УДК 616.831 – 005.1:681.3

МЕТОДИКА ФОРМИРОВАНИЯ СЛОВАРЯ ИНФОРМАТИВНЫХ ПРИЗНАКОВ ПРИ РАСЧЕТЕ ВЕРОЯТНОСТИ ПОВТОРНОГО ИНСУЛЬТА НА ОСНОВЕ КРИТЕРИЯ ИНФОРМАТИВНОСТИ КУЛЬБАКА

И.Я. Львович, Н.А. Гладских

В статье представлена методика формирования словаря информативных признаков при расчете вероятности повторного инсульта на основе критерия информативности Кульбака и расчета диагностических коэффициентов

Ключевые слова: информативность по Кульбаку, диагностические коэффициенты, формирование словаря информативных признаков, расчет вероятности повторного инсульта

Цереброваскулярная патология на сегодняшний день является одной из наиболее значимых медико-социальных проблем во всем мире. Особое значение придается острым формам нарушениям мозгового кровообращения – инсультам. Значимость проблемы инсульта не ограничивается высокой заболеваемостью и смертностью. Во всех странах инсульт – лидирующая причина первичной инвалидизации: более половины пациентов, перенесших мозговую катастрофу, нуждаются в той или иной степени ухода за собой. Вторичная профилактика – гораздо более специализированная область лечения инсультов. Многочисленные клинические испытания подтвердили эффективность мероприятий стратегии высокого риска в профилактике повторных инсультов. Основная направленность современных исследований - разработка индивидуализированной вторичной диагностики, обеспечивающей коррекцию факторов риска развития инсультов и увеличение продолжительности и качества жизни пациента.

Любому врачу в его работе необходимо «вероятностное мышление» и, в частности, понимание вероятностного подхода к диагностике. По-видимому, на таком подходе в значительной мере основан тот подсознательный процесс, который лежит в основе установления диагноза опытным врачом, учитывающим патогномические симптомы, частые симптомы, симптомы, не характерные для данного заболевания или не встречающиеся при нем никогда. Вероятностный подход придает диагностическим построениям строгую количественную форму**,** дает в руки врача хорошо разработанный и, вместе с тем, не слишком сложный математический аппарат, но это не означает, что врачу необходимо его применять каждый раз у постели больного. Все эти вычисления можно осуществить один раз при определении плана лечения.

Львович Игорь Яковлевич – ВИВТ, д-р техн. наук, профессор, тел. 8(4732)727398

Гладских Наталья Александровна – ВГМА им. Н.Н. Бурденко, канд. техн. наук, ассистент, тел. 89192320285, E-mail ngladskikh@rambler.ru

В процессе проведенных исследований применялись клинические методы, методы математической статистики, математического моделирования.

На предварительном этапе при расчете вероятности повторного инсульта необходимо получить перечень наиболее информативных признаков. В медицинских исследованиях понятие информативности признака связывают с его диагностической ценностью в задачах дифференциальной диагностики.

Формирование словаря признаков, используемого для расчета вероятности повторного инсульта, является важной и достаточно сложной задачей.

При разработке словаря признаков приходится сталкиваться с некоторыми ограничениями. Одно из них состоит в том, что в словарь могут быть включены только те признаки, для которых имеется априорная информация, достаточная для описания классов на языке этих признаков. Другое ограничение заключается в том, что некоторые из признаков нецелесообразно включать в априорный словарь ввиду того, что они малоинформативны.

В рабочем словаре следует использовать лишь те признаки, которые, с одной стороны, наиболее информативны и, с другой – могут быть в принципе определены имеющимися или специально созданными средствами наблюдения.

Построенный таким образом словарь признаков должен явиться информативной базой для расчета вероятности повторного инсульта.

Определение словаря признаков возможно с использованием следующих подходов:

1. Игровой подход к построению словаря признаков.
2. Метод, основанный на сравнении апостериорных вероятностей.
3. Метод, основанный на сравнении вероятностных характеристик признаков.
4. Метод, основанный на определении количества информации.
5. Метод, базирующийся на определении информативности Кульбака.

