта 99–01–00029

ВВЕДЕНИЕ

Пневмония при острых экзогенных отравлениях является одним из наиболее опасных осложнений, значительно ухудшающих течение заболевания. Привлечение средств прикладной математики для раннего прогноза пневмонии может способствовать повышению эффективности ее лечения.

Нами исследованы истории болезни двух выборочных совокупностей больных с острыми отравлениями психотропными и снотворными средствами (ПСС). Первую выборку (*A*) составили 129 пациентов 1986–92 гг. Из них 79 без осложнений (I группа) и 50 с развившейся не позднее 3-х суток пневмонией (II группа). Состояние каждого больного из этой выборки оценивали по 63 клинико-лабораторным признакам в 1-е сутки болезни. Вторая выборка (*B*), сформированная, исходя из результатов исследований выборки *A*, состоит из 15 больных с пневмонией (1996–99 гг.). В ней рассматривали всего два параметра, анализ соотношения между которыми и является целью данной работы.

Рассмотрим выборку *A*.

Исходя из трактовки прогноза пневмонии как статистической задачи распознавания образов ясно, что первым этапом ее решения должен быть поиск признаков, существенных для прогноза. Алгоритм поиска таких признаков может быть основан, в частности, на следующих исходных положениях:

1. В медицинских задачах распознавания важнейшей является дихотомия, когда требуется отнести предъявленный объект к одной из *двух* групп (классов).
2. Для каждого осложнения существует свой набор клинико-лабораторных показателей, на которые врач прежде всего обращает внимание при его диагностике.
3. Формирование диагноза базируется на последовательном переборе врачом клинико-лабораторных показателей и фиксации их отклонений от нормы.
4. В предположении, что характер корреляционной связи инвариантен по отношению к классу, к которому принадлежит предъявленный больной (хотя, как будет показано ниже, это не всегда справедливо), диагностическое суждение является наиболее эффективным, если существенные для прогноза показатели статистически независимы.

Поэтому один из естественных (хотя, конечно, не наилучший) алгоритмов отбора прогностически важных признаков можно сформулировать следующим образом: среди всех характеризующих больного признаков путем последовательного их перебора следует найти такие, которые имеют достоверно различающиеся внутригрупповые выборочные распределения.

Все признаки можно разделить на два типа: измеренные в [квази]непрерывных шкалах (лабораторные показатели) и дискретные (симптомы). В данном исследовании непрерывные признаки сначала оценивали на соответствие их выборочных распределений нормальному закону, после чего для выявления групповых различий применяли два критерия – Стьюдента и ω^2. Сравнение внутригрупповых частот дискретных признаков проводили посредством построения таблиц сопряженности 2×2 и вычисления χ^2. В зависимости от принятого уровня достоверности можно получить различные наборы прогностически важных признаков.

Сначала сравнения проводили при (p<0.1). В результате был сформирован следующий набор из 9 прогностически важных признаков: длительность глубокой комы, центральное венозное давление, лейкоциты (L), сегментоядерные нейтрофилы (S), напряжение кислорода в крови, иммуноглобулин-G, индуцированный НСТ-тест, наличие трахеобронхита и экспозиция яда. Следует подчеркнуть, что полученный набор полностью согласуется с интуицией врачей-токсикологов.

Следующим этапом была проверка статистической независимости полученной совокупности признаков в группах I и II. Для этого были вычислены две внутригрупповые матрицы парных корреляций размерностью 9×9. При анализе полученных матриц выяснилось, что выборочные корреляционные моменты (r) достоверно (p<0.05) отличны от нуля только в трех случаях. Два из них не представляют, на наш взгляд, интереса, т.к. в них $|r| \le 0.35$, т.е. уровень корреляции достаточно низок. Третий случай является основной темой настоящего исследования.

Согласно выборочным данным в I группе наблюдается высокая корреляционная связь ($r \approx 0.7$) между переменными L и S. В то же время во II группе корреляция между ними достоверно отсутствует.

