УДК 004.942: 616.037

МЕТОДОЛОГИЯ РАЗРАБОТКИ ТЕХНОЛОГИЙ ИЗВЛЕЧЕНИЯ ЗНАНИЙ ИЗ КЛИНИЧЕСКИХ ДАННЫХ ИНДУКТИВНЫМИ МЕТОДАМИ

Савченко Е.А., Поднебесная Г.А., Долгополов И.Н.

Международный научно-учебный центр информационных технологий и систем НАН и МОН Украины,
savchenko@irtc.org.ua, pidnebesna@irtc.org.ua, dolgigo@ukr.net

На прикладі розв'язання задач клінічної діагностики описана методологія розробки технології аналізу даних та видобування знань за вибіркою експериментальних даних індуктивними методами. Ця технологія включає задачі попередньої обробки вибірки даних, класифікації та кластеризації, а також побудови моделей для керування та прогнозу. З метою керування стану хворого на ішемічну хворобу серця запропонована побудова узагальненої моделі стану хворого.
Ключеві слова: клінічна діагностика, класифікація, кластеризації, керування станом, МГУА.

Development of technology of data analysis and knowledge extraction from an experimental data sample by inductive methods is described on the example of decision of tasks of clinical diagnostics. This technology includes the tasks of the preliminary data processing, classification, clusterization, management and prognosis models construction. The generalized model construction of the patient state with ischemic heart trouble for the patients state management is offered.
Keywords: clinical diagnostic, classification, clusterization, management and prognosis models construction, GMDH.

На примере решения задач клинической диагностики описана методология разработки технологии анализа данных и извлечения знаний из выборки экспериментальных данных индуктивными методами. Эта технология позволяет решать задачи предварительной обработки данных, классификации и кластеризации, а также построения моделей для управления и прогноза. С целью управления состоянием больного ишемической болезнью сердца предложено построение обобщенной модели.
Ключевые слова: клиническая диагностика, классификация, кластеризация, управление состоянием, МГУА.

Введение. С каждым годом число людей с сердечно-сосудистыми заболеваниями неуклонно возрастает. Только в Киеве проживает более миллиона человек с сердечно-сосудистыми заболеваниями, у 500 тысяч человек - ишемическая болезнь сердца. Инфаркты и инсульты "помолодели" и все чаще случаются в возрасте 30 - 40 лет.

Развитие информационных технологий дает возможность врачу использовать компьютер в качестве помощника и советчика на всех стадиях процесса лечения [1]: обработки данных медицинского исследования, постановки диагноза, прогноза течения болезни.

На примере решения всех этапов задачи клинической диагностики [2] разрабатывается технология анализа данных и извлечения знаний в сложных системах на основе индуктивных методов самоорганизации.

Существующие в современной медицине методы оценки состояния больного решают только отдельные этапы задачи клинической диагностики. Они носят выборочный характер, что позволяет говорить об их специфичности от выборки данных конкретного больного.

Данная технология разрабатывается на примере решения задачи постановки диагноза и лечения больного ишемической болезнью сердца (ИБС). Конечная цель данного исследования - построение комплексной модели состояния кардиологического больного, которая включает последовательные стадии развития ИБС. Такая модель даст возможность врачу анализировать состояние больного, чтобы определить факторы, способствующие улучшению его состояния.

Состояние больного ИБС характеризуется рядом показателей, которых измеряется и рассчитывается более 100. Поэтому одной из первых задач, возникающих в работе врача-аналитика, является обработка выборки данных наблюдения за пациентами. Из всех имеющихся данных необходимо извлечь факторы, которые будут информативны для текущего состояния болезни.

Следующим этапом является построение моделей для поиска взаимосвязей в данных, кластеризации данных на отдельные множества, классификации стадий болезни, управления текущим состоянием больного и прогноза его в дальнейшем.

Некоторые из медицинских исследований являются дорогостоящими и не всегда достаточно информативными. Учитывая то, что часто точно не известно, какие именно факторы влияют на состояние больного, найти математическое описание его состояния посредством дедуктивных методов практически невозможно.

