УДК 004.855

РЕАЛИЗАЦИЯ АЛГОРИТМА ID3 НА ЯЗЫКЕ ПРОГРАММИРОВАНИЯ PYTHON

Габриельчик П.В. и Ермакович В.А., студенты гр.250702

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Вашкевич М.И. – доктор техн. наук

Аннотация. Данная научная работа посвящена изучению основ машинного обучения. В ней исследуются принципы машинного обучения, а также представляется реализация алгоритма ID3 как способа построения дерева решений для классификации данных. Реализация алгоритма производится на языке Python. Результаты эксперимента показывают эффективность и точность алгоритма ID3 при решении задач классификации. Исследование позволяет расширить понимание основ машинного обучения и применять полученные знания на практике.

Ключевые слова. Машинное обучение, Искусственный Интеллект, теория информации, алгоритм ID3, энтропия, информационное обучение, прирост информации, дерево решений, классификация данных, переобучение и недообучение.

1. Введение

Машинное обучение (machine learning) — это раздел искусственного интеллекта, который позволяет компьютерным системам извлекать знания и предсказывать результаты на основе анализа большого количества данных. Машинное обучение используется для решения широкого круга задач, включая классификацию, регрессионный анализ, кластеризацию, обнаружение аномалий.

Существуют различные типы машинного обучения, такие как обучение с учителем (supervised machine learning), обучение без учителя (unsupervised machine learning) и обучение с подкреплением (reinforcement machine learning). В обучении с учителем модель обучается на основе размеченных данных, то есть данных, которые содержат правильные ответы. В обучении без учителя модель пытается найти структуру в данных без разметки. В обучении с подкреплением модель обучается на основе взаимодействия с окружающей средой и получает награду за правильные решения.

Машинное обучение нашло широкое применение в различных отраслях, таких как медицина, банковское дело, транспорт, производство, маркетинг. С помощью машинного обучения можно сократить время и затраты на анализ данных, повысить точность предсказаний, автоматизировать процессы и многое другое.

Одним из основных принципов машинного обучения является создание моделей, которые могут обучаться на данных и применять полученные знания для решения новых задач. В настоящее время существует множество алгоритмов машинного обучения, включая деревья решений, нейронные сети, метод опорных векторов, случайные леса и другие.

Цель и задачи исследования. Цели данного исследования: изучить основы машинного обучения, рассмотреть алгоритм ID3, используемый для построения деревьев решений, а также разработать собственный алгоритм для классификации данных.

Для достижения этих целей были поставлены следующие задачи:

1) Изучить основные понятия машинного обучения, такие как энтропия, информационное обучение, прирост информации и деревья решений.

2) Изучить алгоритм ID3 и его реализацию для построения деревьев решений.

3) Проанализировать примеры использования алгоритма ID3 для классификации данных.

4) Разработать код реализации алгоритма ID3 на языке Python.

5) Протестировать реализацию алгоритма ID3 на реальных данных и оценить ее точность.

6) Сделать выводы об эффективности и применимости алгоритма ID3 для построения деревьев решений в различных областях.

Обоснование выбора алгоритма ID3 для классификации данных. Алгоритм ID3 является одним из самых простых и эффективных алгоритмов для построения деревьев решений на основе данных с категориальными признаками. Его основное преимущество заключается в том, что он может автоматически извлекать наиболее значимые признаки из данных и использовать их для построения оптимального дерева решений.

Алгоритм ID3 использует информационный критерий прироста информации для выбора наилучшего разделения данных на каждом узле дерева. Он также способен обрабатывать отсутствующие данные и имеет возможность обработки больших объемов данных.

В связи с этим, выбор алгоритма ID3 для реализации дерева решений имеет ряд преимуществ:

1. Простота реализации и использования. ID3 легко реализуется на любом языке программирования, в том числе и на Python, и не требует специальных знаний в области машинного обучения.

2. Высокая точность. Алгоритм ID3 обладает высокой точностью и может давать хорошие результаты при правильном выборе параметров.

3. Универсальность. ID3 может быть использован для построения деревьев решений в различных областях, включая банковское дело, медицину, транспорт и т.д.

Таким образом, алгоритм ID3 является оптимальным выбором для реализации деревьев решений на основе категориальных данных.

