УДК 681.518

С.В. МИРОНОВ, Д.В. СПЕРАНСКИЙ

Саратовский государственный университет им. Н.Г. Чернышевского, Россия

ДЕРЕВЬЯ РЕШЕНИЙ В ЗАДАЧАХ СОКРАЩЕНИЯ ДИАГНОСТИЧЕСКОЙ ИНФОРМАЦИИ*

Исследуется задача сокращения диагностической информации (ДИ), используемой при локализации неисправностей дискретных устройств. Для ее решения предложена модификация алгоритма Д. Квинлана, первоначально предназначенного для построения деревьев решений. Этот алгоритм выполняет поиск близкой к оптимальной маски, с помощью которой осуществляется сокращение ДИ. Приводятся статистические данные, подтверждающие эффективность предложенного алгоритма.

техническая диагностика, дискретные устройства, диагностическая информация, деревья решений

Введение

Один из подходов к диагностике дискретных устройств (ДУ) – диагностика с использованием предварительно подготовленной диагностической информации (ДИ). ДИ состоит из реакции исправного устройства на диагностический тест, а также набору реакций на тот же тест всех неисправных модификаций этого устройства. Сам же процесс диагностики заключается в сравнении реакции на тест исследуемого устройства с реакциями, составляющими диагностическую информацию.

Основной недостаток такого подхода заключается в том, что для современных дискретных устройств объем ДИ очень велик. Для сокращения объема ДИ были предложены различные методы [1 – 5]. Некоторые из них связаны с применением так называемых масок [2 – 5]. Маска определяет, какие позиции полной реакции на диагностический тест будут сохраняться в составе ДИ. Таким образом, вместо полных реакций на диагностический тест ДИ будет содержать некоторое подмножество этих реакций, полученное из полных реакций «наложением» маски. Маску называют единой (общей), если при составлении ДИ она применяется к реакциям на тест всех неисправных модификаций ДУ.

Идеальная общая маска максимально сокращает объем ДИ без потери полезной информации, т.е. без потери глубины диагностирования. Но задача нахождения такой маски относится к числу переборных, и, следовательно, при больших объемах полной ДИ требует неприемлемых временных затрат.

Возможным выходом из данной ситуации является применение различных эвристик, доставляющих хотя и не оптимальное, но достаточно близкое к идеалу решение. В настоящей работе предлагается адаптировать для нахождения общей маски модификацию алгоритма построения дерева решений, применяемого для решения задач классификации [6] и являющегося по сути «жадным» алгоритмом.

Основные определения и обозначения

Пусть рассматриваемое ДУ f_0 имеет m выходов и множество $F = \{f_i | 0 \le i \le N\}$ есть множество его неисправных модификаций. Предполагаем, что на каждом выходе упомянутых ДУ может появиться только сигнал 0 или 1.

Пусть $T = t_1, t_2, \ldots, t_k$ – диагностическая последовательность (тест) для рассматриваемого устройства, где t_j, $0 \le j \le k$ – входные вектора. Тогда полную реакцию устройства f_i на тест T можно представить в виде бинарного вектора длины mk в

* Исследования выполнены при поддержке РФФИ (гранты №05-08-18082, №05-08-49999)

котором на s-ом месте стоит значение сигнала на выходном полюсе с номером $s-\lfloor s/m \rfloor$ после подачи входного вектора с номером $\lceil s/m \rceil$.

Представим ДИ для ДУ f_0 с множеством неисправностей F в виде матрицы D порядка $(N+1)\times mk$, где строка с номером i, $0 \le i \le N$, представляет полную реакцию устройства f_i на диагностический тест T. Не теряя общности будем предполагать, что все строки матрицы D различны, т. е. нет ни одной пары устройств f_i и f_j, неразличимых с помощью представленной ДИ. В противном случае каждый класс неразличимых устройств можно заменить одним представителем.

