История развития ансамблевых методов классификации в машинном обучении

Research · June 2015

DOI: 10.13140/RG.2.1.3933.2007

1 author:

Yury Kashnitsky

National Research University Higher School of Economics

8 PUBLICATIONS **1** CITATION

История развития ансамблевых методов классификации в машинном обучении

Кашницкий Ю. С.

Национальный Исследовательский Университет
Высшая Школа Экономики
Москва
ykashnitsky@hse.ru

Аннотация Рассматривается история развития машинного обучения - прикладного направления искусственного интеллекта, изучающего системы, способные обучаться на базе данных и делать на этой основе прогнозы. Приводится описание основных направлений, задач и алгоритмов машинного обучения. Отдельно рассматриваются ансамблевые методы классификации - методы, объединяющие несколько менее качественных моделей или алгоритмов классификации в одну более качественную модель.

1 Введение

Основная задача науки и реальной жизни – получение правильных предсказаний о будущем поведении сложных систем на основании их прошлого поведения. Многие задачи, возникающие в практических приложениях, не могут быть решены заранее известными методами или алгоритмами. Это происходит по той причине, что нам заранее не известны механизмы порождения исходных данных или же известная нам информация недостаточна для построения модели источника, генерирующего поступающие к нам данные. Как говорят, мы получаем данные из «черного ящика». В этих условиях ничего не остается, как только изучать доступную нам последовательность исходных данных и пытаться строить предсказания совершенствуя нашу схему в процессе предсказания. Подход, при котором прошлые данные или примеры используются для первоначального формирования и совершенствования схемы предсказания, назы- вается методом машинного обучения (Machine Learning).

Артур Самуэль был пионером в области компьютерных игр и искусственного интеллекта [1]. Его программа Checkers-playing - одна из первых самообучающаяся программ в мире и одна из первых демонстраций искусственного интеллекта. Основа обучения машины была дерево поиска положений, достижимых от текущего состояния. Самуэль развивал функцию выигрыша, основанную на положении в любой момент времени. Эта функция попыталась измерить шанс победы для каждой стороны в данном положении. То что Самуэль назвал «зубрежкой» - когда программа помнила

каждое положение фигур, которое уже видело, наряду с предельной ценностью функции награды - как раз и есть основа машинного обучения. Его метод изучения через игры продолжался при длительной работе над шашками и в других играх, таких как шахматы, а теоретические основы игры дали толчок к развитию теории машинного обучения.

Машинное обучение имеет дело с обучаемыми алгоритами, то есть, как определено изначально Самуэлем [2], алгоритмами, которые могут обучаться не будучи явно запрограммированными. Впоследствии Том Митчелл [3] формализовал это определение:

Алгоритм *обучается* на основе опыта E по отношению к некоторому классу задач T и меры качества P, если качество решения задач из T улучшается (по мере P) с приобретением опыта E.

Обучаясь таким образом на данных, машинное обучение помогает обнаружить зависисмости в данных и превратить данные в информацию. Сегодня машинное обучение активно используется во многих областях нашей жизни в таких задачах как распознавание речи, жестов, рукописного текста и образов, техническая диагностика, медицинская диагностика, обнаружение мошенничества и спама, категоризация документов, финансовый надзор, кредитный скоринг, предсказание ухода клиентов, ранжирование в информационном поиске и во многих других.

2 Основные задачи машинного обучения

2.1 Общая постановка задачи

Имеется множество объектов (ситуаций) и множество возможных ответов (откликов, реакций). Существует некоторая зависимость между ответами и объектами, но она неизвестна. Известна только конечная совокупность прецедентов — пар «объект, ответ», называемая *обучающей выборкой*. На основе этих данных требуется восстановить зависимость, то есть построить алгоритм, способный для любого объекта выдать достаточно точный ответ. Для измерения точности ответов определённым образом вводится функционал качества.

Далее перечислены наиболее часто встречающиеся типы задач машинного обучения.