2. Классификация данных и информационное обучение

Классификация данных — это процесс определения к какому классу относится определенный объект или набор данных. Классификация является одной из основных задач машинного обучения.

В задаче классификации, на основе известных данных, необходимо построить алгоритм, который может определять класс, к которому относится новый объект данных. Классификация данных может быть двуклассовой, когда необходимо отнести объект к одному из двух классов, или многоклассовой, когда необходимо отнести объект к одному из нескольких классов.

Для классификации данных могут использоваться различные алгоритмы машинного обучения, такие как деревья решений, нейронные сети, метод опорных векторов и другие. Каждый из этих алгоритмов имеет свои преимущества и недостатки, и выбор конкретного алгоритма зависит от характеристик данных и требований к точности и скорости работы.

Классификация данных находит широкое применение в различных областях, таких как биология, медицина, экономика, финансы, маркетинг и многие другие. Примеры задач классификации данных включают определение злокачественных опухолей, распознавание рукописных цифр, категоризацию товаров в интернет-магазинах, определение кредитоспособности заемщиков.

Информационное обучение — это подход к машинному обучению, основанный на теории информации, который использует понятие энтропии для описания степени неопределенности данных. Информационное обучение стремится минимизировать энтропию данных, то есть уменьшить их неопределенность путем построения модели, которая наилучшим образом описывает данные.

Энтропия — это мера неопределенности данных. В теории информации энтропия используется для описания количества информации, содержащейся в некотором сообщении. Чем больше энтропия, тем больше неопределенность в данных. В машинном обучении энтропия используется для оценки неопределенности данных в задачах классификации.

В контексте алгоритма ID3, энтропия используется для оценки неопределенности различных атрибутов и выбора наилучшего атрибута для разделения данных на подмножества. Идея заключается в том, что при выборе атрибута, который имеет наименьшую энтропию, можно разделить данные на наиболее однородные группы. Это позволяет построить более точную модель для классификации новых данных.

Информационное обучение и энтропия являются важными понятиями в машинном обучении и широко используются в различных алгоритмах, таких как деревья решений, наивный Байесовский классификатор и другие. Они позволяют учитывать неопределенность данных и создавать более точные модели для классификации и прогнозирования.

3. Деревья решений: принцип обучения

Дерево решений — это метод машинного обучения, который использует древовидную структуру для принятия решений на основе последовательного применения набора правил. Оно может быть использовано для классификации или регрессии данных.

Принцип построения деревьев решений заключается в последовательном разбиении данных на более мелкие подмножества, до тех пор, пока каждое подмножество не будет однородным или достигнет заданного критерия остановки. Это позволяет построить дерево решений, где каждый узел представляет собой проверку значения некоторого признака, а каждый лист соответствует конечному результату, который можно использовать для принятия решений.

Процесс построения дерева решений может быть формализован с помощью алгоритма, который выбирает лучший признак для разбиения данных на каждом шаге. Для этого можно использовать различные критерии, такие как энтропия, прирост информации или индекс Джини, которые оценивают качество разбиения данных.

Построение дерева решений для прогнозирования по тестовому экземпляру начинается с проверки значения признака в корне дерева. Результат этого теста определяет, в какой из дочерних узлов корня необходимо перейти. Затем проверка значения признака и переход вниз по дереву

повторяются до тех пор, пока процесс не достигнет листа, на котором может быть сделано предсказание.

Важно учитывать, что построение деревьев решений может привести к переобучению, когда модель слишком точно подстраивается под обучающие данные и теряет способность обобщать новые. Для решения этой проблемы можно использовать различные методы, такие как ограничение глубины дерева, сокращение или отбор признаков, а также использование ансамблей деревьев решений, таких как случайный лес или градиентный бустинг.

Проблемы переобучения и недообучения являются ключевыми проблемами при построении моделей машинного обучения, в том числе и при использовании деревьев решений.

Переобучение возникает, когда модель слишком хорошо подстраивается под тренировочные данные и с высокой точностью предсказывает их значения, но при этом плохо работает на новых, ранее неизвестных данных. Это происходит, когда модель слишком сложна и адаптируется к шуму в тренировочных данных вместо того, чтобы обобщать закономерности.

Недообучение происходит, когда модель недостаточно обучена на тренировочных данных и не в состоянии уловить сложные зависимости в данных.

