УДК 681.3

О.Н. Чопоров, С.В. Болгов, И.И. Манакин

ОСОБЕННОСТИ ПРИМЕНЕНИЯ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ И МНОГОУРОВНЕВОГО МОНИТОРИНГА ПРИ РЕШЕНИИ ЗАДАЧИ РАЦИОНАЛИЗАЦИИ МЕДИЦИНСКОЙ ПОМОЩИ

Воронежский институт высоких технологий
Орловский государственный университет
Воронежская государственная медицинская академия им. Н.Н. Бурденко

Рассматривается вопрос использования методов интеллектуального анализа данных при исследовании медицинских систем различного уровня. Сформулированы задачи, возникающие при решении вопроса рационализации медицинской помощи. Определена роль многоуровневого мониторинга при решении поставленных задач и представлен обзор рекомендуемых методов построения прогностических и классификационных моделей. Обоснована необходимость использования процедур предварительной обработки информации.

Ключевые слова: интеллектуальный анализ данных, многоуровневый мониторинг, кластерный анализ, прогнозирование заболеваемости.

На сегодняшний день для решения управленческих задач в медицине все чаще используется методы математического моделирования, системный подход и системный анализ [1-17], помогающие получить возможные варианты решения, прогнозировать последствия принятых решений и оценить их с медицинской и социальной точки зрения.

При рационализации медицинской помощи целесообразно использование комплексного многоуровневого подхода к решению данной проблемы, при котором можно выделить следующие задачи [14]:

1) анализ и прогнозирование заболеваемости населения на федеральном, региональном и муниципальном уровнях;
2) исследование медико-социальных факторов риска развития исследуемого заболевания на индивидуальном уровне;
3) прогнозирование течения исследуемого заболевания и выбор адекватной тактики лечебно-профилактических мероприятий.

Решением первой задачи является комплексный мониторинг с оценкой заболеваемости и качества медицинского обслуживания больных по различным территориальным единицам; краткосрочное и долгосрочное прогнозирование развитие ситуации. В результате полученной оценки можно выделить территориальные единицы с благоприятной и неблагоприятной текущей ситуацией и прогнозом ее развития, что является основой для принятия адекватных административных решений органами управления здравоохранением региона.

Результатом решения второй задачи является проведение медико-социального исследования, включающего анализ факторов риска развития исследуемого заболевания и построение прогностических моделей, позволяющих оценить вероятность развития заболевания на индивидуальном уровне. Полученный прогноз является основой для принятия решения о потребности в проведении профилактических мероприятий и может быть использован при формировании диспансерных групп.

Задача прогнозирования течения заболевания и выбора адекватной тактики лечебно-профилактических мероприятий появляется при диагностировании рассматриваемой патологии. Разработка соответствующих моделей для оценки тяжести заболевания и прогнозирования его течения, а также алгоритмов выбора адекватной тактики лечения, оказывает существенную помощь при принятии обоснованного управленческого решения.

Основой решения первых двух сформулированных задач является проведение комплексного мониторинга на региональном и индивидуальном уровне с использованием при обработке полученных результатов адекватных методов математической статистики, и разработка соответствующих классификационно-прогностических моделей.

Прогностические модели можно использовать для прогнозирования динамики контролируемых показателей, прогнозирования развития заболевания, а также для проведения имитационного эксперимента с целью проигрывания различных ситуаций и выбора оптимального управляющего воздействия.

Еще одной из основных задач, возникающих при проведении мониторинга, является задача классификации (выделение территориальных единиц со схожей ситуацией, групп больных с одинаковым диагнозом и др.).

Для решения задач классификации и прогнозирования широкое распространение получили следующие методы моделирования [14]:

- регрессионный анализ;
- анализ временных рядов;
- кластерный анализ;
- дискриминантный анализ;
- нейросетевое моделирование;
- «деревья решений» и др.

Регрессионный анализ – статистический метод исследования зависимости между зависимой переменной и одной или же несколькими независимыми переменными.

