УДК 004.048:004.622

Е.Я. Сиротюк, Т.А. Васяева
Донецкий национальный технический университет, г. Донецк
кафедра автоматизированных систем управления
E-mail: evgeniy_sirotiuk@mail.ru, vasyaeva@gmail.com

ОТБОР ОПТИМАЛЬНОЙ РЕГРЕССИОННОЙ МОДЕЛИ ДЛЯ ПРОГНОЗИРОВАНИЯ ПЕРИНАТАЛЬНОГО РИСКА С ПОМОЩЬЮ МЕТОДА ГРУППОВОГО УЧЕТА АРГУМЕНТОВ

Аннотация

Сиротюк Е.Я., Васяева Т.А. Отбор оптимальной регрессионной модели для прогнозирования перинатального риска с помощью метода группового учета аргументов. *Разработан метод выбора оптимальной регрессионной модели для прогнозирования перинатального риска. Рассматривается сопутствующая задача отбора факторов риска, влияющих на перинатальный риск у беременных.*

Ключевые слова: *факторы риска, перинатальный риск, оптимальная регрессионная модель, метод группового учета аргументов*

Введение. Ежедневно в мире от осложнений, связанных с беременностью и родами, умирает 1500 женщин. По оценкам экспертов большинство этих случаев можно было предотвратить. В настоящее время в медицине особое значение приобретает направление, связанное со снижением перинатальной смертности. Перинатальным периодом называется период, начинающийся с 28-й недели внутриутробного развития, когда масса плода достигает 1000 г и более, и продолжающийся до 8-го дня (168 ч) жизни новорожденного. При всей своей относительной непродолжительности перинатальный период является важнейшим этапом в жизни человека, так как смертность в этот период такая же, как смертность в возрасте человека от 8 дней и до 40 лет, а опасность тяжелых неврологических нарушений в этот период даже превышает таковую в последующие десятилетия жизни человека.

Наиболее действенный путь в снижении перинатальной смертности лежит в разработке программ прогнозированиях перинатального риска [1]. Сложность их разработки 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 [2]. Поэтому, кроме задачи прогнозирования, не менее важной является задача отбора факторов риска, так как анализ всей доступной информации, как правило, вызывает существенные затруднения при разработке и реализации методов прогнозирования при создании аналитической системы.

Постановка проблемы. Достаточно часто для решения задачи прогнозирования используются регрессионные модели. Для построения регрессионной модели необходимо выполнить отбор необходимых факторов и затем рассчитать коэффициенты уравнения. При выборе того или иного набора параметров можно получать различные регрессионные модели, причем многие из них будут показывать хорошие результаты. Всегда при построении математической функции классификации или регрессии основная задача сводится к выбору наилучшей модели из всего множества вариантов.

Однако может существовать множество функций, одинаково классифицирующих одну и ту же обучающую выборку (рис. 1).

Теория множественности моделей [3] утверждает, что по экспериментальным данным принципиально нельзя найти единственную модель. Например, в качестве полинома регрессии можно взять полином любого вида и любой степени, и для каждого из них регрессионный анализ укажет значения коэффициентов. В любом достаточно сложном уравнении подбираются оценки коэффициентов так, чтобы ошибка на интервале наблюдения (интерполяции) была мала или даже равна нулю. Отсюда следует, что для каждого объекта, рассматриваемого как некоторый «черный ящик», можно составить не одну единственную, а бесконечное множество моделей, имеющих одинаковые или почти одинаковые внешние проявления. Решение вопроса о выборе единственного уравнения регрессии оптимальной сложности дает принцип внешнего дополнения. Только внешний критерий приводит к единственной модели оптимальной сложности.

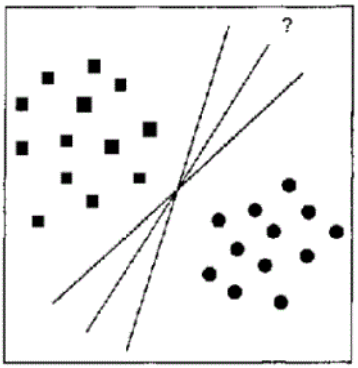

Рисунок 1 – Варианты линейного разделения обучающей выборки

Ошибка, измеренная на всех экспериментальных точках, не является внешним дополнением. Поэтому достаточно часто применяется разделение таблицы исходных данных на две части, называемые обучающей и проверочной, причем ошибка (например, среднеквадратическая),

определяемая на проверочных данных, может являться критерием для выбора структуры модели, т.е. внешним дополнением.

Разработка алгоритма. В большинстве случаев, в медицинских задачах, результат прогнозирования зависит от большого количества неодинаковых по значимости факторов, которые к тому же могут быть взаимосвязаны. Этот факт значительно усложняет этап отбора данных, исключая возможность использовать большую часть известных методов.

Отбор данных для анализа выполняется врачом по следующему принципу: сначала осуществляется выделение факторов риска, относящихся к определенной патологии, 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 выделение факторов риска является не единственной задачей, также необходимо оценить роль каждого из них. Из этого следует, что значимость каждого фактора на риск развития различных акушерских осложнений будет различна [4]. Тем не менее, отбор факторов риска является одним из самых важных этапов построения прогнозирующей модели и в значительной степени определяет ее качество.

