Ансамбли моделей [M.190]

Введение в ансамбли моделей

При создании алгоритмов машинного обучения разработчики сталкиваются с такими проблемами, как вычислительные затраты на реализацию алгоритма, прозрачность построенных моделей для пользователя, а также точность результатов. Большинство исследователей сосредотачиваются на повышении точности классификации и предсказания, поэтому производительность новых систем часто рассматривается именно с этой точки зрения. И это легко понять: точность играет решающую роль во всех приложениях машинного обучения и может быть легко оценена, в то время как прозрачность для пользователя субъективна. Что касается вычислительных затрат, то с прогрессом вычислительной техники они во многих случаях вообще отошли на задний план.

В последние несколько лет значительно возрос интерес к вопросу увеличения точности моделей Data Mining, основанных на обучении, за счет создания и агрегирования набора классификаторов. В результате появились новые подходы к анализу на базе ансамблей классификаторов, которые применимы к широкому кругу систем машинного обучения и основаны на теоретическом исследовании поведения составных классификаторов.

Комбинирование решений

Принимая важное решение, опытный руководитель не только полагается на собственные знания и интуицию, но и старается привлечь экспертов в конкретных предметных областях. Считается, что выводы экспертов, полученные на основе анализа данных, связанных с решаемой задачей, позволят сделать оптимальный выбор.

Например, когда предприятие планирует выпуск нового продукта, взвешиваются все «за» и «против», привлекаются специалисты в сферах экономики, маркетинга, производства, рекламы и т. д., и каждый из них высказывает свое мнение. Выслушав всех, руководитель приходит к окончательному решению.

Однако выводы, сделанные разными экспертами, могут противоречить друг другу. Поэтому неизбежно встает вопрос: как скомбинировать несколько экспертных оценок, чтобы на их основе принять правильное решение? В простейших случаях руководитель может, не прибегая к формальным методам, просто воспользоваться своими опытом и интуицией.

Существуют методы, которые позволяют принять более обоснованное решение. Допустим, в качестве оценки определенного решения эксперт указывает число баллов: 100 баллов соответствуют положительному решению («выпускать продукт»), 0 баллов — отрицательному («не выпускать продукт»). Указывая промежуточные значения, эксперты оценивают вероятность успеха или провала проекта. Полученные оценки суммируется, и если сумма превысит заданный порог, то решение будет положительным, а если нет — отрицательным. При желании можно произвести усреднение оценок, то есть разделить сумму баллов на число экспертов.

Хорошие результаты часто дает взвешивание оценок: в зависимости от важности предметной области, профессионального уровня эксперта и т. д. экспертной оценке присваивается определенный вес. Чем опытнее эксперт и важнее оценка, тем больше ее вес.

Наконец, самым простым, хотя и не всегда эффективным методом является голосование: выбирается решение, принятое большинством голосов.

Остановимся еще на одном важном моменте — доверии к эксперту. Действительно, эксперты отличаются уровнем подготовки, опытом и т. д. В этой связи их можно разделить на сильных и слабых. Например, слабым может считаться эксперт, который участвовал в 10 проектах, причем в 7 случаях его мнение оказалось ошибочным.

Используются два варианта экспертных оценок: параллельные и дополняющие. В первом случае каждый эксперт высказывает мнение по всему спектру проблем, связанных с решаемой задачей. Во втором каждый эксперт дает заключение только по одному аспекту задачи. Тогда выводы экспертов дополняют друг друга, вместе покрывая весь спектр проблем.

Таким образом, выбирая методы извлечения экспертных оценок и комбинируя полученные результаты, можно найти наилучшее решение.

Аналогичная ситуация складывается при использовании моделей, основанных на машинном обучении, — деревьев решений, нейронных сетей и т. д. Эти модели фактически играют роль экспертов и в явном или неявном виде предоставляют информацию, необходимую для обоснованного принятия решений.

Можно ограничиться результатами, полученными единственной моделью. Но, также как и эксперты, модели бывают слабыми и сильными. Если обученной модели хорошо удается разделить классы и она допускает мало ошибок классификации, то такая модель может рассматриваться как сильная. Слабая модель, напротив, не позволяет надежно разделять классы или давать точные предсказания, допускает в работе большое количество ошибок.

