УДК 004.9

АНАЛИЗ МЕТОДОВ ОТБОРА ФАКТОРОВ РИСКА РАЗВИТИЯ ПАТОЛОГИЙ В АКУШЕРСТВЕ И ГИНЕКОЛОГИИ

И. В. Соков, А. С. Сокова, Т. А. Васяева, Д. Е. Иванов
Донецкий национальный технический университет
Институт прикладной математики и механики НАН Украины
E-mail: sokov_ivan@mail.ru, vasyaeva@gmail.com, ivanov@iamm.ac.donetsk.ua

Соков И. В., Сокова А. С., Васяева Т. А. Иванов Д.Е. Анализ методов отбора факторов риска развития патологий в акушерстве и гинекологии. *Рассматривается задача отбора факторов риска развития осложнений при беременности и родах. Выполнен анализ данных, предоставленных медицинскими сотрудниками. Проанализированы методы отбора и оценивания факторов риска с учетом особенности медицинской информации.*

Введение. Ежедневно в мире от осложнений, связанных с беременностью и родами, умирает 1500 женщин. По оценкам экспертов большинство этих случаев можно было предотвратить. Около 80% смертей обусловлено прямыми причинами: кровотечения, инфекции, гестозы, патологические роды. При этом даже если удается предотвратить гибель женщины, то следствием перенесенного критического состояния является развитие какой-либо тяжелой хронической патологии − вплоть до инвалидности. Таким образом данная проблема носит не только медицинский, но и выраженный социальный характер.

Не вызывает сомнения, что наиболее действенный путь снижения, к примеру, частоты кровотечений лежит в разработке программ прогнозирования индивидуального объема потери крови при родах. Сложность разработки таких программ заключается в необходимости научного анализа большого количества клинических и лабораторных показателей, которые находятся в сложной зависимости друг от друга и не всегда поддаются количественной оценке.

Наиболее перспективное направление для решения задач медицинского прогнозирования [1] основано на интеллектуальном анализе данных с применением современного программного обеспечения. На первом этапе которого необходимо определить факторы риска и оценить их информативность.

Факторы риска. Как правило, отбор данных для анализа выполняется врачом по следующему принципу: сначала осуществляется выделение факторов риска относящихся к определенной патологии и в основном по формальному принципу: отягощенный семейный анамнез, перинатальные факторы, заболевания в анамнезе, неблагоприятная макросоциальная среда, социально-гигиенические факторы и т. д. Затем группы факторов риска определяются временем их воздействия, видом (биологические, средовые и т. д.) и количеством воздействующих факторов. При анализе данных, предоставленных различными врачами для прогнозирования тех или иных заболеваний в области гинекологии можно сделать вывод, что перечень собранных факторов, является относительно стабильным, причем одинаковым для анализа большинства гинекологических проблем и очень большим. В него входят медицинские и социально-демографические факторы. Перечень таких факторов риска частично представлен:

- возраст моложе 18 или старше 35 лет;
- рост менее 155 см и вес до беременности на 20% ниже или выше нормы для данного роста;
- пятая и последующая беременность, особенно если беременная старше 35 лет;

- злоупотребление курением;
- многоплодная беременность;
- отсутствие прибавки в весе или минимальная прибавка;
- срок беременности более 42 недель
- и многие другие.

Однако главная задача заключается не только в выделении факторов риска, но и в оценке роли каждого из них. Значимость каждого фактора на риск развития различных акушерских осложнений будет различна.

Постановка задачи. Предположим, существует набор факторов риска S, содержащий *m* примеров (1). Каждый пример, S_i набора данных *S* состоит из *n* определяющих параметров $X_i = (x_1, x_2, \ldots, x_n)$ и параметра – результата Y_i (2):

$$S = \{S_1, S_2, ..., S_m\}, \quad (1)$$

$$S_i = \{X_{i,1}, X_{i,2}, ..., X_{i,n}, Y_i\}, i \in [1, m]. \quad (2)$$

Таким образом, каждый *i*-й пример набора данных *S* представлен набором значений факторов риска *X* и результата *Y*. При этом факторы риска *X* набора данных *S* могут включать в себя как факторы, содержащие полезную информацию, так и факторы, частично или полностью неинформативные. Кроме этого, информативные факторы тоже могут содержать шум.

