А.А.Венедиктов, доктор экономических наук, профессор
В.И.Стеклов, кандидат медицинских наук

Инструментальные методы прогнозирования результатов эксперимента в военной медицине в условиях многомерных исходных данных и малой выборки

В статье рассматриваются проблемы инструментальной реализации авторского метода прогнозирования исхода эксперимента при небольшом количестве наблюдений, многомерных исходных данных и широком диапазоне значений каждого из наблюдаемых параметров. Приводится пример применения разработанных методов для автоматизированной обработки результатов экспериментов в целях прогнозирования развития некоторых заболеваний сердца.

Введение

Решение задач прогнозирования результатов эксперимента в военной медицине в большинстве случаев сводится к следующей модели. Имеется набор N наблюдаемых параметров (p_i, $i=\overline{1,N}$), значения которых применительно к конкретному пациенту или их группе могут быть получены на основе объективных исследований. Набор таких значений будем называть вектором исходных данных эксперимента. Для j-го эксперимента он будет иметь вид $V_j=(v_{j1},v_{j2},\ldots,v_{jN})$. Компоненты v_{ji} могут быть как непрерывными, так и дискретными величинами. Как было показано в [1], применительно к результатам прогнозирования эксперимента в военной медицине, как правило, более применимы дискретные показатели. Далее в настоящей статье мы будем считать компоненты вектора V_j целыми неотрицательными числами. Натуральные числа будут представлять собой код реального значения параметра: для компонент, имеющих дискретную природу, код присваивается каждому из их возможных значений; для непрерывных показателей код задается для определенного диапазона. Нулевое значение компонента v_{ji} будет означать, что значение i-го параметра в j-м эксперименте по тем или иным причинам не было зафиксировано в ходе эксперимента и восстановить его по медицинским документам в настоящее время не представляется возможным.

Кроме того, имеется набор из M параметров, которые в рассматриваемой модели интерпретируются как результаты эксперимента: r_i, $i=\overline{1,M}$. В простейшем случае результат выражается единственным значением, т.е. $M=1$. В настоящей статье мы будем рассматривать именно такой, наиболее простой, способ представления результата.

В работе [1] авторами был разработан метод прогнозирования исхода эксперимента при небольшом количестве наблюдений, многомерных исходных данных и широком диапазоне зафиксированных значений каждого из наблюдаемых параметров.

Проблемы применения существующего метода

Несмотря на то, что упомянутый метод позволил получить весьма высокий результат (около 94% верных прогнозов), по мнению авторов, он представляет, скорее, теоретический интерес и не может эффективно применяться в лечебной деятельности военно-медицинских учреждений по следующим причинам:

1. Отбор параметров, которые могут быть предположительно отнесены к числу оказывающих влияние на прогнозируемый результат,

производится посредством вычисления доли «положительных исходов» (в рассматриваемом примере – невозникновения фибрилляции предсердий на момент анализа) в общей совокупности наблюдений. Если количество наблюдений, при которых тот или иной показатель имеет конкретное значение, оказалось менее $Q_{min}=20$, то такое значение в ходе последующего анализа не учитывалось, поскольку соответствующие результаты представляются недостаточно репрезентативными.

Таким образом, набор предположительно значащих параметров, полученный на момент написания работы [1], по мере накопления данных об экспериментах может (и должен) уточняться. Однако соответствующая расчетная задача достаточно трудоемка для того, чтобы врач имел возможность уточнять перечень параметров при прогнозировании результата конкретного эксперимента (т.е. в ходе анализа анамнеза каждого пациента). Возникает желание формализовать процесс отбора предположительно значащих факторов в целях обеспечения автоматизированной обработки имеющихся данных экспериментов, что позволит оперативно учитывать при моделировании информацию обо всех наблюдениях.

