РАЗДЕЛ 1

АНАЛИЗ СОВРЕМЕННОГО СОСТОЯНИЯ ВОПРОСА, ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

1.1. Современные научные подходы, направления и методы предсказания в медицине.

В последнее десятилетие стремительно возрастает значение информационного обеспечения современных технологий. Оно становится критическим фактором развития практически во всех областях знания, в частности, в современных медицинских технологиях. Поэтому разработка и внедрение информационных компьютерных систем является в настоящее время одной из самых актуальных задач в медицинских учреждениях.

Анализ применения персональных компьютеров в медицине показывает, что главным образом компьютер используются для обработки текстовой документации, хранения и обработки баз данных, ведения статистики и финансовых расчетов. С другой стороны все большая часть компьютеров используется совместно с различным диагностическим и лечебным оборудованием.

В большинстве этих областей использования компьютеров применяется стандартное программное обеспечение - текстовые редакторы, системы управления базами данных, статистические пакеты и др. Однако на важнейших участках лечебно-диагностических технологий в настоящее время совершенно не достаточно используются возможности современной компьютерной техники. Прежде всего, это относится к диагностированию, назначению лечебных мероприятий, прогнозированию течения заболеваний и их исходов. Причины этого явления носят сложный характер и постоянно дискутируются. Основными из них являются недостаточно развитая техническая база и низкая компьютерная грамотность в отечественных медицинских учреждениях. Большое значение имеет также психологический

аспект применения компьютеров в процессах постановки диагноза и назначения лечения.

Разработка математических методов решения медицинских задач ведется уже несколько десятков лет. Специалистами предложено огромное количество вариантов проверки гипотез и получений правил, широко используются статистические методы анализа, вариационные ряды, моделирование, кибернетический анализ, методы искусственного интеллекта. В настоящее время предсказание в медицине, в основном, выполняется статистическими методами. Это позволяет давать прогностическую оценку в числовом выражении, делая ее информативной, выполнять анализ доказательно, а не интуитивно. Но, к сожалению, не всегда такие методы позволяют учесть все влияющие параметры и получить однозначный ответ. Это связано с особенностями медико-биологической информации. Решения в медицинских и биологических задачах, зависят от большого количества неодинаковых по значимости факторов, которые к тому же взаимосвязаны между собой. Поэтому, даже если удается получить совокупность правил, определяющих порядок обработки информации с целью постановки диагноза, алгоритм хорошо работает только для тех данных, на которых проводились исследования. При использовании алгоритма для других, даже подобных объектов, правила вывода почти всегда приходится полностью разрабатывать заново. Получить универсальные правила методами, которых объединяет наличие явных алгоритмов принятия решений, не представляется возможным. Многолетние исследования, которые проводились с самыми различными явными алгоритмами, показали, что медицинские задачи, имеющие неявный характер, решаются этими методами с недостаточной точностью и удобством, что делает их непривлекательными для широкого использования в практических задачах диагностики, прогнозирования и принятия решений [39].

Умение анализировать и классифицировать данные, а так же приобретать новые знания является специфической чертой практикующего врача. Но эта способность ограничена сравнительно небольшим объемом данных. Недавний прорыв в области компьютерных технологий, основанный на использовании баз данных, привел к тому, что получили возможность записывать и хранить громадные объемы информации, накопленные за многие годы работы. Появилась возможность, а вместе с тем и необходимость анализировать информацию, содержащую тысячи объектов, описанных с помощью сотен параметров. Ситуация осложняется еще и тем, что данные могут содержать ошибки в результате неточных измерений, кроме того некоторые из них вообще могут быть пропущены. Для такого класса задач оправдывает себя применение искусственного интеллекта. Методы искусственного интеллекта не предъявляют строгие требования к исходным данным, позволяют извлекать закономерности, обрабатывая большие объемы информации и получать при этом хорошие практические результаты. Поэтому их целесообразно применять для решения неявных задач медицины и биологии.

Несмотря на то, что в медицине этап получения правил, является наиболее важным, задача не сводится только к методам извлечения закономерностей. Задачи диагностики, прогнозирования и принятия решений в медицине – это комплексный процесс, который начинается с получения и представления данных и заканчивается оценкой качества полученных решений. В целом весь процесс можно разделить на следующие этапы:

- отбор данных;
- предобработка данных;
- редукция данных;
- поиск закономерностей;
- оценка и интерпретация найденных закономерностей;
- использование полученных знаний для решения задачи.

Анализ публикаций о применении методов искусственного интеллекта в медицине показывает, что практически отсутствуют какие-либо методологии разработки медицинских систем, о чем свидетельствует как отсутствие работ такого профиля, так и огромное разнообразие подходов к алгоритмам извлечения знаний и построению экспертных систем. Это подтверждает то, что диагностика в области медицины как наука находится еще, в основном, на стадии накопления фактического материала.

1.2. Современные взгляды на проблему синдрома внезапной смерти грудного ребенка

На протяжении нескольких десятилетий внимание ученых всего мира привлекает одна из самых драматичных и до сих пор не выясненных проблем медицины – синдром внезапной смерти грудного ребенка (СВСГР).

Интерес к данной проблеме не прекращается, прежде всего потому, что число жертв СВСГР не имеет тенденции к снижению. СВСГР является одной из ведущих причин смерти младенцев в развитых странах и ежегодно уносит жизни нескольких тысяч детей грудного возраста, а в ряде зарубежных стран и вовсе выходит на лидирующие позиции, оставляя за собой другие, известные причины детской смертности [2,3,65,80,85,90,93].

В 1969 г. в Сиэтле на 2-й Международной конференции по внезапной детской смерти был принят существующий в настоящее время термин – «SUDDEN INFANT DEATH SYNDROME – синдром внезапной смерти младенцев» и по предложению J.B. Beckwith дано его определение: «Внезапная, неожиданная смерть грудного ребенка с отсутствием адекватных для объяснения причин смерти данных анамнеза и патологоанатомического исследования» [67]. СВСГР является «диагнозом исключения», то есть ставится на основании отрицания других возможных причин смерти. В 1985 г. в Канберре на 1-й Австралийской конференции по внезапной детской смерти обсуждались статистика, эпидемиология, морфология и ряд факторов риска, характерных для СВСГР. Отмечалось, что СВСГР включает случаи

внезапной и необъяснимой смерти в колыбели младенцев, считавшихся здоровыми или легко больными [4].