Для решения проблемы переобучения можно использовать методы регуляризации, такие как обрезание деревьев, добавление штрафов за сложность модели и т.д. Для решения проблемы недообучения можно использовать более сложные модели, более мощные алгоритмы оптимизации, увеличение объема тренировочных данных.

Важно понимать, что выбор модели зависит от конкретной задачи и доступных данных, и не существует универсального способа решения проблем переобучения и недообучения, но в общем случае нужно стараться достичь баланса между сложностью модели и ее способностью обобщать закономерности в данных.

4. Алгоритм ID3

Алгоритм итеративного дихотомизатора 3 (Iterative Dichotomiser 3 – ID3) является одним из классических алгоритмов машинного обучения, который используется для построения деревьев решений. Он был разработан Россом Кузе в 1960-х годах. ID3 является алгоритмом обучения с учителем, который основывается на концепции прироста информации (information gain).

Алгоритм ID3 строит дерево в рекурсивном порядке обхода в глубину, начиная с корня узла и заканчивая листами. Алгоритм начинается с выбора наилучшего признака для тестирования. Этот выбор делается путём вычисления прироста информации благодаря признакам в обучающем множестве. Затем корень добавляется в дерево и помечается выбранным проверяемым признаком. После этого обучающее множество разбивается на части с использованием теста. Для каждого возможного результата теста создаётся одно подмножество, которое содержит обучающие экземпляры, возвращающие этот результат. Для каждого подмножества из узла вырастает новая ветвь. Затем этот процесс повторяется для каждой ветви с использованием соответствующего подмножества обучающего множества после исключения соответствующего признака из дальнейшего тестирования. Этот процесс повторяется до тех пор, пока все экземпляры в подмножестве не будут иметь одинаковое значение целевого признака, и в этот момент создаётся лист, помеченный этим значением.

Особенностью алгоритма ID3 является механизм, используемый для определения того, какой признак является наиболее информативным для тестирования в новом узле. Алгоритм ID3 использует метрику прироста информации, чтобы выбрать лучший признак для тестирования на каждом узле дерева. Следовательно, выбор лучшего признака для разделения множества данных основан на однородности результирующих подмножеств в множествах данных. В результате прирост информации для определённого признака может отличаться на разных узлах в дереве. Одним из следствий этого является то, что признак с низким приростом информации в корне может иметь высокий показатель прироста информации в одном из внутренних узлов, поскольку он является прогнозирующим в подмножестве рассматриваемых экземпляров в этом внутреннем узле.

Алгоритм ID3 может использоваться для решения задач классификации и прогнозирования. Он может быть эффективным при работе с данными, которые имеют небольшое количество признаков и где признаки являются категориальными. Однако, если признаки являются числовыми, то необходимо использовать алгоритмы, способные работать с такими данными, например, алгоритм C4.5, который является усовершенствованной версией ID3.

Алгоритм ID3 может быть применен на различных типах данных, включая категориальные и числовые данные. Рассмотрим несколько примеров его применения:

1. Классификация пациентов на основе медицинских данных:

В этом примере, алгоритм может быть использован для классификации пациентов на группы в зависимости от их состояния здоровья. Входными данными являются медицинские показатели, такие как пульс, кровяное давление, уровень сахара в крови и т.д. Алгоритм использует эти данные для построения дерева решений, которое помогает врачам принимать более точные решения в отношении диагностики и лечения пациентов.

2. Классификация посетителей веб-сайта на основе действий:

В этом примере, алгоритм может быть использован для классификации посетителей веб-сайта на группы в зависимости от их действий, например, покупок или регистраций на сайте. Входными данными являются действия посетителей, такие как клики на определенные элементы, время нахождения на страницах, данные о покупках и т.д. Алгоритм использует эти данные для построения дерева решений, которое помогает владельцам веб-сайтов принимать более эффективные маркетинговые решения.

3. Классификация рисков при инвестировании на основе финансовых данных:

В этом примере, алгоритм может быть использован для классификации инвестиционных возможностей на группы в зависимости от их риска. Входными данными являются финансовые показатели, такие как прибыль, уровень инфляции, стоимость акций и т.д. Алгоритм использует эти данные для построения дерева решений, которое помогает инвесторам принимать более обоснованные решения по выбору инвестиционных возможностей.