2.2 Обучение с учителем

Обучение с учителем (supervised learning) — наиболее распространённый случай. Каждый прецедент представляет собой пару «объект, ответ». Требуется найти функциональную зависимость ответов от описаний объектов и построить алгоритм, принимающий на входе описание объекта и выдающий на выходе ответ. Функционал качества обычно определяется как средняя ошибка ответов, выданных алгоритмом, по всем объектам выборки. Типы задач в обучении с учителем[1]:

[1] http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf

– Классификация отличается тем, что множество допустимых ответов конечно. Их называют метками классов. Класс — это множество всех объектов с данным значением метки.
– В задаче регрессии допустимым ответом является действительное число или числовой вектор.
– Задача ранжирования отличается тем, что ответы надо получить сразу на множестве объектов, после чего отсортировать их по значениям ответов. Может сводиться к задачам классификации или регрессии. Часто применяется в информационном поиске и анализе текстов.
– Задача прогнозирования характеризуется тем, что объектами являются отрезки временных рядов, обрывающиеся в тот момент, когда требуется сделать прогноз на будущее. Для решения задач прогнозирования часто удаётся приспособить методы регрессии или классификации, причём во втором случае речь идёт скорее о задачах принятия решений.

2.3 Обучение без учителя

Обучение без учителя (unsupervised learning). В этом случае ответы не задаются, и требуется искать зависимости между объектами.

Типы задач в обучении без учителя:

– Задача кластеризации заключается в том, чтобы сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов. Функционалы качества могут определяться по-разному, например, как отношение средних межкластерных и внутрикластерных расстояний.
– Задача поиска ассоциативных правил. Исходные данные представляются в виде признаковых описаний. Требуется найти такие наборы признаков и такие значения этих признаков, которые особенно часто встречаются в признаковых описаниях объектов.
– Задача фильтрации выбросов — обнаружение в обучающей выборке небольшого числа нетипичных объектов. В некоторых приложениях их поиск является самоцелью (например, обнаружение мошенничества). В других приложениях эти объекты являются следствием ошибок в данных или неточности модели, то есть шумом, мешающим настраивать модель, и должны быть удалены из выборки.
– Задача построения доверительной области — области минимального объёма с достаточно гладкой границей, содержащей заданную долю выборки.
– Задача сокращения размерности заключается в том, чтобы по исходным признакам с помощью некоторых функций преобразования перейти к наименьшему числу новых признаков, не потеряв при этом никакой существенной информации об объектах выборки. В классе линейных преобразований наиболее известным примером является метод главных компонент.
– Задача заполнения пропущенных значений — замена недостающих значений в матрице объекты–признаки их прогнозными значениями.

2.4 Частичное обучение

Частичное обучение (semi-supervised learning) занимает промежуточное положение между обучением с учителем и без учителя. Каждый прецедент представляет собой пару «объект, ответ», но ответы известны только на части прецедентов. Пример прикладной задачи — автоматическая рубрикация большого количества текстов при условии, что некоторые из них уже отнесены к каким-то рубрикам.

3 Ансамблевые методы

3.1 Сравнительная модель прогнозирования

Задача принятия правильного рационального решения является центральной в науке и практике. Решение принимается на основе некоторых наблюдаемых данных. Правильный прогноз или правильное решение ведут к меньшим потерям, чем неправильные. При традиционном статистическом подходе мы оцениваем потери при наших прогнозах в сравнении с некоторой идеальной моделью принятия правильных решений, которая обычно основана на некоторой статистической модели, описывающей наблюдаемые данные. При традиционном подходе сначала оцениваются параметры статистической модели на основе наблюдений, а потом производится прогноз на основе этой модели при оцененных параметрах.

При сравнительном подходе вместо одной идеальной модели рассматривается набор возможных моделей, которые называются конкурирующими экспертными стратегиями, или просто, экспертами[2]. Множество таких экспертных стратегий может быть конечным или бесконечным и даже несчетным. Используя исходы, поступающие в режиме онлайн, экспертные стратегии производят прогнозы будущего исхода. Прогнозирующий алгоритм может наблюдать прогнозы конкурирующих экспертных стратегий и оценивать их эффективность в прошлом. После этого алгоритм делает свой прогноз.

3.2 Ансамблевые методы классификации

В задаче классификации алгоритм, или классификатор, называется *слабым*, если его ошибка на обучающей выборке меньше 50%, но более 0%. В случае бинарной классификации (то есть когда классов всего два) можно сказать, что классификатор слабый, если он не намного лучше, чем простое случайное «угадывание». Если ошибка классификатора на обучающей выборке может быть уменьшена до значения, сколь угодно близкого к 0%, за полиномиальное время, то тогда классификатор называется *сильным*. [4]

Применительно к задачам классификации, парадигма сравнительного прогнозирования воплощается в ансамблевых методах классификации. Общая идея: с помощью нескольких слабых классификаторов создать такое

[2] http://www.iitp.ru/upload/publications/5759/vyugin1.pdf

решающее правило, с помощью которого можно повысить точность предсказания и сделать таким образом один сильный мета-классификатор.