Необходимо выполнить отбор значимых факторов и оценить их. При вычислении меры значимости каждого фактора необходимо найти такие весовые коэффициенты $W = (w_1, w_2, ..., w_n)$ для факторов риска $X = (x_1, x_2, ..., x_n)$, при которых выполняется (3):

$$E(X \cdot W) \le \varepsilon, \quad (3)$$

где $E(X \cdot W)$ – критерий оценивания, ε – допустимая погрешность. В качестве критерия оценивания можно использовать формулу (4).

$$E = \frac{1}{M} \cdot \sum_{i=1}^{M} (F_i - Y_i)^2 \quad (4)$$

где *M* – количество обучающих примеров, *F* – полученный результат, *Y* – желаемый результат.

Методы определения бальной оценки факторов риска. В медицине достаточно часто используют статистические методы анализа и обработки данных [2, 3]. Для использования таких методов необходимо предварительно подготовить данные [4]. Препроцессорная обработка заключается в первую очередь в кодировании нечисловой информации. Для анализа значимости числового фактора при различных значениях, необходимо разбить диапазон значений этой переменной на интересующие интервалы. Например, такой фактор риска, как возраст матери на момент родов, целесообразно анализировать в следующих промежутках: до 18 лет, от 18 до 25 лет, от 25 до 35 года и после 35 года. Таким образом, вместо одного параметра следует анализировать полученные 4.

Для оценки зависимостей между параметрами наиболее часто применяют корреляционный анализ. Строятся корреляционные портреты, которые показывают тесноту связей между различными факторами. Таким образом, при оценивании влияния на прогностическую величину каждой входной переменной, определяется коэффициент парной корреляции. Если коэффициент корреляции между переменной, входящей в модель, и результатом > 0,7÷0,8, то делается вывод, что фактор значим, и оценивается его значимость. В соответствии со значимостью фактора ему сопоставляется вес (определенное число баллов). Задача выбора весовых коэффициентов неоднозначна. Коэффициенты могут выбираться полностью экспертным путем или формальным методом. Один из вариантов получения весовых коэффициенты представлен формулой (5)

$$w_i = \frac{r_{x_i y}}{\sum_{j=1}^{n} r_{x_j y}} \cdot 100 \qquad (5)$$

.

Альтернативным методом отбора информативных факторов может быть, например, метод группового учета аргументов [5]. В многорядном алгоритме метода группового учета аргументов к каждой *i*-й переменной оптимальным образом подбирается наилучший для данного случая набор из остальных аргументов. Следовательно, можно построить структурную таблицу размером (*n* x *n*), в которой *i*-я строка соответствует *i*-й модели, а *j*-й столбец – *j*-му аргументу (*i, j* = 1, ..., n). Значение *i, j*-той ячейки этой таблицы равно *j*, если *j*-й аргумент участвует в формировании *i*-й модели, или 0, если не участвует. Эта таблица показывает, какие исходные аргументы участвуют в формировании каждой *i*-й модели. Можно определить значение «индекса полезности» *j*-го аргумента ($kolx_j$) как число, показывающее, сколько раз данный аргумент участвует в формировании всех моделей (частота использования *j*-го аргумента). Очевидно, что эти числа лежат в интервале от 1 до *n*. Если $kolx_j = 1$, то можно утверждать, что *j*-й аргумент несуществен для данного случая, так как присутствует только в модели, где он включен в соответствии со спецификой алгоритма. Если $kolx_j = n$, то можно утверждать, что *j*-й аргумент существен для данного случая. Следует ввести некоторый порог для определения значимости факторов: можно считать аргумент значимым, если $kolx_j > 0{,}5n$. Кроме того, их можно упорядочить по убыванию величины $kolx_j$ и, пользуясь дополнительными критериями, выбрать необходимое количество факторов в зависимости от конкретной задачи.

На сегодняшний день разработано множество различных шкал факторов риска для прогнозирования наступления неблагоприятного исхода. Существуют изолированные шкалы риска (шкала оценивания риска развития одной определенной патологии), которые наиболее чувствительны и точны в определении вероятности наступления неблагоприятного исхода, но вследствие узкой специфичности имеют меньшую практическую значимость чем универсальные, такие как, например, шкала Апгар. Но, к сожалению, не всегда такие методы позволяют корректно учесть все влияющие параметры и получить достоверный ответ. Это связано с особенностями медицинской информации. Решения в медицинских и биологических задачах зависят от большого количества неодинаковых по значимости факторов, которые, как правило, взаимосвязаны между собой.