Таким образом, при построении оптимальной модели выполним – отбор факторов перинатального риска, из параметров первоначально предложенных врачами, при этом будем учитывать взаимосвязанные между собой переменные.

Полный перебор регрессионных моделей, даже в пределах заданной опорной функции, при достаточно большом наборе входных параметров на практике реализовать не представляется возможным. Для достаточно сложных задач моделирования (например, большой набор обучающих данных) применяются многорядные алгоритмы метода группового учета аргументов

(МГУА) [3]. Многорядный алгоритм МГУА исключает из перебора некоторые модели благодаря наличию порогов.

Предварительно в многорядном (пороговом) алгоритме МГУА на вход подается некоторый вектор входных переменных $x = x_1, x_2, ..., x_n$. На первом ряду селекции образуются «частные описания» (1)-(3), объединяющие входные переменные по две:

$$y_1 = f_{11}(x_1, x_2) = a_{1,0} + a_{1,1}x_1 + a_{1,2}x_2 \quad (1)$$

$$y_2 = f_{12}(x_2, x_3) = a_{2,0} + a_{2,1}x_2 + a_{2,3}x_3 \quad (2)$$

…

$$y_s = f_{1s}(x_{n-1}, x_n) = a_{s,0} + a_{s,n-1}x_{n-1} + a_{s,n}x_n \quad (3)$$

Из них выбирается некоторое число моделей, наиболее удовлетворяющих внешнему критерию селекции. В нашем случае в качестве такого критерия будет среднеквадратичная ошибка (4) на проверочных данных.

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

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

На втором ряду образуются «частные описания» второго ряда:

$$z_1 = f_{21}(y_1, y_2) = b_{1,0} + b_{1,1}x_1 + b_{1,2}x_2 + b_{1,3}x_3, \quad (5)$$

…

$$z_i = f_{2i}(y_i, y_k) = b_{2,0} + b_{2,l}x_l + b_{2,k}x_k + b_{2,m}x_m + b_{2,n}x_n, \quad (6)$$

…

$$z_1 = f_{2p}(y_{S-1}, y_S). \quad (7)$$

Из них также выбирается некоторое количество наилучших для использования в следующем, третьем ряду и т.д. Для каждого ряда находится наилучшая (по критерию селекции) модель (рис. 2). Ряды селекции наращиваются, пока оценка критерия уменьшается («правило останова»). На последнем ряду лучшая модель будет оптимальной. Коэффициенты в регрессионных моделях рассчитываются методом наименьших квадратов (МНК).

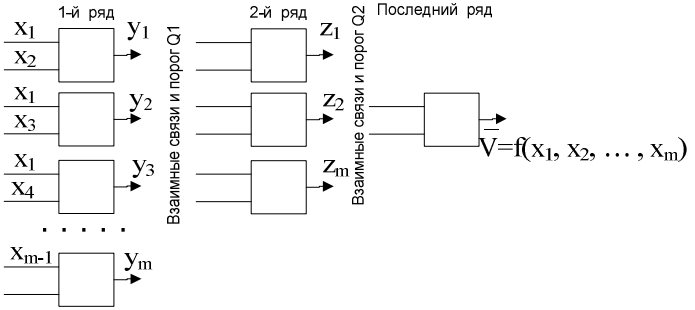

Рисунок 2 – Многорядный МГУА

Выводы. Рассмотрена актуальная задача отбора оптимальной регрессионной модели, для прогнозирования перинатального риска, включая задачу отбора факторов риска. В дальнейшем планируется реализовать рассмотренный математический аппарат и протестировать на реальных медицинских данных, предоставленных сотрудниками центра материнства и детства. Также планируется разработка и внедрение системы прогнозирования перинатального риска.

Список литературы

1. Радзинский В.Е. Акушерский риск. Максимум информации минимум опасности для матери и младенца. / В.Е. Радзинский, С.А. Князев, И.Н. Костин. – Изд.: Эксмо. – 2009 г. – С. 285

2. Т.А. Васяева Применение метода группового учета аргументов для отбора оптимальной регрессионной модели прогнозирования потери крови при родах// Т.А. Васяева/ Вестник Херсонского национального технического университета. – 2012. – № 1(44). – С. 374.

3. Ивахненко А.Г. Самоорганизация прогнозирующих моделей / Ивахненко А.Г., Мюллер И.А. – К.: Техника, 1985. – 223 с.

4. Т.А. Васяева Анализ методов отбора факторов риска развития патологий в акушерстве и гинекологии / Т.А. Васяева, Д.Е. Иванов, И.В. Соков, А.С. Сокова // Збірка матеріалів II Всеукраїнської науково-технічної конференції студентів, аспірантів та молодих вчених.ІУС КМ-2011 11–13 квітня 2011р., Донецьк: ДонНТУ, 2011. – № 1. – С. 209 –