В рамках данного исследования был выбран подход, основанный на определении информативности признаков по Кульбаку.

Данный метод по сравнению с другими методами минимизации информативной избыточности наиболее прост и доступен для алгоритмизации.

Методика расчета информативности признаков по Кульбаку базируется на определении диагностических коэффициентов, рассчитанных для основной и контрольной групп пациентов.

Под наблюдением находился 191 больной, перенесший один и более инсульт. За 38 пациентами осуществлялось ретроспективное наблюдение, 153 пациента наблюдались с первого момента развития заболевания. Контрольную группу составили 80 больных, перенесших один инсульт, 111 больных с повторными нарушениями мозгового кровообращения – основную группу. Их обследование и лечение проходило на базе неврологического отделения для больных с нарушениями мозгового кровообращения Воронежской областной клинической больницы №1, с 2000 года по 2007 год.

В нейрососудистом отделении проводился отбор и наблюдение за больными, поступившими с диагнозом ишемический инсульт. В результате проспективного 5 летнего наблюдения, сформировалось основная группа перенесших повторное нарушение мозгового кровообращения – 91 человек и контрольная группа без повторных инсультов – 62 человека. Наблюдение осуществлялось как в поликлинических условиях или на дому, так и при госпитализации в стационар при повторном инсульте или на очередном курсе лечения. Параллельно отбирались пациенты, поступившие с диагнозом повторный ишемический инсульт с интервалом менее 5 лет (20 человек) – они так же вошли в основную группу и с повторным эпизодом в период более 5 лет (18 человек) – контрольная группа. Конечной точкой являлся повторный ишемический инсульт. Критерием исключения – геморрагический характер первого или повторного нарушения мозгового кровообращения.

Диагностический коэффициент представляется в виде логарифма отношения вероятностей проявления данного признака в основной и контрольной группе (*p(Xij|A1)* и *p(Xij|A2)* соответственно) и умноженный на 100

$$ДК = 100 \lg \frac{P(x_j / A_1)}{P(x_j / A_2)} \qquad (1)$$

Диагностические коэффициенты представляют собой чаще всего двузначные или однозначные положительные или отрицательные числа. Положительными они являются в случае преобладания вероятности *p(Xij|A1),* находящейся в числителе, отрицательными — в случае преобладания вероятности *p(Xij|A2).* То есть диагностические коэффициенты со знаком « + » говорят о большем правдоподобии гипотезы А1 (о принадлежности к основной группе) со знаком «—» — о большем правдоподобии гипотезы А2 (о принадлежности к контрольной группе). Очевидно, коэффициенты с положительным знаком несут положительную информацию, приближая сумму диагностических коэффициентов к порогу, который для *A1* является положительным. Коэффициенты с отрицательным знаком, наоборот, «отдаляют» сумму от порога. Для гипотезы *A2*, наоборот, коэффициенты с отрицательным знаком приближают сумму к порогу, а коэффициенты с положительным знаком — отдаляют ее от порога, так как порог является величиной отрицательной.

Следует отметить, что чем больше величина диагностического коэффициента, тем больше дифференциально-диагностической информации, т. е. информации о преобладании вероятности одного из диагнозов, он несет. Однако информативность каждого значения признака зависит также от частоты, с какой встречается это значение при каждом из заболеваний, т. е. от величин *p(Xij|A1)* и *p(Xij|A2)*. Если диагностический коэффициент значения признака x_i^j велик, но больные с этим значением встречаются сравнительно редко, то в процессе диагностики роль такого значения признака x_i мала.

