Васяева Т.А., Скобцов Ю.А.

Разработка экспертных систем медицинской диагностики с явным представлением продукционных правил на основе ГП с учетом неопределенности данных

Донецкий национальный технический университет, г. Донецк, vasyaeva_tanya@tr.dn.ua, skobtsov@kita.dgtu.donetsk.ua

Введение

Формирование базы знаний является одной из наиболее трудоемких задач при разработке экспертных систем (ЭС). Один из подходов формирования знаний заключается в разработке программ, способных обучаться под руководством эксперта-учителя. Данная работа является развитием [1], где для извлечения знаний в виде системы продукций используется аппарат генетического программирования. В отличие от предыдущих работ, где фактически используется двоичная логика, в настоящей работе применяется троичная логика, которая позволяет учитывать неопределенность (или отсутствие) значений некоторых параметров, как на этапе обучения, так и в процессе эксплуатации ЭС.

При работе с медицинскими данными, достаточно часто возникает ситуация, когда некоторые параметры неизвестны. Как правило, эти данные собираются по карточкам пациентов, которые находились на лечении несколько лет назад. Поэтому при отсутствии некоторой информации практически не возможно ее восстановить. Классические автоматизированные методы формирования знаний на базе машинного обучения (machine learning) работают, если известны все выделенные факторы риска для каждого пациента. Так как в большинстве случаев у разных пациентов отсутствуют данные о различных факторах риска, формирование обучающей выборки в этом случае выполняется с существенной потерей данных.

Целью проектируемой системы в данной работе является получение продукционных правил для диагностирования заболевания в условиях неопределенности некоторых входных данных (на примере определения высокой степени риска синдрома внезапной смерти грудных детей – (СВСГД) - одного из малоизученных и загадочных заболеваний).

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.

Генетическое программирование

Для решения поставленной задачи предложено использовать генетическое программирование (ГП) [2]

Предобработка входных данных

Входное обучающее множество должно быть представлено в виде булевых переменных. Для этого исходные данные были преобразованы 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)

Терминальное множество составляют перечисленные ранее 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). Значение фитнесс-функции для особей с правильным диагнозом принимает значение 1, а для особей с неправильным диагнозом принимает значение 0.

Предлагается следующий метод 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 3-мя правилами. Такое представление особи значительно упрощает интерпретацию результата. В данном примере расшифровка будет следующей:

ЕСЛИ правило 1 ИЛИ правило 2 ИЛИ правило 3, ТО результат 1, ИНАЧЕ результат 2.

С целью минимизации потери данных при обучении и расширения возможностей диагностирования при неизвестных значениях некоторых факторов риска предлагается использовать

троичную логику. При этом переменные могут принимать три логические значения {0,1,*}, где ‘*’ представляет неопределенное значение (это 0 или 1, но неизвестно, что именно). Подобный подход применяется во многих отраслях науки и техники, например при проектировании цифровых систем с использованием логического моделирования в троичной (или многозначной) логике [3].

В таблицах 1-3 приведены таблицы истинности для следующих логических функций: И, ИЛИ и НЕ.

Применение системы, которая оперирует с неизвестными состояниями, позволит выполнять диагностику даже при отсутствии некоторых параметров, что не приведет к невозможности функционирования разработанной системы. На этапе обучения, такой подход позволит сформировать оптимально полный набор входных параметров, и не упустить важные, информативные параметры.

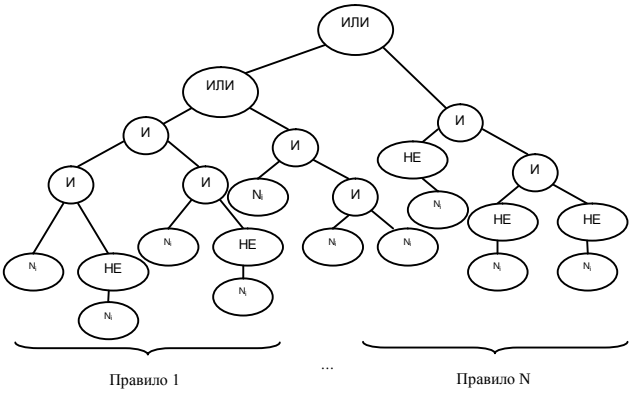

Рисунок 1.- Пример дерева в дизъюнктивной нормальной форме.

Таблица 1

N_1	N_2	И
0	0	0
0	1	0
1	0	0
1	1	1
*	0	0
*	1	*
*	*	*

Таблица 2

N_1	N_2	ИЛИ
0	0	0
0	1	1
1	0	1
1	1	1
*	0	*
*	1	1
*	*	*

Таблица 3

N_1	НЕ
0	1
1	0
*	*

Решение задачи на основе ГП можно представить 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.

