УДК 004.048:004.622

Васяева Т.А., Скобцов Ю.А.

ИЗВЛЕЧЕНИЕ ЗНАНИЙ НА ОСНОВЕ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ И ГЕНЕТИЧЕСКОГО ПРОГРАММИРОВАНИЯ

Рассмотрены этапы разработки медицинских экспертных систем. Рассмотрен и реализован метод определения информативной информации из состава факторов риска на основе нейронных сетей и генетических алгоритмов. Разработана архитектура нейронной сети, подобраны генетические операторы, разработана фитнесс-функция. Разработан аппарат генетического программирования для прогнозирования СВСГР. Проведены исследования и приведены результаты использования методов на реальных медицинских данных.

Введение. Современным направлением диагностики и прогнозирования являются методы, основанные на извлечении знаний. Формирование знаний очень важный этап, который в значительной степени определяет качество получаемой системы принятия решений. Как правило, основную ценность представляет собой явная формализация правил вывода, но иногда достаточно компьютерной системы – метод черного ящика – как инструмента предсказания.

Один из подходов формирования знаний заключается в разработке программ, способных обучаться под руководством эксперта-учителя. Так учитель предъявляет программе примеры реализации некоторого концепта, а задача программы состоит в том, чтобы извлечь из предъявленных примеров набор атрибутов и значений, определяющих этот концепт.

Естественно, наиболее впечатляющим примером обучаемой системы является организм человека или животного, который эволюционировал вместе с окружающим миром. Этот подход к обучению, основанный на адаптации, отражен в генетических алгоритмах и генетическом программировании.

Целью работы является разработка метода прогнозирования с явной формализацией правил вывода на примере синдрома внезапной смерти грудных детей (СВСГД).

Задачи диагностики, прогнозирования и принятия решений в медицине – это комплексный процесс, который охватывает шаги, начиная от получения и представления данных до оценки 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 поставленной задачи.

Подготовка данных. Отбор данных выполняется врачом. Для решаемой задачи использовались реальные данные полученные при обследовании 120 детей, которые умерли в Донецкой области от СВСГД, 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. Собрана максимально полная информация о возможных параметрах, которые в той или иной степени могут влиять на СВСГД. К возможным факторам риска выделили следующую информацию:

– информация о матери: место жительства – город или село; вредные условия труда; образование; состоит ли в браке; бытовые условия и количество м2 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.

Разработка представления обучающих данных - 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 возникает необходимость корректного представления этой информации в виде чисел, сохраняющих смысл и внутренние взаимосвязи в данных.

Существует огромное количество способов представления информации для различных целей. В нашей задаче использовалось кодирование в булевы переменные для поиска закономерностей и кодирование в числовые переменные для редукции данных.

Анализ всех факторов риска вызывает существенные затруднения при построении правил вывода.

В этом случае, как и в других медицинских задачах, результат прогнозирования зависит от большого количества неодинаковых по значимости факторов, которые к тому же могут быть взаимосвязаны. А это означает, что использование традиционных статистических методов может не привести к желаемому результату, что и привело к необходимости в применении методов искусственного интеллекта.

Рассмотрен метод, который позволяет выявить значимые входные параметры, с помощью нейронных сетей (НС) [1,2] и генетических алгоритмов (ГА) [2].

Для выделения полезных входных переменных с помощью НС нужно перебрать различные варианты их комбинаций. Такая стратегия может эффективно реализована с помощью ГА, которые являются эффективным инструментом поиска решений в комбинаторных задачах. Схема работы ГА: каждый возможный вариант набора входных переменных можно представить в виде битовой маски. Ноль в соответствующей позиции означает, что данная входная переменная не включена во входной набор, единица – что включена. Таким образом, маска представляет собой строку битов – по одному на каждую возможную входную переменную – и ГА оптимизирует такую битовую структуру. Алгоритм следит за некоторым набором таких строк, оценивая каждую из них по контрольной ошибке (ошибка обучения). По значениям ошибки производится отбор лучших вариантов масок, которые

комбинируются друг с другом с помощью искусственных генетических операций: скрещивания и мутации.

Рассмотренный подход реализован в среде программирования C++ Builder 6. Реализованы следующие возможности: можно построить и обучить (НС) типа многослойный персептрон, произвести отбор входных параметров с помощью (ГА). Архитектура сети определяется количеством слоев, количеством нейронов на каждом слое и активационной функцией для каждого слоя. Предусмотрена возможность использования пре- и пост-процессинга входных данных. Входные и выходные данные можно загрузить из файла Microsoft Excel.