Кроме того, алгоритм ID3 может быть использован в любых других областях, где необходимо классифицировать данные на основе их признаков.

Анализ времени работы алгоритма ID3 является важной частью его оценки производительности. Время работы алгоритма зависит от размера обучающей выборки, количества признаков и числа возможных значений каждого признака.

В худшем случае, когда каждый элемент обучающей выборки уникален по значениям всех признаков, алгоритм ID3 может построить дерево решений с глубиной, равной количеству элементов в обучающей выборке. Это может привести к значительному увеличению времени работы алгоритма и к переобучению модели.

Для уменьшения времени работы и избежания переобучения, можно использовать различные методы оптимизации, например, прореживание дерева (pruning) или ограничение глубины дерева. Также можно использовать более эффективные алгоритмы, такие как C4.5 и CART, которые являются усовершенствованными версиями алгоритма ID3.

В целом, алгоритм ID3 показывает хорошие результаты на средних и маленьких обучающих выборках, однако на больших выборках может быть неэффективен. Поэтому, при выборе алгоритма для конкретной задачи, необходимо учитывать размер выборки и характеристики данных.

5. Реализация алгоритма ID3 для классификации данных на языке Python

Для реализации алгоритма ID3 был использован язык программирования Python. В качестве основных инструментов были использованы стандартные библиотеки Python, такие как numpy и pandas.

Библиотека numpy была использована для работы с массивами данных, матрицами и вычислений. Библиотека pandas - для анализа и обработки данных.

В целом, для реализации алгоритма ID3 использовались основные инструменты и библиотеки, доступные в языке Python.

Подготовка данных для обучения. Для подготовки данных для обучения в машинном обучении важно, чтобы данные были в правильном формате и содержали необходимую информацию. В данном случае мы рассмотрим подготовку данных для обучения дерева решений с помощью библиотеки pandas.

При работе с таблицей данных в формате xls, которая содержит результаты обследований голосовых характеристик, можно использовать библиотеку pandas для чтения данных из файла и преобразования их в удобный формат. Для этого используем следующий синтаксис:

```
train_data = pd.read_excel('train.xlsx')
```

После выполнения этой команды, данные из файла Excel будут загружены в DataFrame `train_data`. Теперь мы можем использовать этот DataFrame для анализа данных, включая обработку, визуализацию и обучение моделей машинного обучения.

Затем данные необходимо преобразовать из числового типа в категориальный. Для этого воспользуемся методом .qcut из библиотеки Pandas, применяя его ко всем столбцам таблицы train_data. Например, для столбца ‘Jitter’ реализация метода .qcut будет выглядеть следующим образом:

```
train_data['Jitter']  = pd.qcut(train_data['Jitter'], q=6,
                                labels=['1', '2', '3', '4', '5', '6'])
```

Здесь числовой признак "Jitter" разбивается на 6 интервалов равной длины.

Применяем данный метод ко всем столбцам и получаем измененную таблицу, которую можно использовать для построения дерева решений.

Реализация алгоритма. После чтения и обработки данных можно начать реализацию алгоритма.

Сперва необходимо реализовать функции для выбора лучшего критерия для разделения множества на наиболее однородные подмножества:

1. Вычисление энтропии всего набора данных.

2. Вычисление энтропии для отфильтрованного набора данных.

3. Вычисление прироста информации.

4. Поиск наиболее информативного признака (признак с наибольшим проростом информации).

Затем необходимо добавить узел, помеченный признаком с наибольшим приростом информации. Если любое значение признака представляет только один класс, то можно сказать, что значение признака представляет чистый (однородный) класс. Если значение признака не представляет чистый класс, придется расширять его дальше, пока мы не найдем чистый класс.

После выбора чистого класса мы должны удалить строки из набора данных, соответствующие значению выбранного признака.

Процесс построения дерева решений согласно алгоритму ID3 можно представить, как рекурсивное пошаговое выполнение следующих действий:

1) Поиск наиболее информативного признака

2) Создание узла дерева с именем признака и значениями признака в качестве ветвей.