Отметим теперь, что только для двух из девяти существенных для прогноза пневмонии признаков различия внутригрупповых средних получены при 0.05<p<0.1. Это признаки S и иммуноглобулин-G. Таким образом, если исключить их из первоначального набора, то получим множество из 7 статистически независимых (p<0.05) признаков, каждый из которых имеет различающиеся внутригрупповые выборочные распределения. Это позволяет, свернув получившееся признаковое пространство по схеме "замороженного" последовательного анализа [4], свести задачу прогноза пневмонии к простейшей одномерной задаче распознавания образов. В работе [4] показано, что получающийся в результате байесовский классификатор имеет теоретическую ошибку прогноза менее 10%.

Однако данная схема прогноза имеет ряд недостатков [5]. Главный из них состоит в том, что, несмотря на достаточно большое число наблюдений в выборке A, проверить на ней сформулированный на ее же основе алгоритм невозможно – по причине отсутствия больных, у которых измерены все существенные для прогноза пневмонии признаки. Среднее число имеющихся значений таких признаков у одного больного приблизительно равно 4. При этом нельзя исходить из того, что на этапе практического применения алгоритма на его входе окажется в среднем больше информации чем в фазе исследования.

Таким образом желательно иметь критерий прогноза, базирующийся на легко доступных клинико-лабораторных признаках, количество которых при сохранении приемлемого качества прогноза должно быть сведено к минимуму. Возможность удовлетворительного решения этой задачи следует из отмеченного выше резкого различия характера корреляционной связи между лейкоцитами и сегментоядерными нейтрофилами в группах больных без пневмонии и с пневмонией и того известного факта, что среднее число лейкоцитов у больных без пневмонии достоверно ниже такового при ее наличии.

Введем некоторые обозначения. Пусть ω_1 это класс больных с острыми отравлениями ПСС и неосложненным течением заболевания, а ω_2 – с развившейся в процессе патогенеза пневмонией. Рассмотрим плоскость, на которой введена прямоугольная система координат xOy таким образом, что ось x соответствует лейкоцитам, а ось y – сегментоядерным нейтрофилам (вместо самих признаков L и S будем рассматривать их безразмерные аналоги). Измеренное состояние больного теперь можно изобразить на плоскости в виде точки $\bar{x} = (x, y)$. Введем в рассмотрение случайный вектор $\bar{X}_k = (X_k, Y_k)^*$, реализациями которого будут точки (x, y). Пусть априорные условные плотности вероятности случайных векторов $\bar{X}_k$ (в классах ω_k)

* – Всюду далее индекс k есть номер класса, причем он везде принимает лишь два значения – k=1, 2, и всякое выражение, в которое входит k, справедливо как по отношению к объектам из w_1, так и для объектов из w_2.

представлены соответственно функциями $p_k(\bar{x})$ а вероятности появления в клинической практике больных этих классов равны P_k.. Запишем уравнение границы областей принятия решений байесовского классификатора:

$$\frac{p_2(\bar{x})}{p_1(\bar{x})} = \lambda\,, \text{ где } \quad \lambda = \frac{P_1}{P_2}\,,. \tag{1}$$

АППРОКСИМАЦИЯ НОРМАЛЬНЫМ РАСПРЕДЕЛЕНИЕМ

Допустим, что $\bar{X}_k$ подчинены двумерному нормальному закону распределения вероятности. В таком случае их функции плотности можно записать в виде

$$p_k(\bar{x}) = \frac{1}{2\pi\sqrt{|D_k|}}\ exp\left\{\frac{1}{2}(\bar{x}-\bar{a}_k)'C_k(\bar{x}-\bar{a}_k)\right\}, \tag{2}$$

