

УДК 004.93

ИСПОЛЬЗОВАНИЕ ДЕРЕВЬЕВ РЕШЕНИЙ ДЛЯ ДИАГНОСТИРОВАНИЯ АВТОТРАНСПОРТНЫХ СРЕДСТВ

Гофман Е. А., Олейник А. А., Субботин С. А.

Запорожский национальный технический университет
кафедра программных средств
E-mail: olejnikaa@gmail.com

Аннотация

Гофман Е. А., Олейник А. А., Субботин С. А. Использование деревьев решений для диагностирования автотранспортных средств. Рассмотрена задача диагностирования автотранспортных средств с использованием деревьев решений. Предложен эволюционный метод синтеза деревьев решений. Разработано программное обеспечение идентификации деревьев решений на основе эволюционного подхода.

Введение

При изготовлении автотранспортных средств важным этапом является проведение комплекса работ по техническому диагностированию с целью исследования и установления состояния изготовленных автомобилей (или их частей) без разборки по выбранным признакам [1]. В настоящее время широкое распространение получили методы неразрушающего диагностирования, позволяющие определять места локализации неисправностей и прогнозировать состояние объекта без необходимости проведения исследований, требующих выведения объекта из работы либо его демонтажа. Такое диагностирование связано с необходимостью построения моделей, описывающих состояние объекта в зависимости от некоторых его характеристик (признаков) [1, 2].

Построение диагностических моделей выполняется с помощью методов и средств регрессионного анализа, теории искусственных нейронных сетей, нечеткой логики и др. Однако регрессионные модели и модели, построенные на основе аппарата нечеткой логики, не обладают высоким уровнем обобщения, а нейронные сети являются трудно интерпретируемыми специалистами-прикладниками, что усложняет их применение на практике.

Поэтому для синтеза диагностических моделей автотранспортных средств целесообразно использовать деревья решений [3–5], представляющие собой графовые интеллектуальные модели, во внутренних узлах которых расположены функции принятия решений на основе значений входных переменных, а во внешних узлах содержатся значения выходной переменной, соответствующие условиям внутренних узлов. Такие модели обладают высокими обобщающими способностями и хорошо интерпретируются людьми-специалистами в прикладных областях, которые, как правило, не знакомы с методами и моделями искусственного интеллекта.

Для синтеза деревьев решающих правил в настоящее время используют методы CART, C4.5, ID3 и др. [3]. Однако такие методы применяют жадную стратегию поиска и в случае неудачного выбора атрибута для разбиения не возвращаются на предыдущую итерацию, вследствие чего в результате часто синтезируются деревья решений, не обеспечивающие приемлемый уровень аппроксимации [3–5].

Для построения деревьев решений целесообразно использовать эволюционные методы [6–8], представляющие собой стохастические, эвристические методы многомерной безградиентной оптимизации.

Целью работы является создание эволюционного метода построения деревьев решений, а также решение задачи диагностирования автотранспортных средств на основе предложенного метода.

Общая постановка проблемы

Предприятие ЗАО "АЗ" производит автомобили марки Chevrolet. Актуальной задачей для данного предприятия является диагностирование кузовов Т-150 для автомобилей модели Lanos. Диагностирование проводится с целью выявления дефектных кузовов на ранней стадии, поскольку выявление дефектов на начальных этапах производства позволяет сократить затраты на их устранение. Процесс диагностирования кузова в соответствии с терминологией, принятой на ЗАО "АЗ", именуется "процедурой АВС". Буквами А, В и С обозначаются группы измеряемых параметров.

К группе измеряемых параметров А относятся контрольные точки, расположенные на "черном" кузове. Под "черным" понимается неокрашенный кузов, на который еще не устанавливались другие узлы (двигатель, трансмиссия, ходовая часть, двери, элементы салона и др.). На "черном" кузове располагается 1400 контрольных точек. Базовые точки кузова совмещаются с базовыми точками измерительного стенда. В процессе работы стенда изменяются отклонения от номинальных значений по трем координатам (XYZ) для каждой контрольной точки. По статистике, собранной на ЗАО "АЗ", в среднем на один кузов в область допуска попадает 85% контрольных точек группы А.

