Жадный алгоритм формирования решающих ансамблей

А. В. Фомченко[1] Д. В. Парфенов[2]

Рассмотрен подход к совершенствованию методов решения задач машинного обучения на основе ансамблей алгоритмов на примере задачи классификации. Предложен метод выбора слабых решателей на основе жадного алгоритма и построения избирательного ансамбля. Этот подход является достаточно общим и может найти применение в системах поддержки принятия решений и в других экспертных системах.

Ключевые слова: машинное обучение, слабые решатели, решающие ансамбли, бустинг

1. Постановка задачи

Задача обучения по прецедентам $\langle X, Y, y^*, X', Y' \rangle$, где X – пространство объектов; Y – множество ответов; $y^* : X \to Y$ – неизвестная целевая зависимость; X' – обучающая выборка; Y' – вектор ответов на обучающих объектах, заключается в том, чтобы построить алгоритм $r : X \to Y$, аппроксимирующий целевую зависимость с заданной точностью α. Это значит, что вероятность верного ответа $P(r(x) = y) \geq \alpha$, то есть вектор результатов алгоритма должен совпадать с вектором ответов не менее чем на $\alpha \cdot 100\%$ [1].

2. Особенности ансамблевого подхода

Одним из методов построения эффективного решателя является ансамблирование, предполагающее использование нескольких слабых решающих правил (далее для краткости решателей) для формирования сильного правила требуемого качества. Такой метод имеет следующие преимущества:

1) Ансамбль способен обеспечить значительно лучший результат для разнородных данных по сравнению с отдельными его частями;

[1] *Фомченко Александр Валерьевич* — аспирант каф. высшей математики Института искусственного интеллекта РТУ МИРЭА, e-mail: fomchenko@mirea.ru.

Fomchenko Aleksandr Valerevich — Ph.D. student, Russian Technological University (MIREA), Institute of Artificial Intelligence, Department of Higher Mathematics.

[2] *Парфенов Денис Васильевич* — к.т.н., доцент каф. высшей математики Института искусственного интеллекта РТУ МИРЭА, e-mail: parfenov@mirea.ru.

Parfenov Denis Vasilevich, Ph.D. — associate professor, Russian Technological University (MIREA), Institute of Artificial Intelligence, Department of Higher Mathematics.

2) Ансамблированный решатель гораздо проще дообучать на новых данных в случае необходимости, поскольку дообучение можно проводить локально;

3) Решатель хорошо поддаётся декомпозиции в силу способа его построения, что облегчает его анализ и расширяет возможности параллельной программной реализации;

4) Есть возможность вычислительной оптимизации для принятия решения через выбор части ансамбля или даже только одного слабого решающего алгоритма ситуативно [2].

При построении ансамбля одновременно используют конечное множество предварительно обученных решателей, выходные сигналы которых объединяются в более качественный ответ. Возникает ряд вспомогательных проблем, которые рассмотрим детальнее.

Проблема 1: агрегирование результатов.
Одним из простейших способов объединения является метод голосования, где результат выбирается простым большинством по совокупности ответов всех слабых решателей. Развитием этой идеи служит использование весовых коэффициентов для каждого решателя в голосовании, но возникает задача отыскания оптимальных весов. Общий подход состоит в обучении еще одного дополнительного алгоритма, входами которого будут прогнозы всех алгоритмов ансамбля, выходом – итоговый прогноз. После завершения этапа обучения для каждого решателя определяется коэффициент его компетентности на совокупности данных [1]. Также популярно применение алгоритма бустинга – процедуры последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов [3].

Проблема 2: обеспечение разнообразия ансамбля.
Гарантия различия индивидуальных слабых решателей является фундаментальной задачей при построении ансамблей [4]. Очевидно, агрегация схожих решателей в ансамбле затрудняет существенное повышение качества решения задачи. При этом индивидуальные слабые решатели обучаются для решения одной задачи по одной обучающей выборке и, как следствие, достаточно сильно коррелированы.