Целью регрессионного анализа является: 1) определение наличия и характера (математического уравнения, описывающего зависимость) связи

между переменными; 2) определение степени детерминированности вариации критеральной переменной предикторами; 3) определение вклада независимых переменных в вариацию зависимой.

В отличие от корреляционного анализа, который только отвечает на вопрос, существует ли связь между анализируемыми признаками, регрессионный анализ дает и ее формализованное выражение. При этом если корреляционный анализ изучает любую взаимосвязь факторов, то регрессионный – одностороннюю зависимость, т.е. связь, показывающую, каким образом изменение факторных признаков влияет на признак результативный.

Общая формула уравнения регрессии представляется в виде полинома [12]

$$y = b_0 + \sum_{j=1}^{k} b_j x_j + \sum_{l,j=1}^{k} b_{lj} x_l x_j + \sum_{j=1}^{k} b_{jj} x_j^2 \qquad (1)$$

где x_j — независимые переменные (k — число переменных, включенных в модель);

y — зависимая переменная (моделируемая величина);

b_{ij} — коэффициенты уравнения регрессии.

В случае, когда зависимая переменная (Y) имеет только два возможных значения («0» или «1») рекомендуется использовать логистическое регрессионное уравнение, принимающее вид [16]:

$$Y = \frac{e^{b_0 + b_1 x_1 + \ldots + b_i x_i + \ldots + b_K x_K}}{1 + e^{b_0 + b_1 x_1 + \ldots + b_i x_i + \ldots + b_K x_K}}, \qquad (2)$$

где Y — моделируемая величина;

$x_1, x_2, \ldots, x_k$ — независимые переменные;

$b_0, b_1, \ldots, b_k$ — коэффициенты регрессионного уравнения.

Использование методов регрессионного анализа требует выполнения ряда условий: 1) все опыты должны быть проведены независимо друг от друга; 2) статистическая природа случайных составляющих должна оставаться неизменной во всех опытах; 3) показатели, включаемые в уравнение регрессии в качестве независимых переменных, не должны быть связаны друг с другом.

Построение уравнений множественной регрессии в большинстве случаев осуществляется путем шагового (многошагового) анализа. Одной из важнейших процедур регрессионного анализа является проверка адекватности модели, так как исследователь должен удостовериться в положительном результате при практическом использовании полученной модели. Выбирая структуру модели, необходимо стремиться к тому, чтобы она была как можно проще, т.е. включала как можно меньше коэффициентов. Сокращение числа коэффициентов ведет к облегчению, как процедуры оценивания, так и использования модели. Для выбора

оптимального набора показателей, включаемых в модель рекомендуется к использованию метод «дискретных корреляционных плеяд» [15, 16].

Для данных представленных в виде временных рядов используются методы адаптивного моделирования и прогнозирования, основой которых является модель экспоненциального сглаживания [12], суть которого состоит в том, что временной ряд сглаживается при помощи взвешенной скользящей средней, в которой веса распределяются по экспоненциальному закону.

Основным преимуществом методов, основанных на экспоненциальном сглаживании, является возможность учета временной ценности информации и адаптация к изменяющимся условиям, что имеет большое значение при нестабильном протекании процессов.

При работе с *методами кластерного анализа* построение процедур классификации основывается на минимаксном критерии. Сущность данного метода заключается в интуитивном представлении понятия класса. Объединение объектов в классы происходит по следующему признаку: объекты внутри класса более «похожи» (более близки), чем объекты из различных классов [14, 16].

Критерий качества кластеризации в той или же другой мере отображает следующие неформальные требования:

а) необходима тесная связь между объектами внутри;

б) объекты разных групп должны быть далеки друг от друга;

в) при других равных условиях распределение объектов по группам должно быть равномерным.

Узловым моментом в кластерном анализе считается выбор метрики (или меры близости объектов), от которого решающим образом зависит результат разбиения объектов на группы при заданном алгоритме разбиения [16].

При определении степени близости между объектами для различных типов данных используются следующие показатели [16, 17].

