УДК 004.93
А. А. Олейник, Е. А. Гофман, С.А Субботин
Запорожский национальный технический университет

ИДЕНТИФИКАЦИЯ ДЕРЕВЬЕВ РЕШЕНИЙ С ИСПОЛЬЗОВАНИЕМ ТЕОРИИ ФУНКЦИЙ ДОВЕРИЯ

Запропоновано новий метод ідентифікації дерев рішень, що використовує теорію функцій довіри для роботи в умовах невизначеності параметрів при розв'язанні завдання автоматичної класифікації за ознаками. Розроблений метод дозволяє ідентифікувати структуру та параметри дерев рішень.

Дерево рішень, ідентифікація, класифікація, невизначеність, функція довіри.

1. Введение. Постановка задачи исследования

Деревья решений – один из наиболее широко используемых методов классификации в искусственном интеллекте [1]. Их популярность объясняется в основном способностью представлять знания в формальном виде, который легче интерпретируется как экспертами, так и обычными пользователями. Несмотря на свои преимущества в условиях точных и однозначных данных, классические методы построения деревьев решений [2] не в состоянии обработать неопределённые данные при решении задач классификации. Результаты работы методов однозначны и не обрабатывают неопределенность, которая может наблюдаться в значениях признаков или в экземплярах [3].

Для преодоления этих ограничений разработаны вероятностные деревья решений [4], главная цель которых состоит в том, чтобы работать с экземплярами, которые характеризуется отсутствием или неоднозначностью значений признаков. Однако в разработанных методах учитывается только статистическая неопределенность, вызванная информацией, полученной в результате наблюдений.

Таким образом, актуальной является разработка методов, позволяющих решать задачу классификации в условиях неопределённости и неполноты данных. Целью данной статьи является разработка метода классификации с помощью деревьев решений, позволяющего преодолевать проблему неопределенности и неполноты данных.

Предлагаемый метод использует теорию функции доверия, представленной в рамках передаваемой доверительной модели (ПДМ) [5, 6]. Такая модель позволяет создавать подходящую систему классификации благодаря способности представления неоднозначности. Кроме того, ПДМ позволяет экспертам выражать частичные представления более гибким способом, чем это можно делать при помощи функций вероятности. Такой подход также позволяет обрабатывать частичное или даже полное незнание о параметрах классификации.

2. Функции доверия

Пусть q – фрейм распознавания, представляющий конечное множество элементарных гипотез, связанных с проблемной областью. Множество всех подмножеств q обозначим как 2^q. Чтобы представить степени доверия, Shafer [7] ввёл так называемые основные распределения доверия (basic belief assignments). Эти распределения определяют часть доверия, которая покрывает подмножество гипотез, не покрывая однозначного подмножества общего множества при нехватке соответствующей информации [5]. Основное распределение доверия (ОРД) – это функция m, которая принимает значение в пределах [0, 1] для каждого подмножества A из q:

$$m : 2^q \longrightarrow [0,1],$$

при этом:

$$m(\varnothing) = 0 \text{ и } \sum_{A \subseteq q} m(A) = 1.$$

Подмножества A фрейма распознавания Θ, для которых $m(A)$ – однозначно положительны, называются фокальными элементами ОРД.

Вероятность Bel и достоверность Pl рассчитываются следующим образом:

$$Bel(A) = \sum_{B \subseteq A} m(B),$$

$$Pl(A) = \sum_{A \cap B \neq \varnothing} m(B).$$

Величина $Bel(A)$ выражает полное доверие, которое полностью передается подмножеству A из Θ. $Pl(A)$ представляет собой максимальное доверие, которое может покрывать подмножество A.

В рамках теории функции доверия легко выразить состояние полного незнания. Это достигается за счёт так называемой пустой доверительной функции, в которой единственным фокальным элементом является сам фрейм распознавания Θ [7]:

$$m(\Theta) = 1 \text{ и } m(A) = 0,$$

при условии:

$$A \neq \Theta.$$

Одним из важных понятий теории функции доверия – совмещение. Пусть Bel_1 и Bel_2 – две функции доверия, покрывающие две различные части знания. Пусть m_1 и m_2, определяют их ОРД, соответственно.