где $\bar{a}_k = (a_{X,k}\,,\, a_{Y,k})$ – вектор средних значений, D_k – ковариационная матрица, C_k обратная к ней. Подставив функции (2) в уравнение (1) и прологарифировав обе части, его несложно привести к виду (см.[3])

$$\frac{1}{2}ln\frac{|D_2|}{|D_1|} - \frac{1}{2}tr[C_1(\bar{x}-\bar{a}_1)(\bar{x}-\bar{a}_1)'+\frac{1}{2}tr[C_2(\bar{x}-\bar{a}_2)(\bar{x}-\bar{a}_2)' = ln\ \lambda \tag{3}$$

Последнее уравнение в случае двух измерений легко преобразуется к общему уравнению кривой второго порядка на плоскости $Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0$, коэффициенты которого определяются компонентами векторов $\bar{a}_k$ и элементами матриц D_k и C_k (см. [5]). Как обычно, после переноса начала координат в точку O'=(p ,q), задаваемую условиями $Ap + Bq = -D,\ Bp + Cq = -E$ и последующего поворота осей координат вокруг нового начала на угол $\alpha = 0.5\ arctg[2B/(A-C)]$, будем иметь в координатах $x'O'y'$:

$$A'x'^2 + C'y'^2 + F' = 0\,. \tag{4}$$

Здесь $A' = A\cos\alpha + 2B\sin\alpha\cos\alpha + C\sin\alpha$, $C' = A\sin\alpha - 2B\sin\alpha\cos\alpha + C\cos\alpha$,

$F' = Ap^2 + 2Bpq + Cq^2 + 2Dp + 2Eq + F$.

Приведя (4) к каноническому виду, окончательно запишем уравнение (1) в новых координатах $x'O'y'$ в виде

$$\frac{x'^2}{a^2} \pm \frac{y'^2}{b^2} = 1\,, \text{ где } a^2 = \frac{-F'}{A'} \text{ и } b^2 = \frac{-F'}{C'}\,. \tag{5}$$

Тип определяемой последним уравнением кривой (парабола, гипербола, эллипс и т.д.) остается открытым до интерпретации модели.

Визуализация аппроксимаций внутригрупповых выборочных распределений показателей L и S групп больных I и II выборки A представлена на рис.1. Более размытая плотность соответствует группе с пневмонией.

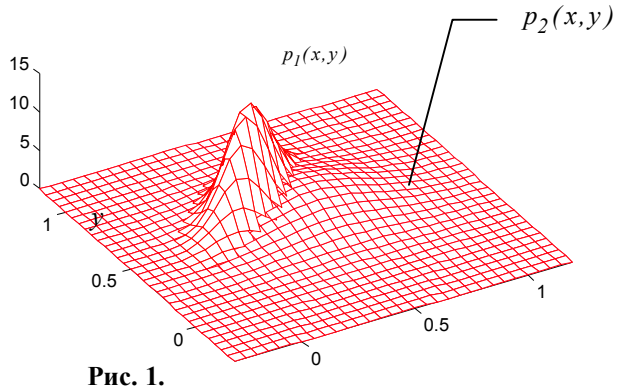

Рис. 1.

На рис.2 представлена проекция линии пересечения двух нормальных плотностей, изображенных на рис.1, на плоскость xOy. Эллипсы I и II (пунктир) суть линии уровня аппроксимаций распределений вероятностей соответственно для классов больных ω_1 è ω_2. Размеры полуосей в обоих случаях равны соответствующим среднеквадратическим отклонениям в системах координат, связанных с главными осями.

Третий эллипс (сплошная линия, обозначен буквой $\mathfrak{I}$) есть решение уравнения (1) при $P_1=P_2=0.5$. Он является разделяющей границей между областями решений байесовского классификатора A_1 (часть плоскости, ограниченная эллипсом) и A_2 (часть плоскости вне эллипса). Если предъявляемый образ попадает в область A_1, то принимается решение о его принадлежности классу ω_1 (пневмония не прогнозируется), в противном случае он должен быть отнесен к классу ω_2, что говорит о большой вероятности наличия пневмонии у больного, послужившего источником этого образа.