Интуитивно это можно представить так: при наличии пяти классификаторов с точностью предсказания 60% можно назначать классифицируемому объекту ту метку, которая предсказывается большинством из этих пяти слабых классификаторов. Это ансамблевый метод классификации простым голосованием (или голосованием большинством). В таком случае, как это легко показать, точность мета-классификатора повышается до 68%.

Простое голосование большинством - далеко не единственный способ объединения предсказания базовых классификаторов. В работе [5] перечисляются 18 типов таких алгоритмов, среди которых также голосование методом Борда (Borda count), нахождение «средней» и «серединной» метки класса, адаптивное взвешивание «мнений» более точных классификаторов, ранжирование с последующей логистической регрессией на множестве рангов, бэггинг, стекинг, бустинг, смесь локальных экспертов (Mixture of Local Experts, MLE) и прочие.

Среди множества ансамблевых методов классификации рассмотрим более подробно три типа:

1. бэггинг (bagging, bootstrap aggregating)
2. бустинг (boosting)
3. стекинг (stacking)

3.3 Бэггинг

Исследователи, занимающиеся статистикой уже давно используют метод, называемый «bootstrap sampling», что условно может быть переведено на русский как «вариация загрузочной выборки». Одно из воплощений этой идеи в машинном обучении - «bootstrap aggregating», или сокращенно «bagging», то есть «объединение результатов при различных загрузках» [6].

Идея бэггинга в том, что при отсутствии большой обучающей выборки можно создавать много случайных выборок из исходной простым выбором с замещением. Хотя элементы в выборках могут пересекаться или дублироваться, на практике все же результаты объединения по многим выборкам оказываются точнее, чем только по одной начальной. Метод так называется, поскольку он объединяет результаты предсказания различных классификаторов, обученных на случайных подмножествах.

Бэггинг оказывается полезен только в случае разных классификаторов и нестабильности, когда малые изменения в начальной выборке приводят к существенным изменениям классификации [7].

3.4 Бустинг

В работе 1984 года [8] были представлены теоретические основы PAC-модели вероятно почти корректного обучения (Probably Approximately Correct),

при котором рассматривалась возможность улучшить алгоритм классификации с помощью нескольких слабых классификаторов.

В 1989 году Шапир первым придумал такой алгоритм с полиномиальной сложностью и опубликовал в статье с ярким названием «Сила слабого обучения» [9], а через год Фрейнд разработал более эффективную реализацию [10], которая стала основой алгоритма AdaBoost, представленного в 1995 году Шапиром и Фрейндом уже вместе.

Бустинг (boosting, улучшение) — это процедура последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов. Бустинг представляет собой жадный алгоритм построения композиции алгоритмов. Изначально понятие бустинга возникло в работах по вероятно почти корректному обучению в связи с вопросом: возможно ли, имея множество плохих (незначительно отличающихся от случайных) алгоритмов обучения, получить хороший.

В течение последних 10 лет бустинг остаётся одним из наиболее популярных методов машинного обучения, наряду с нейронными сетями и машинами опорных векторов. Основные причины — простота, универсальность, гибкость (возможность построения различных модификаций), и, главное, высокая обобщающая способность.

Бустинг над решающими деревьями считается одним из наиболее эффективных методов с точки зрения качества классификации. Во многих экспериментах наблюдалось практически неограниченное уменьшение частоты ошибок на независимой тестовой выборке по мере наращивания композиции. Более того, качество на тестовой выборке часто продолжало улучшаться даже после достижения безошибочного распознавания всей обучающей выборки. Это перевернуло существовавшие долгое время представления о том, что для повышения обобщающей способности необходимо ограничивать сложность алгоритмов. На примере бустинга стало понятно, что хорошим качеством могут обладать сколь угодно сложные композиции, если их правильно настраивать.