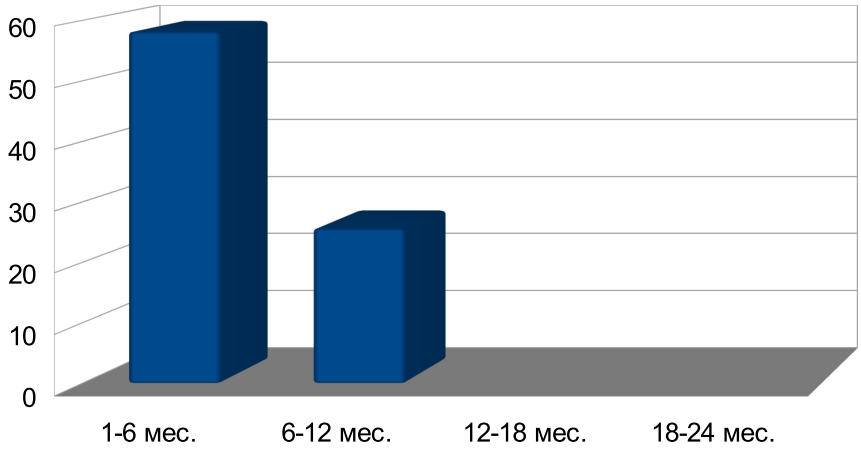

Рисунок 1 – Распределение больных с диагнозом «фибрилляция предсердий» по срокам диагностирования с момента операции радиочастотной абляции кавотрикуспидального перешейка (%)

2. Если «отрицательный» исход эксперимента (диагностирование фибрилляции предсердий у больного, прооперированного по поводу типичного трепетания предсердий) не вызывает сомнений в его достоверности на любом этапе послеоперационного наблюдения, то «положительный» результат может лечь в основу дальнейших выводов лишь по истечении некоторого срока с момента операции. В 84% случаев фибрилляция предсердий диагностируется в течение первых 12 месяцев после операции и лишь в 16% – в последующий период (рисунок 1). На основании этого, а также с учетом клинической практики второго автора, был сделан вывод о том, что даже в случае отсутствия диагноза «фибрилляция предсердий» до истечения года с момента операции нет оснований считать исход эксперимента положительным в целях применения данного результата для прогнозирования развития болезни у иных пациентов.

Это обусловливает необходимость динамически отслеживать те содержащиеся в базе данных результаты наблюдений, для которых с момента операции прошло более года, и учитывать их при прогнозировании, временно исключая из анализа те результаты, для которых данный срок еще не истек. Очевидно, что эта

проверка может и должна производиться в автоматическом режиме, а врач должен быть избавлен от соответствующих рутинных расчетов.

Методика автоматической обработки статистических данных

В целях устранения негативных последствий перечисленных проблем авторами была разработана формализованная методика, позволяющая в автоматическом режиме, т.е. без участия специалиста, обработать данные, содержащиеся в базе результатов наблюдений, и, на основе сопоставления информации об имеющихся исходах эксперимента с анамнезом пациента, сделать вероятностный прогноз развития его болезни. Ниже приводятся этапы данной методики, ориентированной на ее реализацию в виде программы для ЭВМ.

Этап 1. Сведения об имеющихся результатах наблюдений считываются из таблицы в формате Excel., куда они были предварительно введены специалистом, проводящим исследование. Проводится их контроль на предмет наличия возможных погрешностей, которые могут быть связаны как с неправильным заданием формата ячейки (например, поле, являющееся по своей природе датой, отображается как число либо наоборот), так и с техническими ошибками при вводе (введения показателя в другую ячейку таблицы, неверное проставление десятичной точки, изменяющее значение показателя на порядок, и т.п.). Особое место в анализе входных показателей занимает сопоставление значений наблюдаемых параметров с данными иных экспериментов на предмет выявления «нестандартных» их сочетаний. При этом флуктуационные «выбросы» не исключаются из дальнейшего анализа, но специалист, проводящий исследование, информируется об их обнаружении.

На первый взгляд, значимость данного этапа может показаться несколько преувеличенной. Однако практическое применение методики показало, что доскональный автоматизированный контроль исходных данных позволяет выявить достаточно большое количество ошибочных записей (7-9% от общего числа содержащихся в базе результатов экспериментов).

Этап 2. Производится отбор статистически значимых параметров и их значений.

Параметр считается статистически значимым если среди значений, которые он принимает, есть хотя бы одно статистически значимое. Значение параметра является статистически значимым, если оно удовлетворяет следующим условиям:

1. Не равно нулю.

2. Количество имеющихся в базе данных записей, содержащих данное значение в соответствующем поле, не менее $Q_{min}=20$. Отметим, что данная величина носит эмпирический характер. Применительно к конкретной предметной области этот и другие упомянутые далее эмпирические параметры могут быть изменены пользователем. Инструментарий, предназначенный для уточнения таких значений, будет рассмотрен ниже.

3. Доля положительных (отрицательных) исходов для экспериментов, содержащих данное значение в соответствующем поле, отличается от доли положительных (отрицательных) исходов в общем числе экспериментов более чем на $\varepsilon=0{,}04$. Данная величина также носит эмпирический характер и может уточняться исследователем применительно к конкретной предметной области с использованием описанного ниже инструментария.

