

УДК 519.17: 511.12

АЛГОРИТМ ПОСТРОЕНИЯ «ДЕРЕВА» РЕШЕНИЙ**Гулай Т.А., Жук А.Н., Коробейник А.А.***ФГБОУ ВО «Ставропольский государственный аграрный университет», Ставрополь,
e-mail: inf@stgau.ru*

Методика построения дерева решения основана на рекурсивном разбиении множества объектов из обучающей выборки на подмножества, содержащие объекты, относящиеся к одинаковым классам. В первую очередь выбирается независимая переменная, которая помещается в корень дерева. Из вершины строятся ветви, соответствующие всем возможным значениям выбранной независимой переменной. Множество объектов из обучающей выборки разбивается на несколько подмножеств в соответствии со значением выбранной независимой переменной. Таким образом, в каждом подмножестве будут находиться объекты, у которых значение выбранной независимой переменной будет одно и то же. Метод, рассмотренный в работе лежит в основе многих современных алгоритмов построения деревьев решений, этот метод известен еще под названием разделения и захвата. При использовании данной методики, построение дерева решений будет происходить сверху вниз. Поскольку все объекты были заранее отнесены к известным нам классам, такой процесс построения дерева решений называется обучением с учителем. Процесс обучения также называют индуктивным обучением или индукцией деревьев.

Ключевые слова: дерево принятия решений, листья, узел, оптимальное решение, алгоритм

THE ALGORITHM FOR CONSTRUCTING A «TREE» OF SOLUTIONS**Gulay T.A., Zhuk A.N., Korobeynik A.A.***Federal State Budgetary Educational Institution of Higher Education Stavropol State Agrarian
University, Stavropol, e-mail: inf@stgau.ru*

A method of constructing a decision tree based on recursive partitioning of a set of objects from the training sample into subsets that contain objects belonging to the same class. First selects the independent variable that is placed in the root of the tree. Tops are constructed of branches corresponding to all possible values of the selected independent variable. Many objects from the training sample is divided into multiple subsets in accordance with the selected value of the independent variable. Thus, each subset will contain objects, where the value of the chosen independent variable will be one and the same. The method considered in the paper is the basis for many modern algorithms for building decision trees, this method is known as separation and capture. When using this technique, building the decision tree will occur from the top down. Since all the objects have been previously assigned to the known classes, this process of building decision tree is called supervised learning. The training process is also called inductive learning or induction trees.

Keywords: decision tree, leaf, node, optimal solution, algorithm

Теория графов – область дискретной математики, особенностью которой является геометрический подход к изучению объектов. Основным объектом теории графов – граф и его обобщения.

Через V обозначим множество, соединенных некоторым образом точек. Назовем V – множеством вершин, а элементы V вершинами. Граф $G = G(V)$ есть некоторое семейство пар точек $E(a; b)$, $a \in V$, $b \in V$, с указаниями, какие точки считаются соединенными.

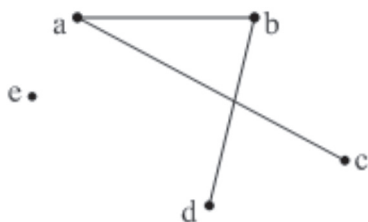


Рис. 1

Например, граф показанный на рис. 1, задан вершинами $V = (a, b, c, d, e)$, а множество E имеет вид $E((a,b), (a,c), (a,d), (b,d))$.

В соответствии с геометрическим представлением графа, каждая конкретная пара соединенных точек (a, b) называется ребром. Далее будем говорить, что вершины a и b инцидентны ребру (a, b) . Вершина e выше построенного графа, не инцидентна никакому ребру. (инцидент от лат. *incidens* – столкновение). Такие вершины будем называть изолированными. В наших рассмотрениях несущественно прямые линии соединяют вершины, или же вершины соединены непрерывными дугами.

Пример.

План города можно рассматривать как граф, в котором ребра представляют улицы, а вершины перекрестки.

Локальные степени:

Пусть $G = G(V)$ – граф и $a \in V$ его вершина. Число $p(a)$ ребер графа инцидентных

$\bar{a}$ называется локальной степенью графа G в вершине a .

Теорема. Сумма всех локальных степеней графа является четным числом.

Доказательство. Каждому ребру соответствуют две вершины, следовательно, каждое ребро учитывается дважды.

Следствие. В графе число вершин a с нечетной степенью $p(a)$ всегда чётно.

Граф называется однородным, если $p(a) = p(b)$ для всех его вершин. В этом случае, если $p(a) = n$, то граф называется однородным степени n .