1. Для количественных шкал - линейное расстояние

$$d_{Lab} = \sum_{i=1}^{I} \left| X_a^i - X_b^i \right|, \tag{3}$$

евклидово расстояние

$$d_{Eab} = \left(\sum_{i=1}^{I} \left(X_a^i - X_b^i \right)^2 \right)^{1/2}, \tag{4}$$

обобщенное степенное расстояние Минковского

$$d_{Pab} = \left(\sum_{i=1}^{I} \left(X_a^i - X_b^i \right)^p \right)^{1/p} \quad (5)$$

2. Для качественных шкал используется коэффициент Хемминга

$$\mu_{ab}^h = S_{ab} / I, \quad (6)$$

где S_{ab} – общее число совпадающих значений свойств

(нулевых и единичных: 1 - наличие свойства, 0 - отсутствие).

При отсутствии предпочтительности той или иной шкалы для каждого показателя из содержательных соображений, осуществляется переход к нормированным данным. При этом возникает необходимость максимального учета качественной специфики показателей и выбора соответствующего способа нормировки. При возможности нормировку необходимо производить по величинам, не зависящим от выборки.

Еще одним узловым моментом является выбор алгоритма кластерного анализа, которых существует достаточно много. Все алгоритмы подразделяются на две группы: иерархические и неирархические. Наиболее распространены иерархические (древовидные) процедуры, среди которых выделяют агломеративные (от слова agglomerate – собирать) и итеративные дивизивные (от слова division – разделять) процедуры [12, 14].

Принцип работы иерархических агломеративных (дивизивных) процедур заключатся в последовательном объединении (разделении) групп элементов сначала самых близких (далеких), а затем все более отдаленных (близких) друг от друга. Большая часть из этих алгоритмов берет за основу матрицы расстоянии (сходства). Из недостатков иерархических процедур стоит отметить сложность их вычислительной реализации. Каждый шаг алгоритма требует вычисления матрицы расстояний, а значит, емкую машинную память и большое количество времени. В связи с этим выполнение таких алгоритмов при числе наблюдений, составляющее больше нескольких сотен, бессмысленна, а в ряде случаев и невыполнима.

Общий принцип работы агломеративного алгоритма состоит в следующем. Первый шаг заключается в том, что каждое наблюдение G_i *(i = 1, 2,..., n)* оценивается как отдельный кластер. Далее на каждом шаге реализации алгоритма выполняется объединение двух наиболее близких друг к другу кластеров, и, ввиду принятого расстояния, по формуле вычисляется матрица расстояний, размерность которой, разумеется, снижается на единицу. Завершение работы алгоритма происходит тогда, когда все наблюдения объединены в один класс. Большая часть программ, которые реализуют данный алгоритм иерархической классификации,

предусматривают графическое представление классификации в виде дендрограммы.

Задачи медицинской диагностики решаются наиболее эффективно агломеративными методами минимальной дисперсии: древовидная кластеризация и двухвходовая кластеризация, а также дивизивный метод k-средних.

Метод древовидной кластеризации предусматривает различные правила иерархического объединения в кластеры: 1) правило «одиночной связи»; 2) правило «полных связей»; 3) правило «невзвешенного попарного среднего»; 4) «взвешенное попарное среднее»; 5) «невзвешенный центроидный»; 6) «взвешенный центроидный»; 7) «метод Уорда» [12].

Эффективность применения представленных правил нормировки и классификации зависит от характера распределения исходных данных, в связи с этим разумно провести классификацию с использованием разных комбинаций нормировок и правил классификации и выбрать наилучший вариант.

Дискриминантный анализ позволяет выявить различия между группами и делает возможным классификацию объектов по принципу максимального сходства. Но существуют некоторые ограничения, которые касаются статистических свойств дискриминантных переменных.

Во-первых, ни одна переменная не может быть линейной комбинацией других переменных, соответственно недопустимым является существование переменных, коэффициент корреляции которых равен 1. Настоящее требование может быть реализовано с использованием методов оптимизации признакового пространства [16].

