РАЗДЕЛ 2

ПОСТАНОВКА ЗАДАЧИ И ПОДГОТОВКА ВХОДНОЙ ИНФОРМАЦИИ ДЛЯ ЭКСПЕРТНОЙ СИСТЕМЫ ОПРЕДЕЛЕНИЯ СТЕПЕНИ РИСКА СВСГР

2.1. Аспекты проектирования медицинских экспертных систем

В подразделе 1.1 было рассмотрено, что медицинские задачи являются сложными и плохо структурированными. Для них не существует единого алгоритма, т.е. точного предписания о выполнении действий в определенном порядке. Естественной целью врача при решении таких задач является нахождение алгоритма для конкретной ситуации. Подобного рода деятельность требует участия интеллекта человека. Задачи, связанные с отысканием алгоритмов решения для определенных типов задач, называются интеллектуальными [], а системы, которые способны самостоятельно (без участия человека) решать интеллектуальные задачи называются интеллектуальными системами (ИС) []. Основным свойством ИС является то, что для решения задач они используют не только данные, но еще и знания, поэтому такие системы иначе называют системы основанные на знаниях.

Наиболее распространенными ИС являются экспертные системы (ЭС). Экспертная система (ЭС) [9,22] — это компьютерная программа, которая содержит знания в определенной предметной области, накопленные в результате практической деятельности эксперта, и использует их для решения задач в некоторой узкой предметной области.

Любая ЭС представляет собой симбиоз человека и компьютера, т.к. без интерактивного взаимодействия человека и системы вряд ли можно решить действительно трудные задачи. Правильное распределение функций между человеком и машиной является одним из ключевых условий эффективного внедрения компьютерных программ, а ЭС, прежде всего, является программным продуктом, назначение которого – автоматизация

деятельности человека. Принципиальным отличием ЭС от других программ является то, что она выступает не только в роли «ассистента», выполняющего за врача часть работы, а и в роли «компетентного партнера» – эксперта – консультанта в данной предметной области. ЭС не предназначена заменить эксперта – врача в его непосредственной деятельности, она расширяет возможную сферу применения его знаний, а также усиливает его умственные способности.

Структуру простейшей ЭС образуют следующие компоненты:

- база знаний (БЗ) и база данных (БД);
- машина вывода;
- интерфейс пользователя;
- компонент приобретения знаний (редактор БЗ);
- модуль обучения (извлечения знаний).

Основными (базовыми) компонентами являются первые три.

БЗ и БД содержат информацию о предметной области. Машина вывода – это механизм (программа) извлечения (вывода) ответа на поставленный вопрос. Стратегия, реализуемая машиной вывода, привязана к конкретной реализации ЭС и модели представления знаний. Интерфейс пользователя обеспечивает взаимодействие человека и системы на различных этапах решения задачи. Компонент приобретения знаний (редактор БЗ) предназначен для добавления новых или редактирования существующих знаний. Модуль обучения (извлечения знаний) позволяет корректировать знания на основе механизма обучения. ЭС в классическом варианте просто копируют знания эксперта-человека, но человеку, присуща способность к обучению, такая способность является существенным фактором, который определяет параметры и возможности любой ИС.

В современных медицинских ЭС можно выделить два основных направления компьютерного использования медицинских знаний. Первый – это классический вариант ЭС, связанный с попыткой заложить в программу формализованные представления правил постановки диагноза. Такие ЭС

разрабатываются при участии ведущих специалистов в соответствующих областях медицины. Второй подход основан на использовании обучающихся программ. Суть обучения заключается в обработке и анализе массивов историй болезней с установленным диагнозом и разработке на основе этого алгоритма, который позволяет поставить диагноз в каждом конкретном случае, даже если он не был рассмотрен при обучения.

Способность ЭС решать поставленную перед ней задачу не ослабеет со временем и не забудется при отсутствии практики. При многократном решении одной и той же задачи ЭС выдаст одно и тоже решение в отличие от врача, который подвержен эмоциональным факторам, этим ЭС и привлекательна в качестве партнера-консультанта.

2.2. Разработка структуры и описание функций ЭС определения степени риска СВСГР

Процесс создания ЭС называется инженерией знаний (knowledge engineering), так как извлечение знаний является одним из наиболее важных этапов разработки любых интеллектуальных систем [30,31,60,61]. Под приобретением знаний, как правило, понимают автоматизированный процесс прямого общения с экспертами. Существуют разработанные программы, которые целенаправленно опрашивают эксперта с целью получения знаний и заносят их в уже существующие структуры базы знаний. В подразделе 1.1 говорилось, что умения анализировать, классифицировать и выделять закономерности ограничены объемом анализируемых данных. А в подразделе 1.5 рассмотрены различные методы машинного обучения, которые позволяют эффективно решать практические задачи, обрабатывая громадные потоки информации. В последнее время используется объединение таких методов с технологией ЭС, что позволяет создавать высокоэффективные гибридные интеллектуальные системы.

В проектируемой ЭС определения степени риска СВСГР будут использованы методы машинного обучения, с целью извлечь знания из

достаточно большого набора историй течения беременности и обеспечить возможность повторного обучения (извлечения знаний) по желанию врача при накоплении других новых данных.

ЭС определения степени риска СВСГР разрабатывается, например, для врачей гинекологов, которые наблюдают течение беременности, или акушеров-гинекологов, а так же для педиатров.

Перечень задач, которые должна решать проектируемая экспертная система, включает:

- ведение учета пациентов, включая каждую беременность отдельно;
- определение степени риска СВСГР на различных этапах беременности и сразу после рождения ребенка;

Конечное решение о предполагаемой степени риска СВСГР и о планируемых мероприятиях направленных на ее снижение принимает врач гинеколог.

Для определения степени риска СВСГР экспертная система должна выполнять следующие функции:

- ввод общей информации о пациенте, течении текущей беременности и исходах предыдущих;
- обработка информации: представление ее в виде пригодном для дальнейших действий (обучение или эксплуатация системы);
- оценка значимости каждого фактора в отдельности для всего общего набора входной информации;
- выделение информативных факторов риска из общего набора входной информации;
- обучение экспертной системы, система должна позволять получать новые знания при накоплении информации;
- определение степени риска СВСГР, на различных сроках беременности и после рождения ребенка;

– хранение исходных данных, информации, получаемой в процессе обработки, а также результатов определения степени риска СВСГР.

Экспертную систему целесообразно реализовать, используя модульный принцип. Использование модульного построения позволяет получить максимальную гибкость настройки, высокую универсальность, а также возможность расширения функциональности системы в процессе эксплуатации.

На основании выше сказанного впервые разработана структура гибридной экспертной системы определения степени риска СВСГР, укрупненная структура которой представлена на рисунке 2.1.