Проблема 3: обоснованный выбор количества решателей в ансамбле.
Начиная с некоторого момента, увеличение их количества перестает значимо способствовать улучшению точности решения задачи, но приводит к существенному росту требуемых вычислительных ресурсов [5].

3. Применение жадного алгоритма для генерации ансамбля

Рассмотрим предлагаемый жадный алгоритм построения ансамбля на примере задачи классификации и то, как он справляется с описанными проблемами.

Основные шаги алгоритма:

1) Создание множества n различных слабых решателей и их обучение на исходной совокупности данных $A = \langle X', Y' \rangle$.

2) Проверка каждого решателя на выборке A, выбор одного из них с наилучшим результатом. Пусть он обеспечил правильные ответы на некоторой выборке $B \in A$.

3) Дальнейшее дообучение выбранного слабого решателя на выборке B с целью улучшения его надёжности, что важно для противостояния шумам и погрешностям в реальных данных. При этом решатель специализируется на качественной работе с определённым подклассом данных.

4) Переход к шагу 2 на усеченной выборке $A = A - B$, если не выполнено хотя бы одно из двух условий:

   а) Достигнута заданная точность (мощность новой выборки составляет заданную малую часть от мощности изначальной $|A| < (1 - \alpha)\ |\langle X', Y' \rangle|$). Это означает, что далее гарантированно будут происходить еще меньшие отсечения от основной выборки, и оставшиеся решатели нет смысла включать в ансамбль. Это нормальное завершение алгоритма.

   б) Исчерпано множество слабых решателей. Это свидетельствует о недостаточности ансамбля для решения задачи с заданной точностью и необходимости его пополнения, либо усиления разнообразия.

Рассмотрим алгоритм построения ансамбля подробнее. При создании и обучении слабых решателей важно, чтобы они давали различные результаты на одном наборе данных. Если ответы двух решателей совпадают сильнее заданной для задачи величины α, то нет смысла применять оба в ансамбле. Действительно, алгоритм использует ровно один лучший решатель на каждом шаге, что вычислительно экономично. По этой же причине на шаге 3 алгоритма применяется дообучение, а не обучение с самого начала. Этим обеспечиваются как разнородность сравниваемых решателей, так и экономия вычислительных ресурсов.

В самом начале создается набор из n слабых решателей, удовлетворяющих требованию разнообразия ансамбля. Один из методов такой генерации – манипуляции с входными переменными и параметрами обучения [6]. За счёт этого решатели при обучении могут сходиться к разным результатам. Однако это не гарантируется для всех решателей, какие-то могут сходиться к близким, и даже одинаковым параметрам. Такие случаи на шаге 1 алгоритма отфильтровываются для увеличения производительности обучения. Другой подход – генерация уникальных обучающих подмножеств для каждого слабого решателя [6]. Здесь возникает сложность их обоснованного выбора.

После создания набора решателей из него итеративно выбираются наиболее подходящие для итогового ансамбля. Изначальное множество данных A сокращается на каждой итерации, поэтому цикл может завершиться только двумя способами: либо множество данных для обучения сократится до объема $(1 - \alpha)$ от изначального, либо закончится набор самих решателей. Первый случай является успешным завершением алгоритма.

После нахождения и обучения минимально достаточного набора решателей требуется их объединить с помощью агрегирующего решающего правила. Для формирования ансамбля можно использовать голосование с использованием весов, описанное выше. Вес для каждого решателя задается соотношением множества его правильных ответов на обучающей выборке ко всей обучающей выборке. Таким образом, веса автоматически получаются нормированными. Однако при этом никак качественно не учитывается область применимости того или другого слабого решателя. Альтернативный подход – использование отдельного селективного решателя для выбора одного наиболее подходящего по ситуации слабого решателя из ансамбля. Селективный решатель обучается после построения всех решателей в ансамбле. Итоговый алгоритм описывается блок-схемой, приведённой на Рисунке 1.