Другое предположение, принимаемое во многих случаях, состоит в том, что ковариационные матрицы для генеральных совокупностей равны между собой для различных классов. Часто применяемой форме дискриминантного анализа свойственно линейные дискриминантные функции, которые соответствуют простой линейной комбинации дискриминантных переменных. Данный метод наиболее тривиален, т.к. предположение, касаемо одинаковых ковариационных матриц в классах, облегчает формулы вычисления дискриминантных функций, и упрощает проверку гипотез о статистической значимости.

Следующее предположение касается того, что закон распределения для каждого класса является многомерным нормальным, иначе говоря, каждая переменная имеет нормальное распределение при фиксированных остальных переменных. Настоящая гипотеза дает возможность получения точных значений вероятности принадлежности к данному классу и критерия значимости.

Каноническая дискриминантная функция есть линейная комбинация дискриминантных переменных, и имеет следующее математическое представление:

$$f_{jn} = u_0 + u_1 X^1_{jn} + u_2 X^2_{jn} + u_1 X^1_{jn}, \quad (7)$$

где f_{jn} — значение канонической дискриминантной функции для n-го объекта в группе j;

u_i – коэффициенты дискриминантной функции;

X_{ijn} – значение дискриминантной переменной X_i для n-го объекта в группе j.

Коэффициенты u_i для первой функции отбираются так, чтобы ее средние значения для различных классов сильно отличались друг от друга. Коэффициенты второй функции выбираются тем же образом, т.е. необходимо, чтобы соответствующие средние значения максимально отличались по классам, в то время как накладывается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой. Идентично третья функция должна быть некоррелирована с первыми двумя и т.д.

Если заранее известно, сколько классов заключается в исходной выборке, процесс построения моделей прекращается при получении соответствующего числа дискриминантных функций.

Искусственные нейронные сети (ИНС) – это удобный и естественный базис для представления имитационных моделей. Нейросеть можно формально рассматривать [14, 16], как совокупность простых процессорных элементов (часто называемых нейронами), которые обладают полностью локальным функционированием, и объединены однонаправленными связями (называемыми синапсами). После приема некоторого входного сигнала из внешнего мира, сеть пропускает его сквозь себя с преобразованиями в каждом процессорном элементе. Во время прохождения сигнала по связям сети выполняется его обработка, результатом которой является определенный выходной сигнал. В укрупненном виде ИНС делает функциональное соответствие между входом и выходом, и может работать, как информационная модель исследуемой системы.

Деревья решений являются одним из методов автоматического анализа данных. Построение деревьев решений (классификации) – один из наиважнейших способов, которые используется при проведении «добычи данных и разведывательного анализа» (Data Mining), осуществляемый как совокупность методов аналитической обработки больших массивов информации, чтобы обнаружить в них значимые закономерности и/или систематические связи между предикторными переменными, которые после можно употребить к новым совокупностям измерений.

Деревья решений – это последовательные иерархические структуры, которые включает в себя узлы, содержащие правила – логические конструкции вида «если … то …». Конечные узлы дерева – «листья», которые соответствуют найденным решениям и объединяют некоторое количество объектов классифицируемой выборки.

Качество исходного статистического материала сильно влияет на точность моделей и полученных на их основе прогнозов.

Медицинская статистическая информация, обычно, имеет небольшой процент измерений, которые имеют заведомо ложные значения, из-за различных объективных или субъективных причин (ошибки при оценке или измерении показателя, ошибки при записи и т.д.). Эти «выбросы», если их не брать в расчет при рассмотрении, могут оказать значительное влияние на построение математической модели, изменив ее вид. Следовательно, во время построения моделей необходимо применять процедуры фильтрации информации, которое позволяют исключить ошибочные измерения из выбранной статистики [14-16].

Для снижения количества показателей, которые используются при построении модели, реализуется выбор наиболее информативных показателей, а также используются алгоритмы исключения параметрической избыточности, которые позволяют удалить из рассмотрения сильно связанные показатели [14-16].