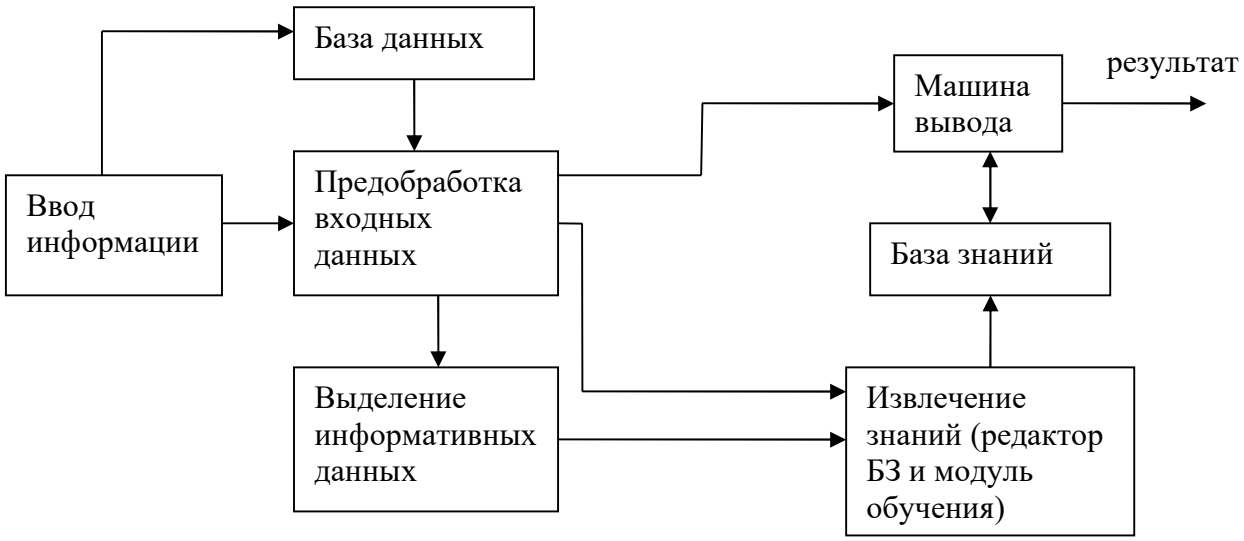

Рисунок 2.1. Укрупненная структура гибридной ЭС определения степени риска СВСГР.

Для разработки ЭС определения степени риска СВСГР необходимо:

– разработать БД, реализовать процессы ввода, хранения и предобработки всей необходимой информации;
– занести в разработанную БД имеющуюся информацию для дальнейшей обработки и преобразований;
– разработать механизм выбора информативных данных, из общего набора факторов;

- разработать механизм извлечения знаний с целью получения возможности определять степень риска СВСГР;
- разработать механизм, реализующий определение степени риска СВСГР.

2.3. Постановка задачи формирования знаний для экспертной системы

При разработке ЭС, приобретение знаний является одной из наиболее трудоемких задач. В разделе 1.5 были рассмотрены методы извлечения знаний, основная цель которых – избавиться от услуг человека-эксперта. Система по методу черного ящика может быть тоже использована, что упрощает работу, особенно для не подготовленного пользователя, учитывая относительную простоту и удобство в использовании. Не менее важна не только возможность распознавать высокую степень риска СВСГР, но и извлекать множества правил из предъявляемых примеров, с наглядным представлением правил вывода. Медицинские данные не всегда являются полными, т.е. некоторые параметры могут быть неизвестны. Это не должно в идеале влиять на результат. Поэтому необходимо предусмотреть возможность работы с неполными данными.

При решении задачи предсказания необходимо определить функцию вида:

$$Y = f(x_1,x_2,...,x_n), \quad (2.1)$$

где Y - исследуемая величина, зависящая от факторов $x_1,x_2,...,x_n$. Искомая функция может быть найдена явно или нет.

Предсказание в зависимости от того какой вид имеет целевая переменная может бать выполнено на основе классификации или регрессии. Мы рассматриваем задачу классификации.

Применительно к нашей задачи Y- классифицируемая степень риска, $x_1,x_2,...,x_n$ – факторы риска СВСГР, функция (2.1) позволяет определить степень риска СВСГР при некоторых определенных факторах $x_1,x_2,...,x_n$. Мы рассматриваем несколько подходов к решению задачи классификации

степени риска СВСГР.

1. Прежде всего, необходимо выполнить классификацию по степеням риска: очень низкая степень риска, низкая степень риска, высокая степень риска, очень высокая степень риска. Формально это задача классификации на l классов, где каждый класс соответствует определенной степени риска СВСГР, и функция для классификации имеет вид, представленный формулой (2.2).

$$Y = \begin{cases} K_1, F \in [0; N_1) \\ K_2, F \in [N_1; N_2) \\ \dots \\ K_l, F \in [N_{l-1}; N_l] \end{cases} \quad (2.2)$$

$$F = \sum_{i=1}^{n} x_i * w_i \quad (2.3)$$

где $K=(K_1,K_2,\dots,K_l)$ - возможные результаты классификации; F – некоторая функция представленная формулой (2.3); $X=(x_1,x_2,\dots,x_n)$ – факторы риска; $W=(w_1,w_2,\dots,w_n)$ - весовые коэффициенты факторов риска; $N=(N_1,N_2,\dots,N_l)$ - пороговые значения. Разработка метода классификации заключается в нахождении весов W, и пороговых значений N, а непосредственно классификация в суммировании весов присутствующих у пациента факторов риска.

2. Для практики самым важным является определение высокого значения степени риска СВСГР с целью назначения соответствующего лечения, направленного на снижение степени риска СВСГР. В этом случае, по сути, необходимо выполнить классификацию на два класса и нет необходимости находить функцию (2.1) явно.

а) примером реализации такого подхода может быть нейронная сеть [27,34,56]. При этом необходимо подобрать архитектуру НС, на входы которой подаются закодированные факторы риска, а на выходе получаем результат: высокая или низкая степень риска СВСГР.

б) для решения этой задачи мы используем математический аппарат булевых функций, которые представляются в виде древообразных структур.

При этом листьями дерева являются закодированные в бинарные значения факторы риска, функциональными узлами – логические функции (И, ИЛИ, НЕ), а корень дерева определяет значение булевой функции, где единица соответствует высокой степени риска, а ноль – низкой.

3. Для улучшения интерпретации решается задача построения классификационных правил, что дает возможность не только определять значение булевой функции (степени риска), но и объяснить, как было получено это значение.

Пусть существует вектор факторов риска $X=(x_1,x_2,...,x_n)$, для каждого i-го фактора существует некоторый вектор допустимых значений $P_i=(p_{i,1},p_{i,2},...p_{i,s})$, где s количество допустимых значений, разное для каждого фактора x_i.