Такой метод решает всю совокупность проблем, описанных выше:

1) Разнородные данные обрабатываются так, что для каждой выборки с особыми свойствами в ходе цикла берётся один решатель, лучше работающий на ней. Таким образом попутно осуществляется кластеризация обучающих данных с позиций их проекции на решатели, что сильно упрощает анализ алгоритма.

2) Разнообразие ансамбля обеспечивается шагом 3 алгоритма, где за счёт дообучения на отдельных непересекающихся выборках данных решатели будут значительно отличаться, даже если исходно их поведение было похожим. Кроме того, изначальная генерация n слабых решателей уже подразумевает их некоторое разнообразие.

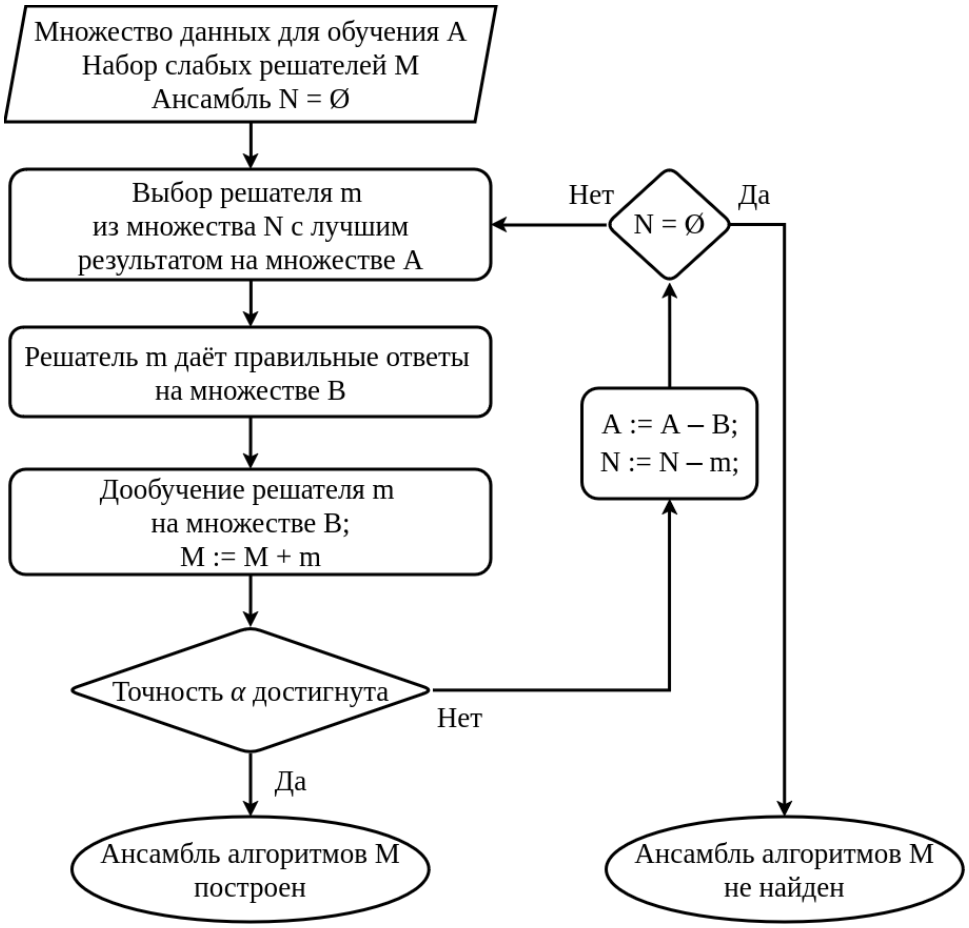

Рисунок 1. Блок-схема алгоритма построения ансамбля решателей.