Если в результате обучения мы получили «слабую» модель, то ее необходимо усовершенствовать, подбирая тип классификатора, алгоритмы и параметры обучения. Но часто встречаются ситуации, когда все возможности совершенствования единственной модели исчерпаны, а качество ее работы по-прежнему неудовлетворительно. Это может быть связано со сложностью решаемой задачи и искомых закономерностей, с низким качеством обучающих данных и другими факторами. Пример такой ситуации представлен на рисунке 1.

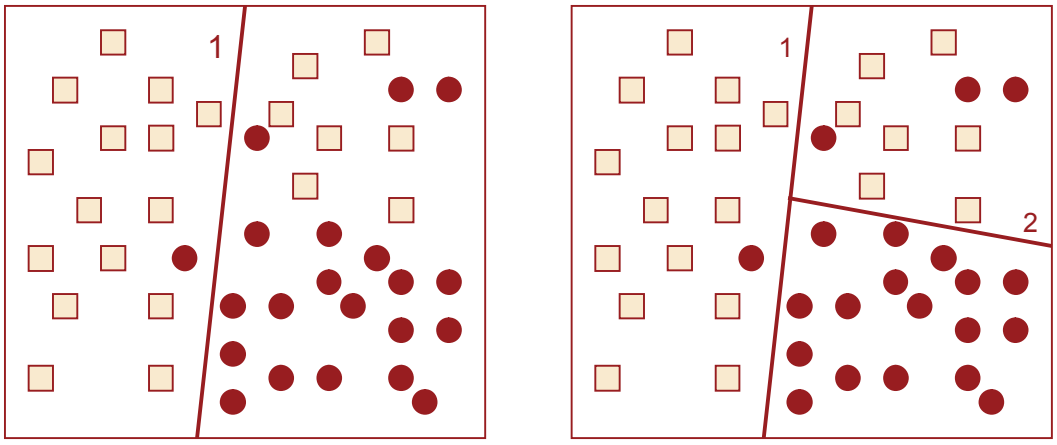

Рисунок 1– Иллюстрация сложной разделимости классов

На рисунке 1а изображено множество объектов, содержащее два класса — круги и квадраты. В результате обучения построена модель, разделившая классы по линии 1. Поскольку классы пересекаются, они являются трудноразделимыми, из-за чего значительное количество квадратов оказалось распознано как круги. Ожидать от такой модели хорошей работы с новыми данными не следует, при этом ее возможности уже исчерпаны: как бы мы ни провели линию 1, всегда будет присутствовать значительная ошибка классификации.

Неизбежно возникает вопрос: как усилить «слабую» модель, что сделать для повышения эффективности классификации? Вполне логичным выходом из ситуации является попытка применить к неудачным результатам работы первой модели еще одну модель, задача которой — классифицировать те примеры, что остались нераспознанными. Результаты работы второй модели представлены на рисунке 1б. Как видно, ей удалось почти полностью разделить круги и квадраты. Если и после этого результаты неудовлетворительны, можно применить третью модель и т. д. до тех пор, пока не будет получено достаточно точное решение.

Таким образом, для решения одной задачи классификации или регрессии мы применили несколько моделей, при этом нас интересует не результат работы каждой отдельной модели, а

результат, который дает весь набор моделей. Такие совокупности моделей называются ансамблями моделей.

Определение

> *Набор моделей, применяемых совместно для решения единственной задачи, называется ансамблем (комитетом) моделей.*

Набор моделей, применяемых совместно для решения единственной задачи, называется ансамблем (комитетом) моделей.

Цель объединения моделей очевидна — улучшить (усилить) решение, которое дает отдельная модель. При этом предполагается, что единственная модель никогда не сможет достичь той эффективности, которую обеспечит ансамбль.

Использование ансамблей вместо отдельной модели в большинстве случаев позволяет повысить качество решений, однако такой подход связан с рядом проблем, основными из которых являются:

- увеличение временных и вычислительных затрат на обучение нескольких моделей;
- сложность интерпретации результатов;
- неоднозначный выбор методов комбинирования результатов, выдаваемых отдельными моделями.