Тогда при решении задачи классификации на два класса классификационное правило может быть представлено, например, следующим образом.

ЕСЛИ $x_1=p_{1,2}$ И $x_2\neq p_{2,7}$ И $x_3=p_{3,4}$ И … И $x_i\neq p_{i,2}$ ТО $Y=K_1$, ИНАЧЕ $Y=K_2$.

Часто имеется несколько таких правил, которые приводят к правильной классификации, например:

ЕСЛИ $x_1=p_{1,2}$ И $x_2\neq p_{2,7}$ И $x_3=p_{3,4}$ И … И $x_i\neq p_{i,2}$ И … И $x_n=p_{i,1}$

ИЛИ

$x_1=p_{1,5}$ И $x_2=p_{2,4}$ И $x_3\neq p_{3,4}$ И … И $x_i\neq p_{i,1}$ И … И $x_n=p_{i,1}$

ИЛИ

…

ИЛИ

$x_1\neq p_{1,1}$ И $x_2=p_{2,3}$ И $x_3\neq p_{3,5}$ И … И $x_i\neq p_{i,8}$ И … И $x_n=p_{i,4}$

ТО $Y=K_1$, ИНАЧЕ $Y=K_2$.

Перечисленные правила представляют собой классификационные правила, каждое из которых самостоятельное правило классификационной системы, выполнение которого отнесет к классу K_1.

В этом случае систему правил теперь можно описать так:

ЕСЛИ *условие 1=да* или *условие 2=да* или … или *условие m=да* ТО $Y=K_1$, ИНАЧЕ $Y=K_2$.

Где *условие i=да* может соответствовать например такое условие: $x_1=p_{1,2}$ И $x_2\neq p_{2,7}$ И $x_3=p_{3,4}$ И … И $x_i\neq p_{i,2}$ И … И $x_n=p_{i,1}$.

Перечисленные правила можно формально описать следующим образом:

ЕСЛИ $f_1(x_1,x_2,...,x_n)=1$ *или* $f_2(x_1,x_2,...,x_n)=1$ *или … или* $f_m(x_1,x_2,...,x_n)=1$ ТО $Y=K_1$, ИНАЧЕ $Y=K_2$.

При такой интерпретации для построения классификационных правил необходимо найти систему булевых функций вида:

$$f_j(x_1,x_2,...,x_n)=1 \qquad (2.4)$$

При этом аргументами каждой из функций не обязательно должен быть полный перечень факторов *X*. К тому же, если рассматривать факторы как бинарные значения (т.е. их присутствие или отсутствие) функцию можно представить булевым выражением. В таком случае функция может быть реализована в базисе: И, ИЛИ, НЕ. А всю конструкцию ЕСЛИ можно представить так:

$$F(x_1,x_2,...,x_n)=\bigvee_{j=1}^{m} f_j(x_1,x_2,...,x_n) \qquad (2.5)$$

В этом случае разработка классификационных правил заключается в построение булевых функций (2.4). Базис логических функций можно изменять, например добавить ИЛИ-НЕ и И-НЕ. Для более удобного восприятия можно наоборот исключить ИЛИ из функций (2.4), в таком случае вся конструкция ЕСЛИ будет представлена в дизъюнктивной нормальной форме (ДНФ) и иметь вид:

ЕСЛИ $F(x_1,x_2,...,x_n)=1$ ТО $Y=K_1$, ИНАЧЕ $Y=K_2$ (2.6)

Где $F(x_1,x_2,...,x_n)$ определена в (2.5), а $f_j(x_1,x_2,...,x_n)= x_1*x_2*...*x_n$, при этом присутствие каждого фактора не обязательно, и каждый присутствующий фактор может быть в прямом или инверсном состоянии (операция НЕ).

Так или иначе, разработка способа извлечения знаний сводится к нахождению определенной функции или системы функций F. Критерием оценки качества найденного решения может быть оценка качества точности классификации. Пусть критерием оценивания является ошибка классификации *E(F)*, тогда задача поиска решения сводится к нахождению функций *F,* так чтобы выполнялось условие:

$$E(F) \to \min \tag{2.7}$$

В качестве ошибки классификации может использоваться:

- доля правильной классификации (2.8);
- среднеквадратичная ошибка (2.9);
- средняя абсолютная ошибка (2.10);
- критерий суммы квадратов ошибки (2.11);

$$E = \frac{1}{M} * (M - \sum_{i=1}^{M} (| F_i - Y_i |)) \tag{2.8}$$

$$E = \frac{1}{M} * \sum_{i=1}^{M} (F_i - Y_i)^2 \tag{2.9}$$

$$E = \frac{1}{M} * \sum_{i=1}^{M} (| F_i - Y_i |) \tag{2.10}$$

$$E = \sum_{i=1}^{M} (F_i - Y_i)^2 \tag{2.11}$$

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

2.4. Постановка задачи предварительной обработки данных

Предварительная обработка данных [36,37,52] предполагает выполнение двух задач: кодирование данных и понижение размерности данных.

Разработка представления обучающих данных (кодирование) - очень важный этап, который в значительной степени определяет качество получаемой экспертной системы. Экспертная система оперирует с информацией, представленной только в виде чисел. Числа подаются на

входы экспертной системы и ответы, снимаемые с выходов, также представляют собой числа. А информация, на основании которой, система должна давать ответ, имеет самый разнообразный вид: термины, описывающие какие-либо заболевания, числа различного вида и величины и т.д. Поэтому необходимо корректно представлять информацию в виде чисел, сохраняющих смысл и внутренние взаимосвязи в данных. В разделе 1.3 рассмотрены методы кодирования и нормировки данных, которые будут использоваться.

Выбор метода понижения размерности зависит от типа экспертной системы, для которой готовятся данные. Если экспертная система предназначена реализовывать цепочки рассуждений, имитирующих анализ ситуации экспертом человеком, то необходимо использовать метод выделения информативных признаков, а если основную ценность экспертной системы будет представлять компьютерная система по методу черного ящика, то можно рассматривать подход, связанный с понижением размерности.

Для системы, имитирующей анализ ситуации экспертом человеком, необходимо подготовить данные с известным набором конкретных информативных признаков.

Предположим, существует набор данных S содержащий m примеров формула (2.12). Каждый пример S_i набора данных S состоит из n определяющих параметров $X_i=(x_1,x_2,...,x_n)$, и параметра – результата Y_i формула (2.13).

$$S = \{S_1, S_2, ..., S_m\} \tag{2.12}$$

$$S_i = \{X_{i,1}, X_{i,2}, ..., X_{i,n}, Y_i\}, i \in [1, m] \tag{2.13}$$

Т.е. каждый i-й пример набора данных S представлен вектором факторов X и результатом Y. При этом факторы $(x_1,x_2,\ldots,x_n)$ набора данных S могут включать в себя как факторы, содержащие полезную информацию о принадлежности к Y, так и факторы частично или полностью

неинформативные. При этом, определенную величину шума могут содержать и информативные факторы.

