Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Компьютерных сетей и систем

Кафедра Информатики

РЕФЕРАТ

по курсу «Машинное обучение»

Деревья решений

Студент:
гр. 758641
Ярош Г.И.

Проверил:
Заливако С. С.

Минск, 2018

СОДЕРЖАНИЕ

ДЕРЕВЬЯ РЕШЕНИЙ

1. Введение

Метод деревьев решений (decision trees) является одним из наиболее популярных методов решения задач классификации и прогнозирования. Иногда этот метод также называют деревьями решающих правил, деревьями классификации и регрессии.

Деревья решений могут быть применены при решении следующих классов задач:

- Описание данных. Деревья решений позволяют хранить информацию о данных в компактной форме, вместо них мы можем хранить дерево решений, которое содержит точное описание объектов.
- Классификация. Деревья решений отлично справляются с задачами классификации, т.е. отнесения объектов к одному из заранее известных классов. Целевая переменная должна иметь дискретные значения.
- Регрессия. Если целевая переменная имеет непрерывные значения, деревья решений позволяют установить зависимость целевой переменной от независимых(входных) переменных. Например, к этому классу относятся задачи численного прогнозирования (предсказания значений целевой переменной).

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.

Пример дерева решений приведен на рисунке 1.

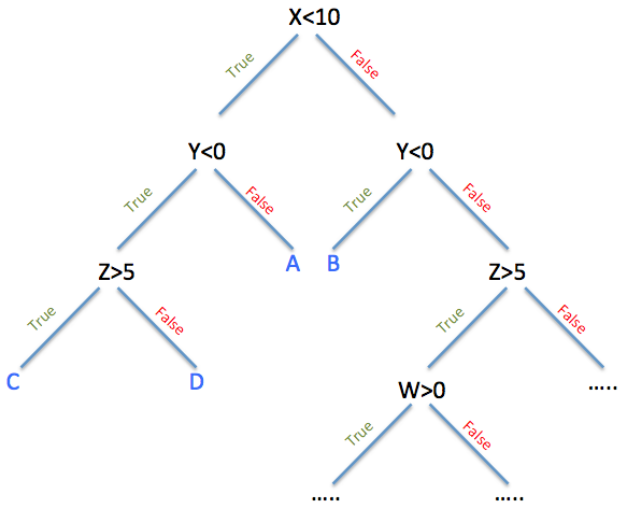

Рис. 1. Пример дерева решений

В наиболее простом виде дерево решений - это способ представления правил в иерархической, последовательной структуре. Узлы дерева содержат некоторое

условия на атрибуты входных данных. Ветви соответствуют результатам применения этих условий к данным. Листья дерева представляют собой классы объектов.

2. Преимущества и недостатки деревьев решений

Деревья решений обладают рядом преимуществ перед другими алгоритмами машинного обучения:

1. Интуитивность деревьев решений. Классификационная модель, представленная в виде дерева решений, является интуитивной и упрощает понимание решаемой задачи.
2. Деревья решений дают возможность извлекать правила на естественном языке либо в графическом виде.
3. Деревья решений позволяют создавать классификационные модели в тех областях, где аналитику достаточно сложно формализовать знания.
4. Алгоритм конструирования дерева решений не требует от пользователя выбора входных атрибутов (независимых переменных). На вход алгоритма можно подавать все существующие атрибуты, алгоритм сам выберет наиболее значимые среди них, и только они будут использованы для построения дерева.
5. Точность моделей, созданных при помощи деревьев решений, сопоставима с другими методами построения классификационных моделей (статистические методы, нейронные сети).
6. Многие классические статистические методы, при помощи которых решаются задачи классификации, могут работать только с числовыми данными, в то время как деревья решений работают и с числовыми, и с категориальными типами данных.

В то же время, деревья решений обладают следующими недостатками:

1. Проблема получения оптимального дерева решений является NP-полной с точки зрения некоторых аспектов оптимальности даже для простых задач. Таким образом, практическое применение алгоритма деревьев решений основано на эвристических алгоритмах, таких как жадный алгоритм, где единственно оптимальное решение выбирается локально в каждом узле. Такие алгоритмы не могут обеспечить оптимальность всего дерева в целом.
2. При обучении дерева решений может быть создано большое количество ветвей и условий, которые будут хорошо подходить для предсказания данных из тестовой выборки, но плохо предсказывать данные, на которых дерево не обучалось. Данная проблема называется переобучением. Для того, чтобы избежать данной проблемы, необходимо использовать методы регулирования глубины дерева.
3. Для данных, которые включают категориальные переменные с большим набором уровней (закрытий), больший информационный вес присваивается тем атрибутам, которые имеют большее количество уровней.

3. Процесс построения деревьев решений

Алгоритмы конструирования деревьев решений состоят из этапов построения дерева (tree building) и сокращения дерева (tree pruning). В ходе построения дерева решаются вопросы выбора критерия расщепления и остановки обучения (если это предусмотрено алгоритмом). В ходе этапа сокращения дерева решается вопрос отсечения некоторых его ветвей.