Для построения моделей в работе используется метод группового учета аргументов (МГУА), в частности, комбинаторный алгоритм МГУА [3, 4]. Это индуктивный метод автоматического поиска наилучшей модели по выборке экспериментальных данных. Он хорошо известен в мире и зарекомендовал себя как метод, дающий возможность открывать зависимости и извлекать новые знания, которые содержатся в выборке данных, но неизвестны человеку – автору моделирования. На основе знаний, полученных в результате моделирования эксперт, врач, может проводить анализ текущего состояния больного, находить факторы, которые влияют на изменение его состояния, то есть иметь возможность корректировать процесс лечения.

Установление зависимостей между отдельными показателями даст возможность оптимизировать процесс исследования состояния больного, уменьшить количество данных, которые необходимы для проведения исследований, а также прогнозировать влияние на состояние больного различных факторов, которые могут привести к ухудшению его состояния, и даже угрожать его жизни.

Автоматическая классификация состояния больного ИБС позволит ускорить процесс постановки диагноза и сделать объективным процесс принятия решений, нацеленных на выбор метода коррекции состояния больного.

Из всего вышесказанного можно сделать вывод, что актуальной является задача автоматизации процесса анализа данных, чтобы сделать его объективным и разработать компьютерную технологию, которая бы позволяла автоматически извлекать из данных наблюдений пациентов новые нетривиальные знания в форме математических моделей.

Предлагаемая технология состоит из следующих блоков, решающих задачи:

- предварительной обработки выборки наблюдений;
- кластеризации наблюдений на отдельные множества;
- построение моделей:
    - классификации стадий заболевания;
    - прогноза течения болезни;
    - управления состоянием больного.

Рассмотрим каждую из этих задач подробнее.

1. Предварительная обработка выборки данных наблюдений

С бурным развитием технических средств особенно актуальной является проблема анализа большого количества данных наблюдений. Но сами по себе эти данные не обеспечивают максимального объема диагностической информации. Необходимость обработки огромного потока информации очень остро ставит задачу максимально полного и целесообразного ее использования.

Одной из решаемых задач является предварительная обработка выборки данных наблюдений, которая включает различные подзадачи. Кроме накопления данных из различных источников, измеряемых и вычисляемых для каждого пациента, это и выделение из них тех, которые характеризируют степень его заболевания, чтобы по ним спрогнозировать дальнейшее его состояние. Например, по показателям, полученным в состоянии покоя пациента спрогнозировать влияние на его состояние физической нагрузки.

Часто возникают ситуации, когда некоторая информация пропущена, недостаточна или некорректна. В этом случае возникает задача восстановления недостающих данных. В [5] для заполнения пропусков в данных использованы модели, построенные по МГУА.

Характерной особенностью медицинских задач часто является большое число измерений, не все из которых информативны для данного заболевания. В таком случае рекомендуется исключать неинформативные измерения и этим уменьшить число измеряемых параметров, что может значительно удешевить

исследования и предоставить врачу инструмент, удобный для анализа данных с целью постановки верного диагноза.

2. Задача кластеризации

Задача кластеризации в общем случае состоит в разбиении имеющейся выборки объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались [6].

В медицинской практике врач располагает определенной статистикой наблюдений за состоянием больных за некоторый интервал времени. Это позволяет воспользоваться средствами кластеризации для определения групп (кластеров) больных одной стадии заболевания (например, пациентов с ИБС) или по схожести признаков течения болезни (например, в одной возрастной группе, по сопутствующим заболеваниям и т.д.), чтобы обобщить эту информацию и отработать некоторые стандартные принципы лечения для каждой группы. Такое автоматическое разбиение всего множества пациентов на некоторые характерные группы может значительно помочь в работе врача-аналитика.

Существует большое число методов кластеризации и их различные комбинации (ассоциация, деревья решений, метод ближайших соседей, нейронные сети, нечеткая логика, регрессионные методы и многие другие). Наиболее популярны методы расщепления (дивизивные методы), непосредственно разбивающие всю совокупность записей на несколько кластеров. Из них наибольшее распространение получили различные модификации метода K-средних, который хорошо работает, если данные по своей естественной природе делятся на компактные, примерно сферические группы.

