УДК 004.85:512.563

Д.В. Виноградов

Оценка степени переобучения алгебраического машинного обучения для случая булевой алгебры

Предлагается оценка вероятности переобучения для ВКФ-метода алгебраического машинного обучения в простейшем случае булевой алгебры без учета контр-примеров. Модель использует идеи В.Н. Вапника и А.Я. Червоненкиса о минимизации эмпирического риска. Асимптотически вероятность переобучения в фиксированной доле тестовых примеров при стремлении длины описания (и объема затребуемого числа гипотез) к бесконечности стремится к нулю быстрее, чем экспонента.

Ключевые слова: *эмпирический риск, переобучение, ВКФ-метод, булева алгебра*

DOI: 10.36535/0548-0027-2022-06-2

ВВЕДЕНИЕ

Вероятностно-комбинаторный формальный метод алгебраического машинного обучения (ВКФ-метод) был нами предложен и исследован ранее [1]. Современная программная реализация использует спаривающую цепь Маркова для вычисления индуктивного обобщения (сходств) обучающих примеров в гипотезы о причинах их целевого свойства. Семейство спаривающих цепей Маркова обеспечивает останавливаемость процесса вычислений с вероятностью единица. Однако имеется монотонный вариант алгоритма поиска сходств (гипотез), который в случае булевой алгебры совпадает с классическим ленивым случайным блужданием на соответствующем гиперкубе. Для случая булевой алгебры получены точные оценки скорости перемешивания (см., например, [2]) порядка $\frac{1}{2} \cdot n \cdot \log n$. В этом случае предельное распределение будет равномерным. Для спаривающей цепи Маркова в работе [2] для булевой алгебры были доказаны теоремы о сильной концентрации длин траекторий около среднего значения, которое имеет порядок $n \cdot \log n$. Так как нижняя компонента спаренной цепи Маркова движется по траектории монотонной цепи Маркова, то эти результаты в совокупности приводят к тому, что результаты выдачи спаривающей цепи Маркова можно считать приближенно равномерно распределенными. Впрочем, если требуется достичь большей равномерности, то можно сделать дополнительные шаги. Поэтому мы в дальнейшем будем предполагать, что гипотезы имеют равномерное распределение на булевом гиперкубе.

Используя вероятностную модель семейств Бернулли для порождения обучающих примеров и контр-примеров, мы доказали [3] неизбежность порождения фантомных (случайных) гипотез, которые могут приводить к переобучению – неправильному предсказанию целевого свойства у тестовых примеров, предъявленных для прогнозирования. Однако такая вероятностная модель требует обоснования и, как продемонстрировала в ходе эмпирических исследований аспирантка ФИЦ ИУ РАН Л.А. Якимова, нуждается в существенном усложнении для получения более точных оценок.

На возможность прямой оценки переобучения с помощью так называемой слабой вероятностной аксиоматики обратил внимание К.В. Воронцов [4]. Ключевая идея – метод минимизации эмпирического риска, восходящий к классической работе В.Н. Вапника и А.Я. Червоненкиса [5]. Цель настоящей работы – получение оценки такого переобучения для ВКФ-метода в особенно простом случае, когда обучающие и тестовые примеры представляют собой коатомы в булевой алгебре, а контр-примеры не учитываются. При этом мы будем считать, что гипотезы имеют равномерное распределение. К сожалению, общие результаты К.В. Воронцова [4] выражаются через кратные суммы гипергеометрических коэффициентов и в интересующем нас случае не могут быть хорошо оценены. Поэтому мы сведем нашу ситуацию к сериям испытаний Бернулли и применим прямой подсчет.

В Первом разделе настоящей статьи мы представим необходимые определения. Второй раздел содержит вывод асимптотической верхней оценки переобучения в фиксированной доли тестовых примеров. Этот вывод опирается на формулу Стирлинга (см., например, [6]). В Заключении мы обсудим некоторые возможные проявления этого факта.