Частота СВСГР на протяжении ряда десятилетий остается в одних странах на одном уровне, в других – имеется тенденция к росту. В развитых странах, где детская смертность ниже 20 на 1000 новорожденных, СВСГР является ведущей причиной детской смертности, выходя на 1-е место [76,85,90,93]. В среднем число случаев СВСГР составляет 2 на 1000 новорожденных – от 0,3‰, в Израиле до 4,4‰ в Тасмании, в Италии – 0,57‰, во Франции – 1,2-1,6‰, в Германии – 1,3-1,6‰, в Англии – 2,3‰, в США – 1,4-2,8‰ [4, 18].

В Украине, по данным статистики, ежегодно около 200 детей умирают по этой причине, при этом есть предположение, что данные могут быть занижены. Существование данной проблемы долгое время просто игнорировалось, и фактические случаи СВСГР регистрировались под другими, чаще всего респираторными или кишечными заболеваниями [4, 18]. Даже сейчас эту проблему и пути ее решения не знают не только родители, но и большая часть медицинских работников.

Несмотря на наличие большого числа факторов риска, относящихся непосредственно к СВСГР, следует отметить, что имеется немало случаев СВСГР, не имеющих ни одного из факторов риска. В то время как младенцы с наличием даже нескольких перечисленных факторов риска не погибают внезапно. В связи с этим логично предположить, что факторы риска СВСГР могут иметь значение при определенном фоновом состоянии [28,70,75,79] или в определенной совокупности.

СВСГР не имеет характерной клинической картины – внешне здоровый ребенок спокойно засыпает в своей кроватке, и через несколько часов его обнаруживают мертвым.

На протяжении нескольких десятилетий существовали самые различные теории, объясняющие случаи СВСГР. Возможными причинами смерти считали перегрев, закрытие дыхательных путей подушкой,

невосприимчивость к коровьему молоку, судороги при рахите, ненормальное развитие ткани легких, увеличение вилочковой железы. В связи с тем, что наибольшая частота случаев СВСГР наблюдается во время сна, особое внимание исследователей привлекал поиск возможной связи данного синдрома с особенностями регуляции функции дыхания у детей во сне. Большое значение в ряду причин, способных вызвать смерть младенца, придавалось инфекционным заболеваниям, особенно пневмонии. Исследования показывают, что в 30% случаев погибшие дети имели инфекции верхних дыхательных путей, отит, бронхит, диарею, рвоту незадолго до смерти [68,78]. Существует концепция, связывающая СВСГР с «рефлексом паралича страха» у детей [4]. Применительно к СВСГР пусковыми факторами чаще всего являются умеренно выраженные острые респираторные вирусные инфекции (ОРВИ). К ним же, вероятно, могут принадлежать и изменения температуры воздуха и резкие и незнакомые звуки, незнакомые объекты и непривычное окружение ребенка, резкое пробуждение от сна, тугое пеленание и ограничение движений, и многие другие. Несколько исследований констатировали важность грудного вскармливания для профилактики СВСГР. Следует заметить, что курение и искусственное вскармливание часто бывают взаимосвязано. Большинство курящих матерей, если им не удается отказаться от курения во время беременности, не кормят детей грудью именно потому, что курят. Какой фактор при этом действует сильнее тоже не ясно.

Таким образом, большинство исследователей соглашаются с мнением о том, что данный феномен может реализоваться у разных младенцев под воздействием различных, конкретных для определенного случая факторов [63,91,92]. Наличие большого числа теорий и гипотез, объясняющих причины СВСГР, по существу подчеркивают его неясность. При этом основная часть теорий сводится к выявлению либо пускового механизма, приводящего к внезапной смерти (вирусы, прививки, нарушения сердечного ритма и др.), либо к выяснению особенностей организма (врожденные,

генетические, фоновые состояния), которые могут предрасполагать к реализации СВСГР. Принципы профилактики синдрома внезапной смерти грудного ребенка в литературе освещены очень противоречиво.

Данные литературы подтверждают актуальность данной проблемы, как в мире, так и в Украине. Они свидетельствуют о том, что снижение процента смертности от СВСГР возможно лишь в том случае, если будет проведено повышение индивидуальной медицинской активности населения по предупреждению возникновения факторов риска. Профилактика должна начинаться еще в период беременности, так как 60-70% случаев смертности от СВСГР связано с факторами риска, возникающими в это время.

1.3 Анализ способов представления обучающих данных

Существует огромное количество способов представления информации для различных целей [41].

Рассмотрим классификацию данных, с которыми столкнулись при создании медицинской экспертной системы, а так же способы их представления в численном виде.

Классификация компонентов входных данных

Можно выделить три основных типа входных данных:

1. Числа.

2. Взаимоисключающие качественные варианты.

3. Совместимые качественные варианты.

1. Числа. Данные такого типа могут принимать любые значения. Примером может быть вес или рост новорожденного ребенка, возраст мамы на момент родов и др. Данные в виде чисел не требуют какой-либо кодировки. Они могут использоваться в готовом виде или с применением масштабирования.

2. Взаимоисключающие качественные варианты. Один из наиболее сложных типов данных, который требует продуманного представления. Информация представлена в виде только одного варианта из заранее определенного набора вариантов. Причем нельзя ввести осмысленное расстояние между состояниями. Примером такого признака может быть состояние больного - тяжелое, среднее, легкое. Нельзя сказать, что расстояние от легкого состояния до среднего больше, меньше или равно расстоянию от среднего состояния до тяжелого. Все взаимоисключающие признаки можно классифицировать так:

- неупорядоченные варианты. Признак неупорядоченный, если никаким двум состояниям нельзя сопоставить естественное в контексте задачи отношение порядка. Часто к этому типу относятся данные, представляемые всего двумя вариантами (да - нет, мальчик - девочка, болел - не болел и т.д.);
- упорядоченные варианты. Признак упорядоченный, если для любых двух состояний одно из них предшествует другому. Такие данные находятся в определенных взаимосвязях друг с другом. Они могут быть выражены отношениями типа «больше», «меньше». Примером может служить степень тяжести заболевания (I,II,III) или (легкое состояние < среднее состояние < тяжелое состояние). В любом случае варианты располагаются в определенном порядке - как правило, сортируются по возрастанию или убыванию;
- частично упорядоченные варианты. У таких признаков способ упорядочивания не очевиден, но его можно найти, если это требуется для решения задачи.