Примером однородного графа степени 2 является многоугольник без самопересечений.

Связность. Расстояния в связном графе:

Цепью называется последовательность ребер, встречающихся не более одного раза, в которой два соседние ребра имеют общую вершину.

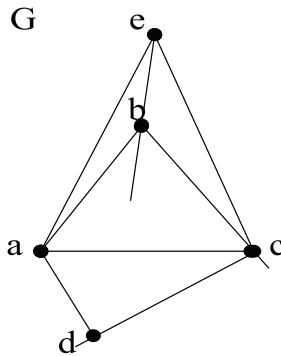


Рис. 2

Например, в графе G последовательность ребер $(a, b), (b, c), (c, d), (d, b), (b, e)$ будет цепью.

Замкнутая цепь называется циклом.

Граф называется связным, если любую его вершину можно соединить с любой другой, цепью, состоящей из ребер графа.

Теорема: Граф с n вершинами и числом ребер большим, чем $\frac{1}{2}(n-1)(n-2)$, связан.

Пусть a и b две вершины графа. Рассмотрим кратчайшую цепь (цепь, содержащую наименьшее число ребер среди всех соединяющих a и b цепей). Число ребер этой цепи обозначим через $d(a: b)$. Это число назовем расстоянием между вершинами a и b .

Функция $d(a: b)$ удовлетворяет всем обычным аксиомам расстояния:

- 1) $d(a: b) = 0$, если $a = b$,
- 2) $d(a: b) \geq 0$,
- 3) $d(a: b) = d(b: a)$,
- 4) справедливо неравенство треугольника $d(a: b) + d(b: c) \geq d(a: c)$.

Граф называется полным, если любые две его вершины образуют ребро. Отметим, что граф будет полным, если $d(a: b) = 1$ для всех несовпадающих вершин графа.

Диаметром графа (D) называется максимальное расстояние между двумя точками графа, точнее $D = \max(d(a, b))$, $a \in V, b \in V$.

Деревья:

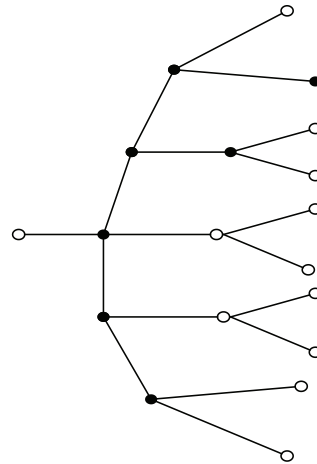


Рис. 3

Связный граф называется деревом, если он не имеет циклов, т.е. замкнутых цепей.

Теорема: В дереве любые две вершины связаны единственной цепью.

Доказательство. В противном случае был бы цикл.

Задача. Необходимо соединить n различных городов линиями электропередач так, чтобы не строить лишних линий. Из теоремы 1 сразу следует, что схема этих линий должна образовать дерево. Возникает вопрос о числе различных деревьев с n — вершинами.

Теорема: Число различных деревьев, которые можно построить на n вершинах, равно $(n-2)!$.

Ориентированный граф:

Определение. Ориентированным графом называется совокупность множества точек и совокупность ориентированных дуг, соединяющих эти точки.

Применение теории графов:

При анализе надежности сетей связи, электрических схем, коммуникационных сетей возникает задача о нахождении количеств непересекающихся цепей, соединяющих различные вершины графа, и она успешно решается с помощью теории графов.

На практике результат одного решения заставляет нас принимать следующее и т.д. Эту последовательность нельзя выразить таблицей доходов, поэтому необходимо использовать другой процесс принятия решений, а именно — «дерево» решений.