Впоследствии феномен бустинга получил теоретическое обоснование. Оказалось, что взвешенное голосование не увеличивает эффективную сложность алгоритма, а лишь сглаживает ответы базовых алгоритмов. Количественные оценки обобщающей способности бустинга формулируются в терминах отступа. Эффективность бустинга объясняется тем, что по мере добавления базовых алгоритмов увеличиваются отступы обучающих объектов. Причём бустинг продолжает раздвигать классы даже после достижения безошибочной классификации обучающей выборки.

К сожалению, теоретические оценки обобщающей способности дают лишь качественное обоснование феномену бустинга. Хотя они существенно точнее более общих оценок Вапника-Червоненкиса, всё же они сильно завышены, и требуемая длина обучающей выборки оценивается величиной порядка

$10^4 \dots 10^6$. Более основательные эксперименты показали, что иногда бустинг всё же переобучается [3].

3.5 Стекинг

Стековое обобщение, или просто стекинг, - еще один способ объединения классификаторов, вводящий понятие мета-алгоритма обучения. В отличие от бэггинга и бустинга, при стекинге используются классификаторы разной природы. Идея стекинга такова:

1. разбить обучающую выборку на два непересекающихся подмножества
2. обучить несколько базовых классификаторов на первом подмножестве
3. тестировать базовые классификаторы на втором подмножестве
4. используя предсказания из предыдущего пункта как входные данные, а истинные классы объектов как выход, обучить мета-алгоритм обучения.

4 История развития ансамблевых методов классификации

Далее рассмотрим ключевые наработки в области ансамблевых методов классификации за последнее время [11], [7], [12].

1988 В 1988 году Уиттнер и Денкер предлагают метод обучения многоуровневой нейронной сети в задаче классификации.

1990 Шапир изобретает бустинг.

Кляйнберг публикует теоретическую статью с описанием общего метода разделения точек в многомерном пространстве с помощью так называемой стохастической дискриминации. Метод оказался многообещающим и позже, в 1998 году, послужил основой для метода случайного подпространства (random subspace method).

Хансен и Саломон продемонстрировали выгоду использования ансамблей схожих нейронных сетей.

1992 Уолперт предлагает метод стекового обобщения (stacked generalization), который мимнимизирует суммарную ошибку нескольких обобщающих функций.

Крыжак и Суен рассматривают объединение нескольких классификаторов и их применение в задаче распознавания рукописного текста. Они заявляют, что по типу объединения результатов отдельных классификаторов, комбинированные алгоритмы делятся на 3 типа:

[3] http://www.machinelearning.ru/wiki/index.php?title=Бустинг

– использующие голосование
– использующие формализм Байеса
– использующие формализм Демпстера-Шафера

Было показано, что точность классификации отдельно взятого алгоритма может быть значительно повышена, и для этого целесообразней использовать формализм Демпстера-Шафера, так как при этом помимо высокой точности также поддерживается высокая полнота и робастность.

1993 Перрон и Купер представляют свои теоретические разработки, позволяющие значительно улучшать точность регрессионныйх моделей с помощью ансамлевых методов.

Джордан и Якобс представляют иерархическую смесь экспертных моделей.

1994 Хо, Халл и Ширари предлагают множественную модель классификаторов на основе ранжирования.

Баттити и Колла обнаруживают, что сочетание небольшого числа нейронных сетей (двух или трех) с плохо коррелирующими ошибками позволяет достичь значительно большей точности, чем у сетей по отдельности.

1995 Чо и Ким используют наработки Баттити и Коллы и нечеткую логику, что позволяет достичь более высокой точности классификации.

Бишоп рассматривает теорию ансамблей нейронных сетей. Если у L сетей ошибки имеют нулевое среднее значение и не коррелированы, то общую среднеквадратичную ошибку можно уменьшить в L раз просто усреднив классификации базовых нейронных сетей. Хотя в реальности ошибки все же коррелированы. Однако, Бишоп также с помощью неравенства Коши показывает, что ошибка ансамбля нейронных сетей теоретически не может увеличиться по сравнению с максимальной ошибкой входящих в него сетей. Если же в голосовании больше веса давать более точным сетям, то общую ошибку можно сократить еще больше.

Лэм и Суен изучили эффективность четырех комбинированных методов: голосования большинством, двух видов формализма Байеса и взвешенного голосования (где веса определяются с помощью генетического алгоритма). Они приходят к важному выводу: при отсутствии репрезентативной обучающей выборки голосование большинством остается наиболле простым и надежным методом голосования из четырех изученных.

