УДК 004.046

МЕТОД ОТСЕЧЕНИЯ ВЕТВЕЙ ДЕРЕВА РЕШЕНИЙ

THE METHOD OF CUT–OFF BRANCHES OF THE DECISION TRE

©Мифтахова А. А.
Поволжский государственный университет телекоммуникаций и информатики
г. Самара, Россия
miftaxovaaa@mail.run
©Miftakhova A.
Volga region state university of telecommunications and informatics
Samara, Russia
miftaxovaaa@mail.ru

Аннотация. При работе с очень большими наборами данных построение дерева решений может быть затруднительно. Даже в случаях, когда это возможно, это может быть не лучший способ использовать данные. В качестве решения проблемы используется метод отсечения ветвей дерева решений на этапе реализации.

Отсечение ветвей или замена некоторых ветвей поддеревом проводится там, где эта процедура не приводит к возрастанию ошибки. Процесс проходит снизу вверх, тем самым он является восходящим. Это более популярная процедура, чем использование правил остановки. Деревья, получаемые после отсечения некоторых ветвей, называют усеченными.

Если такое усеченное дерево все еще не является интуитивным и сложно для понимания, используют извлечение правил, которые объединяют в наборы для описания классов. Каждый путь от корня дерева до его вершины или листа дает одно правило. Условиями правила являются проверки на внутренних узлах дерева.

Статья посвящена описанию метода отсечения ветвей дерева решений для решения задачи классификации.

Abstract. When working with very large data sets to build a decision tree can be difficult. Even in cases where this is possible, it may not be the best way to use the data. As a workaround, use tree branches cut method decisions in the implementation phase.

Pruning branches or replace some subtree branches held where this procedure does not lead to an increase in errors. The process takes place from the bottom up, thus it is rising. This procedure is more popular than the use of rules of a stop. Trees, received after the cut off some branches, called truncated.

If a truncated tree still is not intuitive and hard to understand, using the extraction rules that combine to describe a set of classes. Each path from the root to its top or sheet gives one rule. Terms and Conditions are checks at internal nodes of the tree.

The article describes the method of tree branches cut solutions to solve classification problems.

Ключевые слова: дерево решений, механизм отсечения ветвей, CART.

Keywords: decision tree, mechanism of cut–off branches, CART.

Различные алгоритмы построения деревьев решений способны строить деревья с множеством узлов и ветвей. Ветвистые деревья разбивают обучающее множество на подмножества, состоящие из все меньшего количества объектов, тем самым становясь сложными для восприятия. Дерево, состоящее из малого количества узлов, которым бы соответствовало большое количество объектов из обучающей выборки является более предпочтительным и понятным для восприятия.

Решением проблемы слишком ветвистого дерева является его сокращение путем отсечения (pruning) некоторых ветвей [1].

Отсечение ветвей дерева используется для предотвращения сложных деревьев, трудных для понимания, которые имеют много узлов и ветвей. Точность распознавания дерева решений — это отношение правильно классифицированных объектов при обучении к общему количеству объектов из обучающего множества, а под ошибкой — количество неправильно классифицированных.

Процесс отсечения ветвей происходит снизу вверх, в отличие от построения, начиная с листьев дерева. Узлы отмечаются как листья, либо заменяются поддеревом.

Существует множество алгоритмов, реализующих деревья решений, например, CART, C4.5, NewId, ITrule, CHAID, CN2 и т. д.

В данной работе рассматривается CART (Classification and Regression Tree) — это алгоритм построения бинарного дерева решений – дихотомической классификационной модели [2].

Каждый узел дерева при разбиении имеет только двух потомков.

Механизм отсечения дерева (Minimal cost–complexity tree pruning) — наиболее серьезное отличие алгоритма CART от других алгоритмов построения дерева. CART рассматривает отсечение как получение компромисса между двумя проблемами: получение дерева оптимального размера и получение точной оценки вероятности ошибочной классификации [3].

Для получения оптимального решения выбирают дерево с наименьшей ошибкой обучения.

Отсечение ветвей является более затратной в вычислительном плане процедурой, чем ранняя остановка. Однако ранняя остановка имеет риск потери потенциально интересных правил, до которых алгоритм может просто не «дойти».

В качестве примера приведены деревья решений для классификации студентов вуза по форме обучения.

Для построения дерева использовался алгоритм CART и программная среда Python 2.7.

Для получения решения были выполнены следующие действия:

1 — построено дерево;

2 — отсечены те ветви, которые не приведут к возрастанию ошибки.

Для задания значения отсечения вводится команда prune (tree,n), где n — величина параметра отсечения.

В данном случае пороговое значение — prune (tree,1.0).

На Рисунке 1 представлено дерево до отсечения.