Критерием оценки качества может быть оценка качества всего набора данных *S* или некоторая мера качества (значимости) каждого отдельного фактора. Выбор критерия оценивания зависит от конкретной задачи.

Допустим, необходимо оценить качество всего набора и критерием оценки *E(X)* выбрана точность классификации (например, погрешность). Тогда задача сводится к нахождению такого вектора *X,* при котором достигается допустимая точность классификации при минимальном количестве входных параметров формула (2.14).

$$E(X) \to \min \text{ и } n \to \min \qquad (2.14)$$

где *n,* количество факторов.

При вычислении меры значимости каждого фактора, необходимо найти такой вектор весов $W=(w_1,w_2,\ldots,w_n)$, при котором выполняется (2.15)

$$E(X * W) \to \min \qquad (2.15)$$

Где $E(X * W)$ - критерий оценки классификации. В качестве критерия оценивания можно использовать формулы (2.8)-(2.11).

2.5. Реализация способов представления обучающих данных

Медицинскими сотрудниками городской больницы №3 города Донецка и кафедры акушерства и гинекологии Донецкого национального медицинского университета им. Горького предоставлены для работы данные и определены факторы риска. Информация получена при обследовании 120 детей (71 мальчик и 49 девочек), которые умерли в Донецкой области от СВСГР, и контрольная группа из 120 живых детей на первом году жизни, подобранных по принципу копий-пар в соответствии с возрастом, полом, годом и месяцем рождения, а также географическим распределением в рамках города. Анализировалась информация о матери и ребенке первого года жизни, представленная в приложении А.

Собрана максимально полная информация о возможных параметрах,

которые в той или иной степени могут влиять на СВСГР. К возможным факторам риска [3,4,28,54,69,74,76,84] врачами - экспертами выделена следующая информация:

- информация о матери: место жительства – город или село; вредные условия труда; образование; состоит ли в браке; бытовые условия и количество $м^2$ на человека; рост и вес; возраст на момент первой беременности; чем закончилась первая беременность; возраст на момент первых месячных; регулярность, болезненность, длительность и интервал месячных; возраст на момент беременности; номер беременности; роды по счету; чем закончились предыдущая беременность; количество абортов, самоабортов, мертворождений; плодность текущей беременности; курение, алкоголь, наркотики в течении беременности; перенесенные заболевания; способы контрацепции; TORCH – инфекции; патология беременности; гинекологические заболевания; группа крови и резус фактор.
- информация об отце: возраст; курение; алкоголь; наркотики.
- информация о ребенке: пол, кормили грудью, искусственное питание или смешанное; вес; рост; количество баллов по шкале Апгар; срок гестации; врожденные пороки; сразу после родов находился: в палате интенсивной терапии, в палате, с мамой.

В разделе 1.3 рассмотрены способы представления обучающих данных. Учитывая, что анализируемые данные, имеют самый разнообразный вид, реализованы все перечисленные в разделе 1.3 способы кодирования.

Для корректного представления исходной информации в виде чисел, написана программа в среде визуального объектно-ориентированного программирования C++ Builder 6 [1,10,16], реализующая описанные выше способы кодировки нечисловых факторов риска и методы нормировки для

числовых. Разработанное приложение состоит из следующих файлов: файл проекта, исполняемый файл, заголовочные файлы, файлы кода программы, файлы визуальных форм. Назначение основных файлов проекта представлены в таблице 2.1.

Таблица 2.1.

Назначение основных файлов проекта FeaturesPreProcessing.

Название	Назначение
FeaturesPreProcessing.bpr	файл проекта
FeaturesPreProcessing.exe	исполняемый файл
FeaturesPreProcessing.cpp	Содержит основную функцию WinMain(…), загружает основную форму.
unit_cells_diapazon.cpp unit_cells_diapazon.h	Содержат функции для здания диапазонов чтения данных
main.cpp main.h	Содержат функции обработки меню главного окна проекта, взаимодействие с файлами Microsoft Excel (обмен данными)
unit_preprocessing_classes. cpp unit_preprocessing_classes.h	Содержат описание классов и методы для выполнения кодирования и нормирования входных величин
unit_features_processing.cpp unit_features_processing.h	Содержат функции обеспечивающие диалог с пользователем по выбору метода предварительной обработки

При запуске программы появляется экранная форма представленная в приложении Б на рисунке Б.1, которая позволяет загрузить данные из файла и сохранить обработанные значения в существующий или другой файл.

Программа считывает входное множество признаков, позволяет последовательно кодировать каждый отдельно. Пример экранной формы

представлен на рисунке Б.2 в приложении Б.

Среди способов приведения предусмотрены следующие:

- линейное масштабирование – выполняет линейную нормировку;
- нелинейное преобразование – выполняет нормировку, используя функции активации нейронов. Реализованы следующие функции: гиперболический тангенс и сигмоид;
- интервальный тип преобразования – предусмотрен для преобразования числовых признаков в двоичный код; происходит разбиение входного интервала на отрезки, количество отрезков определяет новую размерность признака; единичное значение будет у того бита, номер которого соответствует интервалу, в который попадает значение;
- кодирование бинарных признаков – предусмотрены варианты кодирования бинарных признаков рассмотренные выше;
- неупорядоченные признаки – предусмотрено кодирование *«n → n»* и *«n → m»*;
- упорядоченные признаки – используется рассмотренные выше варианты *«n → n»* с накоплением, пропорционально встречаемости.

Выбор способа предобработки зависит от реализации экспертной системы. Разработанная программа предусматривает возможность подготовки данных для различных типов экспертных систем. Программа работает с файлами в формате MS Excel, каждый обработанный признак при сохранении дописывается в открытый файл.

Выполнена предварительная обработка данных различным образом, результаты представлены в приложениях В, Г, Д и Е.

2.6. Разработка и реализация методов выбора информативных факторов риска СВСГР

В медицине достаточно часто используют корреляционный анализ [15,19,40], для оценивания зависимостей между параметрами. Строятся корреляционные портреты, которые показывают тесноту связей между различными факторами. Мы будем использовать, рассмотренный в разделе 1.4, корреляционный анализ, для оценивания влияния каждого отдельного фактора риска на выходную величину.