Группу параметров В составляют контрольные точки навесных узлов кузова: 4 двери и капот. Количество контрольных точек на каждой двери – 48, на капоте – 88. Поскольку навесные узлы являются достаточно простыми и преимущественно штампованными изделиями, количество точек, находящихся в допуске, составляет для 1 узла в среднем 95%, т. е. погрешности изготовления данных деталей практически не влияют на качество готового автомобиля.

К параметрам группы С относятся зазоры и сопряжения (выступание или западание одной поверхности относительно другой) между кузовом и навесными узлами в 46 контрольных точках. Для измерения параметров группы С на черный кузов устанавливаются навесные узлы, которые фиксируются магнитами, имитирующими замки, далее с помощью линейки и клиновидного измерителя зазоров измеряются значения этих параметров в контрольных точках. Количество точек группы С, находящихся в допуске, составляет в среднем 80% для 1 кузова. Это оказывает негативное влияние на качество кузова в целом, поскольку согласно плану для одного кузова не менее 88% точек суммарно по всем трем группам (АВС) должны быть в допуске.

Сложность заключается в том, что не всегда из годного (согласно замерам) "черного" кузова получается годный законченный автомобиль. Это связано с деформациями металла при повышении температуры в процессе окраски кузова, с изменением массы навесных узлов в процессе их окончательной сборки и т. д. Наиболее частыми дефектами кузова, ухудшающими внешний вид автомобиля, согласно экспертным оценкам являются: несоответствие зазоров и сопряжений между дверями и порогами, а также между бамперами и кузовом (параметры группы С). Устранение данных дефектов связано с частичной разборкой и, соответственно, повторной сборкой кузова. Также необходимо учитывать, что сборка кузова осуществляется с применением сварочного оборудования, что существенно осложняет и, соответственно, удорожает его разборку и сборку. Кроме того, проведение восстановительных работ может привести к необходимости повторной окраски автомобиля.

Как отмечено выше, характеристики группы В практически не влияют на качество готового автомобиля. Следовательно, основное влияние на состояние автомобиля (пригоден или непригоден для эксплуатации) оказывают параметры групп А и С.

Учитывая, что наиболее частыми являются дефекты параметров группы С, актуальным является построение моделей зависимости параметров группы С от параметров группы А. Прогнозирование с помощью синтезированных моделей по значениям параметров группы А позволит до этапа установления навесных узлов выявлять кузовы, значения параметров С которых будут выходить за пределы допуска. Это, в свою очередь, позволит сократить вероятность необходимости доработки таких кузовов и снизить расходы на производство автомобилей.

Выборка данных формировалась по результатам измерений 172 автомобилей. Из 1400 измеряемых точек группы параметров А экспертами (специалистами в автомобилестроении) выделены те, которые наибольшим образом влияют на качество кузова. Таким образом, имеем 48 входных признаков. Экспертами определено, что основными (наиболее серьезными) параметрами группы С являются: зазор и сопряжение между капотом и крылом, между стойкой и крылом, между крылом и дверью, между передней и задней дверьми, между дверьми и порогом, между задней дверью и боковиной, между боковиной и крышкой багажника. Таким образом, выявлено 16 наиболее критичных параметров, для каждого из которых необходимо построить модель зависимости от основных параметров группы А.

Обозначим полученную выборку данных в виде $\langle X, Y \rangle$, где X – набор значений признаков, характеризующих рассматриваемый объект или процесс; Y – массив значений выходного параметра в заданной выборке.

Необходимо аппроксимировать зависимость отклика Y от набора факторов X в виде модели, представленной деревом решений $T = \{ t_k \}$ таким образом, чтобы значение ошибки прогнозирования или классификации E построенной модели было минимальным, где $t_k = \langle c_k, l_k, r_k \rangle$ – k -ый узел дерева T , представляющий собой структуру, в которой c_k – функция принятия решений на основе значений входных переменных (в случае, если узел является внутренним) или значение выходной переменной (для внешних узлов), l_k и r_k – ссылки на левого и правого потомков k -го узла, соответственно, представляющих собой структуры, аналогичные t_k .

Задача построения дерева решений $T = \{ t_k \}$ по заданной выборке $\langle X, Y \rangle$ заключается в идентификации узлов $t_k = \langle c_k, l_k, r_k \rangle$ так, чтобы $E \rightarrow \min$.