Этап 3. Производится преобразование векторов исходных данных в целях устранения из них сведений, которые не потребуются в ходе прогнозирования. Для этого из них удаляются компоненты, которые соответствуют описательным (фамилия, имя, отчество, адрес пациента и т.п.) и статистически незначащим параметрам. В компонентах, соответствующих статистически значимым параметрам, статистически незначащие значения заменяются нулями. Это равносильно отсутствию сведений о соответствующем параметре в векторе исходных данных.

Этап 4. Вектор исходных данных, соответствующий анализируемому пациенту, попарно сравнивается с преобразованными на этапе 3

векторами (вычисляется мера по «жесткому» варианту в смысле [1], т.е. определяется число компонент, в которых эти вектора различаются). Пары, для которых мера численно превышает половину длины сравниваемых векторов (с учетом ее изменения на этапе 3) не учитываются в ходе дальнейшего анализа. Для остальных подсчитывается количество пар, имеющих меру 0, 1, 2 и далее до максимально возможного значения, т.е. до $N_4=\left[\frac{N_3}{2}\right]$, где N_3 – размерность векторов после выполнения этапа 3, $[x]$ – целая часть числа x. Иными словами, формируются два множества:

$$D=\{d_1, d_2, \ldots, d_{N_4}\},\ D^+=\{d_1^+, d_2^+, \ldots, d_{N_4}^+\}$$

где d_i – количество векторов, отличающихся от анализируемого ровно в i компонентах;

d_i^+ – количество векторов, отличающихся от анализируемого ровно в i компонентах, по которым исход эксперимента является положительным.

Очевидно, что $d_i^+ \leqslant d_i,\ i=\overline{1,N_4}$.