Для реализации данного метода необходимо предварительно подготовить данные. Обработка заключается в первую очередь в кодировании нечисловой информации. Чтобы выяснить влияние одного фактора при различных значениях, необходимо разбить диапазон значений этой переменной на интересующие интервалы, как показано в приложении Ж. Например, такой фактор риска как, возраст матери на момент родов, будем анализировать в следующих промежутках: до 17 лет, от 17 до 25 лет, от 25 до 31 года и после 31 года. Таким образом, вместо одного параметра, для анализа получили 4. Корреляционный анализ выполнялся в пакете Microsoft Office Excel. Вычислялся коэффициент корреляции между переменой, входящей в модель и результатом. В соответствии со значимостью фактора ему сопоставлен вес (определенное число баллов). Задача выбора весовых коэффициентов неоднозначна. Коэффициенты могут выбираться полностью экспертным путем или формальным методом. В данном случае медицинским работникам были предложены весовые коэффициенты, определенные по формуле (2.16), а они откорректировали их на свое усмотрение. Полученные весовые коэффициенты представлены в приложении З.

$$w_i = \frac{r_{x_i y}}{\sum_{j=1}^{n} r_{x_j y}} \bullet 100 \quad (2.16)$$

Для реализации задачи отбора значимых факторов написана программа в среде визуального объектно-ориентированного программирования C++ Builder 6 [1,10,16]. С ее помощью можно построить и обучить нейронную сеть типа многослойный персептрон, произвести отбор входных параметров с помощью генетических алгоритмов. Программа позволяет задавать архитектуру сети: указывать количество слоев, количество нейронов на каждом слое и выбирать активационные функции для каждого слоя (здесь предусмотрены наиболее распространенные для решения подобных задач варианты: линейная, сигмоидальная, гиперболический тангенс, гладкая ступенчатая). Предусмотрена возможность использования пре- и пост-процессинга входных данных. Входные и выходные данные можно загрузить из файла Microsoft Excel.

Разработанное приложение состоит из следующих файлов: файл проекта, исполняемый файл, заголовочные файлы, файлы кода программы, файлы визуальных форм. Назначение основных файлов проекта представлены в таблице 2.2.

Таблица 2.2.

Назначение основных файлов проекта.

Название	Назначение
ChoosingMLP_Inputs_withGA.bpr	файл проекта
ChoosingMLP_Inputs_withGA.exe	исполняемый файл
ChoosingMLP_Inputs_withGA.cpp	Содержит основную функцию WinMain(…), загружает основную форму.
genetic_alg_class.cpp	Содержат описание классов и методы

genetic_alg_class.h	реализации ГА
neural_classes.cpp neural_classes.h	Содержат описание классов и методы реализации и функционирования НС
unit_cells_diapazon.cpp unit_cells_diapazon.h	Содержат функции для задания диапазонов чтения данных (обучающих, проверочных, …)
unit_charts.cpp unit_charts.h	Содержат функции обеспечивающие диалог с пользователем по выбору лучшей эпохи обучения
unit_best_epoch.cpp unit_best_epoch.h	Содержат функции обеспечивающие выбор лучшей эпохи обучения
unit_param_seti.cpp unit_param_seti.h	Содержат функции обеспечивающие выбор архитектуры сети, функций активации, реализация пре- и пост-процессинга
main.cpp main.h	Содержат функции обработки меню главного окна проекта, взаимодействие с файлами Microsoft Excel (обмен данными)
unit_chart_gen_alg.cpp unit_chart_gen_alg.h	Содержат функции обеспечивающие диалог с пользователем по выбору настроек ГА
unit_param_gen_alg.cpp unit_param_gen_alg.h	Содержат функции обеспечивающие выбор настроек ГА
unit_results_GA.cpp unit_results_GA.h	Содержат функции сохранения результатов
unit_stop_conditions.cpp unit_stop_conditions.h	Содержат функции обеспечивающие диалог с пользователем по выбору условия остановки НС

Экранные формы работы с программой представлены в приложении И. При запуске программы появляется окно представленное на рисунке И.1

При выборе пункта «создать новую прогнозирующую сеть» загружается окно показанное на рисунке И.2, в котором задаются параметры новой сети.

Сеть обучается алгоритмом обратного распространения ошибки [26,33,41,56]. Так как в общем случае не существует доказательства сходимости данного алгоритма, так и не существует какого-либо четкого определенного критерия останова. Есть несколько обоснованных вариантов критерия останова, которые можно использовать. Каждый из них имеет как свои практические преимущества, так и недостатки. Одним из таких критериев является малая интенсивность изменений среднеквадратической ошибки в течение эпохи. Интенсивность изменения среднеквадратической ошибки обычно считается достаточно малой, если она лежит в пределах до 1% за эпоху. Иногда используется уменьшенное значение – 0,01%. Такой критерий может привести к преждевременной остановке процесса обучения. Другим критерием может быть достижение определенного значения среднеквадратической ошибки. Недостатком этого критерия является то, что может потребоваться довольно много времени для сходимости данного алгоритма. Так же можно задавать определенное количество эпох обучения. Как правило, нет информации о том, сколько их может потребоваться, поэтому данный критерий тоже не всегда удобен. Если необходимо, чтобы сеть обладала хорошей способностью к обобщению, можно использовать перекрестную проверку [56]. В данном случае обучающее множество делится на подмножество для обучения и проверочное подмножество, которое используется для проверки эффективности различных моделей сети, из которых необходимо выбрать лучшую. В разработанной программе предусмотрены все перечисленные выше критерии останова (рисунок И.3).

После того как будет определена архитектура нейронной сети, при которой сеть успешно обучится, разрабатывается генетический алгоритм

[29,42], который используя архитектуру выбранной сети будет пытаться обучить сеть, используя различные комбинации набора входных переменных.

Предусмотрены различные варианты параметров ГА (рис. И.4): в качестве метода селекции можно выбирать колесо рулетки или турнир (с указанием количества особей в туре); в качестве метода редукции предусмотрена элитарная стратегия, полная замена, частичная замена популяции (с указанием процента заменяемых особей); можно задавать вероятность операции мутации и скрещивания, (предусмотрен одноточечный и двухточечный оператор скрещивания). В качестве критерия останова можно использовать определенное количество итераций или указать количество повторений результата.

Принцип реализации кодирования хромосомы [11,13,43,44] представлен на рисунке 2.2. Каждая хромосома представляет собой последовательность определенного числа бит (определяется максимальным количеством факторов риска). Значение каждого бита равно 1, если фактор с соответствующим номером присутствует в данном наборе, и нулю если этот фактор отсутствует.