Другие методы кластеризации – объединительные (аггломеративные), начинаются с создания элементарных кластеров, каждый из которых состоит ровно из одного исходного наблюдения (одной точки), а на каждом последующем шаге происходит объединение двух наиболее близких кластеров в один. Момент остановки этого процесса объединения может задаваться путем указания требуемого числа кластеров или максимального расстояния, при котором допустимо объединение.

Ассоциация, или *метод корзины покупателя* (market basket analysis) является одним из вариантов кластеризации, используемым для поиска групп характеристик в том случае, если несколько событий связаны друг с другом. Использование этого метода целесообразно как один из первых шагов исследования, например, при исследовании характера жалоб больных, обратившихся в клинику в первый раз. Такие пациенты могут не иметь

истории болезни, результатов специализированных анализов, а характеризуются только данными, выявленными при первичном осмотре.

Использование этого метода также целесообразно при исследовании временных рядов, когда необходимо выявить группы нескольких событий, имеющих тенденцию происходить в строго фиксированной последовательности. Основными недостатками метода являются: резкий (экспоненциальный) рост объема вычислений с увеличением числа параметров, фактически полное неприятие в расчет редко встречаемых параметров и ограниченные возможности метода по учету дополнительных знаний о свойствах параметров.

Деревья решений - эти методы используют подход, который основан на изучении условных вероятностей. Одним из наиболее важных свойств деревьев решений является представление данных в виде иерархической структуры. Существует большое количество вариантов алгоритмов, строящих деревья решений. Используется подход, где критерий расщепления данных реализуется в виде перцептрона. Ввиду того, что медицинское знание не является точным, по-видимому, перспективно использование нечетких критериев расщепления.

Алгоритмы деревьев – один из самых быстрых и эффективно реализуемых в программных продуктах, поэтому они получили широкое распространение. Их вычислительная сложность определяется главным образом типом применяемого критерия расщепления. К сожалению, применяемые на практике системы часто строят очень большие деревья даже для не очень больших баз данных. Кроме того, такие структуры трудно использовать для понимания сути влияния тех или иных независимых параметров на результат.

Нечеткая логика. В реальных задачах, в том числе и медицинских, часто приходится сталкиваться с задачами, имеющими элемент неопределенности: неполная, неточная, противоречивая информация (например, записана со слов больного). Для анализа таких наборов данных, когда невозможно причислить данные к какой-либо группе применяются нечеткая логика и алгебра. Можно отнести данные к какой-либо группе только с некоторой вероятностью находящейся в интервале от 0 до 1, но не принимающей крайние значения. Четкая логика манипулирует результатами, которые могут быть либо истиной, либо ложью. Нечеткая логика применяется в тех случаях, когда необходимо манипулировать степенью “может быть” в дополнении к “да” и “нет”.

Метод Объективной Компьютерной Кластеризации (ОКК) является одним из методов автоматического разбиения выборки наблюдений на кластеры [7, 3]. В алгоритмах объективного кластерного анализа кластеры образуются по внутреннему критерию, а оптимальное их число и состав ансамбля признаков определяются по внешнему критерию. Этот метод

предлагается использовать в технологии анализа данных для решения задачи кластеризации, как метод автоматической кластеризации.

3. Построение моделей

Построенные индуктивными методами модели могут быть использованы в клинической диагностике для решения задач: классификации, управления и прогнозирования.

Задача кластеризации аналогична задаче классификации, но отличается тем, что для классификации требуется задать целевую функцию.

Задача классификации. На основании экспертных знаний и имеющихся статистических наблюдений, описаний объекта определяется фиксированное число классов (групп, таксонов), определяющих различные состояния процесса. Необходимо получить правило, по которому любой новый объект может быть определен в один из указанных классов.

Например, ИБС экспертами разделяется на три стадии. Задача классификации стадии заболевания пациента, состоит в том, чтобы для каждого текущего состояния болезни конкретного больного определить, к какой из трех стадий, определенных врачом относится его болезнь.