Для определения той информации, которую несет признак x_i, сначала необходимо вычислить сумму информации, которую дают значения признаков (x_i^j). Для этого необходимо умножить диагностический коэффициент, полученный для данного признака ДК(x_i^j) на разность вероятностей этого признака при принадлежности к основной группе (гипотеза A_1) и к контрольной группе (гипотеза A_2):

$$ДК(x_i^j)[p(Xij|A1) - p(Xij|A2)] \qquad (2)$$

Следует заметить, что разность [*p(Xij|A1) - p(Xij|A2)*] будет положительной в случае, если ДК положителен. Разность же (2) покажет, насколько в среднем будет приближаться сумма диагностических коэффициентов к порогу в результате обнаружения у больного симптома x_i^j.

Аналогично рассчитываются другие значения этого же признака $x_i^1, x_i^2 \ldots x_i^n$. Информативность признака в целом *I(xi)* будет равна их сумме:

$$I(xi) = \sum ДК(x_i^j)[\, p(Xij|A1) - p(Xij|A2) \qquad (3)$$

Если представить величину ДК в развернутом виде, то формула (3) примет вид, идентичный формуле информационного критерия Кульбака:

$$I(xi) = \sum 100 \lg \frac{p(Xij \mid A1)}{p(Xij \mid A2)} \; [\, p(Xij|A1) - p(Xij|A2)] \qquad (4)$$

Таким образом, алгоритм формирования словаря информативных признаков состоит из следующих этапов:

1. Формирование основной и контрольной группы пациентов;
2. Расчет вероятностей проявления признака в основной и контрольной группах *p(Xij|A1)* и *p(Xij|A2);*
3. Расчет диагностических коэффициентов для признаков ДК(x_i^j)
4. Вычисление информативности для заданного значения признака ДК(x_i^j)[*p(Xij|A1) - p(Xij|A2)*]
5. Вычисление информативности признака $I(xi) = \sum ДК(x_i^j)[p(Xij|A1) - p(Xij|A2)]$

6. Отбор признаков, имеющих наибольшее значение *I(xi)*

Пользуясь предложенной методикой, на основе сформированной базы данных были рассчитаны диагностические коэффициенты и значения информативности по каждому из признаков:

X1 – нарушение сознания
X2 –гемианопсия
X3 – парез в руке
X4 – парез в ноге
X5 – расстройство чувствительности (гемигипостезия)
X6 – симптом отрицания (анозогнозия)
X7 – афазия
X8 – нарушение ритма сердца
X9 – сахарный диабет
X10 – показатели глюкозы крови на момент инсульта
X11 – ультразвуковая допплерография (УЗДГ)
X12 – возраст
X13 – пол
X14 – АД
X15 – холестерин
X16 – ИБС
X17 – Локализация очага по бассейнам
X18 – частота подтипов
X19 – Тяжесть инсульта по Ренкину
X20 - Баллы по Бартелу

Значения информативности признаков X1-X20 приводятся в таблице 1.

Таким образом, в результате анализа информативности признаков было сформировано пространство признаков $X_i \quad \left(i = \overline{1, N}\right)$, позволяющее полностью идентифицировать состояние объекта моделирования.

Таблица 1. Результаты расчета информативности диагностических признаков с использованием критерия Кульбака.

Признак	Значения информативности признаков
X6	0
X2	0,0175
X5	0,13086
X1	0,597
X19	0,822
X9	1,278
X20	1,42
X18	3,0178
X17	3,139
X8	4,295
X13	4,44
X10	5,40892
X7	5,9785
X12	6,39
X15	6,46
X4	7,7263
X16	12,08
X3	12,455
X11	15,557
X14	21,61

В качестве диагностически значимых признаков были отобраны:

X3 – парез в руке
X4 – парез в ноге
X7 – афазия
X8 – нарушение ритма сердца
X10 – показатели глюкозы крови на момент инсульта
X11 – ультразвуковая допплерография (УЗДГ)
X12 – возраст
X13 – пол
X14 – АД
X15 – холестерин

X16 – ИБС

X17 – Локализация очага по бассейнам

X18 – частота подтипов

Построенный таким образом словарь признаков является информативной базой для расчета вероятности повторного инсульта.