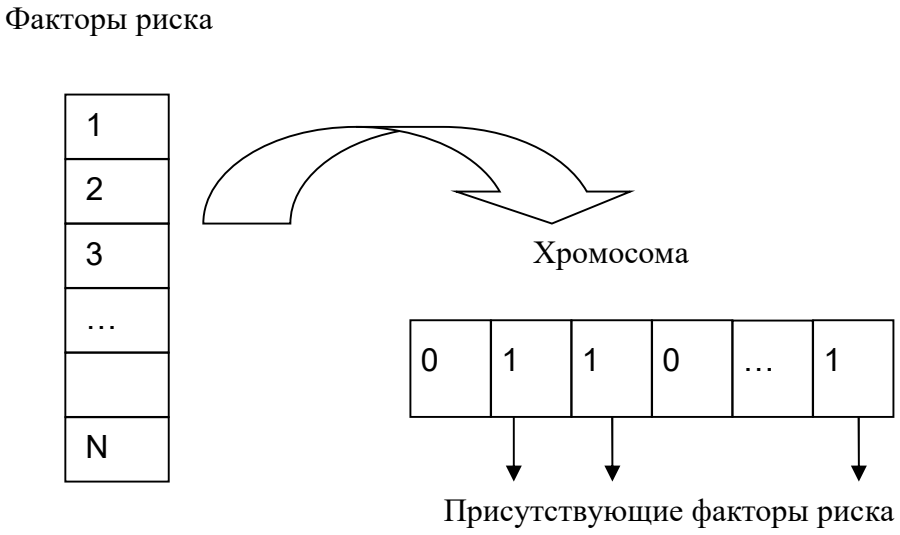

Рисунок 2.2. Кодирование хромосомы.

Разработана фитнесс-функция [11,13,43,44] представленная формулой (2.17). Данная фитнесс функция позволяет выбирать соотношение точности классификации и количества факторов риска.

$$F = \left(\frac{X_i}{X_n} \right) * Q_1 + \left(\frac{E_i - E_n}{E_n} \right) * Q_2 \qquad (2.17)$$

где X_i – количество единиц для i – ой хромосомы, X_n – максимальное количество единиц, E_i - ошибка обучения НС для i – ой хромосомы, E_n – ошибка обучения НС при использовании максимального количества факторов, Q_1 и Q_2 - мера влияния на фитнесс-функцию.

Меру влияния каждого слагаемого можно корректировать вручную. Диапазон допустимых значений – $Q_1, Q_2 \in (0,1)$, и должно выполняться условие 2.18.

$$Q_1 + Q_2 = 1 \qquad (2.18)$$

2.7. Экспериментальные исследования определения информативного набора факторов риска СВСГР

Рассмотрим выбор информативных параметров с помощью нейронных сетей и генетических алгоритмов более подробно. На рисунке 2.3 представлен обобщенный алгоритм работы ГА []. На этапе считывания обучающих данных предполагается, что данные уже предварительно закодированы.

Учитывая, что факторы риска будут подаваться на входы нейронной сети, прежде всего, необходимо закодировать информацию в числовую форму. В разделе 1.3 рассмотрены способы кодирования, учитывая нейросетевую специфику. Наиболее подходящие варианты кодирования представлены в приложениях В и Г.

Проведены исследования по выбору архитектуры и обучению нейронной сети, которые показали, что вариант кодирования, предложенный в приложении В, подходит более, чем тот, что в приложении Г. Поэтому

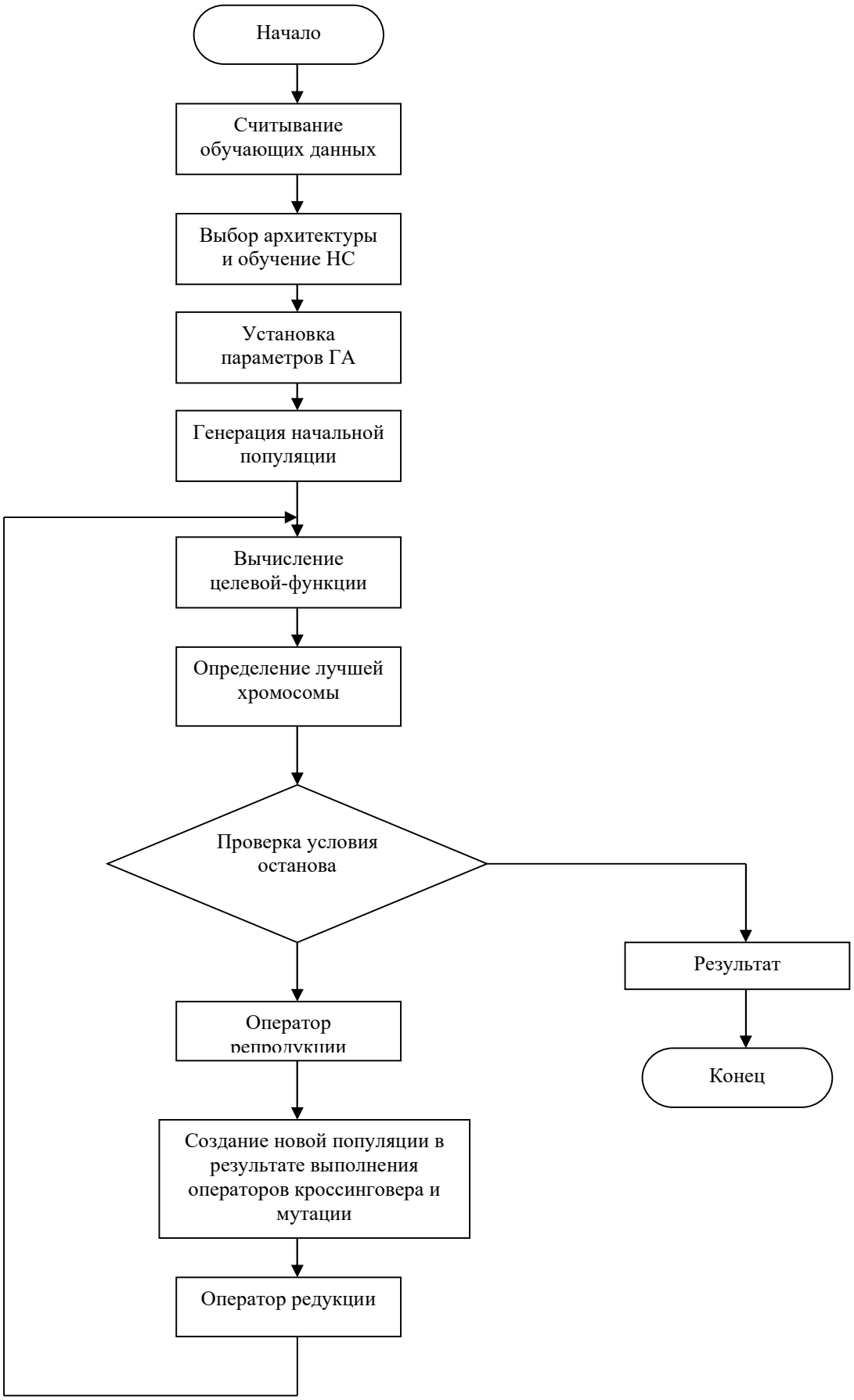

Рисунок 2.3. Обобщенный алгоритм работы ГА.