При использовании большинства методов моделирования одним из требований является полнота информационной базы. Если отсутствует хотя бы один показатель, все данные об очередном больном не рассматриваются, что приводит к снижению объема информационной базы и не позволяет получить адекватные модели. Чтобы решить данную проблему необходимо использовать специальный алгоритм, который позволяет восстановить пропущенные данные с максимальной достоверностью на базе выявленных зависимостей [16].

Вне зависимости от используемого метода моделирования, возникают два вопроса: что является прогнозируемой величиной, и что является входными данными. В большинстве случаев прогнозируемая величина – значение временного ряда на заданном интервале. Иногда цель прогнозирования – не столько сбор значений временного ряда на заданном интервале, а сколько установление вероятности того, что он будет вести себя каким-то образом (возрастать, убывать, находиться в некоторых пределах и т.д.). На индивидуальном уровне прогнозируемая величина это вероятность развития заболевания или динамика его течения.

Одним из определяющих условий рационального планирования и управления медицинской помощью является интерактивный сбор, поиск, накопление разнородной информации, а также возможность получения в

реальном масштабе времени наглядной информации, что достигается посредством применения систем мониторинга.

Необходимо, чтобы система мониторинга представляла собой систему наблюдения, анализа и оценки прогнозных состояний здоровья населения и отдельных больных, определяла причинно-следственные связи между заболеваемостью и факторами, которые на него влияют.

Применение системы мониторинга делает возможным решение следующих задач:

- оценка состояния здоровья населения;
- определение причинно-следственных связей между состоянием здоровья населения и влиянием наиболее значимых факторов на основе системного анализа и оценки риска для здоровья населения;
- выявление причин и определение условий возникновения и распространения заболеваний.
- подготовка возможных предложений для руководителей здравоохранения на различных уровнях, которые необходимы для принятия мер по ликвидации выявленных проблем.

Система мониторинга должна представлять собой базу данных о состоянии здоровья населения, которая формируется в интерактивном режиме при непрерывных системных наблюдений, а также осуществляющей выдачу информации о проведенном анализе и прогнозе заболеваемости в режиме реального времени.

Таким образом, при рационализации медицинской помощи целесообразным является использование комплексного многоуровневого подхода, при котором отдельно рассматриваются задачи, которые связанны с анализом показателей заболеваемости и качества медицинского обслуживания территориальных единиц региона и исследованием индивидуальных медико-социальных факторов риска и прогнозированием на их основе вероятности развития заболевания. На каждом из этапов исследовательского процесса должен выполняться выбор адекватных методов моделирования и принятия решений.

ЛИТЕРАТУРА

1. Гафанович Е.Я. Интеллектуализация выбора медицинских вмешательств при лечении артериальной гипертензии на основе прогностического и оптимизационного оценивания их эффективности / Е.Я. Гафанович, И.Я. Львович // Современные проблемы науки и образования. – 2013. - №4. – С. 132.
2. Гафанович Е.Я. Математические модели и численные методы интеграции диагностических и лечебных технологий / Е.Я. Гафанович,

В.Н. Фролов // Системный анализ и управление в биомедицинских системах. – 2013. – Т.12. - № 4. – С. 976-978.