Закон совмещения Дэмпстера направлен на создание ОРД, которое представляет воздействие объединенного доказательства. Это определено как [7]:

$$\forall A \subseteq \Theta, m(A) = (m_1 \oplus m_2)(A) = K. \sum_{B,C \subseteq q : B \cap C = A} m_1(B).m_2(C),$$

где K – коэффициент нормализации [10], $K^{-1} = 1 - \sum_{B \cap C = \varnothing} m_1(B).m_2$.

Правило совмещения Демпстера – коньюктивное правило. Оно создает ОРД, если обе части правила приняты. Двойственность этого коньюктивного правила определяется дизъюнктивным правилом совмещения [8], которое создает ОРД, представляя взаимодействие двух частей доказательства, когда известно только то, что по крайней мере одна ОРД должна быть принята, но не известно, какая именно:

$$\forall A \subseteq \Theta, m_1 \vee m_2(A) = \sum_{B,C \subseteq \Theta : B \cup C = A} m_1(B).m_2(C).$$

Эти правила совмещения являются коммутативными и ассоциативными. Таким образом, основное распределение доверия, определяющееся комбинацией нескольких частей данных, может быть легко определено путём многократного применения правила в любом порядке их применения. Коньюктивные и дизъюнктивые правила совмещения обобщают операции «ИЛИ» и «И» теории множеств.

Проблема принятия решений в контексте ПДМ была решена в [6].

ПДМ основана на двух уровнях интеллектуальных модулей:

– кредальный уровень, на котором убеждения и выводы описываются функциями доверия;

– пигнистический уровень (pignistic level), на котором убеждения используются для принятия решений и представлены функциями вероятности, названными пигнистическими вероятностями.

Связь между этими двумя функциями достигается путём пигнистического преобразования, которое создает пигнистическую функцию вероятности $BetP$, основанной на функции доверия:

$$BetP(B) = \sum_{A \subseteq q} m(A) \frac{|B \cap A|}{|A|} \text{ для всех } B \subset \Theta.$$

3. Построение деревьев решений с использованием теории функций доверия

Разрабатываемый метод идентификации деревьев решений, основанный на применении математического аппарата теории функций доверия, характеризуется особенностями как на этапе непосредственной идентификации дерева решений, так и на этапе классификации экземпляров с использованием полученного дерева решений.

Вначале на этапе идентификации дерева решений требуется определить основные параметры дерева решений в рамках теории функции доверия, затем разрабатывается метод для построения таких деревьев решений.

Предлагаемый метод для идентификации деревьев решений использует теорию функций доверия и основан на расширенном алгоритме ID3 [2, 3], однако учитывает при этом неопределённость некоторых параметров, связанных с задачей классификации. Таким образом, выделяются некоторые различия при определении гипотез относительно этих параметров при их использовании.

Обучающая выборка в рамках этой неопределенной структуры представляет собой множество, составленное из элементов, представленных по парам (признаки, класс), где для каждого экземпляра, как правило, известны значения каждого из его признаков и уникальный класс, к которому принадлежит данный экземпляр. Такая обучающая выборка (в отличие от тех, что традиционно используются для построения моделей сложных объектов и процессов) может содержать данные, в которых присутствует некоторая неопределённость относительно классов. Другими словами каждый класс обучающей выборки может быть неопределенным или даже неизвестным, тогда как значения признаков, характеризующих каждый экземпляр, однозначно известны.

Предлагается представить неопределённость класса любого экземпляра путем основного распределения доверия, заданном на множестве классов. Это ОРД, традиционно заданное экспертом, представляет собой мнение-убеждение этого эксперта о фактическом значении класса для каждого экземпляра в обучающей выборке.

Среди преимуществ работы с функциями доверия необходимо отметить, что легко можно выразить две граничные ситуации (полное незнание и полное знание):

– если неизвестна никакая информация о классе экземпляра, ОРД будет представлять собой пустую функцию доверия:

$$m(\Theta) = 1 \text{ и } m(C) = 0 \text{ для } C \subset \Theta;$$

– если класс экземпляра известен однозначно, он будет представлен функцией доверия:

$$m(C_i) = 1 \text{ и } m(C) = 0 \text{ для всех } C \neq C_i, C \subseteq \Theta,$$

где C_i – единичный класс.