выполнялись дальнейшие исследования с данными, закодированными первым вариантом. Результаты экспериментов представлены в таблице 2.3.

Таблица 2.3.

Экспериментальные данные по выбору архитектуры сети.

Структура сети	Активационные функции	Среднеквадратичная ошибка для кода 1 (Приложение В)	Среднеквадратичная ошибка для кода 2 (Приложение Г)
99-6-1	Сигмоидная, линейная.	0,0006	0,0127
99-6-1	Гиперболический тангенс, линейная.	0,0000347	-
99-6-3-1	Гиперболический тангенс, гиперболический тангенс, линейная.	0,00059	0,095
99-6-3-1	Сигмоидная, сигмоидная, линейная.	0,0011	0,095
99-8-4-1	Гиперболический тангенс, сигмоидная, линейная.	0,00000264	0,125
99-8-6-3-1	Гиперболический тангенс, сигмоидная, гиперболический тангенс, линейная.	0,00001625	0,0938

Учитывая, что во время выполнения программы происходит многократное обучение НС (на каждом шаге выполнения ГА происходит обучение сети для каждой хромосомы), целесообразно для дальнейшей работы (выбора значимых параметров с помощью генетических алгоритмов) выбирать архитектуру сети с минимальной сложностью, что позволит уменьшить время выполнения программы.

Под сложностью сети будем предполагать вычислительную сложность алгоритма обучения НС. Под вычислительной сложностью алгоритма

обучения будем понимать количество операций за один шаг обучения. Известно, что для алгоритма обратного распространения количество операций связано линейной зависимостью с синаптическими весами (включая пороги) []. Количество весовых коэффициентов для сети с одним выходным нейроном определяется формулой (2.18).

$$N_w = N_{вх} * N_1 + N_1 + \sum_{i=1}^{n-1}(N_i * N_{i+1} + N_{i+1}) + N_n + 1 \qquad (2.18)$$

N_w - число весовых коэффициентов (включая пороги); $N_{вх}$ - число входов; N_i – число нейронов i-го скрытого слоя; n – число скрытых слоев.

Учитывая, что число входов не меняется (определено факторами риска), вычислительную сложность определяет число скрытых слоев и число нейронов на каждом из них.

На основе экспериментальных данных, которые представлены в таблице 2.3 и учитывая сложность сети, выбрана архитектура НС, представленная на рисунке 2.4. Количество входов обусловлено максимальным количеством факторов риска (после кодирования их получили 99). Выбранные активационные функции – гиперболический тангенс и линейная функция.

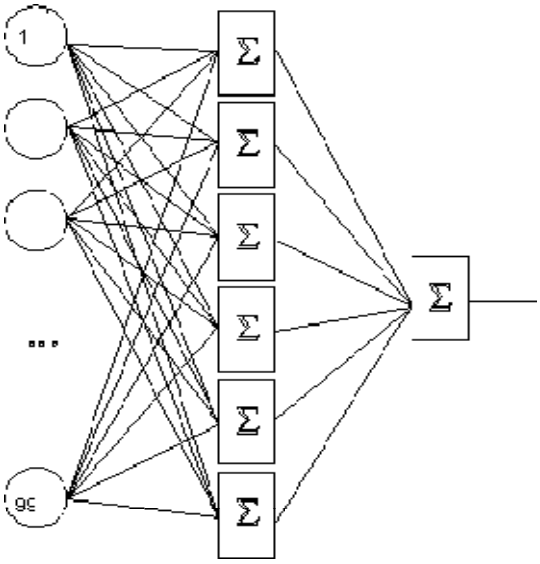

Рисунок 2.4.Архитектура нейронной сети.

Результаты экспериментов по выбору значимых входных параметров представлены в таблице 2.4.

Таблица 2.4.

Экспериментальные данные по выбору значимых параметров.

Желаемое количество факторов, %	Полученное количество факторов, кол-во	Ошибка обучения на обучающей выборке	Ошибка обучения на проверочной выборке
5	10	0,000160618	0,013999935
10	17	5,89474E-05	0,026213435
15	14	0,000131231	0,034513446
20	25	8,17713E-05	0,035583217
25	21	0,000190826	0,064815923
30	31	5,69013E-05	0,028689748
35	20	4,25809E-05	0,027207667
40	45	2,92203E-05	0,034853067
45	46	7,62789E-05	0,020447494
50	57	2,90347E-05	0,025383944
55	53	5,26328E-05	0,054565834
60	55	4,7459E-05	0,044801069
65	65	5,3988E-05	0,032565723
70	66	2,4346E-05	0,022935656
75	74	2,68332E-05	0,021806996
80	77	2,87431E-05	0,019708662
85	87	5,41327E-05	0,021343562
90	88	3,26543E-05	0,022770304
95	95	6,03378E-05	0,016873855
100	98	1,70007E-05	0,017169321

Для дальнейшей работы выберем набор, в котором присутствуют 46 факторов риска, так как данный набор является оптимальным в соотношении с ошибкой обучения (на обучающем множестве и на проверочном множестве) и количеством факторов. В приложении К представлены расширенные результаты экспериментов по выбору информативных факторов с помощью НС и ГА. В данном приложении указаны выбранные факторы риска для каждого набора. На рисунке 2.5 и 2.6 представлены зависимости ошибки обучения от % желаемых факторов.

