УДК 004.048:004.622

ПРОГНОЗИРОВАНИЕ С ПОМОЩЬЮ ГЕНЕТИЧЕСКОГО ПРОГРАММИРОВАНИЯ ПРИ НЕПОЛНЫХ МЕДИЦИНСКИХ ДАННЫХ

Аль-Гбури М.Х , Васяева Т.А.
Донецкий национальный технический университет

Разработан аппарат генетического программирования для прогнозирования на примере акушерских кровотечений. Предложен подход получения дерева для классификации патологической и допустимой потери крови в условиях неизвестных некоторых параметров.

Введение

При работе с медицинскими данными, достаточно часто возникает ситуация, когда некоторые параметры неизвестны. Это затрудняет, как и обучение системы, так и ее тестирование, а также использование. При формировании обучающих данных используются данные, предоставленные медицинскими работниками. 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. Поэтому, если какой-нибудь параметр неизвестен только у одного пациента необходимо, либо удалить пациента из обучающей выборки, либо удалить данный параметр из списка факторов риска. 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 прогнозирование при условии неопределенности некоторых входных данных (на примере определения патологической потери крови при родах).

Описание метода

Для решения поставленной задачи предложено использовать генетическое программирование (ГП) [1-2]. 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.

Предлагается следующий метод кодирования особей для генетического программирования. Каждая особь представляет собой дерево, которому соответствует булева функция. Такое представление удобно для решения задачи классификации (будем классифицировать на патологическое кровотечение и кровотечение в пределах нормы). Пример особи представлен на рис. 1.

Соответственно входные данные должны быть представлены в виде булевых переменных. Для этого выполняется преобразование следующим образом:

– место работы матери, профвредность (да – 0 , нет – 1);
– регулярность месячных (да – 1 , нет – 0);
– болезненность месячных (да – 1 , нет – 0) и т.д.

Терминальное множество состоит из факторов риска, которые после предобработки представляют собой булевы переменные и соответствуют листьям дерева. Функциональное множество состоит из логических операций: И, ИЛИ, И-НЕ, ИЛИ-НЕ, которые представляют внутренние вершины дерева.

Неопределенность данных

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 логике [3].

В таблицах 1–3 приведены таблицы истинности для следующих логических

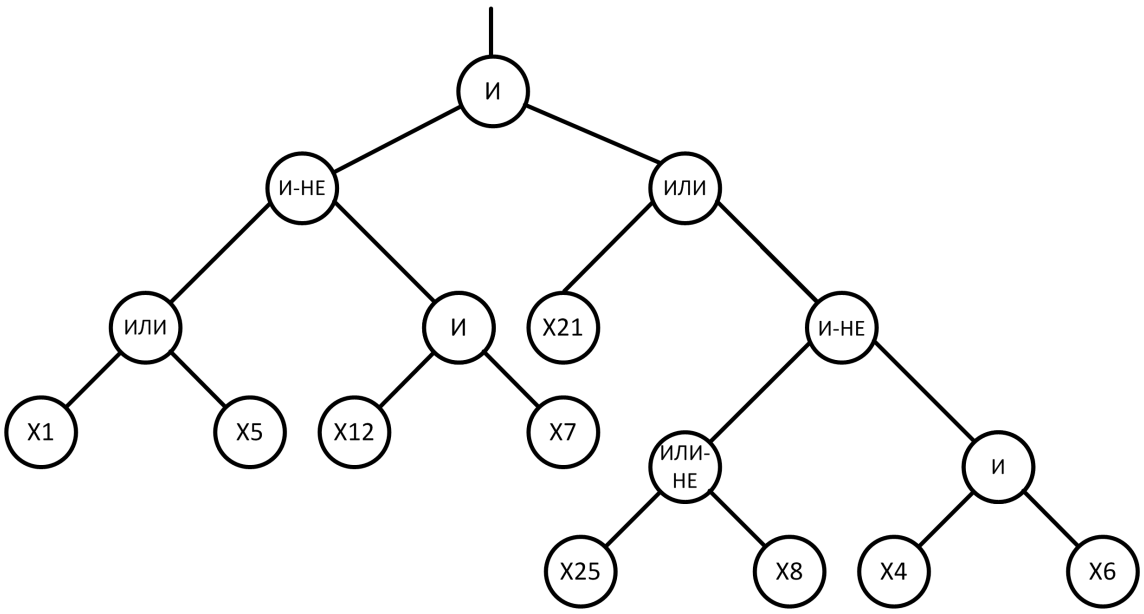

Рисунок 1. Пример структуры дерева для классификации акушерских кровотечений

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
*	*

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 параметры.

Фитнесс-функция

В качестве фитнесс-функции рассматривается доля пациентов с правильно поставленным диагнозом. Переменная диагноза принимает булевы значения 0 или 1. Единица соответствует патологической потери крови и ноль потери крови в пределах нормы. Значение фитнесс-функции для особей с правильным диагнозом принимает значение 1, а для особей с неправильным диагнозом принимает значение 0.

Выводы

Таким образом, получил дальнейшее развитие метод прогнозирования на основе генетического программирования, что позволило получить возможность прогнозировать в условиях неопределенности некоторых параметров. Предложенный подход предлагается для решения задачи определения патологической потери крови при родах, но может быть использован и при решении других задач прогнозирования и классификации.

Список источников

[1] Скобцов Ю.А. Основы эволюционных вычислений. – Навчальний посібник. – Донецьк: ДонНТУ, 2008. – 326 с.

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

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