ГА используют архитектуру подобранной НС и пытаются ее обучить используя различные комбинации набора входных переменных. При этом предусмотрены следующие возможности: в качестве метода селекции можно выбирать колесо рулетки или турнир; в качестве метода редукции предусмотрена элитарная стратегия, полная замена, частичная замена популяции; можно задавать вероятность операций мутации и скрещивания. В качестве критерия останова можно использовать определенное количество итераций или указать количество повторений результата. Предложена следующая функция приспособленности:

$$F = \left(\frac{X_i}{X_n}\right) * W_1 + \left(\frac{E_i - E_n}{E_n}\right) * W_2 \qquad (1)$$

где Xi – количество единиц для i – ой хромосомы, Xn – максимальное количество единиц, Ei - ошибка обучения для i – ой хромосомы, En –ошибка обучения при использовании максимального количества факторов, W1 и W2 - мера влияния на фитнесс-функцию.

Меру влияния каждого слагаемого можно корректировать вручную. Диапазон допустимых значений - $W1,W2 \in (0,1)$, и должно выполняться условие: $W1 + W2 = 1$.

При тестировании на реальных медицинских данных получили следующие результаты.

1. После кодирования в числовые переменные получили 99 входных параметров.

2. Разработана архитектура НС для прогнозирования СВСГР. Сеть состоит из 4 слоев: первый слой - входной, второй и третий - скрытые (3 и 2 нейрона в слое, функции активации - гиперболический тангенс), на выходе один нейрон с линейной функцией активации. Данная НС позволяет прогнозировать с точностью – 0,000017.

3. Результаты экспериментов по выбору значимых входных параметров представлены в таблице 1. Для дальней работы выберем набор в котором присутствуют 46 параметра. Таким образом сократив входной набор практически в два раза.

Таблица 1.

Экспериментальные данные выбора 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

Решение задачи на основе ГП реализуется следующей последовательностью действий.

1. Установка параметров эволюции;
2. Инициализация начальной популяции;
3. T:=0;
4. Оценка особей, входящих в популяцию;
5. T:=T+1;
6. Отбор родителей;
7. Создание потомков выбранных пар родителей – выполнение оператор кроссинговера;
8. Мутация новых особей;
9. Расширение популяции новыми порожденными особями;
10. Сокращение расширенной популяции до исходного размера;
11. Если критерий останова алгоритма выполнен, то выбор лучшей особи в конечной популяции – результат работы алгоритма. Иначе переход на шаг 4.

Для решения задачи с помощью ГП необходимо выполнить предварительные этапы:

- определить терминальное множество;
- определить функциональное множество;
- определить фитнесс-функцию;
- определить значения параметров, такие как мощность популяции, максимальный размер особи, вероятности кроссинговера и мутации, способ отбора родителей, критерий окончания эволюции и т.п.

После этого можно разрабатывать непосредственно сам эволюционный алгоритм, реализующий ГП для конкретной задачи.

Разработка метода. Каждое потенциальное решение в нашем случае представляется деревом, состоящим из функций, которые являются внутренними узлами деревьев, и терминалов, которые формируют листья деревьев. Положительный ответ на выходе соответствует высокой степени риска СВСГР, а отрицательный – соответственно низкой.

Основная идея данного метода заключается в методе кодирования особей для генетического программирования. Каждая особь представляет собой дерево, которое соответствует синтаксическому выражению, представляющее множество правил в дизъюнктивной нормальной форме.

На рисунке 1 представлен пример дерева в дизъюнктивной нормальной форме. Дерево представлено двумя правилами. Данное представление особи

значительно упрощает интерпретацию результата. В рассмотренном примере расшифровка будет следующей:

ЕСЛИ правило 1 ИЛИ правило 2 ТО результат 1, ИНАЧЕ результат 2.

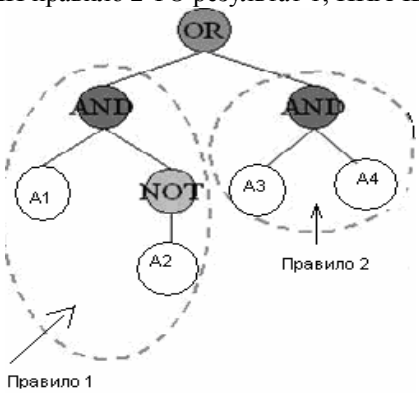

Рисунок 1. Пример дерева в дизъюнктивной нормальной форме.

Определим терминальное множество.