3. Совместимые качественные варианты. Информация может быть представлена одним или одновременно несколькими вариантами из заранее определенного набора вариантов. Примером может быть наличие у

пациента каких-либо заболеваний, перенесенных в период беременности. В таких случаях имеется два различных подхода к кодированию данных:

- для неупорядоченных вариантов, признак разбивается на несколько признаков, количество которых равно количеству вариантов. Затем каждый из них кодируется отдельно.
- для упорядоченных вариантов, можно кодировать с использованием принципа битовой маски.

Числовые признаки

Числовые значения могут быть абсолютно разнородными величинами [41]. Очевидно, что результаты моделирования не должны зависеть от единиц измерения этих величин. Чтобы система трактовала значения единообразно, все входные величины должны быть приведены к единому (как правило, единичному) масштабу. Кроме того, иногда полезно провести дополнительную предобработку данных, выравнивающую распределение значений.

Приведение к единому масштабу обеспечивается нормировкой каждой переменной на диапазон разброса ее значений. В простейшем варианте это – линейное преобразование в единичный отрезок: $\tilde{x}_i \in [0,1]$ - формула (1.1).

$$\tilde{x}_i = \frac{x_i - x_{i,\min}}{x_{i,\max} - x_{i,\min}}, \qquad (1.1)$$

где $x_{i,min}$ – минимальное значение переменной; $x_{i,max}$ – максимальное значение переменной; x_i – текущее значение переменной; $\tilde{x}_i$ – преобразованное значение переменной.

Если значения переменной x_i плотно заполняют определенный интервал, то линейная нормировка оптимальна. При наличии в данных относительно редких выбросов, которые согласно формуле (1.1) определяют масштаб нормировки, основная масса значений нормированной переменной $\tilde{x}_i$ сосредоточится около 0. Поэтому при нормировке лучше ориентироваться

не на граничные значения, а на типичные (1.2), т.е. статистические характеристики данных, такие как среднее (1.3) и дисперсия (1.4).

$$\tilde{x}_i = \frac{x_i - \bar{x}_i}{\sigma_i} \tag{1.2}$$

$$\bar{x}_i \equiv \frac{1}{P}\sum_{\alpha=1}^{P} x_i^{\alpha} \tag{1.3}$$

$$\sigma_i^2 \equiv \frac{1}{P-1}\sum_{\alpha=1}^{P} (x_i^{\alpha} - \bar{x}_i)^2 , \tag{1.4}$$

где x_i – текущее значение переменной; $\tilde{x}_i$ – преобразованное значение переменной, $\bar{x}_i$ – среднее значение переменной, σ_i – дисперсия переменной.

В этом случае нормированные величины не принадлежат гарантированно единичному интервалу, более того, максимальный разброс значений $\tilde{x}_i$ заранее не известен. Если есть необходимость, это тоже можно исправить, используя функциональную предобработку. Например, нелинейное преобразование, представленное формулами (1.5) и (1.6), нормирует данные, одновременно гарантируя что $\tilde{x}_i \in [0,1]$.

$$\tilde{x}_i = f\left(\frac{x_i - \bar{x}_i}{\sigma_i}\right) \tag{1.5}$$

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{1.6}$$

где x_i – текущее значение переменной; $\tilde{x}_i$ – преобразованное значение переменной, $\bar{x}_i$ – среднее значение переменной, σ_i – дисперсия переменной.

Общий случай функциональной предобработки показан формулой (1.7).

$$\tilde{x}_i = f(x_i) \tag{1.7}$$

Функциональная предобработка на примере логистической функции была рассмотрена выше (формулы (1.5) и (1.6)). В функциональной предобработке может использоваться любая функция.

Кодирование бинарных признаков

Бинарные признаки характеризуются наличием только двух состояний – истина и ложь. Но даже такие простые данные можно интерпретировать с различным смыслом, например, для признака «пол ребенка» - можно поставить соответствие значению «мальчик» - истина (для двоичного кодирования 1), а можно и ложь (для двоичного кодирования 0). Если задать диапазон [a,b], возможные способы кодирования бинарного признака приведены в таблице 1.1.

Таблица 1.1.

Кодирование бинарного признака

Входной сигнал	Значение входного сигнала	
	Истина	Ложь
x	a	0
x	b	0
x	b	a
x	a	b

Кодирование неупорядоченных качественных признаков

Поскольку неупорядоченные признаки не связаны друг с другом отношением порядка, то не следует кодировать их разными величинами одного входного сигнала. Для кодирования таких признаков используют столько входных сигналов, сколько состояний этот качественный признак может принимать. Каждый отдельный входной сигнал соответствует одному определенному состоянию признака.

Наиболее часто используется на практике двоичное кодирование типа «$n \rightarrow n$» когда имена n возможных состояний кодируются значениями n бинарных входов, причем первая категория кодируется как (1,0,0,…,0), вторая, соответственно – (0,1,0,…,0) и т.д. вплоть до n-ной: (0,0,0,…,1).

Если входной диапазон брать [a,b], кодировка может быть выполнена следующим образом: первая категория - (b,a,a,...,a), вторая - (a,b,a,...,a), и т.д. n-ная - (a,a,...,a,b).

Такое кодирование будет неоптимальным в случае, если классы представлены существенно различающимся числом примеров. В этом случае, функция распределения значений переменной будет очень неоднородной, что может снизить информативность этой переменной. Тогда имеет смысл использовать более компактный код «$n \rightarrow m$», когда имена n классов кодируются m-битным двоичным кодом. Причем, в такой кодировке активность входов должна быть равномерна.

В качестве примера рассмотрим фактор «кормление ребенка». Значения данного фактора могут быть такими: «кормили грудью», «кормили смесью», «смешанное кормление». При этом значение кормили грудью представлено 192 значениями примеров, что гораздо больше чем остальные. Простое кодирование «$n \rightarrow n$» привело бы к тому, что первый вход активировался бы гораздо чаще остальных. Это можно избежать, закодировав три класса двумя бинарными входами следующим образом: «кормили грудью» – (0,0), «кормили смесью» – (0,1), «смешанное кормление» – (1,1), что обеспечивает более равномерную «загрузку» входов.