3) Количество решателей в ансамбле оптимально подбирается самим алгоритмом. Однако, важна возможность их выбора из исходно достаточно большой и разнородной совокупности. Вообще говоря, чем больше n, тем лучше может оказаться результат ценой увеличения затрат на создание ансамбля. При этом сложность всего ансамблевого решателя не только не увеличивается с ростом n, но и способна снижаться за счёт более удачного выбора его компонент на каждом шаге 2 и, таким образом, сокращения их числа в ансамбле.

Существуют другие жадные алгоритмы для решения задач классификации данных, в том числе разнородных. Так, например, в статье [7] описано итеративное применение одиночных классификаторов на обучающей выборке и учёт в итоговом решении вклада только тех классификаторов,

ошибка которых не превосходит заданный порог. Однако в этом алгоритме не применяется дополнительное дообучение выбранных алгоритмов на данных, где они себя проявляют (шаг 3 алгоритма). В работе [8] приведен алгоритм set covering machine для построения решающих списков. Решающий список закономерностей представляет собой частный случай алгоритмической композиции с голосованием по старшинству. На каждой итерации алгоритма выбирается правило, допускающее наименьшее количество ошибок. В этом алгоритме также нет дообучения, кроме того, есть ограничение на максимальную допустимую долю ошибок на обучающей выборке, подбираемое экспериментально.

4. Экспериментальное сравнение с другими алгоритмами

В наших экспериментах в качестве решателей используются свёрточные нейронные сети прямого распространения, в частности, потому что для них относительно легко обеспечить разнообразие ансамбля, выбирая веса перед обучением случайным образом. Выборкой данных служит разнородная совокупность из 100000 объектов 5 классов; данные в ней распределены по классам равномерно. Использовался один и тот же набор однотипных трёхслойных нейросетей по 5 нейронов в каждом слое. Далее в тексте такие нейросети будут называться малыми. В качестве решателей взяты:

1) Одна нейросеть 6 слоёв по 15 нейронов. Такой размер выбран исходя из приближенного равенства общего числа нейронов в ней и в 6 нейросетях из ансамбля пункта 2).

2) Ансамбль с голосованием – 6 малых нейросетей.

3) Ансамбль с избирательной нейросетью. Избирательная сеть такая же по размерам, как и малая.

4) Ансамбль, сформированный из 10 малых нейросетей со случайной инициализацией весов с применением жадного алгоритма. Для ансамбля использовалась селективная нейросеть такого же размера, как и малая.

Точность алгоритма (вероятность принятия правильного решения) приведена в Таблице 1. В экспериментах жадный алгоритм формирования ансамбля завершался по достижении заданной точности $\alpha = 95\%$. При увеличении α до 100% алгоритм продолжил работу до 5 нейросетей и достиг точности 0.9628. Предлагаемый метод экспериментально

Таблица 1. Сравнение точности алгоритмов.

Метод	Параметры	Точность
Нейросеть	1 нейросеть (90 нейронов)	0.8651
Ансамбль с голосованием	6 нейросетей (суммарно 90 нейронов)	0.7457
Ансамбль с селективной нейросетью	5 нейросетей + селективная (суммарно 90 нейронов)	0.9171
Применение жадного алгоритма	Изначально 10 нейросетей, в результате выбраны 4 + селективная (суммарно 75 нейронов)	0.9513

Таблица 2. Сравнение алгоритмов по площади ROC-AUC.

Метод	Параметры	ROC-AUC
Нейросеть	1 нейросеть (90 нейронов)	0.848
Ансамбль с голосованием	6 нейросетей (суммарно 90 нейронов)	0.864
Ансамбль с селективной нейросетью	5 нейросетей + селективная (суммарно 90 нейронов)	0.935
Применение жадного алгоритма	Изначально 10 нейросетей, в результате выбрано 2 + селективная (суммарно 45 нейронов)	0.966

превосходит сравниваемые, имеющие даже превосходящую сложность. В Таблице 2 представлен результат для бинарной классификации из 100000 объектов теми же алгоритмами. В качестве метрики выбрана площадь под кривой ошибок (ROC-AUC, см. Рисунок 2). В случае безошибочного распознавания площадь под кривой равна единице. Если же использование решателя не отличается от случайного угадывания, значение площади под кривой стремится к 0.5. Использование этой метрики позволяет дать наглядную оценку для случая разнородных данных [9].