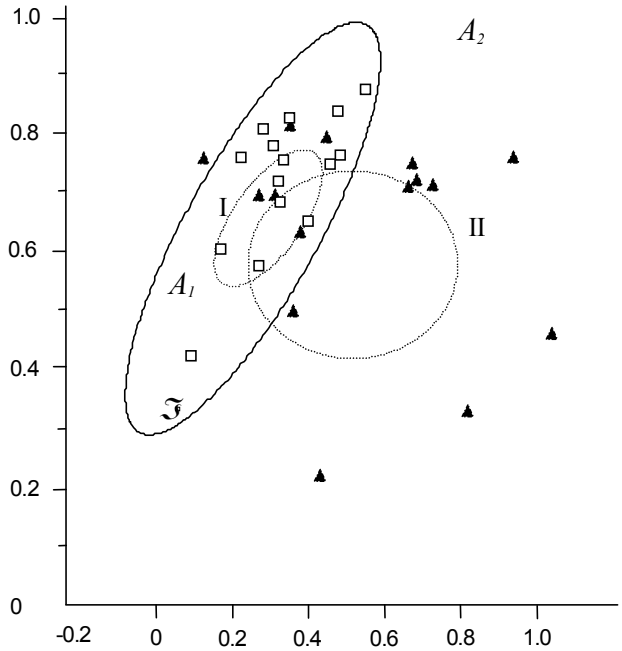

Рис. 2.

На рисунке отображены также 30 наблюдений (по 15 из каждой группы больных; белые прямоугольники – больные без пневмонии, черные треугольники – больные с пневмонией). Видно, что *все* экспериментальные точки, характеризующие больных, относящихся к классу ω_1, попали внутрь области A_1. Что касается класса ω_2, то здесь результат гораздо хуже, т.к. из 15 точек только 10 находятся вне области A_1 (существенно, что одна из них расположена левее и выше границы A_1). Таким образом, если в классе ω_1 экспериментальная ошибка прогноза равна нулю, то в классе ω_2 она равна 1/3. Взятая сама по себе, последняя дробь характеризует качество прогноза во втором классе на первый взгляд как слишком низкое. Однако, при условии высокой точности прогноза внутри класса ω_1, невысокая точность прогноза в классе ω_2 является вполне допустимой. В самом деле, отсутствие ошибок при опознавании объектов класса ω_1 означает малую вероятность ложной тревоги. Следовательно, если алгоритм все таки отнес предъявленного к распознаванию больного к классу ω_2, то вероятность наличие пневмонии является очень высокой. Ясно, что такой подход в любом случае увеличивает количество информации, имеющейся в распоряжении врача, и позволяет начать лечебные мероприятия в более ранние сроки

АППРОКСИМАЦИЯ ПРОИЗВЕДЕНИЕМ КРИВЫХ ПИРСОНА

Нормальное распределение, являясь приближенным представлением выборки, не исчерпывает всех возможных способов ее описания, хотя, при корректном подходе, позволяет судить о характеристиках наблюдаемого выборочного распределения. Важным атрибутом двумерного нормального закона являются главные оси эллипса рассеивания. Не исключено, что их допустимо рассматривать как присущие самому выборочному распределению – разумеется, если статистическая проверка показывает, что в задаваемой ими системе координат, его одномерные компоненты являются независимыми.

Исходя из этого, можно предложить следующий алгоритм аппроксимации для функций плотности выборочных распределений. Рассмотрим в каждом из двух классов ω_k систему координат $x_kO_kh_k$ с центром в точке $(a_{X,k}, a_{Y,k})$, повернутую относительно исходной xOy на угол φ_k, равный углу наклона эллипса рассеивания, полученному при аппроксимации двумерным нормальным распределением для данного класса. В этих координатах будем искать приближенное представление выборочного распределения в виде произведения

$$f_k(\xi_k,\eta_k)=g_k(\xi_k)e_k(\eta_k), \qquad (6)$$

причем сомножители правой части последнего выражения зададим в форме кривых Пирсона [2]. Тогда:

$$f_k(\xi_k,\eta_k)=c_{\xi k}\left(1+\frac{\xi_k-d_{\xi k}}{b_{1\xi k}}\right)^{m_{1\xi k}}\left(1-\frac{\xi_k-d_{\xi k}}{b_{2\xi k}}\right)^{m_{2\xi k}}\times c_{\eta k}\left(1+\frac{\eta_k-d_{\eta k}}{b_{1\eta k}}\right)^{m_{1\eta k}}\left(1-\frac{\eta_k-d_{\eta k}}{b_{2\eta k}}\right)^{m_{2\eta k}} \qquad (7)$$