Перечисленные проблемы аналогичны тем, что возникают при работе нескольких людей-экспертов. Действительно, приходится собирать группу экспертов, предоставлять им необходимую информацию, обсуждать задачу и т. д. — все это отнимает намного больше времени, чем принятие решения одним человеком. Сложность интерпретации результатов экспертных оценок также имеет место, ведь каждый эксперт оперирует терминами своей предметной области, формулирует выводы на уровне своего понимания проблемы. И наконец, выбор метода обобщения отдельных заключений экспертов, позволяющего получить наилучшие результаты, не является однозначным.

Виды ансамблей

В последнее десятилетие ансамбли моделей стали областью очень активных исследований в машинном обучении, что привело к разработке большого числа разнообразных методов формирования ансамблей.

Первым вопросом при формировании ансамбля является выбор базовой модели (base model). Ансамбль в целом может рассматриваться как сложная, составная модель (multiple model), состоящая из отдельных (базовых) моделей. Здесь возможны два случая.

1. Ансамбль состоит из базовых моделей одного типа, например только из деревьев решений, только из нейронных сетей и т. д. (рисунок 2).

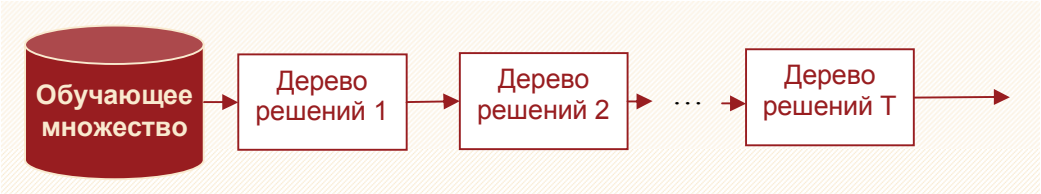

Рисунок 2 – Однородный ансамбль

2. Ансамбль состоит из моделей различного типа — нейронных сетей, деревьев решений, регрессионных моделей и т. д. (рисунок 3).

Рисунок 3 – Ансамбль, состоящий из моделей различного типа

Каждый подход имеет свои преимущества и недостатки. Использование моделей различных типов дает классификатору дополнительную гибкость. Но, поскольку выход одной модели применяется для формирования обучающего множества для другой, возможно, потребуются дополнительные преобразования, чтобы согласовать входы и выходы моделей.

Второй вопрос: как использовать обучающее множество при построении ансамбля? Здесь также существуют два подхода.

1 *Перевыборка* (resampling). Из исходного обучающего множества извлекается несколько подвыборок, каждая из которых используется для обучения одной из моделей ансамбля. Данный подход иллюстрируется с помощью рисунка 4.

Если ансамбль строится на основе моделей различных типов, то для каждого типа будет свой алгоритм обучения.

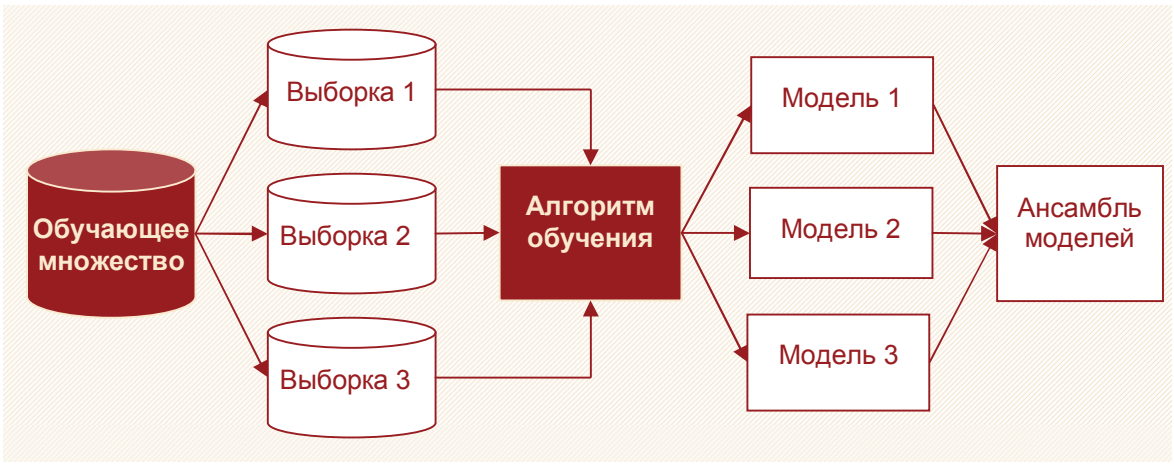

Рисунок 4 – Перевыборка

2 *Использование одного обучающего множества для обучения всех моделей ансамбля* (рисунок 5).