МЕТОД МИНИМИЗАЦИИ ЭМПИРИЧЕСКОГО РИСКА

Любую совокупность гипотез о причинах наличия целевого свойства можно рассматривать как алгоритм классификации: если тестовый пример включает в себя хотя бы одну гипотезу, то алгоритм предскажет наличие целевого свойства; если ни одна из гипотетических причин не вкладывается в тестовый пример, то он классифицируется отрицательно.

Метод минимизации эмпирического риска, предложенный В.Н. Вапником и А.Я. Червоненкисом [5], состоит в выборе таких алгоритмов, для которых классификация обучающих примеров содержит минимальное число ошибок (эмпирический, или наблюдаемый, риск). В нашем случае всегда будут алгоритмы (совокупности гипотез), эмпирический риск которых равен нулю. Ими мы и будем ограничиваться в дальнейшем. С другой стороны, имеется риск совершить ошибку на предсказании тестовых примеров.

Мы будем случайным образом разбивать, следуя К.В. Воронцову, объекты на две группы: обучающие и тестовые примеры. Для простоты допустим, что число объектов четно, а разбиение производится пополам. Это допущение не снижает общности, так как средний биномиальный коэффициент является наибольшим.

Наши алгоритмы (совокупности гипотез) будут порождаться с помощью траекторий спаривающей цепи Маркова [1]. Каждая гипотеза (компонента алгоритма) порождается одной траекторией цепи. Состоянием спаривающей цепи Маркова является упорядоченная пара кандидатов в гипотезы. Траектория обрывается, когда оба кандидата первый раз совпадают. Можно заметить, что меньший (нижний) кандидат в гипотезы движется по траектории монотонной цепи Маркова.

В случае булевой алгебры, которым мы и ограничимся, монотонная цепь Маркова обладает свойством быстрого перемешивания, причем, стационарное распределение будет равномерным. Этот факт составляет вариант известного результата П. Дьякониса для случайных блужданий на булевом гиперкубе. В нашем случае, возникает ленивое случайное блуждание, для которого точная оценка (с удвоенным по сравнению с классическим случайным блужданием коэффициентом) доказана в работе [2].

Для $2n$-мерной булевой алгебры нам потребуется обучающая выборка, содержащая все коатомы. Пусть $O = \{o_1, o_2, \ldots, o_{2n}\}$ будет множеством коатомов (обучающих и тестовых примеров), каждый из которых описывается признаками из списка $F = \{f_1, f_2, \ldots, f_{2n}\}$, и $o_i \, I \, f_j \leftrightarrow j \neq i$ (матрица 1).

Матрица 1

$O\backslash F$	f_1	f_2	…	f_{2n}
o_1	0	1	...	1
o_2	1	0	...	1
⋮	⋮	⋮	⋱	⋮
o_{2n}	1	1	...	0

Ясно, что $o_{j1} \cap o_{j2} \cap \ldots \cap o_{js} = F\backslash\{f_{j1}, f_{j2}, \ldots, f_{js}\}$, так как добавление в сходство примера o_k с номером k удаляет из фрагмента признак f_k с тем же самым номером k. Более того, пример o_k будет предсказан положительно (правильно), если и только если признак f_k отсутствует хотя бы в одном из порожденных сходств.

Обозначим эмпирический риск через η и введем риск прогнозирования как долю $\theta = r/n$ неправильно предсказанных тестовых примеров. Нас интересует вероятность $\boldsymbol{P}[\eta = 0, \theta = \delta]$ при равномерном разбиении объектов на обучающую и тестовую выборки пополам. Ясно, что для различных разбиений вероятность будет одинакова.

АСИМПТОТИЧЕСКАЯ ОЦЕНКА НА ПЕРЕОБУЧЕНИЕ

Можем считать, что первые n объектов попали в обучающую выборку, а последние n объектов образуют тесты, так как вероятности одинаковы для каждого разбиения.