Данные должны быть предварительно обработаны, основное назначение предобработки преобразовать входное обучающее множество в булевы переменные. Для этого исходные данные преобразуем следующим образом:

- место жительства (город – 1, село – 0)
- возраст матери на момент родов (полных лет) ≤ 17
- возраст матери на момент родов (полных лет) ≤ 25
- возраст матери на момент родов (полных лет) ≤ 30
- возраст матери на момент родов (полных лет) > 31
- место работы матери, профвредность (да – 0 , нет – 1)
- и др.

Наличие каждого фактора принято за единицу, отсутствие за ноль.

Терминальное множество в данном случае составляют перечисленные выше параметры, которые после предобработки представляют собой булевы переменные.

Функциональное множество состоит из логических операций: AND, OR, NOT.

В качестве фитнесс-функции рассматривается доля пациентов с правильно поставленным диагнозом. Переменная диагноза принимает булевы значения 0 или 1. Единица соответствует положительному диагнозу (высокой степени риска СВСГР) и ноль отрицательному (низкой степени риска СВСГР).

Реализация и апробация метода. Для реализации поставленной задачи написана программа в среде C++ Builder 6, которая выполняет рассмотренный алгоритм.

Генерация начальной популяции. На данном этапе происходит генерация начальной популяции, в соответствии с заданными параметрами. Популяция состоит из набора деревьев, сгенерированных случайным образом. Генерация каждого дерева происходит рекурсивно, начиная с генерации первым функционального узла ИЛИ и его аргументов. В качестве аргументов на первом шаге может быть только узел ИЛИ. Далее для каждого дочернего узла случайным образом определяется тип и значения его аргументов по следующим принципам:

- после узла ИЛИ может быть только функциональный узел (значениями которого могут быть – ИЛИ или И);
- после узла И может быть функциональный узел (значениями которого могут быть – И или НЕ) или терминальные узлы;
- после узла НЕ может быть только терминальный узел.

Процесс выполняется по левой ветви до тех пор, пока не будет выбран дочерним терминальный узел. Затем генерируются правые ветви.

Вероятность функционального и терминального узлов меняется по следующему принципу: чем ниже вершина, тем больше вероятность терминального узла и меньше функционального. Для функционального узла на каждом последующем шаге увеличивается вероятность узла И и уменьшается вероятность узла ИЛИ.

При формировании дерева в одной ветви ИЛИ (т.е. для одного правила) не используется один и тот же терминальный символ более одного раза.

Предусмотрены методы создания деревьев: полный, растущий и комбинированный.

Применение генетических операций:

Отбор родителей. Предложено использовать отбор пропорционально значению целевой функции реализованный методом рулетки или турниром. При этом если два или более потомка имеют одинаковую фитнесс-функцию, то выбирается дерево минимальной сложности.

Кроссинговер. Для древообразной формы представления используются следующие три основных операторов кроссинговера:

- узловой кроссинговер;
- кроссинговер поддеревьев;
- смешанный.

Учитывая строго определенное представление дерева необходимо модифицировать операторы кроссинговера. Модификация заключается в выполнении оператора кроссинговера для худшего правила и в поиске оптимальной точки разрыва.

Мутация. Для деревьев используются следующие операторы мутации:

- узловая;
- усекающая;
- растущая.

Как и в случае с оператором кроссинговера оператор мутации должен быть модифицирован. Модификация заключается в определении вероятности мутации в соответствии с ошибкой обучения.

Редукция. Предлагается использовать элитную стратегию.

Критерий останова. В качестве критерия останова можно выбирать указание определенного числа итераций или указание определенного числа повторения лучшего результата.

Выводы. При тестировании на реальных медицинских данных получили 95,71% правильно распознанных диагнозов. Таким образом, результат можно считать положительным. Разработанный аппарат ГП создан и протестирован на примере прогнозирования СВСГР, но может быть использован и при решении других задач медицинской диагностики и прогнозирования.

Список литературы

1. Саймон Хайкин Нейронные сети: полный курс, 2-е издание. : Пер. с англ. – М. : Издательский дом «Вильямс», 2006. – 1104 с.
2. Рутковская Д., Пилинский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы: Пер. с польск. И.Д. Рудинского. - М.: Горячая линия – Телеком, 2006. – 452 с. : ил.
3. W. Banzhaf et all. Genetic Programming – an Introduction. – Morgan Kaufman, Heidelberg:San-Francisco, 1998.
4. Skobtsov Y.A., Vasyaeva T.A. Diagnosis of SIDS using Genetic Programming. Advanced Computer Systems and Networks: Design and Application. Proceedings of 3-st International Conference ACSN-2007. 20-22 September, 2007, Lviv, Ukraine 92-93c