Московский государственный университет имени М. В. Ломоносова

Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

«Метод повышения эффективности обучения, основанный на ансамбле промежуточных решений»

Выполнил:

студент 3 курса 317 группы

Королев Николай Сергеевич

Научный руководитель:

д.ф-м.н., в. науч. сотр. ВЦ РАН

Сенько Олег Валентинович

Москва, 2019

Содержание

Аннотация

Данный документ является образцом оформления дипломной работы для студентов кафедры Математических методов прогнозирования ВМК МГУ. Приведённые ниже рекомендации взяты из статьи «Написание отчётов и статей (рекомендации)» на вики-ресурсе `www.MachineLearning.ru`. Студенты, готовящие дипломную работу к защите, могут найти много полезной информации также в статьях «Научно-исследовательская работа (рекомендации)», «Подготовка презентаций (рекомендации)», «Защита выпускной квалификационной работы (рекомендации)» на том же ресурсе.

Аннотация обычно содержит краткое описание постановки задачи и полученных результатов, одним абзацем на 10–15 строк. Цель аннотации — обозначить в общих чертах, о чём работа, чтобы человек, совершенно не знакомый с данной работой, понял, интересна ли ему эта тема, и стоит ли читать дальше. Аннотация собирается в последнюю очередь путем легкой модификации наиболее важных и удачных фраз из введения и заключения.

1 Введение

Во введении рассказывается, где возникает данная задача, и почему её решение так важно. Вводится на неформальном уровне минимум терминов, необходимый для понимания постановки задачи. Приводится краткий анализ источников информации (литературный обзор): как эту задачу решали до сих пор, в чем недостаток этих решений, и что нового предлагает автор. Формулируются цели исследования. В конце введения даётся краткое содержание работы по разделам; при этом отмечается, какие подходы, методы, алгоритмы предлагаются автором впервые. При упоминании ключевых разделов кратко формулируются основные результаты и наиболее важные выводы.

Цель введения: дать достаточно полное представление о выполненном исследовании и полученных результатах, понятное широкому кругу специалистов. Большинство читателей прочтут именно введение и, быть может, заключение. Во введении автор решает сложную оптимизационную проблему: как сообщить только самое важное, потратив минимум времени читателя, да так, чтобы максимум читателей поняли, о чём вообще идёт речь.

Введение лучше писать напоследок, так как в ходе работы обычно происходит переосмысление постановки задачи. Если же введение писать, когда работа еще не готова, задача усложняется вдвойне. В конце обычно приходит понимание, что всё получилось совсем не так, как планировалось в начале, и исходный вариант введения всё равно придётся переписывать. Кстати, к таким «потерям» надо относиться спокойно — в хорошей работе почти каждый абзац многократно переделывается до неузнаваемости.

Введение имеет много общего с текстом доклада на защите, поэтому имеет смысл готовить их одновременно.

1.1 Постановка задачи

$x_1, x_2, \ldots, x_N$ – точки в некотором векторном пространстве; $y_1, y_2, \ldots, y_N$ – значения, соответствующие этим точкам. При этом $y_i = y(x_i) = f(x_i) + \varepsilon$, где ε – случайная величина с нулевым математическим ожиданием и дисперсией σ^2. Пусть также задана функция $\hat{f}_1(x)$, которая приближает функцию $f(x)$. Стоит задача нахождения $M-1$ функции $\hat{f}_2(x), \hat{f}_3(x), \ldots, \hat{f}_M(x)$, а также функции-ансамбля $\hat{f}(x) = a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_M(x))$, приближающей функцию $f(x)$ лучше чем любая из функций $\hat{f}_i(x),\ i = 1, 2, \ldots, M$.

1.2 Теоретическая часть

Из [3] известно разложение математического ожидания квадратичной ошибки на смещение и дисперсию:

$$\mathbb{E}\left[\left(\hat{f}(x) - y\right)^2\right] = \left(\mathbb{E}\,\hat{f}(x) - f(x)\right)^2 + \left(\mathbb{E}\,\hat{f}^2(x) - \left(\mathbb{E}\,\hat{f}(x)\right)^2\right) + \sigma^2$$