Для определения вероятностных оценок рецидива инсульта целесообразно использовать формулу Байеса, которую иногда называют теоремой об обратной вероятности или теоремой гипотез. Это вполне применимо к задачам диагностики: формула позволяет выбрать одну из нескольких возможных диагностических гипотез, основываясь на вычислении вероятностей болезней по вероятности обнаруженных у больного симптомов. С помощью этой формулы на основе следующих данных:

1. *P(Ak)* - априорная вероятность симптома *p(Xij)* представляет собой вероятность симптома *Xij,* во всей группе, т. е. вероятность для любого больного в группе независимо от того, какой болезнью он страдает, иметь симптом *Xij.* Эта величина является отношением числа больных, имеющих симптом *Xij* к общему числу больных в группе.

2. *P(xij|Ak)* - условная вероятность симптома *Xij,* при возможности повторного инсульта (гипотеза *A1*) *p(Xij|A1)* представляет собой вероятность иметь симптом *Xij,* при условии принадлежности к основной группе. Эта величина равна отношению числа больных с повторным инсультом, имеющих симптом *Xij* к общему числу больных, страдающих этой болезнью.

Формула Байеса имеет следующий вид:

$$P(A_1 / x_1) = \frac{P(A_1)P(x_1 / A_1)}{\sum_{л} P(A_k)P(x_1 / A_k)} \qquad (5)$$

Однако у больного могут быть обнаружены одновременно симптомы $x_1, x_2, ..., x_n$. Как вести расчет вероятности рецидива инсульта в этом случае?

Если мы располагаем данными о числе больных, у которых имеется комплекс симптомов $x_1, x_2, ..., x_n$ при болезнях *A1* и A_2, то расчет вероятностей повторного инсульта при наличии указанных симптомов может быть рассчитана на основе использования формулы Байеса.

$$P(A_1/x_1x_2..x_n) = \frac{P(A_1)P(x_1/A_1)P(x_2/A_1)..P(x_n/A_1)}{\sum_k P(A_k)P(x_1/A_k)P(x_2/A_k)..P(x_n/A_k)}, \qquad (6)$$

В данном исследовании *p(A1)*- априорная вероятность появления повторного инсульта, *p(A2)*- априорная вероятность отсутствия повторного инсульта, *p(Xij|Ak)* – условная вероятность (частость) появления признака при принадлежности к основной или контрольной группе.

По существу задача диагностики состоит в том, чтобы установить диагноз, используя тот минимум доступной диагностической информации, который достаточен для достижения необходимой надежности диагноза. Это обычно требует использования не одного симптома, а набора симптомов (симптомокомплекса).

Такой подход может быть назван «многомерным» подходом к установлению диагноза, так как при нем одновременно используют много признаков.

Таким образом, методика расчета вероятности повторного инсульта состоит из следующих этапов:

1. Формирование пространства признаков $X_i \ \left(i = \overline{1, N}\right)$, которые позволяют полностью идентифицировать состояние объекта моделирования.
2. Формирование словаря информативных признаков на основе критерия Кульбака.
3. Расчет вероятности повторного инсульта на основе формулы Байеса (6).
4. Формирование рекомендаций по дальнейшему лечению и профилактике.

Литература

1. Гублер Е.В. Вычислительные методы распознавания патологических процессов / Е.В. Гублер. – Л.: Медицина, 1970. – 320с.

2. Горелик А.Л. Некоторые вопросы построения систем распознавания / А.Л. Горелик, В.А. Скрипкин. – М.: Советское радио, 1974. – 224с.

Воронежский институт высоких технологий

Воронежская государственная медицинская академия им. Н.Н. Бурденко

METHODIC FORMING THE INFORMATIVE SET BY USING KULBAC RULE ON CALCULATION THE PROBABILITY OF RELAPSE STROKE

I.Ya. Lvovich, N.A. Gladskikh

The methodic forming the informative set by using Kulbac rule on calculation the probability of relapse stroke is presented at the article

Key words: the informative set, Kulbac rule, probability of relapse stroke