УДК 519.769 + 519.765

ОБЗОР МЕТОДОВ ОЦЕНКИ ЕМКОСТИ МОДЕЛИ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ МАШИННОГО ОБУЧЕНИЯ

И.Д. Кудинов

Институт проблем управления им. В.А. Трапезникова РАН

Россия, 117997, Москва, Профсоюзная ул., 65

E-mail: ilja@kdsli.ru

Ключевые слова: машинное обучение, обучение по прецедентам, обучение с учителем, задача классификации.

Аннотация: При обучении по прецедентам моделей машинного обучения исследователи сталкиваются с двумя явлениями: недообучением и переобучением. В первом случае недостаток свободных параметров алгоритма не позволяет воспроизвести генерирующее распределение вероятности, которому следуют примеры обучающей выборки. Во втором случае при избытке свободных параметров модель начинает воспроизводить не только восстанавливаемую зависимость, но и ошибки наблюдения, что приводит к потере способности модели к обобщению. Из этого делается вывод, что для всякой задачи существует оптимальная сложность модели, называемая емкостью, при которой достигается наилучшее качество обобщения. Начиная с 60-х годов было предложено несколько теорий оценки емкости различных задач. В данном обзоре рассматривается история данного вопроса, современное состояние области и основные достигнутые результаты.

Постановка решения задачи классификации методами машинного обучения (с учителем) звучит следующим образом. Пусть дано распределение P над множеством пар $(x, y) \in \mathcal{X} \times \{0, 1\}$ объектов $x \in \mathcal{X}$, размеченных одним из двух классов $y \in \{0, 1\}$. Алгоритм машинного обучения пытается решить задачу одноклассовой классификации путем построения классификатора $h_n \colon \mathcal{X} \to \{0, 1\}$ с целью минимизации вероятности ошибки классификации:

$$\mathrm{er}(h_n) = P\{(x, y) \mid h_n(x) \neq y\}.$$

На практике, помимо самой вероятности ошибки, исследователей интересует и скорость сходимости ошибки в ходе процесса обучения модели. Характер кривой зависимости вероятности ошибки $\mathrm{er}(h)$ от числа взятых моделью примеров n в идеальном случае представляет собой экспоненциальную кривую порядка отрицательной степени n, или в редких случаях линейный график.

В общем случае алгоритм обучения не обладает информацией о характере распределения P, однако он может выбрать n независимых одинаково распределенных пар из P. Однако, такая рассмотрение совершенно произвольного P не позволяет дать универсальный метод построения классификатора h_n, дающий оптимальную ошибку $\mathrm{er}(h_n)$ на всей возможной совокупности задач классификации –

этот результат известен как теорема **об отсутствии бесплатных завтраков** (no free lunch theorem) [1].

В ином случае для оптимальной стратегии определения алгоритма обучения необходимо обладать какой-либо информацией о P. Для этого вводят множество $\mathcal{H}$ классификаторов $h : \mathcal{X} \to \{0, 1\}$ называемые **классом концептов классификаторов** (concept class of classifier). Наличие $\mathcal{H}$ позволяет строить предположения о распределении P. Простейшее из них заключается в **реализуемости** P, то есть в гипотезе что найдется такой классификатор $h \in \mathcal{H}$, на которых ошибка $\mathrm{er}(h)$ стремится к нулю при $n \to \infty$:

$$\inf_{h \in \mathcal{H}} \mathrm{er}(h) = 0.$$

В качестве моделей машинного обучения в случае задачи классификации почти всегда рассматривались перцептроны – многослойные полносвязные нейронные сети. Опуская нелинейную функцию активации нейронов, результатом обучения перцептрона из n слоев является полином n-й степени от компонент входного сигнала. Этот полином выражает $(n-1)$-мерную поверхность в пространстве $\mathcal{X}$, которая отделяет точки разных классов друг от друга. В таком случае функции $h \in \mathcal{H}$ представляют собой сечения этой поверхности: $f(x_i) = 0$.

Различные теории машинного обучения пытаются дать числовую оценку сложности задачи классификации, называемой **размерностью** или **емкостью** [2]. Емкость определяется как для задачи, так и для алгоритмов обучения, причем в последнем случае она вычисляется на основе выбранного класса концептов $\mathcal{H}$. Емкость используется главным образом для выбора подходящего для данной задачи класса концептов $\mathcal{H}$.