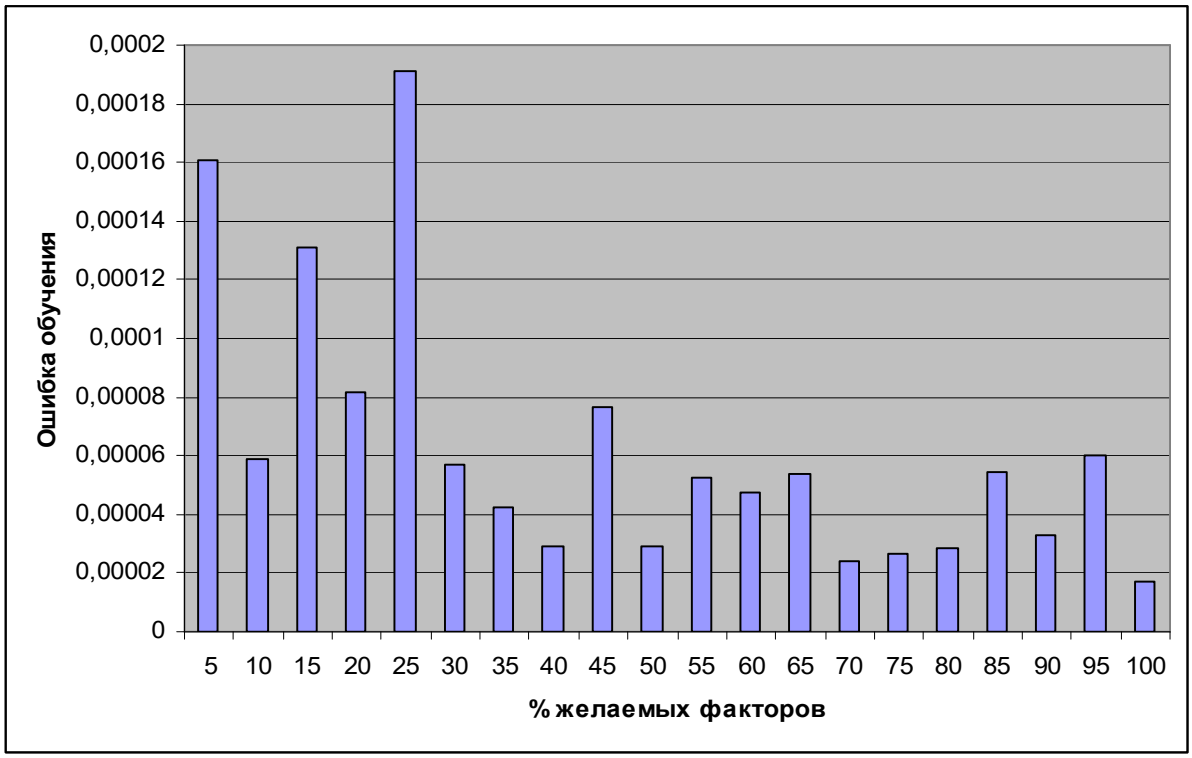

Рисунок 2.5. Зависимость ошибки обучения на обучающих данных от количества желаемых факторов.

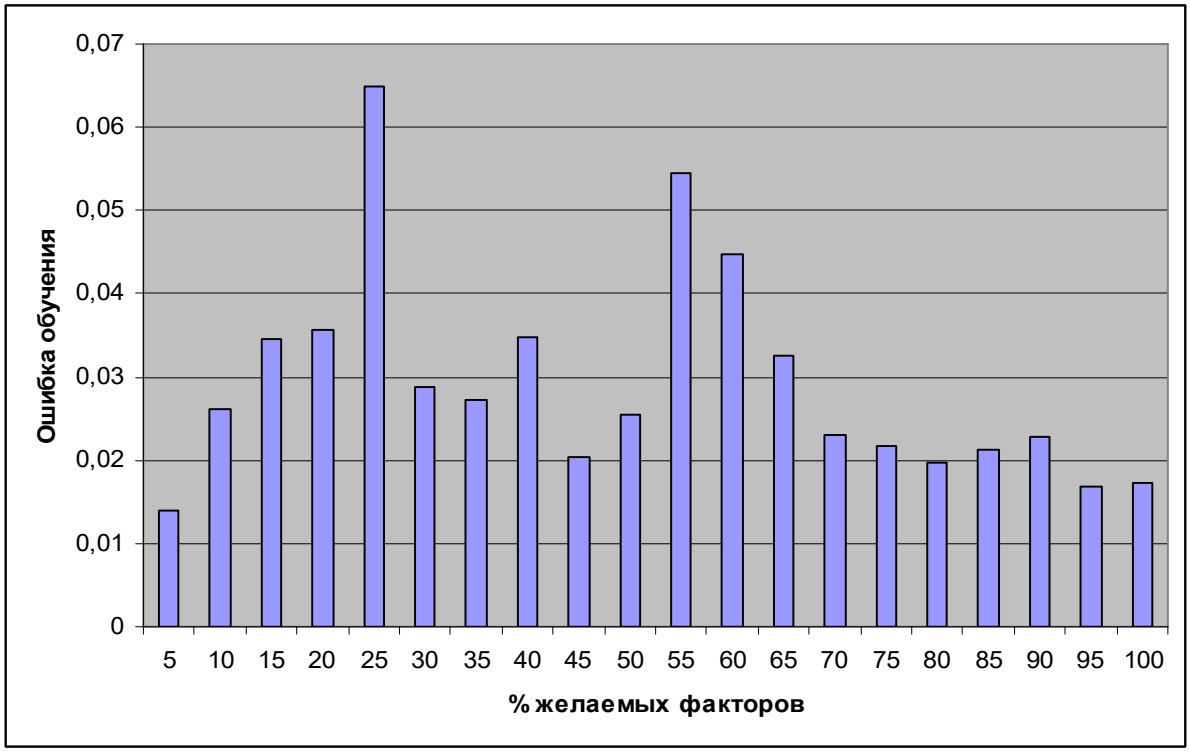

Рисунок 2.6. Зависимость ошибки обучения на проверочных данных от количества желаемых факторов.

Таким образом, выбрав набор информативных параметров состоящий из 46 факторов риска, удалось сократить обучающие данные практически в два раза. Для дальнейшей работы (извлечения знаний для экспертной системы) будем использовать в качестве обучающей выборки выбранный набор, приняв его за базовый.

2.8. Выводы

1. Разработана структура экспертной системы определения степени риска СВСГР. Используется комбинирование классических технологий создания ЭС и методов машинного обучения. Структура ЭС разработана в соответствии с рассмотренными разделе 1.1 этапами диагностики и прогнозирования в медицине.

2. Выполнена математическая постановка задачи к предварительной обработке данных.

3. Выполнена реализация методов предобработки, которые рассмотрены в разделе 1.3. Предобработка предполагает кодирование нечисловой информации и нормировку числовых данных. Предусмотрены различные варианты кодирования нечисловых характеристик, учитывая их тип, для числовой информации реализованы различные способы нормировки.

4. Выполнено кодирование данных различными способами, получена возможность работать с различными вариантами представления знаний ЭС. Экспериментально выбраны оптимальные варианты предварительной обработки для работы с нейронными сетями.

5. С помощью корреляционного анализа определена количественная мера значимости каждого отдельного фактора риска СВСГР.

6. Разработана фитнесс-функция для генетического алгоритма, реализующего выбор информативных факторов риска, что позволило достигнуть высокой эффективности выделения информативных факторов риска в медицине за счет регулирования соотношения количества факторов и

ошибки классификации.

7. Выполнена реализация метода отбора информативных признаков с помощью генетических алгоритмов и нейронных сетей. Выполнены эксперименты по выбору информативных параметров на примере факторов риска СВСГР. Для дальнейшей работы, выбран информативный набор данных, в котором присутствуют 46 факторов риска, что практически в два раза меньше исходного. Таким образом, разработанный метод позволил сократить входной набор данных практически в два раза.

8. Выполнена математическая постановка задачи к методам извлечения знаний, применительно к определению степени риска СВСГР. Поставлена задача к формированию базы знаний для различных вариантов ее представления.