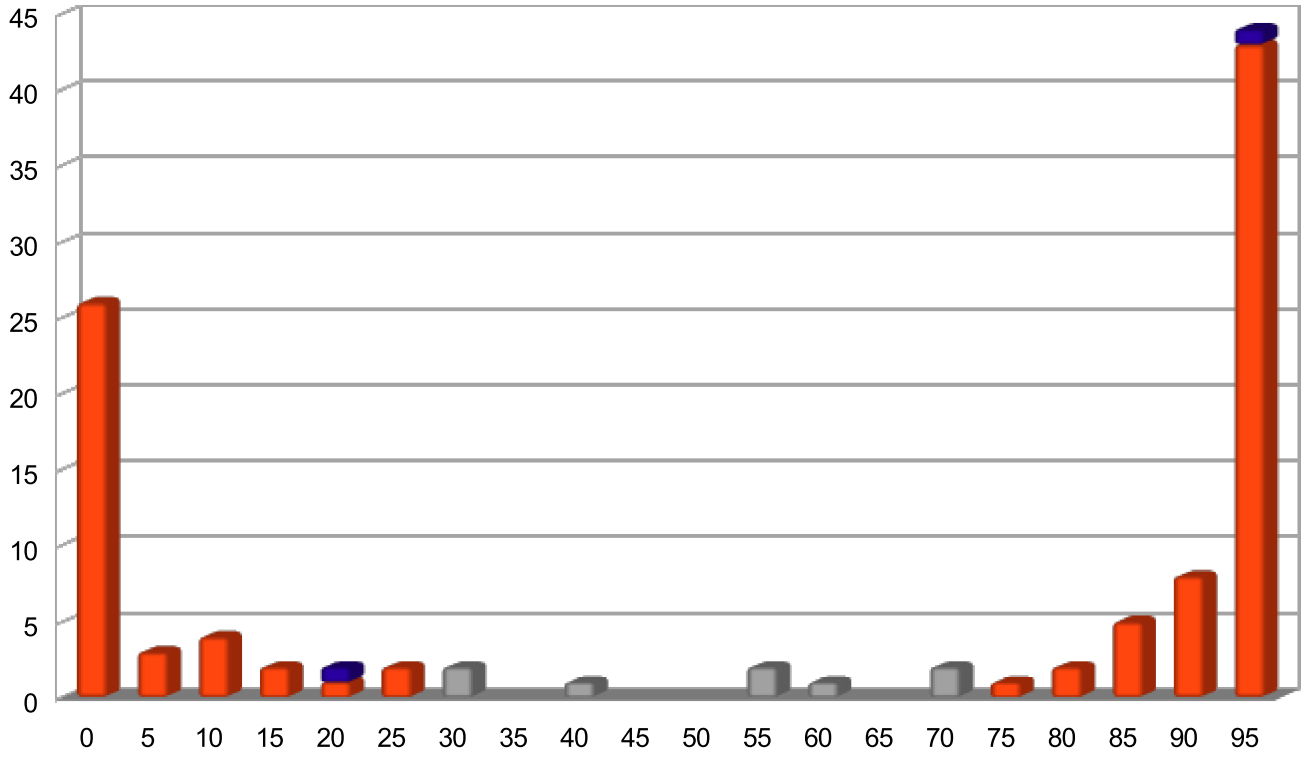

Рисунок 2 – Распределение прогнозов по 5-процентным интервалам оценочной вероятности положительного исхода эксперимента (невозникновения фибрилляции предсердий). Красный цвет – подтвердившиеся прогнозы, синий – ошибочные, серый – недостоверные.

Этап 5. На основании данных, полученных на этапе 4, делается оценка вероятности возникновения положительного исхода в анализируемом эксперименте $P_{оц.}$ по следующей формуле:

$$P_{оц.}=\frac{\sum_{i=0}^{N_4-1}\frac{d_i^+}{\alpha^i}}{\sum_{i=0}^{N_4-1}\frac{d_i}{\alpha^i}}, \qquad (1)$$

где основание степени α определяется эмпирически. В рассматриваемом примере $\alpha=2$.

В случае попадания данной оценки в диапазон $0{,}5-P_{недост.}<P_{оц.}<0{,}5+P_{недост.}$ прогноз считается недостоверным. Конкретное значение $P_{недост.}$ определяется эмпирически с использованием описанного ниже аппарата.

Отметим, что для данной задачи значение $P_{недост.}$ задается равным 0,25. Несмотря на весьма жесткое, на первый взгляд, ограничение

по величине диапазона недостоверных значений (половина диапазона возможных значений оценки вероятности объявляется недостоверным прогнозом), всего 7% экспериментов попадают в этот интервал (рисунок 2).

На гистограмме высота каждого столбца пропорциональна количеству прогнозов, для которых оценочная вероятность положительного исхода болезни попадает в соответствующий 5-процентный интервал. Красным цветом отображены подтвердившиеся прогнозы, синим цветом – ошибочные, серым – недостоверные. Отметим, что различная высота симметричных относительно середины интервала столбцов гистограммы (например, крайнего левого и крайнего правого) объясняется тем, что количество положительных исходов болезни существенно (примерно в два раза) превышает число отрицательных исходов. Соответствующим образом распределяется и количество положительных (правая часть гистограммы) и отрицательных (левая часть) прогнозов. Как видно из рисунка, для рассмотренного примера около 7% прогнозов отнесены к недостоверным. Среди прогнозов, которые в соответствии с рассмотренной методикой отнесены к достоверным, 98% подтвердились и лишь 2% оказались ошибочными.

Инструментальные средства уточнения значений эмпирических параметров модели

Как отмечалось по ходу изложения методики, в модели имеется ряд параметров, значения которых определяются эмпирически:

$Q_{min}=20$ – минимальное число результатов наблюдений, при котором их значение учитывается при прогнозировании;

$\varepsilon=0{,}04$ – минимальное отклонение частости успешного исхода для конкретного наблюдаемого признака, при котором это отклонение принимается во внимание;

$\alpha=2$ – основание степени для мультипликатора, характеризующего «вклад» каждого результата в зависимости от числа совпадений;

$T=365$ – период (в днях), по истечении которого с момента операции ее последствия учитываются при расчетах;

$P_{недост.}=0{,}25$. При отклонении оценочной вероятности от 0,5 менее, чем на эту величину прогноз считается недостоверным.

Значения данных параметров приведены здесь для рассматриваемого примера. Применительно к иным предметным областям у исследователя может возникнуть желание изменить значение данных эмпирических величин.

Для облегчения задачи выбора иных величин в качестве эмпирических параметров авторами разработано инструментальное средство, позволяющее оценить, каким образом изменение их значений может отразиться на точности прогноза. На языке программирования Phyton 3.2 создана программа для ЭВМ, которая осуществляет последовательный перебор имеющихся в базе данных результатов наблюдений, при этом каждый из них поочередно исключается из соответствующего набора и рассматривается в качестве анализируемого вектора.