Одной из ранних попыток определить понятие емкости для задач классификации является теория Вапника-Червоненикса [4] и ее развитие в виде теории Валианта, где вводится понятие **размерности Вапника-Червоненикса** (VC-размерность, VC-dimension), которая определяется как такое наибольшее число d, что d объектов из $\mathcal{X}$ можно разбить на два класса функциями из $\mathcal{H}$ всеми возможными способами. Так, теория вероятно-приближенного корректного обучения (Probability approximately correct learning), [3, 5] связывает оптимальную ошибку обучения, возможную при выбранном классе концептов $\mathcal{H}$, с VC-размерностью $\mathrm{VCdim}(\mathcal{H})$ следующим выражением:

$$\inf_{h \in \mathcal{H}} \sup_{P} \mathrm{er}(h) = O\left(\min\left\{\frac{\mathrm{VCdim}(\mathcal{H})}{n}, 1\right\}\right).$$

Емкость используется не только для оценки оптимальной ошибки обучения. Чем больше соответствие между емкостью задачи и емкостью используемого класса концептов $\mathcal{H}$, тем наиболее подходящим $\mathcal{H}$ считается для задачи. Если емкость задачи больше емкости $\mathcal{H}$, то «сложности» алгоритма обучения не позволяет воспроизвести распределение P исходных данных, что приводит большой ошибке. Если же емкость задачи меньше емкости $\mathcal{H}$, то модель начинает воспроизводить не только восстанавливаемую зависимость класса от объекта $x \in \mathcal{X}$, но и ошибки наблюдения: погрешности и шумы. На практике это приводит к потере модели способности к обобщению – такая модель обучается только тому распределению, которому следуют n выбранных примеров, но не общему распределению P. И

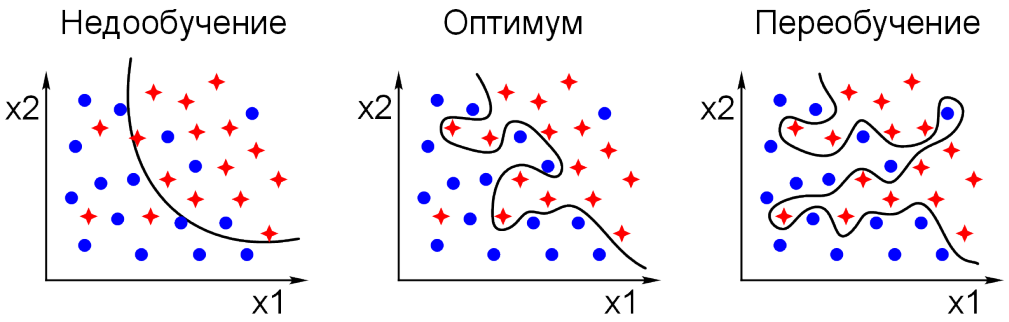

Рис. 1. Демонстрация явлений недообучения и переобучения на примере задачи классификации. Пусть d_1 это емкость задачи, а d_2 – емкость используемого класса концептов $\mathcal{H}$. Тогда при $d_1 \gg d_2$ наблюдается явление недообучения (левый график); при $d_1 \ll d_2$ наблюдается явление переобучения (правый график); при $d_1 \approx d_2$ задача решается оптимальным способом

только примерное равенство емкости задачи и емкости $\mathcal{H}$ позволяет решить задачу удовлетворительным способом. При всем при этом во всех трех случаях цель обучения достигается – вероятность ошибки $\mathrm{er}(h)$ достигает оптимального значения. Демонстрация этой связи приводится на рис. 1.

Однако, теория Вапника-Червоненикса и производные от нее имеет серьезный недостаток. Они фокусируются на худшем из возможных случаев. В качестве верхней оценки вероятности ошибки они приводят значение, значительно превышающее наблюдаемое на практике, а предсказываемая кривая обучения оказывается ограничивающей сверху для всех возможных кривых обучения в рамках задачи и асимптотически равна $\frac{1}{n}$.

Со временем было предложено несколько теорий, призванных приблизить теоретические верхние оценки вероятности ошибки $\mathrm{er}(h)$ к наблюдаемым на практике, и таким образом приблизить вычисляемое значение емкости к действительности. Данный обзор посвящен современному состоянию вопроса и основным достигнутым результатам в этом направлении.

Список литературы

1. Wolpert D.H., Macready W.G. No free lunch theorems for optimization // IEEE Transactions on Evolutionary Computation. 1997. Vol. EC-1. No. 1. P. 67-82.
2. Bengio Y., Goodfellow I., Courville A. Deep learning. Cambridge, MA, USA: MIT press, 2017. Vol. 1.
3. Bousquet O. et al. A theory of universal learning // Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing. 2021. P. 532-541.
4. Goldman S. Computational Learning Theory. 1999.
5. Ehrenfeucht A. et al. A general lower bound on the number of examples needed for learning // Information and Computation. 1989. Vol. 82, No. 3. P. 247-261.
6. Hanneke S. Learning whenever learning is possible: Universal learning under general stochastic processes // 2020 Information Theory and Applications Workshop (ITA). IEEE, 2020. P. 1-95.