Определим маску H диагностической информации как некоторое множество номеров столбцов матрицы D, т.е. $H \subseteq \{i \mid 0 \le i \le N\}$. Объемом маски H будем называть количество элементов в ней.

Обозначим через D_H матрицу, полученную из D удалением всех столбцов, кроме тех, чьи номера содержатся в маске H. Матрицу D_H назовем результатом применения маски H к ДИ D. Эта матрица и представляет сокращенную диагностическую информацию.

В большинстве случаев для произвольной маски H ее применение может привести к потере глубины диагностирования: могут появиться пары устройств, неразличимых с помощью сокращенной ДИ. Это будет выражено в появлении в матрице D_H одинаковых строк.

Таким образом, для максимального сокращения ДИ требуется найти такую маску H минимального объема, чтобы в матрице D_H все строки были различными. В нашем случае оптимальный объем такой маски может варьироваться от $\log_2(N+1)$ до N.

Деревья решений для задач классификации

В задачах классификации каждый объект (класс) из заданного множества характеризуется некоторой совокупностью признаков (атрибутов), имеющих конкретные значения. Поскольку упомянутая совокупность признаков может иметь большую мощность, то для идентификации предъявленного объекта, характеризуемого конкретным набором значений признаков, потребуется хранить большой объем информации. Одним из наиболее популярных подходов к решению проблемы сокращения исходной информации является использование механизма деревьев решений.

Деревья решений представляют собой иерархическую структуру классифицирующих правил типа «если ... то …», имеющую вид дерева. Для того чтобы решить, к какому классу отнести некоторый объект, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Формулировка такого вопроса обычно заключается в проверке значения некоторого атрибута предъявленного объекта. Дуги, исходящие из узла дерева к его потомкам, помечаются вариантами ответов на вопрос, заданный в данном узле. Каждый лист такого дерева соответствует классу со значениями атрибутов, приписанными дугам пути от корня до этого листа.

Таким образом, имея построенное дерево решений, отпадает необходимость в хранении всей совокупности признаков: можно хранить информацию только о тех атрибутах, которые упоминаются в узлах дерева решений.

Для построении деревьев решений часто используется алгоритм C4.5, предложенный Р. Квинланом (R. Quinlan) [6] в 1993 г. Напомним основные моменты этого алгоритма.

Пусть задано множество объектов обучающей выборки P, где каждый объект описывается множеством атрибутов X. Пусть объекты обучающей выборки являются элементами классов из множества C.

Построение дерева происходит сверху вниз. Сначала создается корень дерева, затем потомки корня и т.д. Каждому узлу дерева ставится в соответствие некоторое подмножество множества P.

Корню дерева ставится в соответствие само множество P.

Процесс построения дерева заключается в следующем. Первоначально мы имеем дерево, состоящее только из корня. Следующие шаги необходимо выполнить для каждого листа уже построенного дерева, которому соответствует множество объектов более чем одного класса.

Пусть данному узлу дерева соответствует множество объектов $P' \subseteq P$ классов из $C' = \{c_{i_1}, \ldots, c_{i_n}\}$, а $freq(c_j, P')$ – количество элементов из множества P', принадлежащих классу c_j. Тогда величина

$$I(P') = -\sum_{j=1}^{n} \frac{freq(c_{i_j}, P')}{|P'|} \cdot \log_2\left(\frac{freq(c_{i_j}, P')}{|P'|}\right), \quad (1)$$

где $|P'|$ – количество элементов в множестве P', дает оценку среднего количества информации, необходимой для идентификации объекта из множества P'.

Если произвести разбиение множества P' в зависимости от значений атрибута $x \in X$ на z_x подмножеств, то ту же оценку, но уже после разбиения дает выражение

$$I_x(P') = \sum_{j=1}^{z_x} \frac{|P'_j|}{|P'|} \cdot I(P'_j). \quad (2)$$

В принятых обозначениях алгоритм C4.5 состоит из следующих шагов:

1. Необходимо выбрать такой атрибут $x \in X$, который доставляет максимум величины

$$G(x, P') = I(P') - I_x(P'). \quad (3)$$

2. Для данного узла дерева создать z_x потомков.