Таким образом, представленный метод не уступает ближайшему сопернику – ансамблю с избирательной нейросетью и дополнительным шагом кластеризации. Важно подчеркнуть существенное отличие предложенного подхода от избирательной нейросети. Дело в том, что такая нейросеть представляет собой «чёрный ящик» – нельзя точно сказать, почему был выбран конкретный решатель или совокупность результатов нескольких слабых решателей. Из этого следует дополнительная сложность донастройки ансамбля в случае небольшого изменения исходных

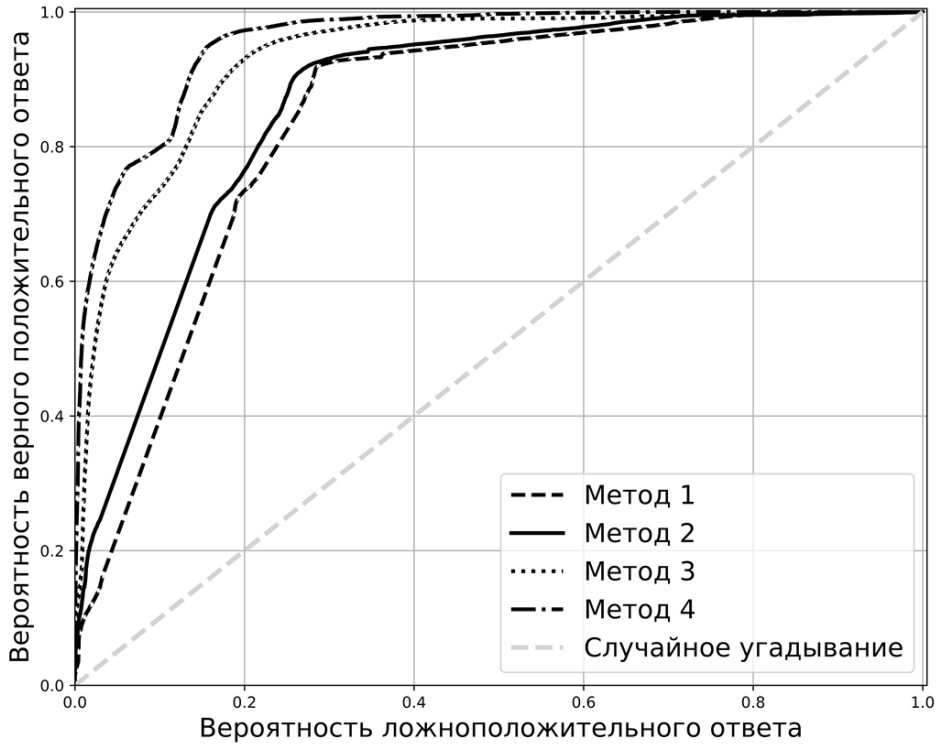

Рисунок 2. ROC-кривые четырёх методов.

данных, а также неприменимость подхода с избирательной нейросетью в задачах, где нужна доказательность решения.

Выводы:

1) В результате тестирования предложенный подход продемонстрировал превосходящую эффективность по обоим применявшимся критериям в сравнении с рассмотренными альтернативами.

2) Новый метод универсален, хорошо модифицируем и применим для решения широкого класса задач интеллектуального анализа данных. Возможно использование решателей различного типа в рамках одного гетерогенного ансамбля.

3) Рассмотренный алгоритм имеет доказательное построение и лишён подстроечных гиперпараметров, что облегчает его анализ, программную реализацию и обучение.