Кодирование упорядоченных качественных признаков

Один из вариантов кодирования таких переменных это поставить в соответствие номерам категорий числовые значения так, чтобы они сохраняли существующую упорядоченность. Естественно, при этом имеется большая свобода выбора - любая монотонная функция от номера класса представляет свой способ кодирования. Кодирование переменных

числовыми значениями должно приводить, по возможности, к равномерному заполнению единичного интервала закодированными примерами, включая при этом и этап нормировки. При таком способе все примеры будут нести примерно одинаковую информационную нагрузку. Таким образом, можно переменные кодировать следующим образом: единичный отрезок разбивается на n отрезков (n равно числу классов), с длинами пропорциональными числу примеров каждого класса в обучающей выборке. Например, переменная, которая содержит информацию о курении во время данной беременности может принимать значения: «мало», «средне», «много». Пусть x_1 соответствует значению «мало», x_2 – «средне» и x_3 – «много». Центр каждого такого отрезка будет являться численным значением для соответствующего класса (рисунок 1.1).

$$\Delta x_n = \frac{P_n}{P} \tag{1.8}$$

где P_n - число примеров данного класса n, а P - общее число примеров.

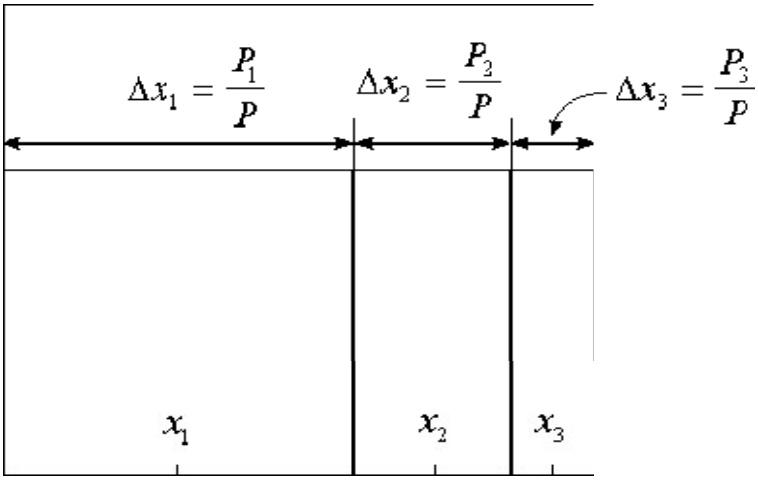

Рисунок 1.1. Иллюстрация способа кодирования одной переменной упорядоченного типа.

Кодирование таким способом, т.е. разными значениями одного входного сигнала не всегда эффективно из-за того, что расстояние между состояниями не определено, а данный способ кодирования эти расстояния задает явно. Поэтому, упорядоченные признаки можно, как и

неупорядоченные, кодировать в виде стольких входных сигналов, сколько состояний у признака. Но, в отличие от неупорядоченных признаков, лучше накапливать число сигналов с максимальным значением. Если все состояния обозначены как $x_1 < x_2 < ... < x_n$, рекомендуемая таблица кодировки приведена в таблице 1.2.

Кодирование совместимых признаков

В качестве примера рассмотрим фактор «TORCH инфекции», значения которого могут быть такими: «хламидии», «микоплазма», «бактериальный вагиноз», «герпес», «краснуха», «токсоплазмоз». Очевидно, что наличие одного из них не гарантирует отсутствия других заболеваний. Поэтому можно применить кодирование аналогичное простому кодированию «$n \rightarrow n$». Например так: «хламидии» – (1,0,0,0,0,0), «микоплазма» - (0,1,0,0,0,0), «бактериальный вагиноз» (0,0,1,0,0,0), «герпес» - (0,0,0,1,0,0), «краснуха» - (0,0,0,0,1,0), «токсоплазмоз» - (0,0,0,0,0,1).

Таблица 1.2.

Кодирование упорядоченного качественного признака

Состояние	Вектор входных сигналов
x_1	(b,a,a,...,a)
x_2	(b,b,a,...,a)
x_n	(b,b,b,...,b)

1.4. Анализ методов понижения размерности обучающей выборки

Врач, при постановке диагноза, старается анализировать весь комплекс сведений о пациенте. При этом одна часть параметров имеет принципиальное значение для принятия решения, другая не столь важна и может игнорироваться. Достаточно часто для диагностики и прогнозирования больному проводят сложные и дорогостоящие методы обследования, порой

небезвредные для здоровья, когда во многих случаях представляется возможным получить ответ и без них. В данной задаче, как и во многих других, нет никакой дополнительной информации о том, какие входные переменные действительно нужны для решения поставленной задачи, а именно, определения степени риска СВСГР. Анализ всех факторов риска вызывает существенные затруднения при построении и обучении экспертной системы, поэтому необходимо решить задачу сокращения размерности входных данных.

Корреляционный анализ

Корреляционный анализ [15,19]– один из широко распространенных методов оценки статистических связей. Он оценивает влияние определенной входной величины на выходную и определяет степень (тесноту) связи между величинами.

При решении задач предсказания необходимо построить статистическую функцию вида:

$$Y = f(x_1,x_2,...,x_n), \qquad (1.9)$$

где Y - исследуемая величина, зависящая от факторов $X = \{x_1, x_2, ..., x_n\}$. Из выше сказанного предположение о такой зависимости для некоторых факторов может оказаться неверным. Одним из методов, по которому можно определить, какие из факторов $X = \{x_1, x_2, ..., x_n\}$ влияют существенно, и оценить количественную меру этого влияния является корреляционный анализ.

Задачей корреляционного анализа является получение корреляционной матрицы по случайной выборке. На ее основе вычисляются частные и множественные коэффициенты корреляции и детерминации.

Парный коэффициент корреляции показывает степень линейной зависимости между двумя переменными при влиянии всех остальных показателей. Для универсальности выражений включим переменную y в состав факторов X, тогда информационной базой для анализа будет являться матрица размерности (m x n)

$$X = \begin{Vmatrix} x_{11} & x_{1j} & x_{1n} \\ x_{i1} & x_{ij} & x_{in} \\ xm1 & x_{mj} & x_{mn} \end{Vmatrix} \qquad (1.10)$$

где i-я строка характеризует i-е наблюдение по всем n показателям $j \in [1, n]$.

Парные коэффициенты корреляции определяются выражением (1.11):

$$r_{jk} = \frac{K_{jk}}{\sigma_j \cdot \sigma_k}, \quad (1.11)$$

где K_{ik} - корреляционный момент, r_{jk} - выборочный парный коэффициент корреляции, который характеризует степень линейной связи между показателями x_j и x_k и изменяется в пределах от -1 до +1. Чем ближе коэффициент корреляции к значениям ±1, тем сильнее зависимость между переменными. Если $r_{ij} > 0$, связь положительная, если $r_{ij} < 0$, то - отрицательная.