3. Каждому потомку данного узла поставить в соответствие определенное подмножество объектов из множества P', имеющих одно и то же значение атрибута x, а дугу от узла к потомку пометить соответствующим значением.

Процесс построения дерева заканчивается, когда всем листьям дерева будут соответствовать множества объектов одного класса.

Деревья решений для задач сокращения диагностической информации

Проводя аналогию между задачей классификации объектов по совокупности значений их признаков и задачей сокращения ДИ, можно сказать, что совокупность признаков объектов в 1-й задаче идентична полной реакции на диагностический тест для конкретной неисправности во второй, а полная информация об объектах 1-й задачи идентична полной ДИ во 2-й задаче.

Таким образом, для задачи сокращения ДИ признаки, проверяемые в узлах дерева решений, могут быть взяты в качестве элементов маски ДИ.

С учетом обозначений, введенных ранее, множество F выступает в роли множества P для алгоритма C4.5. Это же самое множество F представляет набор классов C, т.е. в нашем случае каждый класс представлен лишь одним объектом в обучающей выборке. Матрица D представляет исходные данные для построения дерева решений: каждый столбец этой матрицы выполняет роль атрибута объектов, а каждая строка этой матрицы описывает значения атрибутов.

Поскольку каждый класс представлен лишь одним объектом в обучающей выборке, то формула (1) примет следующий вид:

$$I(P') = \log_2 |P'|. \quad (4)$$

В связи с тем, что возможных значений атрибутов всего два (0 или 1), получаемое дерево решений будет представлять бинарное дерево, где в каждом конкретном узле множество P' разделяется на два подмножества P'_0 и P'_1 в соответствии со значениями атрибута x доставляющего максимум величине

$$G(x, P') = \log_2 |P'| - \frac{|P'_0|}{|P'|} \cdot \log_2 |P'_0| - \frac{|P'_1|}{|P'|} \cdot \log_2 |P'_1|. \quad (5)$$

Формула (5) получена из (3) с учетом (4). Оче-

видно, что $G(x, P')$ в (5) будет иметь максимум в том случае, если множество P' делится по значениям атрибута x на два подмножества с равным числом элементов.

Так как количество листьев в результирующем дереве равно $(N+1)$, а само дерево является бинарным, можно сделать вывод, что количество внутренних узлов дерева, т.е. узлов, в которых производится проверка значений атрибутов, не превышает N. При наилучшем исходе алгоритм C4.5 строит идеально сбалансированное дерево, в котором на любом пути от корня до листа проверяются значения одного и того же набора атрибутов, в котором $\lceil \log_2(N+1) \rceil$ элементов. Эти крайние значения как раз и соответствуют объему оптимальной маски. Таким образом, с помощью алгоритма C4.5 можно получить маску, по объему близкую к оптимальной.

В табл. 1 приведены результаты работы алгоритма получения маски ДИ с помощью алгоритма C4.5 для диагностической информации ДУ из набора схем каталога ISCAS'89.

Заметим, что данные, приведенные в табл. 1 – 3, получены для вероятностных тестов, содержащих 100 входных наборов, и число обнаруженных такими тестами неисправностей представлено во втором слева столбце этих таблиц. Как показывает опыт, значительное увеличение длины вероятностного диагностического теста приводит к довольно малому росту числа обнаруженных с его помощью неисправностей. По этой причине увеличение длины теста, следствием которого является возрастание объема ДИ, по-видимому, может только улучшить показатели в последнем столбце таблиц, но не ухудшить их.

В алгоритме C4.5 выбор атрибута, доставляющего максимальное значение величины (5), происходит независимо в каждом узле дерева. Возможны ситуации, когда сразу несколько атрибутов доставляют максимум этой величине. В такой ситуации данный алгоритм выбирает тот из них, который доставил этот максимум первым.