После того, как обучающая выборка описана, следует задать второй важный параметр разрабатываемого метода – меру выбора признака, которая будет использоваться для выбора тестового признака в каждом узле решения дерева.

Эта мера позволяет определить количество силы дифференциации каждого признака относительно каждого класса. За счёт такого подхода достигается оптимизация дерева. Часто в качестве такой меры используется информационный критерий Квинлана [1, 9].

В рамках разрабатываемого метода мера выбора признака расширяется, и она позволяет работать с неопределённостью, используя теорию функции доверия.

Пусть Bel_j – функция доверия, заданная на множестве возможных классов, и описывающая убеждения экспертов о фактическом значении класса, к которому принадлежит объект I_j. Пусть S – подмножество экземпляров обучающей выборки, из которого случайным образом выбирается один экземпляр. Функция доверия, которая описывает убеждения о конкретном классе, к которому принадлежит этот случайно отобранный экземпляр, является средней функцией доверия, принимающей экземпляр в S. Тогда:

$$Bel_s(C) = \frac{\sum_{I_j inS} Bel_j(C)}{|S|},$$

для всех C подмножеств из $\Theta = \{C_1, ...C_n\}$.

Следует отметить, что ОРД и пигностические вероятности, связанные с этой средней функцией доверия, пропорциональны основному распределению доверия и пигностическим вероятностям объектов из S (для любых подмножеств С из Θ):

$$m_s(C) = \frac{\sum_{I_j inS} m_j(C)}{|S|},$$

$$BetP_s(C) = \frac{\sum_{I_j inS} BetP_j(C)}{|S|}.$$

Предлагается выполнять следующие этапы для определения информационной значимости признаков.

1. Рассчитать функцию средней пигностической вероятности $BetP_T$, исходя из обучающей выборки T. Затем рассчитать энтропию распределения классов в T:

$$Info(T) = -\sum_{i=1}^{n} BetP_T(C_i) \log_2 BeP_T(C_i).$$

Основываясь на полученных данных, рассчитать прирост информации, обеспеченный каждым признаком A:

$$Gain(T, A) = Info(T) - Info_A(T).$$

2. Для достижения выполнения предыдущего этапа необходимо рассчитать $Info_A(T)$ для каждого признака. Предлагается применить тот же подход, что используется для расчета $Info(T)$, но ограничиваясь множеством экземпляров, которые характеризуются одинаковым значением атрибута A, и усредняя эти условные информационные меры.

Таким образом, для каждого значения признака a_m выделяется подмножество T_m полученное из экземпляров T, для которых значение соответствующего признака равно a_m . Далее рассчитывается средняя функция доверия Bel, после чего применяется пигностическое преобразование для расчёта пигностической вероятности $BetP$. После полученных преобразований можно рассчитать $Info(T_m)$, при этом T_m представляет собой обучающую выборку, в которой значение признака A равно a_m .

3. Искомое $Info_A(T)$ будет равно взвешенной сумме $Info(T_m)$ относительно рассматриваемого признака. При этом предлагается, чтобы $Info(T_m)$ были нагружены пропорционально значениям каждого признака в обучающей выборке:

$$Info_A(T) = \sum_{m=1}^{k} \frac{|T_m|}{|T|} Info(T_m) = -\sum_{m=1}^{k} \frac{|T_m|}{|T|} \sum_{i=1}^{n} BetP_{Tm}(Ci).\log_2 BetP_{Tm}(Ci).$$

4. Как только вычислены информационные значимости признаков, выбирается признак с наибольшей информационной значимостью.

Кроме выборочной меры признака, также должны быть определены два других важных параметра работы метода:

– стратегия разделения: необходимо создать ветви для каждого значения признака.

– критерий останова: позволяет прекращать расширение дерева и определять узел как лист. Таким образом, определяется, необходимо ли разделять обучающее подмножество дальше. В контексте разрабатываемого метода предложены следующие варианты критерия останова.