Корреляционный момент определяется формулой (1.12)

$$K_{jk} = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k) \quad (1.12)$$

где X_{ij} - значение i-го наблюдения j-го фактора; $\overline{X}$ - средние значения факторов (формула 1.13); σ - среднеквадратичные отклонения (формула (1.14))

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad (1.13)$$

$$\sigma_j = \sqrt{\frac{1}{n} * \sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2} \quad (1.14)$$

Из парных коэффициентов корреляции может быть составлена корреляционная матрица (1.15).

$$R = \begin{Vmatrix} 1 & r_{12} & \dots & r_{1m} \\ r_{21} & 1 & \dots & r_{1m} \\ r_{31} & r_{32} & \dots & r_{3m} \\ \dots & \dots & \dots & \dots \\ r_{m1} & \dots & \dots & 1 \end{Vmatrix} \quad (1.15)$$

Матрица R является симметричной ($r_{ij} = r_{ji}$) и положительно определенной.

Частный коэффициент корреляции характеризует степень линейной зависимости между двумя переменными при исключении влияния остальных переменных, которые входят в модель. Например, точечная оценка частного коэффициента корреляции *m-2* - го порядка между факторами x_1 и x_2 определяется выражением (1.16)

$$r_{12/3,4,\ldots,m} = \frac{R_{12}}{\sqrt{R_{11}R_{22}}} \quad (1.16)$$

где R_{jk} - алгебраическое дополнение элемента r_{jk} корреляционной матрицы R определяется формулой (1.17).

$$R_{jk} = (-1) \cdot j + k \cdot M_{jk} \quad (1.17)$$

где M_{jk} - минор, являющийся определителем матрицы, полученной из *R* путем вычеркивания *j*-ой строки и *k*-го столбца.

Множественный коэффициент корреляции показывает степень связи между одной (результативной) переменой и остальными, входящими в модель. Множественный коэффициент корреляции изменяется в пределах от 0 до 1, имеет порядок *m-1* и, например, для результативного признака x_1 определяется по формуле (1.18)

$$r_{1/2,3,\ldots,m} = r_1 = \sqrt{1 - \frac{|R|}{R_{11}}} \quad (1.18)$$

где $|R|$ - определитель матрицы *R*.

Таким образом, если оценивается влияние на выходную величину одной входной, то определяется коэффициент парной корреляции. При оценке одновременного влияния нескольких входных величин на выходную находится коэффициент множественной корреляции. Если коэффициент корреляции между переменой, входящей в модель, и результатом $> 0{,}7 \div 0{,}8$, то делается вывод, что фактор значим и оценивается его значимость.

Значимость коэффициента корреляции проверяется с применением критерия Стьюдента.

Методы искусственного интеллекта вычисления информативных признаков

Существует несколько методов, позволяющих выявить относительную значимость входных параметров, с помощью нейронных сетей [20,26,27]. Для выделения информативных входных переменных с помощью нейронных сетей можно использовать так называемый метод проб и ошибок, придерживаясь одной из двух стратегий: наращивания или отсечения.

При наращивании мы начинаем с одной переменной и по одной добавляем другие переменные. Если результат от этого улучшается, то комбинация запоминается. Такой подход имеет существенный недостаток – игнорируется то обстоятельство, что две или более переменных могут быть взаимосвязаны (т.е. необходимо сразу добавить несколько переменных).

При отсечении мы берем все имеющиеся переменные и начинаем их по одной убирать. Если результат от этого ухудшается, то возвращаемся к предыдущей комбинации переменных. Такой подход, как правило позволяет сохранить взаимосвязанные переменные, но он достаточно медленный и не рациональный.

Подобную стратегию можно реализовать с помощью генетических алгоритмов [5,6,42], они являются очень эффективным инструментом поиска в комбинаторных задачах. Схема работы генетического алгоритма такова: каждый возможный вариант набора входных переменных можно представить в виде строки бит. 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 (ошибка обучения НС). По значениям ошибки производится отбор лучших

вариантов наборов, которые комбинируются друг с другом с помощью искусственных генетических операторов: скрещивания и мутации.

1.5. Анализ существующих экспертных систем

Система ДИАГЕН - система для диагностики наследственных болезней База знаний системы ДИАГЕН включала 1200 синдромов моногенной и хромосомной природы, проявляющихся клинически в детском возрасте и описываемых более, чем 1500 признаками.

В системе ДИАГЕН механизм учета в процессе диагностики одной из характеристик интуиции – ассоциативных знаний о симптомах заболевания, был реализован следующим образом. В базе знаний над пространством симптомов введены отношения, определяющие их основные свойства: частотные, причинные, фенотипические (по внешним проявлениям) и другие. Строилась функция близости, описывающая, по существу, отношение принадлежности симптома некоторому множеству признаков. Исходной информацией для вычисления функции близости являлось описание врачом-экспертом корреляционных, структурных и других свойств симптомов. На этой основе были построены графы: структурный – взвешенный древовидный граф (включает морфо-физиологические отношения), причинно-следственный (отражает патогенетические механизмы заболеваний) и фенотипического сходства – взвешенный симметрический граф (рассмотрение близких в содержательном смысле понятий, например, дисплазия и гипоплазия почек). В процессе формирования базы знаний, путем дополнения формировался единый "результирующий" граф, который также являлся взвешенным и двунаправленным. И, наконец, путем замыкания результирующего графа получали граф парных связей со своими соответствиями между каждой парой признаков. В результате замыкания каждая опосредованная связь заменялась на непосредственную. Вес новой продуцированной дуги рассчитывался по способу, близкому к схеме Шортлифа. Таким образом, получали матрицу отношений над пространством признаков. Введение этих отношений как раз и позволило проводить уточнение и некоторое расширение входной последовательности признаков