1996 Соллих и Кроу обнаружили, что в больших ансамблях лучше использовать слабо регуляризованные классификаторы, которые фактически переобучены на обучающей выборке. Зато это позволяет по максимуму использовать эффект сокращения дисперсии при использовании ансамбля классификаторов.

Фрейнд и Шапир представляют алгоритм AdaBoost. Брейман изобретает бэггинг.

1997 Вудс, Кегельмайер и Бойер представляют метод объединения классификаторов, при котором учитываются локальная точность каждого из классификаторов в отдельно взятой области признакового пространства. Авторы верят, что их метод работает лучше, чем отдельно взятый хороший классификатор, на больших выборках со сложным распределением.

1998 Хо представляет метод случаного подпространства для построения лесов решений. Метод показал неплохую точность на практике и лучше всего работал на выборках данных с большим числом признаков и не совсем малым числом примеров.

1999 Шапир развивает теорию, лежащую в основе бустинга. Опиц и Маклин сравнивают бэггинг с двумя алгоритмами бустинга: AdaBoost и арчинг и приходят к выводу, что в условиях небольшого шума бустинг справляется с классификацией лучше, чем бэггинг, который в свою очередь лучше, чем отдельный классификатор.

2000 Наиболее обширная на тот момент работа по комбинированию классификаторов выполнена Джейном, Дьюном и Мао [5]. Они перечисляют причины, по которым можно использовать ансамбли классификаторов в надежде увеличить общую точность классификации. Это могут быть различные наборы признаков, различные обучающие выборки, способы классификации или обучающие сессии. Авторы предлагают таксономию ансамблевых методов классификации. Они также проводят эксперименты с четырьмя комбинированными методами, задействующими двенадцать классификаторов на шести наборах данных рукописных цифр. Из их опыта видно, что нет смысла использовать разные классификаторы на одном и том же множестве признаков, зато целесообразно использовать один классификатор на разных подмножествах признаков.

Кляйнберг наконец реализует алгоритм на основе теории стохастической дискриминации и показывает, что в большинстве тестов он дает лучшие результаты, чем бустинг и бэггинг.

Дитрих сравнивает эффективности рандомизации, бэггинга и бустинга при учучшении производительности дерева решений C4.5. Как показывают эксперименты, в случае незначительного шума рандомизация сравнима с бэггингом (или даже чуть лучше), но хуже бустинга. А в случае значительных шумов бэггинг намного лучше бустинга и иногда лучше, чем рандомизация.

2001 В своей кандидатской диссертации Скурицына изучает проблему стабилизации слабых классификаторов и сравнивает бэггинг, бустинг и метод случайного подпространства. Она приходит к выводу, что бэггинг подходит для слабых нестабильных классификаторов с неубывающей кривой обучения (learning curve). Бустинг хорош в тех же условиях, но для больших размеров обучающей выборки. Метод случайного подпространства показывает

неплохие характеристики в случае нестабильных слабых классификаторов с убывающей кривой обучения и малых размеров обучающей выборки.

2002 Кунчева выводит формулы для ошибки классификации для следующих методов комбинации классификаторов: усреднение, минимум, максимум, серединное значение и голосование большинстовом. Для равномерно распределенной вероятности принадлежности к разным классам лучше всего себя проявили методы минимума и максимума, а в случае нормального распределения все методы показали себя примерно одинаково. При этом в поставленных экспериментах преимущество по сравнению с отдельно взятым классификатором было пренебрежимо малым.

Очень качественные обзоры существующих на тот момент ансамблевых методов делают Валентини с Мазулли и Дитрих.

2003 Кунчева выводит теоретические верхнюю и нижнюю границы точности классификации при голосовании большинством и приходит к выводу, что идеальна ситуация, когда между всеми парами классификаторов отрицательная зависимость.

2004 Дероски и Зенко рассматривают построение ансамблей классификаторов разной природы с помощью стекинга и показывают, что это работает сравнимо с выбором лучшего классификатора методом скользящего контроля. Они также предлагают два новых способа стекинга, используя функцию распределения вероятности принадлежности к разным классам, и показывают, что их алгоритмы лучше, чем выбранный методом скользящего контроля классификатор.