Предлагаемая нами модификация алгоритма C4.5 заключается в том, что на шаге 1 алгоритма при равенстве величины (5) для различных атрибутов предпочтение отдается тому атрибуту, деление по которому уже производилось в каком-то другом (уже рассмотренном ранее) узле дерева. За счет этого множество проверяемых в дереве атрибутов не расширяется, а, следовательно, не увеличивается и объем маски, получаемой в результате. Табл. 2 показывает результаты работы этой модификации на тех же исходных данных, что и в предыдущей таблице.

Таблица 1

Результат применения алгоритма C4.5 для поиска маски ДИ

Наименование ДУ	Число неисправностей в ДУ	Длина полной реакции на диагностический тест, (в битах)	Объем ДИ, (в битах)	Объем маски	Размер сокращенной ДИ по отношению к полной ДИ
S298	92	600	55200	61	10,17%
S344	184	1100	202400	111	10,09%
S349	184	1100	202400	114	10,36%
S386	136	700	95200	86	12,29%
S510	447	700	312900	205	29,29%
S641	253	2400	607200	169	7,04%
S713	252	2300	579600	171	7,43%
S820	159	1900	302100	107	5,63%
S832	159	1900	302100	108	5,68%
S8381	86	100	8600	30	30,00%
S953	607	2300	1396100	287	12,48%
S1423	121	500	60500	62	12,40%
S1488	325	1900	617500	166	8,74%
S1494	322	1900	611800	165	8,68%

Таблица 2

Результат применения МОДИФИКАЦИИ алгоритма C4.5 для поиска маски ДИ

Наименование ДУ	Число неисправностей в ДУ	Длина полной реакции на диагностический тест, (в битах)	Объем ДИ, (в битах)	Объем маски	Размер сокращенной ДИ по отношению к полной ДИ
S298	92	600	55200	48	8,00%
S344	184	1100	202400	79	7,18%
S349	184	1100	202400	73	6,64%
S386	136	700	95200	76	10,86%
S510	447	700	312900	100	14,29%
S641	253	2400	607200	145	6,04%
S713	252	2300	579600	141	6,13%
S820	159	1900	302100	89	4,68%
S832	159	1900	302100	88	4,63%
S8381	86	100	8600	25	25,00%
S953	607	2300	1396100	216	9,39%
S1423	121	500	60500	59	11,80%
S1488	325	1900	617500	133	7,00%
S1494	322	1900	611800	132	6,95%

Как видно из сравнения результатов, модификация алгоритма C4.5 привела к сокращению объема маски до 30% по отношению к результату исходного алгоритма.

Как алгоритм C4.5, так и его приведенная модификация ориентированы на построение идеально-сбалансированного дерева, т.е., на сокращение среднего числа атрибутов, необходимого для идентификации предъявленного объекта. Таким образом, для каждого класса выбирается индивидуальная совокупность атрибутов, необходимая для идентификации его объектов. Но для большего сокращения информации целесообразно найти некое единое множество атрибутов, по которому можно было бы идентифицировать любой объект. Естественно, объединение упомянутых индивидуальных совокупностей для всех классов может иметь существенно большую мощность, чем такое единое множество.

Авторами доклада разработан новый алгоритм, базирующийся на идеях алгоритма C4.5, но направленный на нахождение упомянутого единого множества атрибутов.

Отличие предложенного алгоритма от алгоритма C4.5 состоит в том, что построение дерева производится по уровням, с выбором одного атрибута для проверки во всех узлах уровня. На начальном уровне присутствует только один узел – корень дерева. На каждом последующем шаге рассматривается очередной уровень дерева и генерируется уровень его потомков. Для получения нового уровня выбирается атрибут x, который доставляет максимум величине

$$G'\left(x,\widetilde{P}\right)=I'\left(\widetilde{P}\right)-I'_x\left(\widetilde{P}\right), \qquad (6)$$