Пусть вне интервала [d_{xk} - b_{1xk}, d_{xk}+b_{2xk}] выполняется условие $g_k(\xi_k)\equiv 0$ и вне интервала [d_{hk} - b_{1hk}, d_{hk}+b_{2hk}] – условие $e_k(\eta_k)\equiv 0$. Тогда коэффициенты c_{xk} и c_{hk} определяются условиями нормировки (интеграл от функции плотности должен быть равен единице). При этом остальные параметры подлежат подбору на основании выборочных данных, приведенных (отдельно для каждого класса) к координатам $x_kO_kh_k$, раздельно по каждой переменной. В данной работе с этой целью минимизировали величину χ^2 [1]. Опыт показывает, что при аппроксимации $m_{1\xi\kappa}$, $m_{2\xi\kappa}$ и $m_{1\eta\kappa}$, $m_{2\eta\kappa}$ можно ограничиться целыми числами.

Итак, для каждого из классов ω_k, в связанной с ним системе координат $\xi_kO'\eta_k$, имеется аппроксимация соответствующего внутриклассового выборочного распределения. Исходя из того, что (ξ_k,η_k) получаются из (x,y) результате сдвига на $(a_{X,k}, a_{Y,k})$ и поворота осей на угол φ_k, путем подстановки в формуле (7), определим аппроксимации выборочных внутриклассовых функций плотности в координатах xOy:

$$p_k(x,y)=f_k((x-a_{X,k})\sin\varphi_k+(y-a_{Y,k})\cos\varphi_k,\ ((x-a_{X,k})\cos\varphi_k-(y-a_{Y,k})\sin\varphi_k). \qquad (8)$$

Чтобы получить условия, задающие границу областей решений, перепишем уравнение (1) в виде:

$$p_1(x,y)-\lambda\, p_2(x,y)=0. \qquad (9)$$

По сравнению с уравнением (1), последнему уравнению удовлетворяют также все точки на плоскости, в которых $p_1(x,y)=p_2(x,y)=0$, что делает задачу нахождения решающей границы, на первый взгляд, неопределенной. Чтобы преодолеть эту неопределенность, рассмотрим условия, задающие области решения. Как известно, возможны три реакции байесовского классификатора на предъявление образа: 1 – при попадании образа в некоторую область A_1, предъявленный объект (больной) квалифицируется как объект класса ω_1; 2 – при попадании образа в область A_2 – как объект класса ω_2; 3 – в случае, если образ оказывается принадлежащим области A_3, ответ классификатора не определен. Пусть область A_1 определяется условием $p_1(x,y)-\lambda\, p_2(x,y)>0$, область A_2 – условием $p_1(x,y)-\lambda\, p_2(x,y)<0$, и область A_3 – условием $p_1(x,y)-\lambda\, p_2(x,y)=0$. Каждая из функций $p_k(x,y)$ по определению отлична от нуля лишь в ограниченной области плоскости. Поэтому область A_3 не есть линия (как, например, в случае двух нормальных распределений на плоскости). Таким образом, граница каждой из областей решений будет состоять из участков двух типов – в любой точке участка границы I типа $p_1(x,y)=p_2(x,y)\neq 0$, (там, где A_1 граничит с A_2), на участках II типа $p_1(x,y)=p_2(x,y)=0$, (где A_1 или A_2 граничит с A_3). Очевидно, что участки границы второго типа области A_k суть отрезки границы области ненулевых значений функции $p_k(x,y)$.