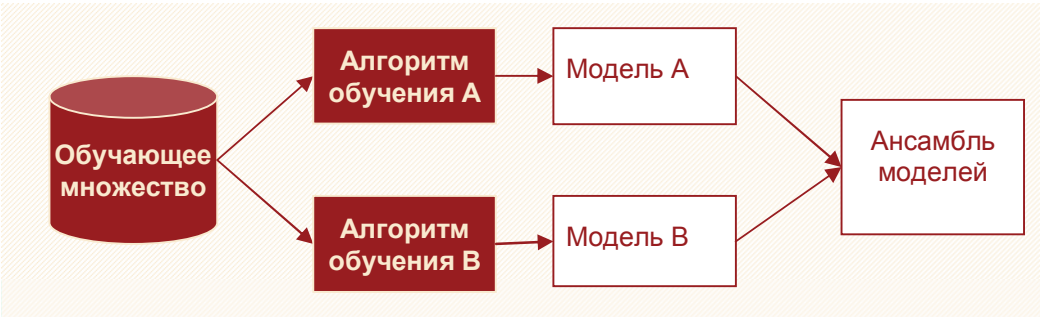

Рисунок 5 – Использование одного обучающего множества для всех моделей ансамбля

И наконец, третий вопрос касается метода комбинирования результатов, выданных отдельными моделями: что будет считаться выходом ансамбля при определенных состояниях выходов моделей? Обычно используются следующие способы комбинирования.

1. *Голосование.* Применяется в задачах классификации, то есть для категориальной целевой переменной. Выбирается тот класс, который был выдан простым большинством моделей ансамбля. Пусть, например, решается задача бинарной классификации с целевыми переменными Да и Нет, для чего используется ансамбль, состоящий из трех моделей. Если две модели выдали выход Нет и только одна — Да, то общий выход ансамбля будет Нет.
2. *Взвешенное голосование.* В ансамбле одни модели могут работать лучше, а другие — хуже. Соответственно, к результатам одних моделей доверия больше, а к результатам других — меньше. Чтобы учесть уровень достоверности результатов, для моделей ансамбля могут быть назначены веса (баллы). Например, в случае, рассмотренном в предыдущем пункте, для моделей, выдавших результат Нет, установлены веса 30 и 40, указывающие на невысокую достоверность этих результатов. В то же время единственная модель, которая выдала Да, имеет вес 90. Тогда голосование будет производиться с учетом весов моделей: $30\ (\text{Нет}) + 40\ (\text{Нет}) = 70\ (\text{Нет}) \leq 90\ (\text{Да})$. Таким образом, модель с выходом Да перевесила обе модели с выходом Нет, и общий выход ансамбля будет Да.
3. *Усреднение (взвешенное или невзвешенное).* Если с помощью ансамбля решается задача регрессии, то выходы его моделей будут числовыми. Выход всего ансамбля может определяться как простое среднее значение выходов всех моделей. Например, если в ансамбле три модели и их выходы равны y_1, y_2 и y_3, то выход ансамбля будет $Y = (y_1 + y_2 + y_3)/3$ (для произвольного числа моделей $Y = (y_1 + y_2 + \ldots + y_K)/K$, где K — число моделей в ансамбле). Если производится взвешенное усреднение, то выходы моделей умножаются на соответствующие веса.

Исследования ансамблей моделей в Data Mining стали проводиться относительно недавно. Тем не менее к настоящему времени разработано множество различных методов и алгоритмов формирования ансамблей. Среди них наибольшее распространение получили такие методы, как бэггинг, бустинг и стэкинг.