РАЗРАБОТКА ЭКСПЕРТНЫХ СИСТЕМ МЕДИЦИНСКОЙ ДИАГНОСТИКИ С ЯВНЫМ ПРЕДСТАВЛЕНИЕМ ПРОДУКЦИОННЫХ ПРАВИЛ НА ОСНОВЕ ГП

Васяева Т.А., Скобцов Ю.А.
Донецкий национальный технический университет 83000, Донецк, ул. Артема 58, vasyaeva_tanya@tr.dn.ua, skobtsov@kita.dgtu.donetsk.ua

При разработке ЭС, приобретение знаний является одной из наиболее трудоемких задач. Общий подход состоит в разработке и использовании программ, способных обучатся 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.

Автоматизированные методы формирования знаний на базе машинного обучения (machine learning) [1] применительно к проблематике экспертных систем используются для:

- извлечения множества правил из предъявляемых примеров;
- анализ важности отдельных правил;
- оптимизация производительности набора правил.

Главной задачей проектируемой экспертной системы является выполнение диагностики - в нашем случае определения высокой степени риска синдрома внезапной смерти грудного ребенка (СВСГР) и извлечения множества правил из предъявляемых примеров, с наглядным представлением правил вывода.

В данной задаче в качестве обучающего множества используются реальные данные обследования 240 пациентов, (120 детей, которые умерли в Донецкой области от СВСГД, и контрольная группа из 120 живых детей на первом году жизни). Данные составляют информацию общего характера и образа жизни беременных, а так же перенесенные заболевания и результаты некоторых анализов.

Для решения поставленной задачи предложено использовать генетическое программирование (ГП) [2]. Предлагается модификация способа кодирования особей для генетического программирования. Каждая особь представляет собой дерево, которое соответствует синтаксическому выражению, представляющее множество правил в дизъюнктивной нормальной форме. Фрагмент дерева прогнозирования высокой степени СВСГР представлен на рисунке 1.

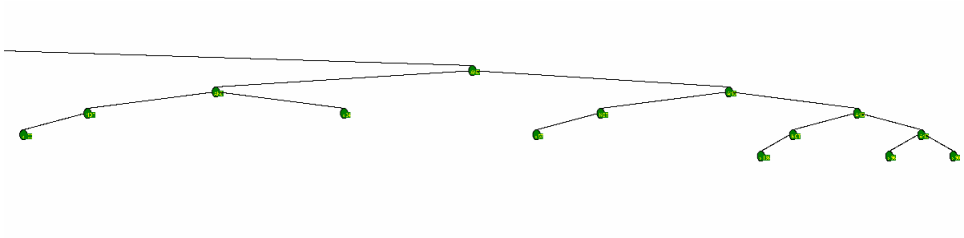

Рисунок 1. Фрагмент дерева прогнозирования высокой степени СВСГР.

Терминальное множество состоит из факторов риска, которые после предобработки представляют собой булевы переменные и соответствуют листьям дерева. Функциональное множество состоит из логических операций: AND, OR, NOT, которые представляют внутренние вершины дерева.

Популяция состоит из набора деревьев, сгенерированных случайным образом. Генерация каждого дерева происходит рекурсивно, начиная с первого 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 правилам:

- 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 генерации функционального или терминального узлов меняется по следующему правилу: 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.

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). Значение фитнесс-функции для особей с правильным диагнозом принимает значение 1, а для особей с неправильным диагнозом принимает значение 0.

Применение генетических операций [3].

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.

Учитывая строго определенное представление дерева необходимо модифицировать операторы кроссинговера и мутации, чтобы не нарушить структуру хромосомы при выполнении этих операторов.

Кроссинговер. Для древообразной формы представления используются следующие три основных операторов кроссинговера: узловой кроссинговер; кроссинговер поддеревьев; смешанный.

В узловом операторе кроссинговера обмен возможен только для терминальных узлов. В кроссинговере поддеревьев родители могут обмениваться только поддеревьями ветви И. При смешанном операторе кроссинговера для некоторых узлов выполняется узловой оператор кроссинговера, а для других - кроссинговер поддеревьев.

Так же предлагается выполнять оператор кроссинговера для худшего правила в дереве. Правило считается худшим, если целевая функция имеет минимальное значение. Каждое правило можно рассматривать как отдельное дерево способное решать поставленную задачу, поэтому вычисление фитнесс функции для каждого правила в отдельности логически обосновано.