1. Если созданный узел покрывает только один экземпляр, то данный узел объявляется как лист, который характеризуется тем же ОРД, который уже определен в обучающей выборке.

2. Если уже нет дальнейшего признака для анализа или если критерий информационной значимости для оставшихся признаков меньше нуля, тогда узел объявляется листом, где его ОРД будет результатом конъюнктивного совмещения ОРД экземпляров, относящихся к данному листу, рассчитанного по закону Демпстера.

В отличие от традиционного дерева решений, в котором каждый лист определяет уникальный класс, предлагаемый метод определяет каждому листу ОРД, выражая, таким образом, множество убеждений о различных классах структуры распознавания.

Пусть T – обучающая выборка, состоящая из экземпляров, характеризуемых признаками $(A_1, A_2, ..., A_m)$, и которые могут принадлежать множеству классов $\Theta = \{C_1, C_2, ..., C_n\}$. Каждому объекту $I_j (j = 1, ..., p)$ из обучающей выборки будет соответствовать основное распределение доверия, выражающее количество убеждений, относящихся к подмножеству классов.

Таким образом, разработанный метод содержит этапы, описанные ниже.

1. Генерация корневого узла дерева решений, включающего все объекты обучающей выборки.

2. Проверка удовлетворения текущего узла критерию останова:

– если проверка дала положительный результат, то объявить узел листом и рассчитать его ОРД так, как было упомянуто ранее;

– в противном случае – найти признак с наибольшей информационной значимостью. Этот признак будет рассматриваться как корень дерева решений, связанный со всей обучающей выборкой.

3. Применение стратегии разделения с целью создания ветви для каждого значения признака, выбранного как корень. Это разделение ведет к нескольким обучающим подмножествам.

Этапы 2 и 3 повторяются для каждого обучающего подмножества до тех пор, пока узел не будет объявлен как лист.

4. Останов построения дерева решений происходит, когда все узлы последнего уровня дерева – листья.

Важно отметить, что предложенный метод получает такие же результаты как метод ID3, если все ОРД однозначно определены. Такая ситуация возникает, если класс каждого экземпляра из обучающей выборки уникален и однозначно известен.

После того, как была произведена идентификация дерева решений, можно выполнять классификацию экземпляров, не относящихся к обучающей выборке.

С одной стороны предложенный метод в состоянии обеспечить традиционную классификацию, при которой предполагается, что неопределённые значения признаков будут определены. При таком подходе, классификацию следует выполнять следующим образом: начиная с корневого узла и повторяя проверку признака на каждом узле, происходят соответствующие переходы до тех пор, пока не будет достигнут лист. Однако вопреки традиционному дереву решений, в котором каждому листу соответствует уникальный класс, в дереве решений, построенном по предложенному методу, неопределённые классы экземпляров определяются за счёт основного распределения доверия, которое соответствует достигнутому листу. Чтобы получить решение и определить вероятность каждого отдельного класса, предлагается применить пигнистическое преобразование к основному распределению доверия, и использовать распределение вероятности, чтобы рассчитать ожидаемую эффективность, необходимую для принятия оптимального решения.

С другой стороны, поскольку работа выполняется с неопределенной выборкой данных, разработанный метод классификации позволяет также классифицировать

неопределённые экземпляры, которые характеризуются неопределенностью в значениях их признаков. В предложенном методе предполагается, что новые экземпляры, которые необходимо классифицировать, описаны не только определенными значениями признаков, а также могут характеризоваться неопределёнными значениями для некоторых признаков. Кроме того, даже могут быть признаки с неизвестными значениями. Классификация таких объектов состоит в поиске листьев, которые могут принадлежать к рассматриваемому экземпляру, отслеживая все возможные пути, вызванные различными значениями признака. В случае неизвестных значений, принимаются во внимание все ветвления относительно рассматриваемого признака.

Как следствие рассматриваемый экземпляр может принадлежать нескольким листьям, каждый из которых характеризуется функцией основного распределения доверия. Полученные ОРД должны быть объединены, чтобы получить убеждения относительно возможного класса экземпляра. Дизъюнктивное правило объединения, разработанное Сметсом [8], является подходящим, поскольку оно предполагает, что как минимум один путь является верным. Так, при упрощенном случае, в котором было выделено только два листа для рассматриваемого экземпляра, и класс экземпляров в первом листе *A*, а во втором – *B*, единственным выводом из этого может быть то, что класс рассматриваемого экземпляра является либо *A*, либо *B*, то есть это и есть дизъюнктивное правило.

ОРД, полученное на основании дизъюнктивного правила, может быть преобразовано в функцию вероятности путём применения пигностического преобразования. Это позволяет вычислить вероятность принадлежности рассматриваемого экземпляра отдельному классу рассматриваемой проблемной области.

4. Эксперименты и результаты

Предложенный метод построения деревьев решений на основе теории функций доверия был программно реализован в среде пакета Matlab 7.0.

Для экспериментов использовалась выборка, характеризовавшая испытания, проводимые для определения работоспособности кузова автомобилей [10]. Выборка характеризовалась 46 признаками, которые описывали состояние 38 экземпляров. Признаки характеризуют значения зазоров и сопряжений в 46 контрольных точках, расположенных по всему кузову автомобиля. При этом в качестве выходных откликов рассматривалось 16 параметров, влияющих на состояние кузова автомобиля.

Фрагмент обучающей выборки представлен в таблицах 1 и 2.

Таблица 1

Значения признаков для экземпляров

№	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	…	x_{39}	x_{40}	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}
1	0,3	0,2	1,0	-0,6	-0,1	1,3	0,1	0,3	0,3	-0,3	…	1,2	-0,2	0,8	0,4	1,2	0,3	0,5	0,2
2	1,0	-0,3	1,4	-0,3	-0,4	1,5	-0,1	0,0	0,3	-0,3	…	1,4	0,1	1,4	0,4	1,2	0,1	0,2	0,2
3	-0,3	-0,4	0,3	-0,1	-0,4	0,3	-0,8	0,1	-0,2	-1,0	…	1,0	-0,5	0,0	0,5	0,1	0,3	-1,5	-0,5
4	-1,4	-0,4	-0,1	-1,0	-0,1	0,4	-1,1	1,0	-0,4	-1,2	…	1,7	-1,0	0,8	0,4	1,3	0,3	0,4	-0,8
5	-1,4	-0,2	-1,2	-1,0	0,2	-0,6	-0,9	0,2	-1,5	-1,0	…	0,9	-1,8	0,9	-0,8	1,2	-0,1	-0,5	-1,6
…	-1,2	-0,7	0,7	-1,7	-0,9	0,5	-1,3	-0,1	-0,6	-1,3	…	1,0	-1,2	0,8	0,5	1,3	1,4	-2,2	-1,0
37	-0,7	-0,1	0,4	-0,3	-0,6	0,1	-0,8	0,2	-0,7	-0,7	…	0,7	-1,0	0,9	0,6	1,2	1,1	-1,2	-0,7
38	-1,2	-0,3	-1,1	-0,8	-0,2	-0,8	-1,2	0,0	-1,3	-1,7	…	1,1	-1,6	0,3	0,1	0,9	0,5	-2,4	-2,0

Таблица 2

Значения выходных откликов для экземпляров

№	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	y_{16}
1	5	5	6,5	6	6	5,5	6	6	6	6	5,5	5	6	7	6,5	6
2	6	6	5,5	6	5	6	5,5	5	5,5	5,5	6	5,5	6	6	6	5
3	5,5	5,5	6	6	5	5	6,5	6	6	6	5	6,5	4,5	5	7	5,5
4	6	6,5	5,2	5,6	5	5	5,6	5	6	5	5,6	5	5	5,5	6,5	5
5	6,5	5,5	5	5	5	5	5,8	6	6	6,5	5,5	5	6	6	5,5	5

…	5,7	6	5,7	5,8	5	5	6	5	5,5	5,5	5	5	5	4,5	5	6
37	6	5,1	5,8	3,3	6,2	4,1	5,8	3,7	5,6	4,8	5,2	3,5	5,5	3,6	4,9	3
38	4	4	5,5	3,5	4,6	4	5,9	3,5	5,3	3	5,5	2,5	6,1	5,2	5,8	4,3

Таким образом, для каждого из выходных откликов были построены деревья решения с использованием как предложенного метода, так и с использованием метода ID3. После этого построенные деревья решений использовались для прогнозирования значений выходных откликов на тестовой выборке с размерностью, аналогичной размерности обучающей выборки, которая характеризовалась неопределённостью значений признаков. На основе полученных значений выходных параметров для тестовой выборки были рассчитаны параметры работы методов (усредненные значения) при прогнозировании значений выходных параметров для тестовой выборки, представленные в табл. 3.

Таблица 3

Результаты работы методов построения деревьев решений

№	Метод идентификации дерева решений	Усреднённые характеристики работы синтезированных деревьев		
		Ошибка прогнозирования, %	Время работы, с.	Количество узлов дерева, шт.
1	ID3	6,7	54,2	38,7
2	Предложенный метод на основе теории функций доверия	1,2	52,3	32,3

Как видно из таблицы 3, деревья решений, получаемые с использованием предложенного метода характеризуются меньшей ошибкой прогнозирования и меньшей сложностью самого дерева, что способствует лучшей интерпретабельности дерева.

5. Выводы.

В работе решена актуальная задача автоматизации синтеза деревьев решений по прецедентам в условиях неполноты данных.

Научная новизна работы заключается в том, что предложен новый метод идентификации деревьев решений, в котором рассчитываются пигнистические вероятности отнесения экземпляров к классам на основании теории функций доверия, что позволяет выполнять классификацию экземпляров в условиях неопределенности или неполноты данных.

Разработанный метод отличается от существующих методов идентификации деревьев решений фазой построения дерева решений, поскольку учитывается неопределенность, характеризирующая классы обучающих экземпляров за счёт использования функций доверия. Предложенный метод характеризуется особой процедурой классификации новых экземпляров, значения признаков которых могут быть неопределены, что позволяет выполнять классификацию неоднозначно заданных экземпляров.

Практическая ценность полученных результатов заключается в том, что на основе предложенного метода разработано программное обеспечение, позволяющее решать задачи классификации в условиях неопределённости или неполноты исходных данных.

Литература

1. Quinlan J. R. Induction of decision trees / J. R. Quinlan // Machine Learning. – 1986. – № 1. – P. 81–106.
2. Субботін С.О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень : навч. посібник / С. О. Субботін. – Запоріжжя: ЗНТУ, 2008. – 341 с.
3. Classification and regression trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. – California : Wadsworth & Brooks, 1984. – 368 p.

4. Quinlan J. R. Decision trees as probabilistic classifiers / J. R. Quinlan // Fourth International Workshop on Machine Learning. – Irvine : Morgan Kaufmann, 1987. – P. 31–37.
5. Smets P., Kennes R. The transferable Belief Model / P. Smets, R. Kennes // Artificial Intelligence. – 1994. – № 66. – P. 191–234.
6. Smets P. The Transferable Belief Model for Quantified Belief Representation / P. Smets // Handbook of Defeasible Reasoning and Uncertainty Management Systems. – 1998. – № 1. – P. 267–301.
7. Shafer G. A mathematical theory of evidence / G. Shafer. – New Jersey : Princeton University Press, 1976. – 246 p.
8. Smets P. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem / P. Smets // International Journal of Approximate Reasoning. – 1993. – № 9. – P. 1–35.
9. Quinlan J. R. C.4.5: Programs for machine learning / J. R. Quinlan. – San Mateo : Morgan Kaufmann, 1993. – 312 p.
10. Гофман Е. А. Использование деревьев решений для диагностирования автотранспортных средств / Е. А. Гофман, А. А. Олейник, С. А. Субботин // Информационные управляющие системы и компьютерный мониторинг : II Международная научно-техническая конференция ИУС и КМ-2011, 11–13 апреля 2011 г. : материалы конференции. – Донецк, 2011. – С. 159–163.