(введенного пользователем симптомокомплекса) до диагностической последовательности, включающей ассоциированные признаки, которая уже поступала на механизм логического вывода. Вычисление новых «весов» признаков осуществлялось с помощью функции "близости". При первоначальной оценке диагностической ценности симптомов экспертами учитывалась частота признака в популяции и возможность экзогенной (внешнесредовой) его обусловленности, например, плоскостопие было оценено в 5 баллов, в то время как врожденный признак фокомелия – в 95 баллов. Кроме того, принималась во внимание сложность обнаружения симптомов при осмотре (к примеру, белая прядь волос и ускоренная оссификация позвонков – в первом случае достаточно визуального осмотра, во втором – необходимо рентгенографическое обследование). Полученная экспертным путем оценка значимости признака для распознавания заболевания предлагается при описании клинической картины больного врачу-пользователю, который может изменить "вес" (в интервале от 0 до 99) любого из отмеченных у ребенка симптомов в соответствии с его личной гипотезой о их диагностической ценности в конкретном случае, что позволяет использовать в системе опыт и интуицию лечащего врача, в какой-то степени осуществлять индивидуальную "настройку" компьютерной системы на пользователя, который таким образом принимает участие в диагностическом процессе. Такая операция по присваиванию "весов" признакам может повторяться многократно, что дает возможность проверять различные предположения о роли тех или иных признаков, вплоть до их исключения (обнуления) при данном заболевании у конкретного пациента. Указанная врачом степень значимости каждого введенного признака соответствует интегральной оценке его предполагаемой диагностической ценности. Для случая, когда пользователь минимизирует «вес» признака, ближайшим современным аналогом можно считать созданную в конце 90-х годов ЭС ЭСТЕР для диагностики лекарственных отравлений, где были использованы таблицы запрещенных значений «диагностический признак –

класс решений», позволившие повысить эффективность процесса диагностики.

Реализованный в системе ДИАГЕН механизм привлечения дополнительной информации в виде признаков по ассоциации (по сходству и др.) обеспечивал, с одной стороны, учет условно нечетких представлений врача в отношении особенностей наблюдаемых признаков. Наряду с этим, в механизме вывода системы ДИАГЕН степень значимости признака, играющая роль априорной вероятности для байесовского метода, связывает апостериорную вероятность гипотезы с ее априорной вероятностью, а такой способ интерпретации исходных данных позволяет собирать фрагментарную и возможно неточную информацию, для того чтобы сделать более полную оценку клинической картины. Кроме того, интерфейс пользователя, реализованный в виде дерева признаков, позволяет осуществлять их отбор на разных уровнях в зависимости от точности представлений (от уровня уточнения клинических проявлений, наблюдаемых у больного), например, так называемые фенотипические характеристики типа формы носа и т.п. Если проанализировать дальше названный пример, то в первом варианте системы (МГЕ), был перечень изменений формы носа, но затем в ЭС ДИАГЕН это уже была просто возможность указать на факт необычной формы носа, а уже интеллектуальная система рассматривала перечень решений (диагнозов), включающих различные отклонения формы носа. Такая возможность функционирования системы в условиях неопределенных и неточных исходных данных обеспечивала эффективную диагностику у больных со стертой клинической картиной или ранними проявлениями болезней, что особенно важно для прогредиентно протекающих наследственных болезней.

Возвращаясь к ЭС ДИАГЕН, нужно отметить, что основной набор правил, описывающих свойства и взаимосвязь симптомов, содержит также специальный коэффициент, изменение которого позволяет усилить или ослабить значимость соответствующего правила. В процессе опытной

эксплуатации варьированием этих коэффициентов подбирались оптимальные стратегии для различных режимов работы, например, режим поиска синдрома с неполным описанием, что крайне важно ввиду достаточно большого процента случаев с неклассической клинической картиной заболевания (ранние проявления, стертая форма заболевания и т.п.). Это особенно важно ввиду того, что система ДИАГЕН ориентирована на поддержку врачебного решения на долабораторном этапе диагностики, где выделение обычно двух – трех диагнозов, один из которых (чаще первый) в дальнейшем подтверждался (более, чем в 90% случаев) после проведения специальных исследований, было крайне важно ввиду сложности и высокой стоимости специфических лабораторных исследований.

Еще один важный аспект, на который также следует обратить внимание, был назван нами коммуникабельностью системы. Под этим подразумевалось:

во-первых, уже упоминавшаяся в другом контексте возможность для врача-пользователя осуществлять коррекцию диагностического "веса" любого признака, что позволяет проверять предположения о диагностическом значении отдельных клинических проявлений болезни, что особенно важно при идентификации нетипичных случаев;

во-вторых, режим дополнения/изменения отобранных параметров практически на любом этапе диагностической процедуры или даже после архивации данных;

в-третьих, в зависимости не только от факта наличия или отсутствия, но и от "веса" признака, система выдает "премии" и "штрафы", отражающиеся на формировании диагностической последовательности, информацию о чем врач может получить информацию, просмотрев протокол объяснений выбора диагноза системой и, таким образом, скорректировать свои представления о диагностической роли симптомов;

в-четвертых, врач может сужать или расширять выдаваемый системой дифференциально-диагностический ряд, осуществляя "настройку" системы путем уменьшения или увеличения порога для включения нозологических форм в диагностическую последовательность. В тот же временной период в ЭС SPHINX [19Lesmo L. et al. 1984] был реализован механизм влияния пользователя на процессы управления системой путем модификации промежуточных результатов. А в недавно созданной в МЦНИТ МНИИПиДХ системе ДИАНЕС [17Подольная, Таперова 2002а] введено понятие параметра настройки P ($0 \le P \le 1$), позволяющего ограничить список всех возможных диагнозов. В соответствие со значением P в список рабочих гипотез попадают

только те диагнозы, у которых часть характеризующих их облигатных и часто встречающихся симптомов отмечена пользователем. Причем отношение числа отмеченных симптомов ко всем симптомам, характеризующим диагноз, должно быть не меньше Р. При Р = 0 в рассмотрение принимается весь список диагнозов, при Р = 1 – только те, все симптомы которых (кроме исключающих), должны быть отмечены у пациента).

ДИН. Система для диагностики неотложных состояний у детей ДИН создавалась с учетом необходимости принятия решений по неполному списку диагностических критериев, т.е. при стертой клинической картине, не полностью развившемся синдроме и при ограничениях на проведение специальных исследований, обусловленных тяжестью состояния или недостатком аппаратуры. (Здесь необходимо сразу отметить, что выбор оптимального плана обследования больного с учетом критерия альтернативы, включающего риск предполагаемого исследования, обусловленный тяжестью состояния, квалификацией врача, характеристиками медицинской аппаратуры и другими параметрами был реализован несколько ранее в системе MEDAS.