Вычисление фитнесс функции не только для каждого правила в отдельности, но и каждого узла И также имеет смысл. При выполнении оператора кроссинговера поддеревьев предлагается осуществлять поиск точки разрыва следующим образом: вычисляется фитнесс функции для каждого узла И начиная с первого снизу. Если значение фитнесс функции для узла И находящегося выше, хуже чем на предыдущем шаге то обмену подлежит один из узлов аргументов данного узла И.

Мутация. Для деревьев используются следующие операторы мутации: узловая; усекающая; растущая. Узловая мутация выполняется для терминального узла или первой снизу вершины ИЛИ. Усекающая мутация выполняется только для узлов И или НЕ. При растущей мутация ветви наращиваются согласно правилам инициализации деревьев.

Предлагается выполнять оператор мутации для худшего правила в дереве. Так же предложено использовать изменение вероятности мутации в зависимости от целевой функции.

Редукция. Предлагается использовать выполнения следующих вариантов редукции: элитная стратегия; чистая замена; равномерная случайная замена (с указанием количества заменяемых особей в %).

Для реализации поставленной задачи написана программа в среде C++ Builder 6, которая выполняет рассмотренные выше действия.

Пример полученного продукционного правила:

ЕСЛИ

<<Рост матери < 160>> И (НЕ<<Курение матери - нет>>) И <<Срок ЖК ≥ 20>> И (НЕ <<ОТЕК = нет>>) И <<Эндометриоз>> И << Патология орг. зрения >> И <<Патология орг. пищеварения>>

ИЛИ

(НЕ <<Роды по счету ≥3>>) И (НЕ <<Чем закончилась предыдущая беременность - роды >>) И (НЕ <<Масса ребенка ≥ 4000>>)

ИЛИ

(НЕ <<кормили - грудь >>) И <<Апгар = [4; 8)>> И (НЕ <<Возраст отца < 20>>) И (НЕ << Патология орг. пищеварения >>) И <<Кол. кесарево ≥0>>

ИЛИ

(НЕ << Чем закончилась предыдущая беременность - роды >>) И << Перинатальная патология = да>>

ИЛИ

(НЕ <<Бытовые условия = квартира>>) И <<Алкоголь отца = средне>>

ИЛИ

<<Срок к груди ≥2 >> И (НЕ <<Номер беременности = 2>>)

На рисунке 2. Представлены результаты экспериментов: зависимость правильной классификации от мощности популяции.

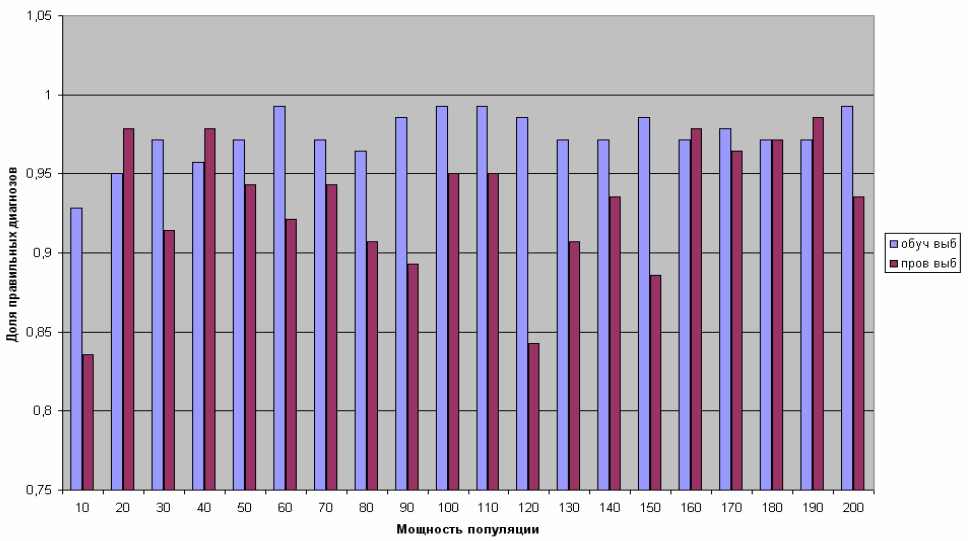

Рисунок 2. – Результаты экспериментов.

Разработанный аппарат ГП создан и протестирован на примере прогнозирования СВСГР, но может быть использован и при решении других задач диагностики и прогнозирования.

Литература

1. Goldberg D.E. Genetic Algorithms in Search, Optimization and Machine Learning.-Addison-Wesley, reading,MA.-1989.

2. Koza J.R. Genetic Programming.Cambridge:MA:MIT Press,1992.

3. 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.