Чаула и коллеги предлагают платформу для построения ансамблей сотен и тысяч классификаторов на небольших выборках данных в распределенной среде. Они показывают, что их подход позволяет классифицировать быстро, точно и с возможностью масштабирования.

Евгениу, Понтил и Елисеев сравнивают эффективность использования ансамблей машин опорных векторов (Support Vector Machines, SVM) и приходят к выводу, что это не приводит к улучшению точности по сравнению с одной правильно настроенной машиной, однако ансамбли более стабильны. Также ансамбли SVM-классификаторов предлагаются Валентини и Дитрих.

2005 Мелвилл и Муни представляют новый алгоритм DECORATE (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples, построение ансамбля разнообразных классификаторов за счет создания искусственных обучающих выборок), который побеждает отдельно взятые классификаторы, бэггинг и случайные леса, а также превосходит AdaBoost на малых выборках и сравним на больших.

2006 Рейзин и Шапир показывают, что улучшение маржи бустинга (boosting margin) также приводит и к увеличению сложности самого классификатора. Маржа бустинга для отдельного примера - это разность между количеством голосов тех классификаторов, которые правильно его классифицировали, и максимальным числом голосов, относящих его к неправильному классу. Авторы приходят к выводу, что максимизация маржи желательна, но не за счет других факторов, таких как, например, сложность базовых классификаторов.

Кунчева в разных работах с разными авторами рассматривает ансамблевые методы кластеризации, что выходит за рамки данного обзора.

2007 Кануто с коллегами исследует, как выбор базовых классификаторов влияет на производительность некоторых методов, разделяющихся на два класса: методы на основе выбора и на основе смешивания классификаторов. Один из выводов таков, что наилучшую точность показывают гибридные алгоритмы.

Бульман и Хоторн представляют статистические основы бустинга. Они делают обзор теоретических аспектов использования бустинга для статистических моделей, а также обращаются к методологии с практической точки зрения.

2008 Клескенс и Хорт публикуют книгу «Выбор модели и усреднение моделей», в которой объясняются, обсуждаются и сравниваются многие критерии выбора моделей [13].

5 Заключение

По мере развития теории машинного обучения и накопления практического опыта применения различных алгоритмов стало понятно, что не существует идеального метода классификации, который был бы лучше всех остальных при всех размерах и обучающей выборки, при любом проценте шума в данных, при любой сложности границы разделения объектов на классы и так далее. Поэтому в настоящее время активно развиваются ансамблевые методы классификации, объединяющие в одной модели множество разных классификаторов, обученных на разных выборках данных. Один из наиболее точных методов, существующих на сегодняшний день, - это бустинг над решающими деревьями. Также очень эффективным оказался алгоритм DECORATE. Однако для многих задач, особенно когда требуется высокая скорость работы алгоритма, все еще имеет смысл использовать отдельный классификатор, такой как правильно настроенная машина опорных векторов. В будущем ожидается развитие ансамблевых методов в высокоскоростном онлайн-обучении.

6 Литература

[1] E.A. Weiss, "Biographies: Eloge: Arthur Lee Samuel (1901-90)," *Annals of the History of Computing, IEEE*, vol. 14, no. 3, pp. 55–69, 1992.

[2] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM J. Res. Dev.*, vol. 3, no. 3, pp. 210–229, July 1959.

[3] Thomas M. Mitchell, *Machine Learning*, McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[4] Michael J. Kearns and Umesh V. Vazirani, *An Introduction to Computational Learning Theory*, MIT Press, Cambridge, MA, USA, 1994.

[5] Anil K. Jain, Robert P. W. Duin, and Jianchang Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[6] Leo Breiman and Leo Breiman, "Bagging predictors," in *Machine Learning*, 1996, pp. 123–140.

[7] Peter Bühlmann and Torsten Hothorn, "Boosting algorithms: Regularization, prediction and model fitting," *Statistical Science*, pp. 477–505, 2007.

[8] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, no. 11, pp. 1134–1142, Nov. 1984.

[9] RobertE. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[10] Y. Freund, "Boosting a weak learning algorithm by majority," *Information and Computation*, vol. 121, no. 2, pp. 256 – 285, 1995.

[11] Martin Sewell, "Ensemble learning," 2011.

[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[13] Gerda Claeskens and Nils Lid Hjort, *Model selection and model averaging*, Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, New York, 2008.