Эволюционный метод синтеза деревьев решений

Как отмечено выше, известные методы построения деревьев решений используют жадную стратегию поиска, не позволяющую в процессе поиска заменять атрибуты, по которым уже выполнено разбиение, на другие, более эффективные. Это в некоторых случаях приводит к тому, что синтезированные деревья решений не обладают приемлемыми обобщающими способностями.

Поэтому для синтеза деревьев решений предлагается использовать эволюционный поиск, который на основе стохастического подхода позволяет перебирать различные комбинации, не закливаясь на исследовании субоптимальных областей.

В разработанном эволюционном методе построения деревьев решений предлагается информацию о деревьях решений хранить в виде хромосом, представленных с помощью древовидной структуры данных. Для кодирования хромосомы (представления дерева решения в виде хромосомы) целесообразно использовать структуру данных в виде бинарного дерева $T = \{ t_k \}$, в узлах $t_k = \langle c_k, l_k, r_k \rangle$ которого расположены функции c_k принятия решений на основе значений входных переменных (в случае, если узел является внутренним) или значение выходной переменной (для внешних узлов), а также ссылки l_k и r_k на левого и правого потомков, соответственно, представляющих собой структуры, аналогичные t_k .

Для оценивания хромосом используется функция $f(H_j) = \alpha_1 E(T_j) + \alpha_2 S(T_j) \rightarrow \min$, отражающая информацию как о качестве аппроксимации с помощью текущего дерева решений $E(T_j)$, так и о его сложности $S(T_j)$, что позволяет строить деревья решений с приемлемыми обобщающими и аппроксимационными свойствами, содержащие при этом незначительное количество узлов, что в свою очередь повышает интерпретируемость синтезированных моделей.

Для оценивания сложности дерева предлагается использовать следующую формулу: $S(T_j) = N(T_j) + r(T_j) + \eta(T_j)$, где $N(T_j)$ – количество узлов в дереве T_j , $r(T_j)$ – количество ребер в дереве T_j , $\eta(T_j)$ – глубина дерева T_j .

Для создания нового множества решений с помощью эволюционных операторов скрещивания и мутации предлагается использовать операторы, аналогичные операторам генетического программирования, поскольку такие операторы эффективно генерируют новые решения, представленные в виде древовидных структур.

Основной особенностью реализации генетических операторов скрещивания и мутации при построении деревьев решений в предложенном методе является поддержка синтаксической целостности получаемых решений (хромосом). Оператор скрещивания предлагается выполнять следующим образом. Вначале выбираются случайным образом на каждом из родительских деревьев одна или несколько точек (узлов дерева решений). Точки разрыва выбираются случайным образом для каждого родителя отдельно, что приводит к тому, вследствие выбора разных точек разрыва возможны ситуации порождения одинаковыми родителями разных потомков. После выбора точек скрещивания выполняется генерация потомков. Для этого необходимо обменять у родительских деревьев фрагменты согласно выбранным точкам разрыва.

Для изменения деревьев решений, выбранных для мутации, необходимо случайным образом выбрать мутирующий узел дерева, после чего выполнить мутацию одним из известных способов (мутация с увеличением, мутация со сжатием или мутация с заменой) [7].

Предложенный эволюционный метод синтеза деревьев решений основан на стохастическом подходе и не использует жадную стратегию поиска, позволяет строить деревья решений с приемлемыми обобщающими и аппроксимационными свойствами, содержащие при этом незначительное количество узлов, что в свою очередь повышает интерпретируемость синтезированных моделей.

Эксперименты и результаты

На основе предложенного эволюционного метода построения деревьев решений было создано программное обеспечение для идентификации деревьев решений по заданной выборке данных.

С помощью разработанного программного обеспечения решалась описанная выше задача построения моделей зависимости определяемых в процессе технического диагностирования параметров группы С от параметров группы А. В результате построено 16 моделей зависимостей параметров С от 48 входных признаков (параметров А).

Для сравнения проводился синтез деревьев решений с помощью методов CART, C4.5 и предложенного эволюционного метода. Усредненные результаты, полученные с помощью различных методов, приведены в табл. 1.