Пусть по обучающей выборке для булевой алгебры было порождено m ВКФ-гипотез с помощью монотонной цепи Маркова. Если траектории выбрать достаточно длинными, то распределение гипотез будет равномерным и независимым. Равномерность следует из свойства быстрой перемешиваемости к равномерному стационарному распределению, а независимость – из независимости траекторий цепи Маркова, порождающих ВКФ-гипотезы.

Обозначим эти гипотезы через $H = \{h_1, h_2, \ldots, h_m\}$ и составим соответствующую матрицу 2.

Матрица 2

$O\backslash H$	h_1	h_2	…	h_m
o_1	0	0	...	1
o_2	1	0	...	0
⋮	⋮	⋮	⋮	⋮
o_n	0	1	...	0
o_{n+1}	0	0	...	0
⋮	⋮	⋮	0	⋮
$o_{(1+\delta)\cdot n}$	0	0	...	0
$o_{(1+\delta)\cdot n+1}$	1	0	...	1
⋮	⋮	⋮	⋮	⋮
o_{2n}	0	1	...	1

Здесь единица соответствует тому, что гипотеза вкладывается в пример, т. е. предсказывает его правильно (положительно). Тогда для $\eta = 0$ требуется, чтобы в каждой из первых n строк встречалась хотя бы одна единица. Для $\theta = \delta$ нужно выбрать $\delta \cdot n$ строк из нижней половины (это можно сделать $n!/(\delta \cdot n)!((1-\delta)\cdot n)!$ способами), в которых должны стоять одни нули, а в остальных опять где-то должны встречаться единицы. Пример в матрице 2 соответствует такой ситуации с выбором строк $o_{n+1}, \ldots, o_{(1+\delta)\cdot n}$.

Из-за равномерной распределенности и независимости гипотез соответствующие ячейки образуют серию испытаний Бернулли с вероятностью успеха ½.

Лемма 1. *Для условия* $\lim \mathbf{P}[\eta = 0] = 1$ *при* $n \to \infty$ *достаточно потребовать, чтобы число гипотез* $m \geq (1+\sigma)\cdot \log_2 n$ *для некоторого* $\sigma > 0$.

Доказательство.

$$1 \geq \lim_{n\to\infty}(1-2^{-m})^n = \lim_{n\to\infty}\left[(1-2^{-m})^{2^m}\right]^{n/2^{-m}} =$$

$$= \lim_{n\to\infty}[e^{-1}]^{n/2^{-m}} \geq \lim_{n\to\infty} e^{-1/n^{\sigma}} = 1.$$

Утверждение леммы 1 можно усилить до, например, $m \geq \log_2 n + \log_2(\sigma\cdot\log_2 n)$, но это ослабит окончательную оценку.

Воспроизведем оценку Стирлинга (см., например, [6])

Лемма 2. $n! \sim \sqrt{2\pi n}\cdot n^n \cdot e^{-n}$.

Теорема 1. *При* $n \to\infty$ *и* $m \geq (1+\sigma)\cdot\log_2 n$ *для вероятности переобучения* $\boldsymbol{P}$ $[\eta = 0, \theta = \delta]$ *имеем*:

$$\lim_{n\to\infty} P \leq$$
$$\leq \frac{1}{\sqrt{2\pi\delta(1-\delta)}} exp\left\{-(1+\sigma)\delta\cdot n\cdot lnn + ln2\cdot n - \frac{lnn}{2}\right\}.$$

Доказательство. Из свойств серий Бернулли следует, что $P = \binom{n}{\delta\cdot n}\cdot(1-2^{-m})^{(2-\delta)n}\cdot 2^{-\delta\cdot n\cdot m}$. Второй сомножитель не превосходит единицы (и асимптотически равен единице по лемме 1). Третий сомножитель оценивается так:

$$2^{-\delta\cdot n\cdot m} \leq n^{-(1+\sigma)\cdot\delta\cdot n} = \exp\{-(1+\sigma)\cdot\delta\cdot n\cdot\ln n\}.$$