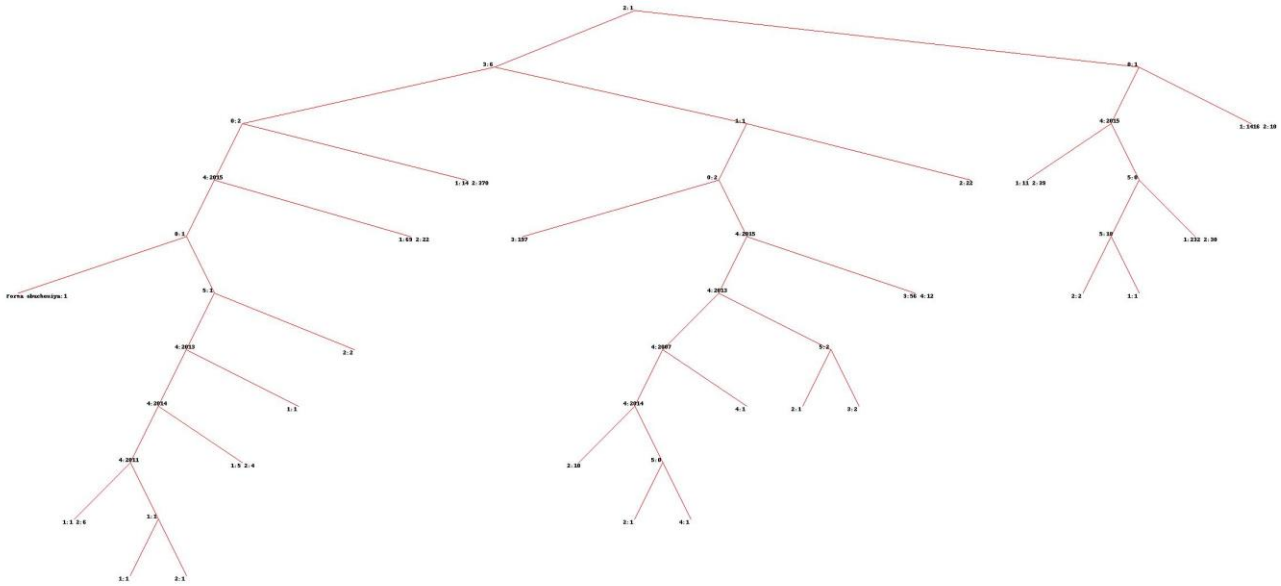

Рисунок 1. Полное дерево решений.

На Рисунке 2 представлено дерево после максимального отсечения.

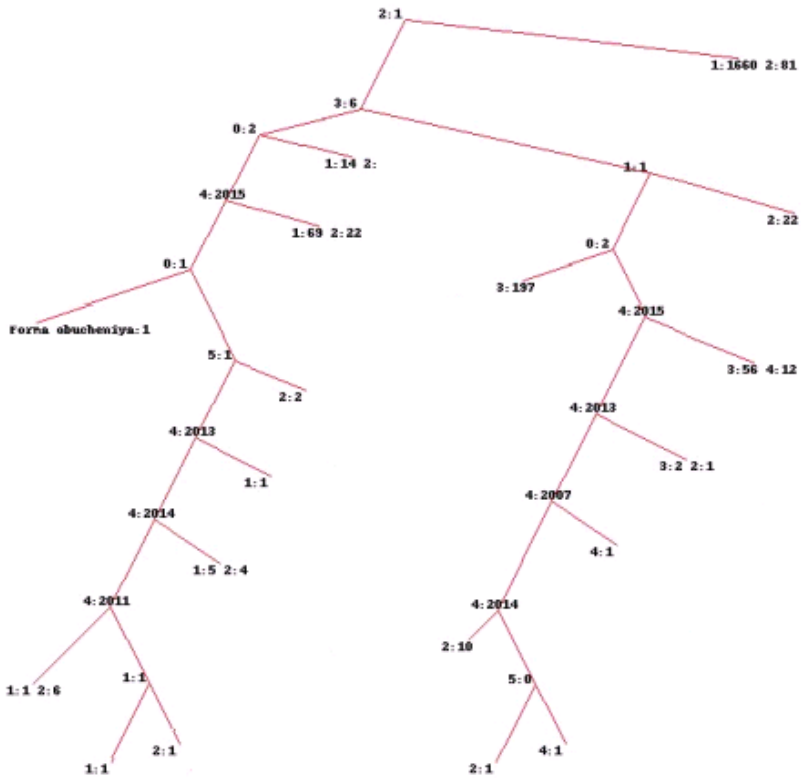

Рисунок 2. Усеченное дерево решений.

Из Рисунка 1 и Рисунка 2 видно, что усеченое дерево в отличие от полного дерева не является сложным для понимания и интерпретации, сохраняя при этом точность распознавания.

Список литературы:

1. Луньков А. Д., Харламов А. В. Интеллектуальный анализ данных: учебно–методическое пособие // Саратовский государственный университет им. Н. Г. Чернышевского. Режим доступа: http://elibrary.sgu.ru/uch_lit/1141.pdf (дата обращения 10.04.2016).

2. Шахиди А. Деревья решений — общие принципы работы // GotAI.NET — Искуственный интеллект — это просто! Режим доступа: http://www.gotai.net/documents/doc-msc-006.aspx (дата обращения 10.04.2016).

3. Калякулина А. И., Юсупов И. И. Классификация объектов городского ландшафта по аэрофотоснимкам высокого разрешения // Нижегородский университетский центр Интернет. Режим доступа: http://www.uic.unn.ru/~zny/ml/Projects/YusipovKalyakulinaFilicheva.pdf (дата обращения 10.04.2016).

References:

1. Lunkov A. D., Kharlamov A. V. Intellektual’nyi analiz dannykh: uchebno–metodicheskoe posobie // Saratovskii gosudarstvennyi universitet im. N. G. Chernyshevskogo. Available at: http://elibrary.sgu.ru/uch_lit/1141.pdf, accessed 10.04.2016.

2. Shakhidi A. Derev’ya reshenii — obshchie printsipy raboty // GotAI.NET — Iskustvennyi intellekt — eto prosto! Available at: http://www.gotai.net/documents/doc-msc-006.aspx, accessed 10.04.2016.

3. Kalyakulina A. I., Yusupov I. I. Klassifikatsiya ob”ektov gorodskogo landshafta po aerofotosnimkam vysokogo razresheniya. Nizhegorodskii universitetskii tsentr Internet. Available at: http://www.uic.unn.ru/~zny/ml/Projects/YusipovKalyakulinaFilicheva.pdf, accessed 10.04.2016.

Работа поступила в редакцию *17.03.2016 г.*

Принята к публикации *21.03.2016 г.*