где $\widetilde{P}$ – совокупность множеств объектов, соответствующих узлам текущего уровня, а величины $I'\left(\widetilde{P}\right)$ и $I'_x\left(\widetilde{P}\right)$ равны соответственно

$$I'\left(\widetilde{P}\right)=\sum_{P'\in\widetilde{P}}\frac{|P'|}{|P|}I\left(P'\right) \qquad (7)$$

и

$$I'_x\left(\widetilde{P}\right)=\sum_{P'\in\widetilde{P}}\frac{|P'|}{|P|}I_x\left(P'\right). \qquad (8)$$

Так же как и в алгоритме C4.5, процесс построения дерева заканчивается, если на очередном уровне каждому узлу дерева будет соответствовать множество объектов одного класса.

Содержательный смысл такого построения состоит в следующем: на каждом уровне дерева выбирается атрибут, проверка которого максимально уменьшает среднее количество информации, необходимой для идентификации любого объекта из множества P.

В табл. 3 приведены результаты работы предложенного алгоритма. Как видно из сравнения данных

Таблица 3

Результат применения алгоритма для поиска единой маски ДИ

Наименование ДУ	Число неисправностей в ДУ	Длина полной реакции на диагностический тест, (в битах)	Объем ДИ, (в битах)	Объем маски	Размер сокращенной ДИ по отношению к полной ДИ
S298	92	600	55200	36	6,00%
S344	184	1100	202400	50	4,55%
S349	184	1100	202400	48	4,36%
S386	136	700	95200	55	7,86%
S510	447	700	312900	71	10,14%
S641	253	2400	607200	107	4,46%
S713	252	2300	579600	105	4,57%
S820	159	1900	302100	66	3,47%
S832	159	1900	302100	66	3,47%
S8381	86	100	8600	18	18,00%
S953	607	2300	1396100	151	6,57%
S1423	121	500	60500	46	9,20%
S1488	325	1900	617500	111	5,84%
S1494	322	1900	611800	110	5,79%

этой таблицы с данными из табл. 1, улучшение объема маски предложенным алгоритмом составляет до 50% по отношению к объему маски, полученной алгоритмом C4.5.

Заключение

Статистические данные, полученные в результате проведенных численных экспериментов с полутора десятками устройств из международного каталога ISCAS'89, подтверждают достаточно высокую эффективность алгоритмов, предложенных в работе.

Так, эти алгоритмы осуществляют поиск решения за приемлемое время: в худшем случае оно не превышало 5 мин. при их реализации на LISP в среде LispWorks PE 4.4 на PC Intel Celeron 1.70GHz, 256MB RAM.

Получаемое решение дает возможность сократить ДИ до объема в диапазоне $4 \div 18\%$ от первоначального, что в практическом плане является хорошим результатом.

Литература

1. Ryan P.G., Fuchs W.K., Pomeranz I. Fault dictionary compression and equivalence class computation for sequential circuits // Proc. IEEE International Conference on Computer-Aided Design (ICCAD'93). – Santa Clara, CA, USA: 1993. – P. 508-511.

2. Барашко А.С., Скобцов Ю.А., Сперанский Д.В. Моделирование и тестирование дискретных устройств. – К.: Наук. думка, 1992. – 320 с.

3. Вознесенский С.С., Раздобреев А.Х. Трудоемкость поиска неисправностей как критерий качества при сокращении объема диагностической информации // Электронное моделирование. – 1980. – № 4. – С. 83-86.

4. Чипулис В.П. Методы минимизации разрешающей способности и диагностической информации // Автоматика и телемеханика. – 1975. – № 3. – С. 133-141.

5. Шаршунов С.Г. Особенности диагноза технического состояния многовыходных объектов с использованием таблиц неисправностей // Автоматика и телемеханика. – 1973. – № 12. – С. 161-168.

6. Quinlan J.R. C4.5: programs for machine learning. – San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

Поступила в редакцию 7.02.2006

Рецензент: д-р техн. наук, проф. А.Ю. Соколов, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков.