УДК 519.6

РЕШЕНИЕ ЗАДАЧ КЛАССИФИКАЦИИ С ПОМОЩЬЮ ДЕРЕВЬЕВ РЕШЕНИЙ

Беляков Д. Е.,
научный руководитель канд. техн. наук Бежитский С. С.
Институт математики и фундаментальной информатики
Сибирского Федерального Университета

Метод деревьев решений (decision trees) является одним из наиболее популярных методов решения задач классификации.

Впервые деревья решений были предложены Ховилендом и Хантом (Hoveland, Hunt) в конце 50-х годов прошлого века. Самая ранняя и известная работа Ханта и др., в которой излагается суть деревьев решений - "Эксперименты в индукции" ("Experiments in Induction") - была опубликована в 1966 году.

В самом простом виде дерево решений - это способ представления правил в последовательной структуре. Ее основа - ответы "Да" или "Нет" на ряд вопросов.

На рис. 1 приведен пример дерева решений, задача которого состоит в том, чтобы ответить на вопрос: "Играть ли в гольф?". Чтобы решить задачу, т.е. принять решение, играть ли в гольф, нужно отнести эту ситуацию к одному из известных классов (в данном случае - "играть" или "не играть"). Для этого требуется ответить на ряд вопросов, находящихся в узлах этого дерева, начиная с его корня.

Первый узел нашего дерева (вопрос "Солнечно?") является узлом проверки, т.е. условием. При положительном ответе на вопрос осуществляется переход к левой части дерева, называемой левой ветвью, при отрицательном - к правой части дерева. Таким образом, внутренний узел дерева является узлом проверки определенного условия. Потом идет следующий вопрос и т.д., пока не дойдем до конечного узла дерева, являющегося узлом решения. Для нашего дерева существует два типа конечного узла: "играть" и "не играть" в гольф.

В результате прохождения от корня дерева (называемого корневой вершиной) до его вершины решается задача классификации, т.е. выбирается один из классов - "играть" и "не играть" в гольф.

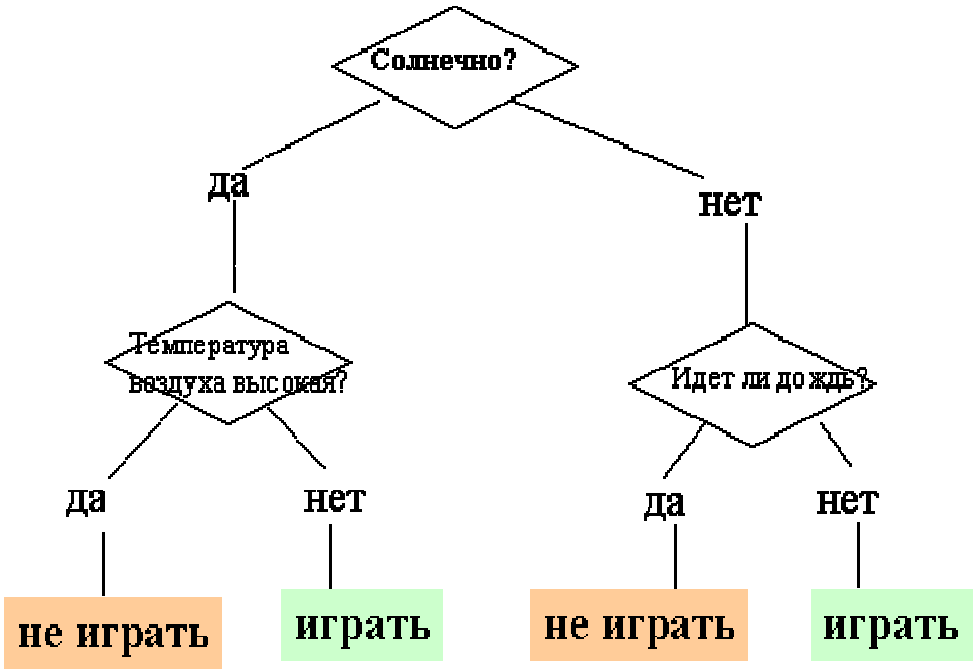

Рис. 1. Дерево решений "Играть ли в гольф?"

Итак, для нашей задачи основными элементами дерева решений являются:

Корень дерева: "Солнечно?".

Внутренний узел дерева (или узел проверки): "Температура воздуха высокая?", "Идет ли дождь?".

Лист, конечный узел дерева, узел решения или вершина: "Играть", "Не играть".

Ветвь дерева (случаи ответа): "Да", "Нет".

В рассмотренном примере решается задача бинарной классификации. Пример демонстрирует работу так называемых бинарных деревьев.

В узлах бинарных деревьев ветвление может вестись только в двух направлениях, т.е. существует возможность только двух ответов на поставленный вопрос ("да" и "нет").

Бинарные деревья являются самым простым, частным случаем деревьев решений. В остальных случаях, ответов и, соответственно, ветвей дерева, выходящих из его внутреннего узла, может быть больше двух.

Рассмотрим более сложный пример. База данных, на основе которой должно осуществляться прогнозирование, содержит следующие данные о клиентах банка, являющиеся ее атрибутами: возраст, наличие недвижимости, образование, среднемесячный доход, вернул ли клиент вовремя кредит. Задача состоит в том, чтобы на основании перечисленных выше данных (кроме последнего атрибута) определить, стоит ли выдавать кредит новому клиенту.

Такая задача решается в два этапа: построение классификационной модели и ее использование.

На этапе построения модели и строится дерево классификации или создается набор неких правил. На этапе использования модели построенное дерево, или путь от его корня к одной из вершин, являющийся набором правил для конкретного клиента, используется для ответа на поставленный вопрос "Выдавать ли кредит?"

Правилом является логическая конструкция, представленная в виде "если , то ".

На рис. 2. приведен пример дерева классификации, с помощью которого решается задача "Выдавать ли кредит клиенту?". Она является типичной задачей классификации, и при помощи деревьев решений получают достаточно хорошие варианты ее решения.

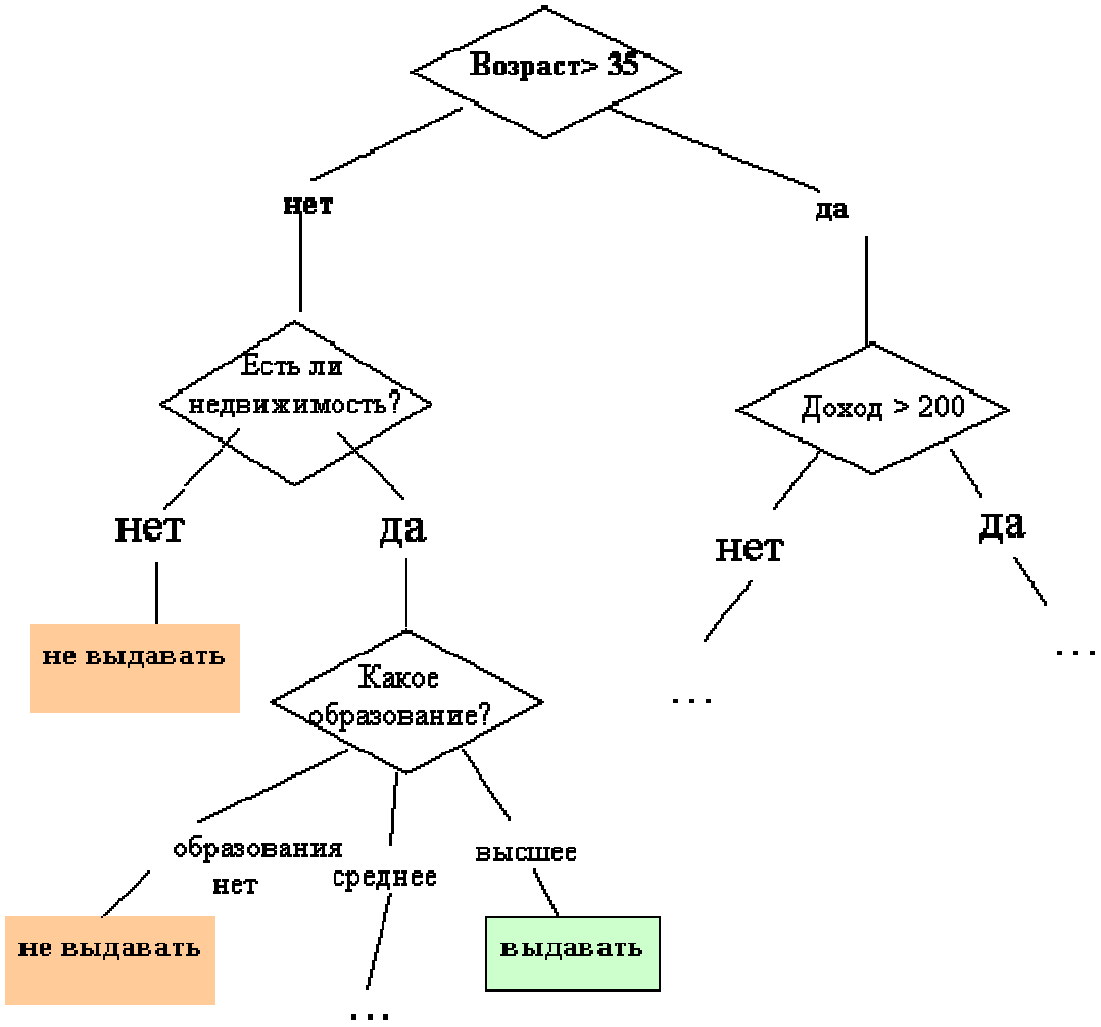

Рис. 2. Дерево решений "Выдавать ли кредит?"

Как мы видим, внутренние узлы дерева (возраст, наличие недвижимости, доход и образование) являются атрибутами описанной выше базы данных. Эти атрибуты называют прогнозирующими. Конечные узлы дерева, или листы, именуются метками класса, являющимися значениями зависимой переменной "выдавать" или "не выдавать" кредит.

Каждая ветвь дерева, идущая от внутреннего узла, отмечена предикатом расщепления. Последний может относиться лишь к одному атрибуту расщепления данного узла. Характерная особенность предикатов расщепления: каждая запись использует уникальный путь от корня дерева только к одному узлу-решению. Объединенная информация об атрибутах расщепления и предикатах расщепления в узле называется критерием расщепления.

На рис. 2. изображено одно из возможных деревьев решений для рассматриваемой базы данных. Например, критерий расщепления "Какое образование?", мог бы иметь два предиката расщепления и выглядеть иначе: образование "высшее" и "не высшее". Тогда дерево решений имело бы другой вид.

Таким образом, для данной задачи (как и для любой другой) может быть построено множество деревьев решений, с различной прогнозирующей точностью.

Метод деревьев решений часто называют "наивным" подходом. Но благодаря целому ряду преимуществ, данный метод является одним из наиболее популярных для решения задач классификации.