3. Гафанович Е.Я. Прогнозирование исходов и выбор рационального лечения артериальной гипертензии с применением математических методов / Е.Я. Гафанович, И.Я. Львович // Вестник Воронежского государственного технического университета. – 2013. – Т.9. - №4. – С. 84-86.
4. Компьютерное моделирование и прогностическое оценивание региональной распространенности сердечно-сосудистых заболеваний / Е.Я. Гафанович, Е.Н. Коровин, И.Я. Львович, И.М. Соколов // Фундаментальные исследования. – 2013. - №9-4. – С. 606-610.
5. Куташов В.А. Вопросы оптимизации лечения и реабилитации пациентов с наркотической зависимостью в Центрально-Черноземном регионе Российской Федерации / В.А. Куташов, Л.А. Куташова // Вестник неврологии, психиатрии и нейрохирургии. – 2013. - № 8. – С. 25-33.
6. Куташов В.А. Оптимизация диагностики и терапии аффективных расстройств при хронических заболеваниях / В.А. Куташов, Я.Е. Львович, И.В. Постникова. – Воронеж: ВГТУ, 2009. – 198 с.
7. Куташова Л.А. Оценка диагностической значимости характеристик больных с хроническими заболеваниями для прогнозирования развития депрессивных расстройств / Л.А. Куташова, В.А. Куташов, Г.Я. Клименко // Молодежь и современные информационные технологии: материалы Всерос. молодежной конференции. – Воронеж, 2011. – С. 281-283.
8. Куташова Л.А. Системный анализ эпидемиологического и экономческого применения некоторых психотропных средств / Л.А. Куташова, В.А. Куташов // Системный анализ и управление в биомедицинских системах: журнал практической и теоретической биологии и медицины. – 2013. – Т.12. №2. – С. 340-345.
9. Львович И.Я. Информационная технология интеллектуализации процесса диагностики физического развития детей / И.Я. Львович, О.В. Минакова, В.П. Ситникова // Вестник Воронежского института высоких технологий. – 2008. – №3. – С. 112-115.
10. Львович И.Я. Использование нормализованных оценок для описания медико-биологических параметров пациентов, неоднородных по полу и возрасту / И.Я. Львович, О.В. Минакова, В.П. Ситникова // Вестник Воронежского института высоких технологий. – 2007. – Т.1. – № 2. – С. 29-34.
11. Львович Я.Е. Моделирование биотехнических и медицинских систем / Я.Е. Львович, М.В. Фролов // Под ред. В.Н. Фролова: учеб. пособие. – Воронеж: Изд-во ВГТУ, 1994.

12. Медик В.А. Математическая статистика в медицине: учеб. пособие / В.А. Медик, М.С. Токмачев. – М.: Финансы и статистика, 2007. – 800 с.
13. Преображенский Ю.П. Применение имитационно-семантического моделирования и полумарковских процессов принятия решений в клинической практике / Ю.П. Преображенский, Н.С. Преображенская // Вестник Воронежского института высоких технологий. – 2010. – №6. – С. 83-89.
14. Химина И.Н. Рационализация управления медицинской помощью больным с заболеваниями желудка и двенадцатиперстной кишки на основе комплексного мониторинга и классификационно-прогностического моделирования / И.Н. Химина, В.Н. Эктов, О.Н. Чопоров О.Н. – Воронеж: изд-во «Научная книга», 2014. – 181 с.
15. Чопоров О.Н. Методы предварительной обработки информации при системном анализе и моделировании медицинских систем / О.Н. Чопоров, Н.В. Наумов, Л.А. Куташова, А.И. Агарков // Врач-аспирант. – № 6.2 (55). – 2012. – С. 382-390.
16. Чопоров О.Н. Оптимизация функционирования медицинских систем на основе интегральных оценок и классификационно-прогностического моделирования: дис. … д-ра техн. наук. – Воронеж, 2001. – 325 с.
17. Юнкеров В.И. Математико-статистическая обработка данных медицинских исследований / В.И. Юнкеров, С.Г. Григорьев. – СПб.: ВМедА, 2002. – 266 с.

O.N.Choporov, S.V.Bolgov, I.I.Manakin

APPLYING DATA MINING TECHNIQUES AND MULTILEVEL MONITORING TO SOLVING A PROBLEM OF MEDICAL AID RATIONALIZATION

Voronezh Institute of High Technologies
Oryol State University
Voronezh State Medical Academy of N.N.Burdenko

The question of using methods of the data intellectual analysis is considered through studies of medical systems of various levels. The authors formulate the problems arising while solving a question of medical aid rationalization. The role of multilevel monitoring while solving application tasks is defined and the review of recommended methods of prediction and classification models construction is presented. The necessity of using procedures of preliminary processing of the information is proved.

Keywords: data intellectual analysis, multilevel monitoring, cluster analysis, disease forecasting.