Формальная постановка задачи классификации состоит в следующем. Пусть $\boldsymbol{X}$ — множество наблюдений, $\boldsymbol{Y}$ — конечное множество классов. Существует неизвестная целевая функция $y^*\colon X \to Y$ — значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1,y_1),\cdots,(x_m,y_m)\}$. Требуется построить алгоритм $a\colon X \to Y$, способный классифицировать произвольный объект. Классификация относит каждый объект к одной из заранее определенных групп.

Задача управления. Целью управления состоянием больного является корректировка (улучшение) его состояния. Для решения этой задачи строится обобщенная модель болезни с тем, чтобы корректировкой одного из показателей, влияющих на состояние болезни, перевести больного из худшей стадии заболевания в лучшую. Для этого необходимо выделить параметры, описывающие каждую стадию болезни, и найти зависимость выходной величины от этих параметров, включая и управляемые параметры.

Задача прогноза. Именно от прогноза зависит решение врача-аналитика о продолжении или смене методов лечения. Например, по данным состояния покоя необходимо спрогнозировать, как будет реагировать организм конкретного больного на испытание с физической нагрузкой, и какая нагрузка может оказаться опасной для его здоровья.

Таким образом, врач по данным, характеризирующим состояние болезни, с помощью компьютерной технологии сможет получить рекомендации системы о стандартном подходе к лечению группы больных, к которой компьютер отнесет текущего больного. Далее, промоделировав возможную реакцию больного на некоторые факторы, например, физическую нагрузку, врач сможет

внести коррективы в схему лечения, а также спрогнозировать, изменение каких параметров приведет к улучшению/ухудшению состояния больного с ИБС.

4. Технология извлечения данных и знаний на основе МГУА

В результате проведенного выше исследования, технология анализа данных и поиска закономерностей может быть представлена следующим образом (рис.1).