Критерий расщепления

Процесс создания дерева происходит сверху вниз, т.е. является нисходящим. В ходе процесса алгоритм должен найти такой критерий расщепления, иногда также называемый критерием разбиения, чтобы разбить множество на подмножества, которые бы ассоциировались с данным узлом проверки. Каждый узел проверки должен быть помечен определенным атрибутом. Существует правило выбора атрибута: он должен разбивать исходное множество данных таким образом, чтобы объекты подмножеств, получаемых в результате этого разбиения, являлись представителями одного класса или же были максимально приближены к такому разбиению. Последняя фраза означает, что количество объектов из других классов в каждом классе должно стремиться к минимуму.

Существуют различные критерии расщепления. Наиболее известные - мера энтропии, мера информационного выигрыша и неопределенность Gini.

В некоторых методах для выбора атрибута расщепления используется так называемая мера информативности подпространств атрибутов, которая основывается на энтропийном подходе и известна под названием "мера информационного выигрыша" (information gain) или мера энтропии.

Другой критерий расщепления реализован в алгоритме CART и называется неопределенность Gini. При помощи этого индекса атрибут выбирается на основании расстояний между распределениями классов.

Проблема слишком ветвистых деревьев

Чем больше частных случаев описано в дереве решений, тем меньшее количество объектов попадает в каждый частный случай. Такие деревья называют "ветвистыми", они состоят из неоправданно большого числа узлов и ветвей, исходное множество разбивается на большое число подмножеств, состоящих из очень малого числа объектов. В результате "переполнения" таких деревьев их способность к обобщению уменьшается, и построенные модели не могут давать верные ответы.

В процессе построения дерева, чтобы его размеры не стали чрезмерно большими, используют специальные процедуры, которые позволяют создавать оптимальные деревья.

Остановка построения дерева

Рассмотрим правило остановки. Оно должно определить, является ли рассматриваемый узел внутренним узлом, при этом он будет разбиваться дальше, или же он является конечным узлом, т.е. узлом решением.

Один из вариантов правил остановки - "ранняя остановка" (prepruning). Она определяет целесообразность разбиения узла. Преимущество использования такого варианта - уменьшение времени на обучение модели. Однако здесь возникает риск снижения точности классификации.

Второй вариант остановки обучения - ограничение глубины дерева. В этом случае построение заканчивается, если достигнута заданная глубина.

Еще один вариант остановки - задание минимального количества примеров, которые будут содержаться в конечных узлах дерева. При этом варианте ветвление продолжаются до того момента, пока все конечные узлы дерева не будут чистыми или будут содержать не более чем заданное число объектов.

Сокращение дерева или отсечение ветвей

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

Качество классификационной модели, построенной при помощи дерева решений, характеризуется двумя основными признаками: точностью распознавания и ошибкой.

Точность распознавания рассчитывается как отношение объектов, правильно классифицированных в процессе обучения, к общему количеству объектов набора данных, которые принимали участие в обучении.

Ошибка рассчитывается как отношение объектов, неправильно классифицированных в процессе обучения, к общему количеству объектов набора данных, которые принимали участие в обучении.

Отсечение ветвей или замену некоторых ветвей поддеревом следует проводить там, где эта процедура не приводит к возрастанию ошибки. Процесс проходит снизу вверх, т.е. является восходящим. Это более популярная процедура, чем использование правила остановки. Деревья, получаемые после отсечения некоторых ветвей, называют усеченными.

Если такое усеченное дерево все еще не является интуитивным и сложно для понимания, при визуализации 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.

4. Алгоритмы построения деревьев решений

Алгоритмы построения деревьев решений различаются следующими характеристиками:

- вид расщепления - бинарное (binary), множественное (multi-way);
- критерии расщепления – информационная энтропия, мера информационного выйгрыша, индекс Gini;
- возможность обработки пропущенных значений;
- процедура сокращения ветвей или отсечения;
- возможности извлечения правил из деревьев.

Для построения деревьев решений используются следующие алгоритмы:

- Алгоритм ID3, где выбор атрибута происходит на основании прироста информации (англ. Gain) либо минимизации информационной энтропии;
- Алгоритм C4.5 (улучшенная версия ID3), где выбор атрибута происходит на основании нормализованного прироста информации;
- Алгоритм CART;
- Автоматический детектор взаимодействия Хи-квадрат (CHAID). Выполняет многоуровневое разделение при построении деревьев;
- MARS: расширяет деревья решений для улучшения обработки цифровых данных.

Далее будут рассмотрены некоторые из них.

Алгоритм ID3

Алгоритм ID3 предложен в 1986 году Россом Квинланом. Алгоритм строит деревья решений со множественными путями, используя меру информационного выигрыша как критерий расщепления.