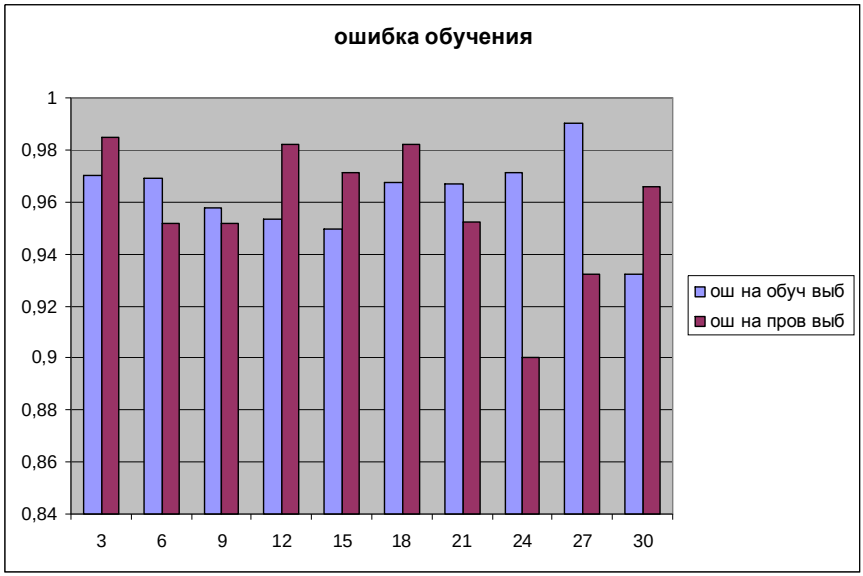

Рисунок 2.- Зависимость правильной классификации от количества неизвестных состояний на входах в %.

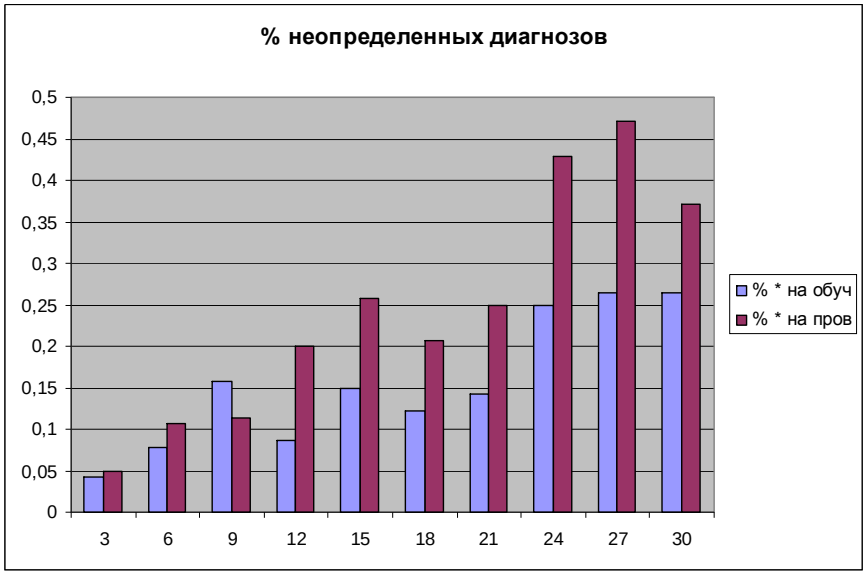

Рисунок 3.- Зависимость не распознанных диагнозов от количества неизвестных состояний на входах в %.

Выводы

Таким образом, получил дальнейшее развитие метод прогнозирования на основе генетического программирования, что позволило получить продукционные правила для прогнозирования высокой степени риска СВСГР в условиях неопределенности некоторых параметров. Предложенный метод протестирован на примере прогнозирования СВСГР, но может быть использован и при решении других задач медицинской диагностики и прогнозирования.

Литература

1. Васяева Т.А., Скобцов Ю.А. Разработка экспертных систем медицинской диагностики с явным представлением продукционных правил на основе ГП. – Тези міжнародної наукової конференції «Інтелектуальні системи прийняття рішень і проблеми обчислювального інтелекту (ISDMCI'2008)». Херсон: ХНТУ, 2008. Т3, Ч1.-192с.

2. Ю.А. Скобцов. Основы эволюционных вычислений.- Навчальний посібник. – Донецьк: ДонНТУ, 2008.- 326с.

Ю.А.Скобцов, В.Ю.Скобцов. Логическое моделирование и тестирование цифровых устройств.-Донецк:ИПММ НАНУ, ДонНТУ, 2005.-436с.