Медицинская постановка задачи в ЭС ДИН сводилась к распознаванию текущего состояния ребенка в терминах как одного, так и нескольких синдромов (в данном случае синдром можно условно считать аналогом заболевания), что крайне важно при критических состояниях. Знания о синдромах охватывали информацию: а) о дополнительных синдромах, состоящих в некоторых отношениях с рассматриваемым (обеспечивая возможность учета фоновых и сопутствующих заболеваний), б) о взаимоисключающих состояниях, в) о дифференцируемых синдромах (синдромы-конкуренты). Реализованные в системе ассоциативные связи (явные, в отличие от скрытых ассоциаций, возникающих в процессе мышления на уровне интуиции, о чем шла речь выше при рассмотрении системы ДИАГЕН) позволяли, таким образом, учитывать: во-первых, на фоне каких состояний может развиться данный синдром, во-вторых, фоном для каких синдромов он может служить, в-третьих, с какими синдромами он может быть совместим, т.е. какие синдромы могут встречаться у пациента одновременно. Такой подход более соответствует многим медицинским ситуациям, так как учитывает наличие сопутствующих заболеваний, клиническая картина которых может пересекаться с признаковым пространством исследуемой ситуации или отягощать состояние больного, что приводит к мнению об ошибочности оценки, сформулированной на основе диагноза экспертной системы.

Гетерогенность заболеваний, определяемая полиморфизмом клинических проявлений (атипичные формы, возрастная динамика, смена состояний в процессе болезни) нашла в

ЭС ДИН отражение в форме «масок» – логических выражений, состоящих из теоретически возможных клинических вариантов болезни. Такой подход несомненно продуктивен для медицинских систем, тем более, что классические формы заболеваний встречаются все реже.

И, несомненно, актуальный и на сегодня вопрос «направления» диагностики – от признаков к диагнозу или от предполагаемого врачом диагноза к подтверждающим признакам. В последнем случае резко ускоряется получение диагностического решения, что особенно важно при неотложных состояниях, а во-вторых, лечащий врач может сразу четко определить свою позицию (не занимаясь последовательным вводом признаков) и получить обоснованное или подтверждение, или опровержение своей гипотезы.

ЭС ВЕСТ-СИНДРОМ. На других принципах была построена экспертная система для диагностики судорожных состояний (эпилепсии). Решение вопроса об этиологии (причине) заболевания, проявляющегося инфантильными спазмами, вызывает в практике врача большие сложности, требует значительного клинического опыта, дополнительных разнообразных методов инструментальной и/или лабораторной диагностики. Учитывая ограниченность информации по редким формам эпилепсии, дебютирующим в возрасте 3 – 7 месяцев жизни, при разработке ЭС ВЕСТ-СИНДРОМ [22Кобринский и др. 1997] была выбрана технология виртуальных статистик (ТВС) [23Марьянчик, 1996]. Виртуальные статистики, в отличие от статистик, получаемых из опыта, формировались с использованием теоремы Байеса и учитывали диагностические оценки экспертов по обобщенным проявлениям болезни.

ТВС позволяет: а) выявлять неявные противоречия в заключениях экспертов по обобщенным примерам заболевания; б) обеспечивать независимость качества извлекаемых знаний от квалификации “посредников-когнитологов”; в) использовать извлеченные знания в форме виртуальных статистик для автоматизации формирования вопросов-рекомендаций по проведению дополнительных исследований.

Инструментами технологии виртуальных статистик являлись:

1) программа, обеспечивающая выявление неявных противоречий в заключениях экспертов, состоящих в придании ими разного веса одним и тем же признакам для одного и того же заболевания, и преобразование диагностических заключений в виртуальные статистики;
2) алгоритм, использующий виртуальные статистики для формирования рекомендаций по проведению дополнительных исследований;
3) оболочка “Алеф” [23Марьянчик, 1996] для разработки экспертных систем, содержащая указанный алгоритм.

При разработке диагностической системы ВЕСТ-СИНДРОМ виртуальные статистики использовались для:

- вычисления текущей вероятности и ранжирования диагнозов в дифференциальном ряду;
- вычисления вероятности диагнозов, которые могут иметь место при последующем исследовании (дообследовании);
- исключения из дифференциального ряда диагнозов, которые ни при каких обстоятельствах не превысят вероятности уже достигнутой другими диагнозами;
- формирования рекомендаций о проведении очередного диагностически наиболее важного исследования.

Экспертные знания формулировались в виде вероятностных заключений. При этом экспертам предоставлялись следующие возможности. Во-первых, допускались нечеткие словесные заключения типа “вероятность этого диагноза незначительна”. Во-вторых, после определения всего комплекса основных симптомов, соответствующих нозологической форме, экспертам представлялся для заключений полный набор возможных их сочетаний, в которых симптомы принимали значение только “да/нет”. Затем экспертные оценки проверялись на соответствие теореме Байеса и экспериментальным данным: порождаемые заключениями виртуальные встречаемости признаков должны были быть независимы от формы проявления одной и той же болезни, представленной в разных портретах, и принадлежать диапазону

экспериментально полученных (из клинического опыта) или известных из литературы встречаемостей признаков, если такой диапазон известен. (Под портретами понимались формальные описания нозологических форм болезней по неполным данным, включавшим основные дифференциально-диагностические симптомы – от 2 до 6). Незначительные изменения в экспертных оценках могли существенно изменить виртуальные значения встречаемостей. Поэтому была предусмотрена проверка заключений, являющаяся принципиально важным моментом технологии виртуальных статистик. ТВС позволяет проверить согласованность экспертных оценок по разным клиническим портретам одного и того же диагноза, характеризующим интервалы неопределенности в экспертных оценках. Затем, руководствуясь, с одной стороны, медицинскими соображениями, а с другой – величиной отклонений, эксперты производили, при необходимости, корректировку своих первоначальных оценок. На рис. 2 и 3 можно видеть совпадение заданных и виртуальных вероятностей по результатам первого и второго этапов экспертной оценки портретов заболеваний.

1.6. Анализ методов извлечения знаний

В настоящее время существует большое количество методов извлечения закономерностей. Наиболее широко для обработки результатов эксперимента используются методы математической статистики и регрессионного анализа. Но такие методы могут быть непонятны неподготовленному пользователю. Лица, принимающие решения в данных областях, как правило, не имеют специальной математической подготовки. Этим обусловлена необходимость представления результатов обработки экспериментальной информации в форме, доступной для специалистов в области медицины. Рассмотрим методы построения экспертных систем, многие из них способны предоставить результат в понятном для человека виде.