Мера информационного выигрыша определяет разницу информационной энтропии до и после разделения набора данных S по некоторому атрибуту A и определяется по формуле:

$$IG(S, A) = \mathrm{H}(S) - \sum_{t \in T} p(t)\mathrm{H}(t) = \mathrm{H}(S) - \mathrm{H}(S|A).$$

где H(S) – мера информационной энтропии, T – подмножества, на которое был разделен набор данных S, p(t) – соотношение количества элементов в множестве t относительно количества элементов во всем наборе данных S, H(t) – мера информационной энтропии подмножества t.

Мера информационной энтропии определяется следующей формулой:

$$\mathrm{H}(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

где S – текущий набор данных, X – множество всех классов, p(x) – соотношение между количеством элементов класса x и всех элементов в наборе данных.

В начале работы алгоритм рассматривает весь набор данных S. На каждой итерации для каждого неиспользованного атрибута подсчитывается мера информационного выигрыша. Далее выбирается атрибут с наибольшим значением меры. Он используется для разделения набора данных на подмножества. Для каждого значения атрибута в текущем наборе данных создается подмножество и

соответствующая ветвь дерева с условием. Далее алгоритм рекурсивно продолжает работу для каждого из подмножеств не учитывая атрибут, использованный на этом шаге.

Алгоритм для текущего множества останавливается в ряде случаев:

- Каждый элемент из текущего множества принадлежит к одному классу. Текущий узел дерева помечается этим классом.
- Больше нет атрибутов, по которым можно разделить множество. В данном случае узел дерева помечается классом, к которому принадлежит наибольшее количество элементов в множестве.
- Нет ни одного элемента в множестве. В таком случае текущий узел дерева помечается классом, наиболее представленным в родительском узле дерева.

В результате работы алгоритма получается дерево решений, где каждый внутренний узел определяет набор условий для конкретного атрибута, а внешние узлы определяют классы.

Алгоритм C4.5

Алгоритм C4.5 является развитием алгоритма ID3. Алгоритм использует нормированную меру информационного выигрыша. Нормировка происходит по количеству всех возможных значений атрибута, для которого подсчитывается мера. Данное усовершенствование позволяет бороться с проблемой переобучения, когда дерево решений чаще использует атрибуты с большим количеством значений, создавая множество ветвей и запоминая данные из исходной выборки.

Также алгоритм C4.5 имеет следующие усовершенствования относительно ID3:

- Возможность обрабатывать непрерывные значения атрибутов. Для обработки непрерывных значений, алгоритм создает пороговое, по которому разделяет выборку на подмножества, где значения атрибута больше, равны, либо меньше порогового.
- Обработка пропущенных значений атрибутов. Пропущенные значения не используются при подсчете мер информационной энтропии и выйгрыша.
- Возможность настройки весов для атрибутов.
- Отсечение ветвей у построенного дерева. После создания дерева, алгоритм проходится по всем узлам дерева и удаляет ветви, которые не ухудшают эффективность классификации, заменяя их на узел с пометкой класса.

Алгоритм CART

Алгоритм CART является алгоритмом построения двоичных деревьев классификации. Он использует неопределенность Gini как критерий расщепления:

$$Gini\,(c) = 1 - \sum_i p_j^2,$$

где pj – соотношение количества элементов класса j и количества элементов в текущей выборке.

Неопределенность Gini показывает вероятность того, что элемент текущего набора данных будет неправильно классифицирован. Алгоритм стремится к минимизации данного значения для всех узлов дерева.

Алгоритм разбиения похож на алгоритм C4.5. На каждой итерации, для каждого атрибута вычисляется наилучшее разбиение его значений на два подмножества основываясь на минимальном значении неопределенности Gini для подмножеств. Выбирается атрибут с наилучшим разбиением.

Для численный атрибутов в узле дерева устанавливаются условия вида меньше либо равно. Для категориальных атрибутов устанавливаются условия принадлежности значения к некоторому множеству.

После того, как алгоритм построил дерево, происходит отсечение лишних ветвей основываясь на проведении кросс валидации между двумя вариантами дерева: с и без ветви.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] Методы классификации и прогнозирования. Деревья решений [Электронный ресурс]. – Электронные данные. – Режим доступа: https://www.intuit.ru/studies/professional_skill_improvements/1210/courses/6/lecture/174?page=1.

[2] ID3 algorithm [Электронный ресурс]. – Электронные данные. – Режим доступа: https://en.wikipedia.org/wiki/ID3_algorithm.

[3] C4.5 algorithm [Электронный ресурс]. – Электронные данные. – Режим доступа: https://en.wikipedia.org/wiki/C4.5_algorithm.

[4] Sklearn. Decision Trees [Электронный ресурс]. – Электронные данные. – - Режим доступа: https://scikit-learn.org/stable/modules/tree.html.

[5] Decision tree learning [Электронный ресурс]. – Электронные данные. – - Режим доступа: https://en.wikipedia.org/wiki/Decision_tree_learning.