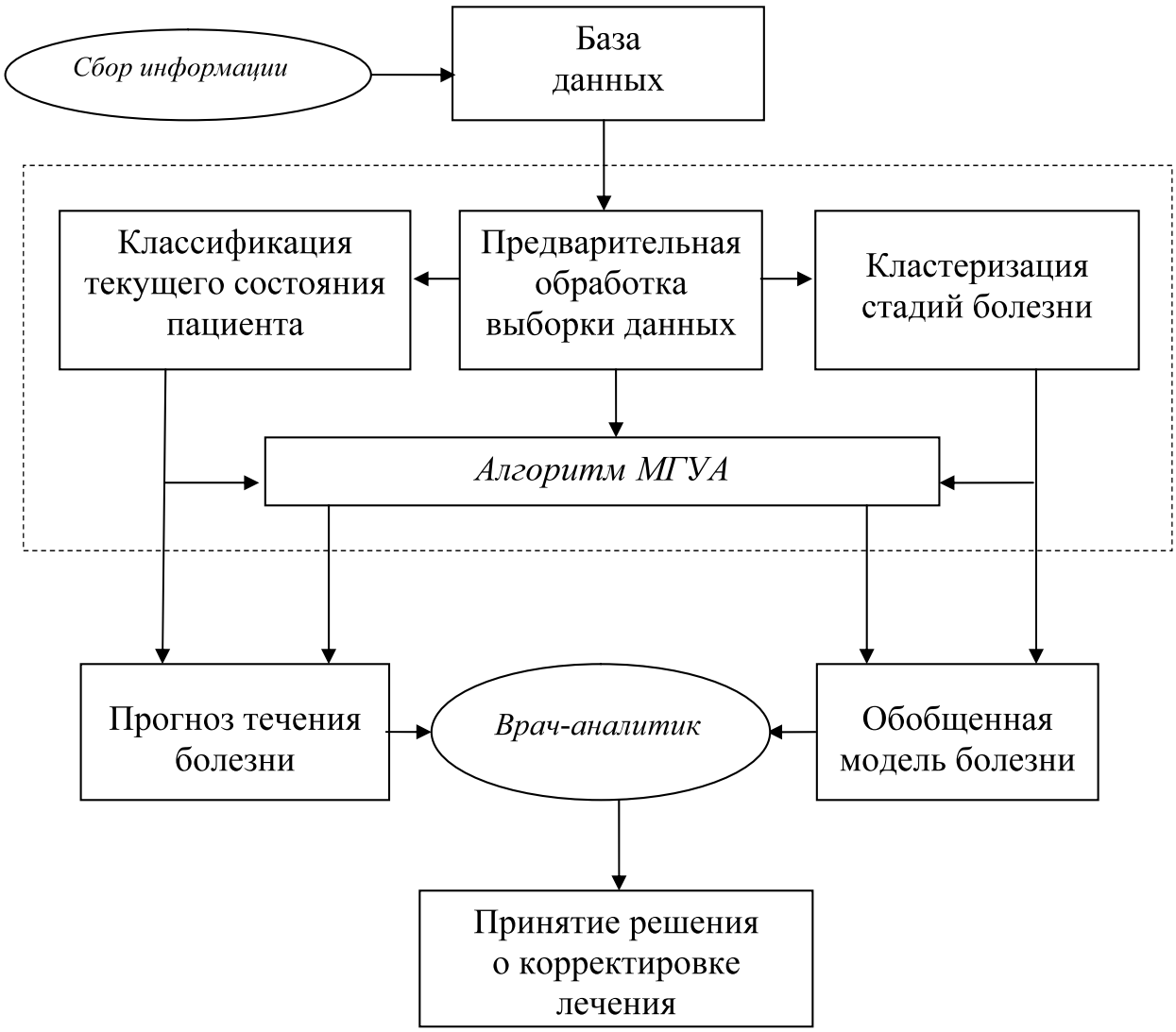

Рис.1 Схема технологии извлечения данных и знаний в клинической диагностике

На основании наблюдений за пациентом формируется база данных, извлекается необходимая информация о состоянии пациента, обрабатывается, т.е. если необходимо, данные восстанавливаются, в них находятся взаимосвязи, если данных много, отбираются информативные переменные. Затем в зависимости от решаемой врачом задачи, строятся модели для классификации,

кластеризации, прогноза или управления состоянием больного. Врач на основе полученных моделей принимает решение о продолжении лечения или его изменении.

Заключение

В настоящее время в существующих технологиях в качестве конечного результата методов извлечения знаний рассматриваются задачи построения математических моделей исследуемых процессов. Предлагается разработка технологии, основанной не только на классификации выявленных знаний и построении модели, но и дающая возможность управления процессом. В клинической информатике управление состоянием больного является значительной проблемой в виду сложности и обусловленности связей параметров, определяющих текущее состояние пациента.

В рамках данного исследования будут разработаны методы и технологии, которые могут быть использованы для построения компьютерной технологии извлечения знаний и выявления закономерностей в сложных системах. Результаты этих исследований могут быть использованы для построения автоматизированного рабочего места врача-аналитика, позволяющего исследовать массивы данных для определения в них кластеров состояний, классификации и извлечения скрытых закономерностей, необходимых для построения обобщенных моделей управления состоянием больного.

Литература

1. Арсеньев С. Извлечение знаний из медицинских баз данных. // Москва, Мегапьютер. – 1999. - http://www.megaputer.ru.

2. Долгополов И.Н. Интеллектуальная система управления процессами клинической диагностики // Проблемы управления и информатики. - 2007. - №3. – С.146 –154.

3. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. – Киев: Наук. думка. – 1986. – 216 с.

4. Ивахненко А.Г. Непрерывность и дискретность. - Киев: Наук. думка. – 1990. – 224 с.

5. Ивахненко А.Г., Савченко Е.А., Ивахненко Г.А., Гергей Т. Применение алгоритмов МГУА для восстановления пропущенных данных и прогноза уровня глюкозы в крови при надомном мониторинге диабета // Проблемы управления и информатики.– 2002.–№3.– С.123-133.

6. Сарычева Л.В. // Проблемы управления и информатики.– 2008.– №2.– С. 86 -104.

7. Ивахненко А.Г., Лу И., Семина Л.П., Ивахненко Г.А.. Объективная компьютерная кластеризация // Автоматика. – 1987. - №1. –С.3-17.