УДК 004.048:004.622
Ю.А. Скобцов, доктор технических наук;
Т.А. Васяева, аспирант;
М.В. Лобачева, магистрант;
Донецкий национальный технический университет
г. Донецк, Украина
skobtsov@kita.dgtu.donetsk.ua
vasyaeva_tanya@tr.dn.ua
lobacheva_m@mail.ru

ФОРМИРОВАНИЕ ЗНАНИЙ ДЛЯ МЕДИЦИНСКИХ ЭКСПЕРТНЫХ СИСТЕМ НА ОСНОВЕ ГЕНЕТИЧЕСКОГО ПРОГРАММИРОВАНИЯ

Введение. Предвидение явлений в медицине является наиболее актуальной научно-практической задачей профессиональной деятельности врача. Совершенствование известных и создание новых методов диагностики и прогнозирования позволяет существенно улучшить качество оказания медицинской помощи населению. 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. Формирование базы знаний (БЗ) для экспертной системы (ЭС) очень важный этап, который в значительной степени определяет качество получаемой экспертной системы.

Постановка задачи. Общий подход большинства методов формирования БЗ состоит 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. Для решения поставленной задачи предложено использовать генетическое программирование (ГП) [1].

Описание метода. Для решения задачи с помощью ГП необходимо выполнить предварительные этапы:

1. Определить терминальное множество;
2. Определить функциональное множество;
3. Определить фитнесс-функцию;
4. 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.

Аппарат ГП позволяет построить эволюционным алгоритмом на основе обучающей выборки программу, которая выполняет прогнозирование заболевания. Рассмотрим реализацию метода на примере прогнозирования синдрома внезапной смерти грудных детей (СВСГД).

В данной задаче в качестве обучающего множества используются реальные данные обследования 240 пациентов, (120 детей, которые умерли за период 1990-1999г. 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.

Предобработка входных данных. Входное обучающее множество представим в виде булевых переменных. 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)
- возраст матери на момент начала месячных (полных лет) ≤ 12
- возраст матери на момент начала месячных (полных лет) ≤ 14
- возраст матери на момент начала месячных (полных лет) > 15
- длительность месячных (кол-во дней) ≤ 3
- длительность месячных (кол-во дней) ≤ 5.5
- длительность месячных (кол-во дней) > 6
- регулярность месячных (да – 1 , нет – 0)
- болезненность месячных (да – 1 , нет – 0)
- и др.

Наличие каждого фактора принято за единицу, отсутствие за ноль.

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

Функциональное множество. Функциональное множество состоит из логических операций: AND, OR, NOT. Так как первые две операции могут иметь два и более входов и один выход, а последняя операция всегда имеет один вход и один выход, то для более удобной программной реализации заменим операцию NOT на AND-NOT и OR-NOT. Такая замена выполнена с целью унифицировать количество входов для всех операций. Таким образом, функциональное множество состоит из 4 логических операций AND, OR, AND-NOT и OR-NOT.

Фитнесс-функция. 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.

Алгоритм. В контексте нашей задачи предлагается использовать ГП для получения дерева, которое будет распознавать высокую степень риска СВСГР. Обобщенный алгоритм ГП:

1. Установка параметров ГП.
2. Генерация начальной популяции. Популяция представляет собой набор хромосом. Каждая хромосома соответствует определенному дереву, представляющее собой решение. Дерево (хромосома), на начальном этапе генерируется случайным образом и состоит из функциональных и терминальных узлов (множество которых описано выше).
3. Оцениваются значения фитнесс-функции особей в популяции.
4. Применение генетических операторов.
5. Проверка критерия останова. При его выполнении переход на шаг 6, иначе шаг 3.
6. Выбор лучшего решения в последней популяции.

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

Установка параметров ГП. В таблице 1 приведены параметры, при использовании которых достигнут лучший результат.

Таблица 1.

Параметры ГП, при которых получен лучший результат

Параметр	Значения
Мощность популяции	200
Максимальная глубина особи	10
Метод генерации начальной популяции	– растущая p_g =50% – полная p_c =50%
Вероятность функционального узла	50%
Вероятность терминального узла	50%
Вероятность кроссинговера	99%
Вероятность мутации	5%
Отбор родителей	– рулетка

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 функционального узла и его аргументов. Для каждого дочернего узла (аргумента) случайным образом определяется будет данный узел функциональным или терминальным, далее в соответствии с типом узла выбирается его значение из соответствующего терминального или функционального множества. Процесс выполняется по левой ветви до тех пор, пока не будет выбран дочерним терминальный узел. Затем генерируются правые ветви.

Предусмотрены следующие методы создания деревьев:

- Полный. При генерации случайного дерева каждая ветвь имеет одинаковую (максимальную) глубину.
- Растущий. При генерации случайного дерева каждая ветвь может иметь различную глубину, но не более чем максимальная.
- Комбинированный. Половина деревьев всей популяции генерируется полным методом, вторая половина – растущим. Предложено использовать данный метод следующим образом: пусть мощность популяции (количество деревьев в популяции) = N; максимальная глубина особи = m; причем N>2*m; Популяция генерируется следующим образом: четвертая часть всей популяции имеет максимальную глубину – m; половина популяции - m/2; и оставшаяся часть – m/4.

Применение генетических операций. Используются последовательно генетические операторы [2] репродукции, кроссинговера, мутации и редукции.

Для древообразной формы представления используются три основных операторов кроссинговера: узловой кроссинговер, кроссинговер поддеревьев, смешанный. А также следующие операторы мутации: узловая, усекающая, растущая.

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. Фрагмент лучшего решения представлен на рисунке 1.

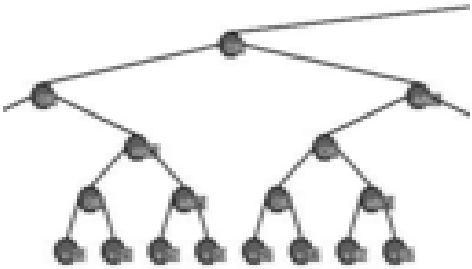

Рисунок 1. Фрагмент дарева

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.

Библиографический список

1. W. Banzhaf et all. Genetic Programming – an Introduction. – Morgan Kaufman, Heidelberg:San-Francisco, 1998.

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.