- Если класс чистый, добавление листового (конечного) узла к узлу дерева
- Если класс нечистый, добавление расширяемого узла к узлу дерева

3) Сокращение/обновление набора данных в соответствии с чистым классом

4) Добавление узла с ветвями в дерево

5) Расширение ветви следующего нечистого класса с обновленным набором данных

Условия выхода из рекурсии:

- Набор данных становится пустым после обновления
- Нет расширяемой ветви (все классы чистые)

Оценка точности и эффективности реализации. Для оценки точности и эффективности реализации алгоритма ID3 можно использовать метрики, такие как точность (accuracy), полноту (recall), точность предсказания положительного класса (precision), F1-меру (F1-score) и ROC-кривую.

Для этого необходимо подготовить данные для тестирования, которые должны быть разделены на обучающую и тестовую выборки. Обучающая выборка используется для обучения модели, а тестовая выборка - для оценки ее эффективности и точности.

Для оценки эффективности и точности модели можно использовать кросс-валидацию, которая позволяет оценить ее работу на различных разбиениях данных.

Также можно провести анализ времени работы алгоритма на различных объемах данных и оптимизировать его для улучшения скорости работы.

В целом, рекомендуется использовать алгоритм ID3 для задач классификации на небольших объемах данных, когда требуется простая и понятная модель, которую легко интерпретировать. Однако, при работе с большими объемами данных или задачами, требующими высокой точности, рекомендуется использовать более сложные и точные алгоритмы, такие как C4.5 или CART.

6. Заключение

Основные результаты исследования. К основным результатам нашего исследования относятся:

1. Реализация алгоритма ID3 для задачи классификации на языке Python.

2. Применение алгоритма на примере реальных данных обследований речевых характеристик.

3. Анализ времени работы алгоритма на различных объемах данных.

4. Обнаружение того, что алгоритм ID3 имеет хорошую точность классификации данных, но при этом может быть неэффективен на больших объемах данных из-за вычислительной сложности.

5. Выявление того, что алгоритм ID3 не всегда лучший выбор для задач классификации, так как существуют другие алгоритмы, которые могут иметь более высокую точность классификации и/или более эффективно работать на больших объемах данных.

Выводы и рекомендации по применению алгоритма ID3.

Выводы:

- Алгоритм ID3 является простым и эффективным инструментом для построения деревьев решений на основе информационного обучения.

- ID3 может быть использован для классификации данных в различных областях, включая медицину, бизнес, финансы и другие.

- Для достижения наилучших результатов при использовании алгоритма ID3 необходимо правильно подготовить данные для обучения, выбрать наиболее информативные признаки и оптимально задать пороговые значения для разбиения данных на подгруппы.

- Важным фактором при использовании алгоритма ID3 является контроль за переобучением модели.

Рекомендации:

- При использовании алгоритма ID3 рекомендуется тщательно подготовить данные для обучения, провести предварительный анализ данных и выбрать наиболее информативные признаки для построения дерева решений.
- Необходимо учитывать, что алгоритм ID3 может быть чувствителен к шуму в данных, поэтому желательно провести их предварительную обработку.
- Для контроля за переобучением модели рекомендуется использовать кросс-валидацию и регуляризацию.
- При работе с большими объемами данных, необходимо учитывать время работы алгоритма ID3, который может быть неоптимальным для больших и сложных данных.
- Рекомендуется проводить дополнительные эксперименты с различными параметрами алгоритма и альтернативными методами классификации данных для достижения наилучших результатов.

Список использованных источников:

1. *Основы машинного обучения для аналитического прогнозирования: алгоритмы, рабочие примеры и тематические исследования: Пер. с англ. / Джон Д. Келлехер, Брайан Мак-Нейми, Аоифе д'Арси – СПб.: ООО «Диалектика», 2019. – 656с: ил.*

UDC 004.855

PYTHON IMPLEMENTATION OF THE ID3 ALGORITHM

Habryielchyk P.V. & Ermakovich V.A.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Vashkevich M.I. – PhD

Annotation. This research paper is devoted to the study of the basics of machine learning. It explores the principles of machine learning and presents an implementation of the ID3 algorithm as a way to build a decision tree for data classification. The implementation of the algorithm is done in Python. The experimental results show the effectiveness and accuracy of the ID3 algorithm in solving classification problems. The research allows us to expand our understanding of the basics of machine learning and to apply the obtained knowledge in practice.

Keywords. Machine learning, Artificial Intelligence, information theory, ID3 algorithm, entropy, information learning, information gain, decision tree, data classification, overtraining and undertraining.