Для уменьшения как смещения, так и дисперсии $\hat{f}(x)$ используются ансамбли:

$$\hat{f}(x) = a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_M(x))$$

В [1] было доказано, что в случае, когда функция $a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_M(x))$ является выпуклой комбинацией своих аргументов:

$$a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_M(x)) = \sum_{i=1}^{M} c_i \hat{f}_i(x),$$

$$\sum_{i=1}^{M} c_i = 1; \quad c_i \geq 0, \quad i = 1, 2, \ldots, M;$$

выполнено

$$\mathbb{E}\left[\left(\hat{f}(x) - y\right)^2\right] = \sum_{i=1}^{M} c_i\,\mathbb{E}\left[\left(\hat{f}_i(x) - y\right)^2\right] - \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} c_i c_j\,\mathbb{E}\left[\left(\hat{f}_i(x) - \hat{f}_j(x)\right)^2\right] \quad (1)$$

Соответственно, для уменьшения среднеквадратичной ошибки необходимо уменьшать среднеквадратичную ошибку каждого предиктора $\hat{f}_i(x)$, а также увеличивать расхождение между прогнозами различных предикторов.

1.3 Существующие методы

В большинстве случаев функция-ансамбль $\hat{f}(x) = a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_M(x))$ является линейной комбинацией функций предикторов $\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_M(x)$. Коэффициенты ищутся при помощи методов регрессионного анализа:

- гребневая регрессия [4]
- метод Лассо [5]
- эластичная сеть [6]
- регрессионная модель, отбирающая признаки, наиболее коррелирующие с откликом [2]

Функции $\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_M(x)$ в таких случаях обычно независимы друг от друга и получены методом оптимизации некоторого функционала.

2 Ансамбль промежуточных решений

Пусть:

X - матрица из N строк, i-ая строка равна x_i;

Y - вектор из N элементов, i-ый элемент равен y_i.

Пусть $\hat{f}_1(x)$ представима в виде функции с параметрами $\tilde{f}(x, \theta)$ и была получена некоторым методом оптимизации функционала среднеквадратичной ошибки $MSE(\hat{f}_1(x), X, Y) = \dfrac{1}{N}\sum\limits_{i=1}^{N}(\hat{f}_1(x_i) - y_i)^2$:

$$\hat{f}_1(x) = \tilde{f}(x, \theta_1)$$

В качестве функции $\tilde{f}(x, \theta)$ может выступать нейронная сеть, тогда параметрами θ будут являться веса нейронной сети.

2.1 Поиск промежуточных решений

Будем искать функции $\hat{f}_2(x), \hat{f}_3(x), \ldots, \hat{f}_M(x)$ в виде $\tilde{f}(x, \theta)$, минимизируя следующий функционал:

$$\mathcal{L}(\hat{f}_k(x), X, Y) = MSE(\tilde{f}(x, \theta_k), X, Y) - \frac{\alpha}{k-1}\sum_{i=1}^{k-1}\left\|\theta_k - \theta_i\right\|^2, \quad k = 2, 3, \ldots, M, \quad (2)$$

где $\alpha \geq 0$ является гиперпараметром.

Данный функционал поощряет функции $\hat{f}_i(x)$ иметь различные параметры θ_i. Таким образом, мы пытаемся добиться максимального расхождения значений функций. Впоследствии это уменьшит среднеквадратичную ошибку функции-ансамбля $\hat{f}(x)$, что следует из уравнения (1).

Стоит отметить, что при больших k вычисление данного функционала может быть вычислительно затратным, так как $\sum\limits_{i=1}^{k-1} \|\theta_k - \theta_i\|^2$ в общем случае требует $O(k)$ времени. Если все θ_i лежат в некотором евклидовом пространстве, данное выражение переписывается в виде:

$$\sum_{i=1}^{k-1} \|\theta_k - \theta_i\|^2 = (k-1)\|\theta_k\|^2 - 2\langle\theta_k, \sum_{i=1}^{k-1}\theta_i\rangle + \sum_{i=1}^{k-1}\|\theta_i\|^2 .$$

Значения выражений $\sum\limits_{i=1}^{k-1} \theta_i$ и $\sum\limits_{i=1}^{k-1} \|\theta_i\|^2$ можно поддерживать в течении всего процесса поиска функции $\hat{f}_k(x)$ и пересчитывать за $O(1)$ времени при переходе к поиску $\hat{f}_{k+1}(x)$.

2.2 Поиск функции-ансамбля

Составим матрицу $\widehat{X}$ размера $N \times M$, где $\widehat{X}_{i,j} = \hat{f}_j(x_i)$. Для поиска функции-ансамбля $\hat{f}(x) = a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_M(x))$ необходимо найти функцию $a(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_M) = a(\hat{x})$, которую будем искать из минимизации функционала $MSE(a(\hat{x}), \widehat{X}, Y)$.