Таблица 1 – Анализ методов синтеза деревьев решений

№ п/п	Метод синтеза дерева решений	Характеристики синтезированного дерева			
		$E(T_j)$	$N(T_j)$	$r(T_j)$	$\eta(T_j)$
1	CART	0,031	36,2	34,8	6,5
2	C4.5	0,028	32,4	31,6	7,2
3	Эволюционный метод	0,021	17,8	16,2	5,6

Из табл. 1 видно, что предложенный эволюционный метод обеспечил построение дерева решений с лучшими аппроксимационными свойствами, при этом обладающего меньшим количеством узлов.

На рис. 1 приведено построенное дерево решений для выходного параметра u_4 .

Как видно, в процессе построения деревьев были выделены наиболее информативные признаки, влияющие на те или иные значения параметров типа С.

Ошибка классификации по построенным деревьям решений, вычисленная по данным тестовой выборки, находится в пределах 1,7–3,2% для различных u , что является допустимым для диагностирования автотранспортных средств.

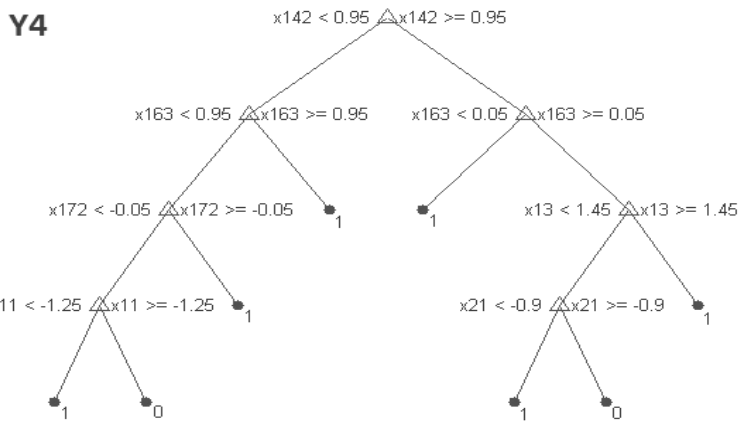


Рисунок 1 – Построенное дерево решений для u_4

Выводы

Решена задача разработки математического обеспечения для диагностирования автотранспортных средств.

Научная новизна работы заключается в том, что предложен эволюционный метод построения деревьев решений, основанный на стохастическом подходе, не использующий жадную стратегию поиска и позволяющий строить деревья решений с приемлемыми обобщающими и аппроксимационными свойствами, содержащие при этом незначительное количество узлов, что в свою очередь повышает интерпретируемость синтезированных моделей.

Практическая ценность результатов работы заключается в том, что: 1) разработано программное обеспечение для идентификации деревьев решений по заданной выборке данных; 2) решена задача технического диагностирования автотранспортных средств.

Литература

1. Биргер И. А. Техническая диагностика / И. А. Биргер. – М.: Машиностроение, 1978. – 240 с.
2. Прогрессивные технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиадвигателей : Монография / А. В. Богуслаев, Ал. А. Олейник, Ан. А. Олейник, Д. В. Павленко, С. А. Субботин. – Запорожье : ОАО «Мотор Сич», 2009. – 468 с.
3. Classification and regression trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. – California : Wadsworth & Brooks, 1984. – 368 p.
4. Rokach L. Data Mining with Decision Trees. Theory and Applications / L. Rokach, O. Maimon. – London : World Scientific Publishing Co, 2008. – 264 p.
5. Субботин С. А. Применение аппарата деревьев решений в задаче диагностирования автомобильных кузовов / С. А. Субботин, Е. А. Гофман // Системний аналіз. Інформатика. Управління : II Всеукраїнська науково-практична конференція САІУ-2011, 10–11 березня 2011 р. : матеріали конференції. – Запоріжжя, 2011. – С. 211–212.
6. Скобцов Ю. А. Основы эволюционных вычислений / Ю. А. Скобцов. – Донецк: ДонНТУ, 2008. – 330 с.
7. Субботін С. О. Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей: монографія / С. О. Субботін, А. О. Олійник, О. О. Олійник ; під заг. ред. С.О. Субботіна. – Запоріжжя: ЗНТУ, 2009. – 375 с.
8. Haupt R. Practical Genetic Algorithms / R. Haupt, S. Haupt. – New Jersey: John Wiley & Sons, 2004. – 261 p