Рис. 3,а служит иллюстрацией к рассуждениям о структуре областей решений байесовского классификатора, когда внутриклассовые плотности заданные произведениями кривых Пирсона (7), аппроксимируют данные выборки A. Как следует из определения, функция $p_1(x,y)$ равна нулю всюду, кроме прямоугольной области $\Re_1$ с центром симметрии в точке $(a_{X,1}, a_{Y,1})$; $p_2(x,y)$ отлична от нуля только в прямоугольной области $\Re_2$ с центром симметрии в точке $(a_{X,2}, a_{Y,2})$. В данном случае область принятия решения A_1 ограничена замкнутой кривой $\Im_1$. Точки M_1 и M_2, принадлежащие $\Im_1$, суть точки пересечения границ областей A_1 и $\Re_2$, они делят $\Im_1$, на два участка. Участок, находящийся внутри области $\Re_2$, является об-

щим для A_1 и A_2, т.е. это участок I типа. Отрезок $\Im_1$, находящийся вне $\Re_2$, совпадает с частью границы $\Re_1$, внутри которой $p_1(x,y)$ отлична от нуля, т.е. является участком II типа. Область $A_2=\Re_2\backslash A_1$, соответствует альтернативному решению (велик риск пневмонии). Граница области решения A_2 состоит также из двух участков: общей границы с A_1 и той части границы области $\Re_2$, которая не пересекается с $\Re_1$. Для сравнения, на рисунке пунктирной линией изображен эллипс, представляющий собой границу решений при аппроксимации условных выборочных распределений нормальным.

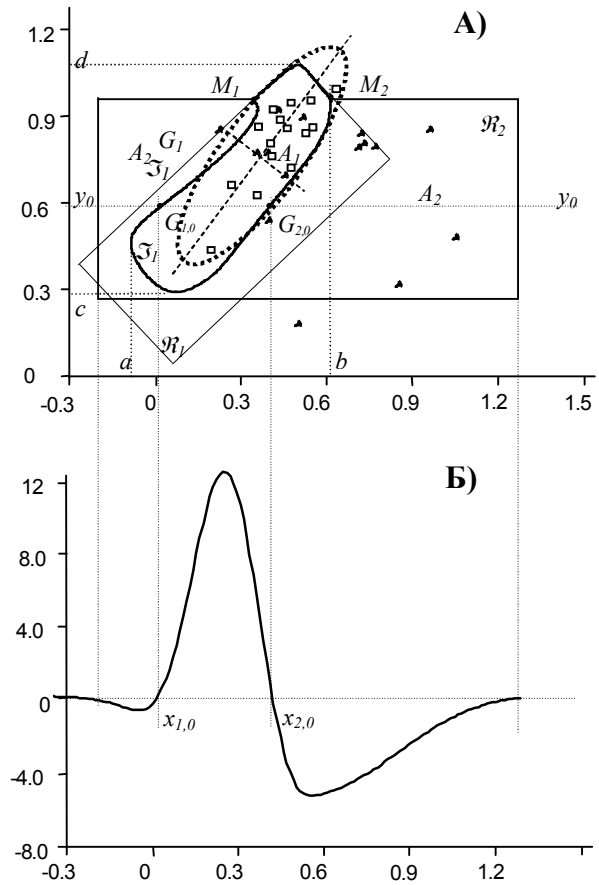

Рис. 3.