3 Вычислительные эксперименты

Для проверки качества работы представленного метода решим задачу классификации на 3 выборках данных с историческим артериальным давлением. Данные заранее разбиты на обучающую и тестовую выборки. Размеры каждой из выборок представлены в таблице 1. Все параметры являются численными.

Номер выборки	Количество объектов		Количество параметров
	Обучающая выборка	Тестовая выборка	
1	718	458	114
2	373	221	114
3	832	495	114

Таблица 1: Размеры выборок данных

В качестве функции $\tilde{f}(x, \theta)$ используется двухслойная нейронная сеть с сигмоидальной функцией активации, которая оптимизируется методом Adam. В качестве функции-ансамбля $a(\hat{x})$ используется точно такая же нейронная сеть, с другим количеством входных параметров и другой размерностью скрытого слоя.

В ходе эксперимента изначально обучается одна нейронная сеть $\hat{f}_1(x) = \tilde{f}(x, \theta_1)$ и оценивается качество её работы. После этого обучается M новых нейронных сетей $\hat{f}_i(x) = \tilde{f}_i(x, \theta_i)$, а также ищется оптимальная функция-ансамбль $a(\hat{x})$ в виде двухслойной нейронной сети. Затем оценивается качество работы ансамбля промежуточных решений $\hat{f}(x) = a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_M(x))$.

3.1 Результаты эксперимента

Наилучших результатов удалось достичь для $M = 4$. Также при построении ансамбля промежуточных решений был замечен эффект улучшения качества предсказания отдельных нейронных сетей, обученных после первой нейронной сети $\hat{f}_1(x)$. Итоговые результаты эксперимента представлены в таблице 2.

Номер выборки	Одна нейронная сеть	Ансамбль промежуточных решений
1	0.78	0.86
2	0.79	0.83
3	0.78	0.81

Таблица 2: Результаты экспериментов (ROC AUC на тестовой выборке)

3.2 Анализ полученных результатов

Полученные результаты показывают, что ансамбль промежуточных решений, построенных с использованием функции потерь (2), способен достаточно серьёзно увеличивать обобщающую способность предсказания в сравнении с отдельными нейронными сетями. Также замечен эффект улучшения качества отдельных нейронных сетей, построенных данным методом. Попробуем объяснить данный эффект. Предполагается, что он вызван тем, что в случае нахождения плохо минимизирующего параметра θ_1 для нейронной сети $\hat{f}_1(x, \theta_1)$, последующие нейронные сети находят параметры θ_i на отдалении от θ_1. Следовательно, θ_i не находится в области θ_1, которая недостаточно минимизирует функцию потерь.

4 Заключение

В процессе выполнения работы был разработан метод повышения эффективности обучения, основанный на ансамбле промежуточных решений. Были проведены вычислительные эксперименты, которые показали возможную применимость данного метода для улучшения качества нейронных сетей, решающих задачу классификации на реальных данных.

Список литературы

[1] *Докукин А. А., Сенько О. В.* Оптимальные выпуклые корректирующие процедуры в задачах высокой размерности // *Ж. вычисл. матем. и матем. физ.* — 2011. — Т. 51. — С. 1751–1760.

[2] *Докукин А. А., Сенько О. В.* Регрессионная модель, основанная на выпуклых комбинациях, максимально коррелирующих с откликом // *Ж. вычисл. матем. и матем. физ.* — 2015. — Т. 55. — С. 530–544.

[3] *Domingos Pedro.* A unified bias-variance decomposition for zero-one and squared loss // Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence. — AAAI Press, 2000. — Pp. 564–569.

[4] *Ng Andrew Y.* Feature selection, l1 vs. l2 regularization, and rotational invariance // Proceedings of the Twenty-first International Conference on Machine Learning. — ICML '04. — New York, NY, USA: ACM, 2004. — Pp. 78–.

[5] *Tibshirani Robert.* Regression shrinkage and selection via the lasso // *Journal of the Royal Statistical Society. Series B (Methodological).* — 1996. — Vol. 58, no. 1. — Pp. 267–288.

[6] *Zou Hui, Hastie Trevor.* Regularization and variable selection via the elastic net // *Journal of the Royal Statistical Society, Series B.* — 2005. — Vol. 67. — Pp. 301–320.