Основываясь на сведениях об остальных имеющихся в базе данных наблюдениях делается прогноз развития болезни для исключенного вектора исходных данных. Прогнозируемый результат сравнивается с имеющимися сведениями об исходе болезни для данного пациента. Сводный результат прогнозирования по всем содержащимся в базе данных записям отображается в графическом виде, который подобен изображенному на рисунке 2. Тем самым исследователь получает возможность оценить, во-первых, какое количество векторов исходных данных при соответствующем наборе эмпирических параметров модели обусловливает достоверный прогноз, во-вторых, какая доля таких прогнозов подтверждается результатами наблюдений, а какая оказывается ошибочной.

Недостатки инструментария и перспективы его развития

По мнению авторов, основной недостаток рассмотренного инструментального средства связан с тем, что рассчитанная в соответствии с изложенной выше методикой оценочная вероятность положительного исхода не всегда позволяет оценить степень достоверности сделанного прогноза. Так рассчитанная по формуле (1) оценочная вероятность положительного исхода в случае наличия 9 положительных исходов из 10 экспериментов, у которых имеется полное совпадение векторов исходных данных с анализируемым, будет равна 0,9:

$$P_{оц.}=\frac{\frac{9}{2^0}}{\frac{10}{2^0}}=0{,}9\,.$$

Если общее число значащих параметров будет равно, например, 30, это будет соответствовать следующим множествам D и D^+:

$$D=\{10,0,0,\ldots,0\}\,,\ D^+=\{9,0,0,\ldots,0\}\,.$$

Однако ровно такое же значение оценочной вероятности будет получено, если найдено столько же векторов с положительным исходом, но среди имеющих от анализируемого вектора ровно 10 отличий. В этом случае

$$D=\{0,0,0,0,0,0,0,0,0,0,10,0,\ldots,0\}\,,$$
$$D^+=\{0,0,0,0,0,0,0,0,0,0,9,0,\ldots,0\}\,.$$

Для таких множеств значение оценочной вероятности также будет равно 0,9:

$$P_{оц.}=\frac{\frac{9}{2^{10}}}{\frac{10}{2^{10}}}=0{,}9\,.$$

Вместе с тем, представляется очевидным, что первый прогноз вызывает существенно большее доверие, чем второй, поскольку базируется на информации о развитии болезни пациентов с полностью совпадающими значениями наблюдаемых параметров.

По мере накопления данных станет возможным не ограничиваться определением оценочной вероятности возникновения положительного (отрицательного) исхода, а с высокой степенью достоверности определять математическое ожидание и среднее квадратическое отклонение времени. При этом должны быть применены полностью аналогичные методические подходы: кластеризация исходных данных по каждому из наблюдаемых параметров, определение близких (в смысле введенной меры) векторов параметров имеющихся наблюдений, статистический анализ их известных исходов (с учетом коэффициента, характеризующего степень близости каждого вектора к анализируемому).

На данном этапе, когда база наблюдений за больными недостаточна для надежного определения перечисленных характеристик случайной величины, в рассмотренное в настоящей статье инструментальное средство добавлен модуль, отображающий исследователю подробную информацию о том, сколько именно векторов с каждым значением меры было обнаружено. Таким образом, специалист получает возможность самостоятельно, экспертным путем, определить, с какой степенью доверия следует относиться к полученным прогнозным показателям. Хотелось бы еще раз подчеркнуть, что по мере накопления статистических данных данная оценка будет производится также автоматически.

Выводы

Таким образом, изложенная методика и созданное на ее основе инструментальные средства (программы для ЭВМ) позволяют:

1. Отслеживать возможные ошибки при вводе результатов экспериментов за счет проверки, попадают ли вводимые показатели в некий разумный диапазон допустимых значений. В случае обнаружения флуктуаций, вызывающих сомнения в корректности вводимых данных, внимание врача обращается на данный показатель. Он имеет возможность исправить либо подтвердить правильность введенного значения.

2. Динамически корректировать перечень значащих для прогноза факторов путем как

уточнения статистики положительных и отрицательных исходов болезни, так и перевода в разряд достоверных отдельных статистических показателей. Отметим, что расширение статистической базы происходит не только за счет накопления данных о результатах экспериментов (т.е. поступления в нее сведений о новых операциях), но и в связи с увеличением времени послеоперационного наблюдения за уже внесенными в нее пациентами и, соответственно, повышения достоверности сведений о количестве «положительных» исходов.

3. Уточнять эмпирические параметры модели применительно к иным предметным областям на основе автоматизированного подбора их значений в целях достижения наилучшего результата прогнозирования.

Список использованных источников

1. Венедиктов А.А., Стеклов В.И. Прогнозирование результатов эксперимента в военной медицине в условиях многомерных исходных данных и малой выборки // Вооружение и экономика. – 2012. – № 5.