Идея приближенного построения границы $\Im_1$, ясна из рис. 3,б., на котором изображено сечение поверхности $p_1(x,y)-\lambda p_2(x,y)$ плоскостью, параллельной плоскости xOz и проходящей через точку y_0. Точки $G_{1,0}$ и $G_{2,0}$ границы $\Im_1$ однозначно определяются нулями $x_{1,0}$ и $x_{2,0}$ функции

$\psi_0(x)=p_1(x,y_0)-\lambda\, p_2(x,y_0)$,

которые легко найти с помощью подходящего численного метода. Беря малое (положительное или отрицательное) приращение Δy, переходя последовательно от точки y_0 к точкам $y_i=y_0 \pm i\cdot\Delta y$, i=1, 2... и поступая аналогичным образом, вычислим нули $x_{1,\pm i}$, $x_{2,\pm i}$ функций $y_i(x)$ и точки $G_{1,\pm i}$, $G_{2,\pm i}$ границы $\Im_1$. В качестве критерия остановки процесса можно использовать условие близости точек $x_{1,\pm i}$ и $x_{2,\pm i}$.

ЗАКЛЮЧЕНИЕ

Посчитаем теперь вероятность ошибки классификации по формуле

$$P(ош.) = P_1\, p(x \in A_2 / \omega_1) + P_2\, p(x \in A_1 / \omega_2). \qquad (10)$$

В силу природы задачи (для построения аппроксимаций используются выборочные данные), вычисление величины $P(ош.)$ с высокой точностью очевидно лишено смысла. Поэтому наиболее простой способ заключается, по всей видимости, в применении метода Монте-Карло. Одним из его преимуществ в данном случае является также то, что он дает единую, не зависящую от типа распределения методику расчета. Опуская, за неимением места, подробности вычислений [5], опишем конечный результат.

При аппроксимации нормальными плотностями, в случае $P_1=P_2=0.5$, вероятность ошибки первого рода (т.е. отнесения объекта класса ω_1 к классу ω_2) равна 0.035, а вероятность ошибки ошибка второго рода (отнесения объекта класса ω_2 к классу ω_1) – 0.135. Полная вероятность ошибки составляет 0.17. При втором способе аппроксимации (произведением кривых Пирсона) вероятность ошибки первого рода равна 0.015, а вероятность ошибки ошибка второго рода – 0.155 и полная вероятность ошибки также равна 0.17. Таким образом, как первый, так и второй способ аппроксимации выборочных внутриклассовых распределений дают почти совпадающие оценки вероятности ошибки классификации. Сравнивая с одномерной классификацией (только по лейкоцитам) [5], видим, что ошибка уменьшилась существенно (в 1.64

раза, от 0.28 до 0.17), при этом ошибка первого рода почти не изменилась, и улучшение качества классификатора обусловлено только падением ошибки второго рода. Отсюда следует весьма важный вывод: *именно совместное рассмотрение количества лейкоцитов (Л) и сегментоядерных нейтрофилов (С) позволяет прогнозировать развитие пневмонии*.

Полученные закономерности подтверждаются исследованием данных контрольной выборки *B,* где рассматривалось соотношение только между показателями формулы крови *L* и *S*. До развития пневмонии выборочное значение корреляции между ними практически совпадает с таковым, вычисленным по данным выборки *A*. Однако корреляция также становится недостоверной за сутки до ее рентгенологического подтверждения.

Таким образом установлено, что возможен ранний (за сутки до рентгенологического подтверждения) и достаточно надежный прогноз пневмонии при острых отравлениях ПСС, если известны значения всего лишь двух легко доступных и дешевых лабораторных показателей – лейкоцитов и сегментоядерных нейтрофилов.

ЛИТЕРАТУРА

1. T.W.Anderson, R.R.Bahadur Classification into two multivariate normal distributions with different covariance matrices. Ann. Math. Stat., 33, 422-431 (1962).

2. Крамер Математические методы статистики. – М., Мир, 1978

3. Дж. Ту, Р. Гонсалес. Принципы распознавания образов. Перевод с англ. – М.: Мир, 1978.

4. А.Н.Ельков Об одном алгоритме распознавания образов для решения задачи прогноза заболевания. Препринт Ин-та прикладной математики им. Н.В.Келдыша РАН, 1998,№34.

5. А.Н.Ельков, К.К.Ильяшенко Прогноз пневмонии при острых отравлениях как задача распознавания образов. Препринт Ин-та прикладной математики им. Н.В.Келдыша РАН, 1999,№65.