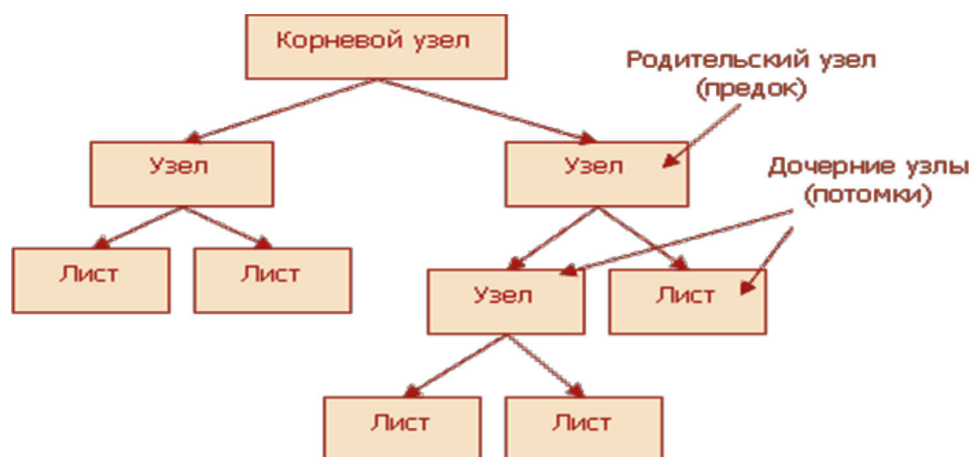


Рис. 4. Общий вид «дерева» решений

«Дерево» принятия решений – это своеобразное средство поддержки принятия решений при прогнозировании, оно широко применяется в статистике и анализе данных.

Прежде чем говорить о конкретных алгоритмах «дерева» решений, необходимо обратить внимание на основополагающие понятия, связанные с данным методом в целом [1].

Итак, «дерево» решений, подобно его «прототипу» из живой природы, состоит из «ветвей» и «листьев». Ветви, а иначе ребра графа, хранят в себе значения атрибутов, от которых зависит целевая функция. Листья – хранят запись о значении целевой функции. Существуют также родительский узел (предок) и дочерний узел (потомок), по которым происходит разветвление (рис. 4).

Цель построения дерева принятия решений – создать модель, по которой можно классифицировать случаи и решать, какие значения может принимать целевая функция, имея на входе несколько переменных.

Например, когда необходимо принять решение о выдаче кредита (целевая функция может принимать значения «да» и «нет») на основе информации о клиенте (несколько переменных: возраст, семейное положение, место работы и т.д.) К примеру, переменная «возраст» с атрибутом «менее 21 года = да» сразу приведет от корневого узла дерева к его листу, при этом целевая функция «выдача кредита» примет значение «нет». Если возраст составляет более 21 года, то ветвь приведет нас к очередному узлу, который, к примеру, «спросит» нас о стабильности дохода

клиента. Таким образом, классификация каждого нового случая происходит при движении вниз до листа, который и укажет нам значение целевой функции в каждом конкретном случае [2].

Рассмотрим на данном примере и общий алгоритм построения «дерева» решений. Вначале необходимо выбрать атрибут Q (в нашем случае, допустим: уровень дохода больше 50 000 руб./месяц) и поместить его в корневой узел.

Далее, из тестовых примеров (или набора данных) для каждого значения атрибута i (в нашем случае их два – «да» и «нет») выбираем только те, для которых $Q=i$. В конечном итоге мы рекурсивно строим «дерево» принятия решений.

На данный момент существует большое количество частных алгоритмов, которые помогают в реализации «дерева» решений CART, C4.5, NewId, IRule, ID3, CN2 и т.д. Но наибольшее распространение и популярность получили только два:

- CART (Classification and Regression Tree) – это алгоритм построения бинарного «дерева» решений – дихотомической классификационной модели. Каждый узел дерева при разбиении имеет только два дочерних узла. Данный алгоритм решает задачи регрессии и классификации.

- C4.5 – это алгоритм построения «дерева» решений, количество дочерних узлов здесь не ограничено. Решает только задачи классификации, так как не умеет работать с непрерывным целевым полем.

Большинство из перечисленных выше алгоритмов являются «жадными алгоритмами». То есть, если один раз был выбран атрибут, по которому было произведено

разбиение на подмножества, то такой алгоритм не может вернуться назад и выбрать иной атрибут, который дал бы лучшее разбиение [3].

Таким образом, «дерево» решений – основной и наиболее популярный метод, при помощи которого можно найти оптимальное решение. Действительно, построение «дерева» решений позволяет наглядно продемонстрировать структуру данных и создать работающую модель классификации данных, какими бы громоздкими они не были.

Список литературы

1. Литвин Д.Б., Гулай Т.А., Долгополова А.Ф. Применение операционного исчисления в моделировании экономических систем // Аграрная наука, творчество, рост. – 2013. – С. 263–265.
2. Смирнова Н.Б., Попова С.В. Использование понятий функции и предела в социально-экономической области человеческой деятельности // Культура и общество: история и современность : материалы III Всероссийской (с международным участием) науч.-практ. конф. Филиал Российского государственного социального университета в г. Ставрополь; под редакцией: О.Ю. Колосовой, Т.В. Вергун, Р.Ф. Гударенко. – 2014. – С. 181–185.
3. Немцова А.В., Попова С.В. Применение средств матричной алгебры для решения задач экономического содержания // Современные наукоемкие технологии. – 2014. – № 5–2. – С. 171–172.