Применим формулу Стирлинга к первому сомножителю $n!/(\delta\cdot n)!((1-\delta)\cdot n)!$. Она после сокращения дает асимптотическую оценку

$$\frac{1}{\sqrt{2\pi\delta(1-\delta)n}} exp\{-\delta\cdot ln\delta\cdot n - (1-\delta)\cdot ln(1-\delta)\cdot n\}.$$

Как в теории информации легко доказать, что $-\delta\cdot\ln\delta - (1-\delta)\cdot\ln(1-\delta) \leq \ln 2$, последнее неравенство завершает доказательство теоремы.

ЗАКЛЮЧЕНИЕ

В настоящей работе доказана суперэкспоненциальная верхняя оценка на вероятность переобучения в частном случае булевой алгебры. Аспирантка ФИЦ ИУ РАН Л.А. Якимова проводила эмпирические исследования переобучения для программной ВКФ-системы. На массиве *Mushrooms* из репозитария данных для тестирования алгоритмов машинного обучения Университета Калифорнии в г. Ирвайн при обучении, когда эмпирический риск был равен нулю, ошибки на тестовой выборке не наблюдались. Это дает надежду на то, что подобный феномен — экспоненциально малое число ошибок предсказания – будет верен и для общего случая произвольной обучающей выборки, когда эмпирический риск равен нулю. Эти результаты резко контрастируют с более ранними эмпирическими исследованиями Л.А. Якимовой и А.С. Опарышевой (выпускницами РГГУ), которые обнаружили возникновение переобучения на основе фантомных гипотез в ДСМ-экспериментах (в которых порождаются все возможные сходства) на реальных данных. Например, на том же самом массиве Mushroom часть поганок была предсказана съедобными, что ставит крест на применении ДСМ-метода автоматического порождения гипотез [7] в критически важных областях.

* * *

Автор благодарит своих коллег по ВЦ им. А.А. Дородницына ФИЦ ИУ РАН за поддержку и полезные дискуссии. Особенная благодарность Л.А. Якимовой за совместную работу, обсуждения и поддержку.

СПИСОК ЛИТЕРАТУРЫ

1. Vinogradov D.V. Machine Learning Based on Similarity Operation // Communication in Computer and Information Science. – 2018. – Vol. 934. – P. 46–59.
2. Виноградов Д.В. Вероятностно-комбинаторный формальный метод обучения, основанный на теории решеток: дис. … д-ра физ.-мат. наук 05.13.17. – Москва: ФИЦ ИУ РАН, 2018. – 131 с.
3. Виноградов Д.В. Скорость сходимости к пределу вероятности порождения случайного сходства при наличии контр-примеров // Научная и техническая информация. Сер. 2. – 2018. – № 2. – С. 21–24; Vinogradov D.V. The Rate of Convergence to the Limit of the Probability of Encountering an Accidental Similarity in the Presence of Counter Examples // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, № 1. – P. 35-37.
4. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. – Москва: Физматлит, 2004. – Т. 13. – С. 5–36
5. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. – Москва: Наука, 1974. – 416 с.
6. Феллер В. Введение в теорию вероятностей и ее приложения. В 2–х томах. Т. 1 / пер. с англ. – Москва: Мир, 1984. – 528 с.
7. ДСМ-метод автоматического порождения гипотез: логические и эпистемологические основания / ред. О.М. Аншаков – Москва: URSS, 2009. – 430 с.

Материал поступил в редакцию 25.04.22,
исправленная версия – 13.05.22

Сведения об авторе

ВИНОГРАДОВ Дмитрий Вячеславович – доктор физико-математических наук, ведущий научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, Москва
e-mail: vinogradov.d.w@gmail.com