4) В нем количество отобранных для ансамбля слабых решателей мало зависит от их количества в изначальном наборе. Оно определяется лишь качеством первичных решателей, их способностью дообучаться на сокращённом наборе данных и требуемой точностью. Даже при большом изначальном количестве решателей в результате получается компактный и эффективный ансамбль.

5) Данный подход прекрасно приспособлен к распараллеливанию вычислений как на стадии обучения, так и при использовании.

6) Его применение является вычислительно экономным в силу использования выбираемого селективным решателем каждый раз одного слабого решателя для выработки полноценного итогового решения. Это возможно благодаря специализации решателей в процессе их дообучения.

Литература

[1] Воронцов К.В., *Математические методы обучения по прецедентам (теория обучения машин)*, `http://www.machinelearning.ru/wiki/index.php`, 2023.

[2] Rokach L., *Artificial Intelligence Review*, **33**:1-2 (2009), 1-39.

[3] Кашницкий Ю.С., Игнатов Д.И., "Ансамблевый метод машинного обучения, основанный на рекомендации классификаторов", *Интеллектуальные системы. Теория и приложения*, **19**:4 (2015), 37-55.

[4] Kuncheva L.I., *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2014, 360 pp.

[5] Кривенко М.П., Васильев В.Г., *Методы классификации данных большой размерности*, ИПИ РАН, М., 2013, 203 pp.

[6] Мангалова Е.С., Агафонов Е.Д., "О проблеме генерации разнообразия ансамблей индивидуальных моделей в задаче идентификации", *Тр. XII Всерос. совещания по проблемам управления ВСПУ-2014*, 2014, 3214-3223

[7] Alsova O.K., Stubarev I.M., "Inhomogeneous ensemble algorithm for classifying different types of data", *News of Samara scientific center of Russian Academy of Sciences*, 2017, 118-123

[8] Marchand M., Shawe-Taylor J., "Learning with the set covering machine", *Proc. 18th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA, 2001, 345-352

[9] Zweig M., Campbell G., "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine", *Clinical Chemistry 1993*, **39**:4 (1993), 561-577

A greedy algorithm for building learning ensembles
Fomchenko A.V., Parfenov D.V.

An approach to improving methods for solving machine-learning problems based on ensembles of algorithms is considered using classification problem as an example. A method based on a greedy algorithm for choosing weak learners and building the selective ensemble is proposed. This approach is general and applicable in decision support systems and other expert systems.

Keywords: machine learning, weak learners, ensembling, boosting

References

[1] Vorontsov K.V., *Mathematical methods of teaching by precedents (machine learning theory)*, `http://www.machinelearning.ru/wiki/index.php`, 2023 (In Russian)

[2] Rokach L., *Artificial Intelligence Review*, **33**:1-2 (2009), 1-39

[3] Kashnitsky Yu.S., Ignatov D.I., "Ensemble machine learning method based on classifier recommendation", *Intelligent Systems. Theory and Applications*, **19**:4 (2015), 37-55 (In Russian)

[4] Kuncheva L.I., *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2014, 360 pp.

[5] Krivenko M.P., Vasilev V.G., *Methods for classifying high-dimensional data*, IPI RAN, M., 2013 (In Russian), 203 pp.

[6] Mangalova E.S., Agafonov E.D., "On the problem of generating diversity of ensembles of individual models in the identification problem", *Proc. of the XII All-Russian meetings on management issues, VSPU-2014*, 2014, 3214-3223 (In Russian)

[7] Alsova O.K., Stubarev I.M., "Inhomogeneous ensemble algorithm for classifying different types of data", *News of Samara scientific center of Russian Academy of Sciences*, 2017, 118-123

[8] Marchand M., Shawe-Taylor J., "Learning with the set covering machine", *Proc. 18th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA, 2001, 345-352

[9] Zweig M., Campbell G., "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine", *Clinical Chemistry 1993*, **39**:4 (1993), 561-577