Отбор информативных факторов риска. Среди статистических методов, направленных в первую очередь на выявление сочетаний факторов риска, следует отметить регрессионный анализ и множественную логистическую регрессию. Применение регрессионных методов в анализе данных активно развивается с середины XX века и является достаточно мощным и универсальным средством. Но, к примеру, логистическая регрессия используется только для задач бинарной классификации да и регрессионная модель не всегда пригодна для качественного предсказания. При решении задач регрессионной моделью на значение прогнозируемой переменной не налагаются никакие ограничения, но на практике они могут быть, причем существенными.

Для повышения уровня эффективности методов отбора информативных данных, расширяя их возможности применяются различные методы искусственного интеллекта. Одним из самых перспективных направлений является использование эволюционных вычислений [6], которые используют алгоритмы поиска, оптимизации или обучения, основанные на некоторых формализованных принципах естественного эволюционного отбора. Эволюционные вычисления используют различные модели эволюционного процесса. Среди них можно выделить следующие основные парадигмы: генетические алгоритмы; эволюционные стратегии; эволюционное программирование; генетическое

программирование. Отличаются они, в основном, способом представления искомых решений и различным набором используемых в процессе моделирования эволюции операторов.

Генетические алгоритмы [7] являются очень эффективным инструментом поиска в комбинаторных задачах, какой и является задача отбора факторов риска. Схема работы генетического алгоритма для решения подобной задачи [8] − каждый возможный вариант набора входных переменных можно представить в виде строки битов. Ноль в соответствующей позиции означает, что данная входная переменная не включена во входной набор, единица – что включена. Таким образом, входной набор представляет собой строку битов – по одному на каждую возможную входную переменную – и генетический алгоритм оптимизирует такую битовую структуру. Алгоритм следит за некоторым набором таких строк, оценивая каждую из них по контрольной ошибке. По значениям ошибки производится отбор лучших вариантов наборов, которые комбинируются друг с другом с помощью искусственных генетических операторов: скрещивания и мутации.

Выводы. В результате проведенного анализа выявлено, что, как правило, медицинские данные содержат большой объем данных, включающий слишком много разнообразных параметров, поэтому необходимо выполнять тщательный анализ и отбор полученных от врачей данных. Особенность медицинских факторов риска состоит в следующем: большое количество неодинаковых по значимости параметров, которые взаимосвязаны между собой. Поэтому применение традиционных методов и подходов, использующих статистический анализ, может не обеспечить желаемый результат. При выборе метода отбора данных для решения конкретной задачи предлагается применять различные гибридные алгоритмы поиска значимых факторов, которые используют как классические (статистические), так и эволюционные методы.

Литература.

1. Інформаційні технології в біології та медицині: Курс лекцій: Навчальний посібник / [Гриценко В.І., Котова А.Б., Вовк М.І. та ін.]. – Київ: Наукова думка, 2007. – 382 с.
2. Платонов А.Е. Статистический анализ в медицине и биологии: задачи, терминология, логика, компьютерные методы. – М.: Издательство РАМН, 2000. 52 с.
3. Юренков В.И. Математико-статистическая обработка данных медицинских исследований / Юренков В.И., Григорьев С.Г. – СПб.: – ВМедА, 2002. – 266 с.
4. Скобцов Ю.А. Подготовка данных при разработке медицинских экспертных систем / Скобцов Ю.А., Васяева Т.А. // Вісник Херсонського національного технічного університету. – 2007. – № 4(27). – С. 49-55.
5. Ивахненко А.Г. Самоорганизация прогнозирующих моделей / Ивахненко А.Г., Мюллер И.А. – К.: Техника, 1985. – 223 с.
6. Скобцов Ю.О. Основи еволюційних обчислень / Скобцов Ю.О. – Донецьк: ДонНТУ, 2009. – 316 с.
7. Рутковская Д. Нейронные сети, генетические алгоритмы и нечеткие системы / Рутковская Д., Пилинский М., Рутковский Л. [Пер. с польск. Рудинского И.Д.]. – М.: Горячая линия – Телеком, 2006. – 452 с.
8. Скобцов Ю.А. Применение искусственного интеллекта для вычисления информативности признаков в медицинских задачах / Скобцов Ю.А., Скобцов В.Ю., Васяева Т.А. // Компьютерные науки и информационные технологии. Тезисы докладов Международной науч. конф., посвященной памяти профессора А.М. Богомолова (Саратов, 2-4 июля 2007). – Саратов, 2007. – С. 110-112.