В последние годы успешно внедряются приложения, реализованные на основе машинного обучения (machine learning) [83]. Область машинного обучения рассматривает вопросы конструирования компьютерных программ, которые автоматически улучшаются с опытом. Все идет к тому, что машинное обучение будет играть все более и более центральную роль в информатике и компьютерных технологиях. Машинное обучение основано на понятиях и происходит от многих областей таких дисциплин как вероятность и статистика, искусственный интеллект, философия, теория информации, нейробиология, вычислительная теория сложности, элементы теории управления и других.

Общий подход машинного обучения состоит в разработке и использовании программ, способных обучатся под руководством эксперта-учителя. Так, учитель предъявляет программе примеры реализации некоторого концепта, а задача программы заключается в том, чтобы извлечь из этих примеров набор атрибутов и значений, которые определяют этот концепт.

Применение машинного обучения к построению экспертных систем используются для:

- извлечения множества правил из предъявляемых примеров;
- анализа важности отдельных правил;
- оптимизации производительности набора правил.

Методы искусственного интеллекта на базе машинного обучения используются как подход к проблеме поиска решения, на основе использования предшествующих знаний обучающих данных. Рассмотрим некоторые из них.

Извлечение знаний с помощью нейронных сетей (НС) [25,32,35,56]. Как правило, нейросети используются как инструмент предсказания, а не понимания. Классический нейросетевой подход – метод черного ящика – предполагает создание модели, без явной формулировки правил принятия решений.

Метод нейросетевого прогнозирования реализуется нейронной сетью, способной осуществлять манипуляции в пространстве признаков большой размерности. Решения задачи с помощью НС можно описать следующим алгоритмом:

1. Формализация обучающей выборки по данным предварительных исследований. Отбор пациентов, информация об исследованиях над которыми включается в обучающую выборку. Выбор важных факторов, характеризующих (влияющих на) данное заболевание.
2. Формирование архитектуры НС. Выбирается тип НС, функции активации и алгоритм обучения.
3. Начальная активация нейронов. Задание начальных весовых коэффициентов.
4. Итерационное обучение НС выбранным алгоритмом.
5. Получение прогноза.

Эволюционные вычисления [66,73,77,81-83] развивались из трех относительно независимых разработок: генетических алгоритмов, эволюционного программирования и эволюционных стратегий. Все три дисциплины начали изучаться в 60-х годах и 70-х годах (первые две в США, третья в Германии), но только в конце 80-х годов они начали признаваться. В настоящее время генетические алгоритмы рассматриваются как один из наиболее успешных методов машинного обучения. По существу, они являются процедурами оптимизации, имитирующими при проектировании модели такие процессы, как генетическая рекомбинация, мутация и отбор, аналогичные тем, что обусловливают естественную эволюцию. Первые генетические алгоритмы были предложены в начале 70-х годов Джоном Холландом (John Holland) с целью имитации эволюционных процессов в живой природе [73].

Реализовать мощные возможности при построении экспертных систем, основанных на правилах, можно с помощью генетического программирования. Генетическое программирование (ГП) является одной из парадигм эволюционных вычислений, которые основаны на формализации принципов естественной эволюции. В настоящее время ГП успешно применяется при решении различных задач. В качестве особи – потенциального решения, в ГП используется компьютерная программа, которая решает определенную задачу. В отличие от других эволюционных методов в ГП особи имеют изменяющиеся размер и форму. Особи строятся на основе множества проблемно-ориентированных элементарных функций (functional set) и констант (terminal set), выбор которых существенно влияет на размерность пространства поиска и качество получаемого решения. В процессе эволюции производится автоматический синтез, в определенном смысле, лучшей программы.

Решение задачи на основе ГП можно представить следующей последовательностью действий:

1. Создание исходной популяции.

2. Выбор родителей для процесса размножения (работает оператор отбора - репродукции).
3. Создание потомков выбранных пар родителей (работает оператор скрещивания - кроссинговер).
4. Мутация новых особей (работает оператор мутации).
5. Расширение популяции новыми порожденными особями.
6. Сокращение расширенной популяции до исходного размера (работает оператор редукции).
7. Если критерий останова алгоритма выполнен, то выбор лучшей особи в конечной популяции – результат работы алгоритма. Иначе переход на шаг 2.

Таким образом, для решения задачи на основе ГП необходимо определить множества элементарных функций и констант, форму представления потенциального решения - особи, операции скрещивания и мутации, задать оценочную функцию. Очевидно, что эффективность решения задачи зависит от целого ряда параметров: размера популяции, стратегии выбора особей из предыдущей популяции, скрещивания и мутации, стратегии сокращения популяции, а также вида оценочной (фитнесс) функции. Применяются различные варианты данных параметров в зависимости от задачи.

1.7. Выводы и постановка задач исследований

По разделу можно сделать следующие выводы:

1. Выполнен анализ применения современных компьютерных технологий в медицине. Показано, что задачи диагностики и прогнозирования это комплексный процесс и рассмотрены его этапы.
2. Проведен анализ методов предварительной обработки входных данных. Учитывая, что факторы риска представляют собой

различные типы данных, включая и нечисловые характеристики, рассмотрена классификация компонентов входной информации и способы их кодирования. Предложены методы кодирования нечисловых характеристик, учитывая их тип. Для числовой информации рассмотрены способы нормировки.

3. Проведен анализ современных методов построения экспертных систем. Показано, что большинство из них имеют существенные недостатки. Поэтому следует обосновать новый метод построения экспертной системы с использованием генетического программирования.
4. Поставлена цель диссертационной работы - разработка экспертной системы определения степени риска синдрома внезапной смерти грудного ребенка, обеспечивающей прогнозирование СВСГР еще на этапе беременности и сразу после рождения ребенка.

Для разработки экспертной системы определения степени риска синдрома внезапной смерти грудного ребенка необходимо решить задачи:

- разработка структуры экспертной системы прогнозирования синдрома внезапной смерти грудного ребенка;
- реализация методов предобработки для подготовки обучающих данных;
- реализация и разработка методов выбора информативных данных;
- разработка методов интеллектуального анализа данных;

Для решения перечисленных задач выбраны следующие направления исследований:

- исследование методов предобработки данных с целью формирования входных данных для обучения экспертной системы;

- исследование методов выбора информативных данных с целью понижения размерности обучающих данных;
- исследование методов построения экспертных систем;
- исследование методов интеллектуального анализа данных с целью извлечения знаний.