DOI: 10.15514/ISPRAS-2019-31(4)-7

Регуляризация Байеса при подборе весовых коэффициентов в ансамблях предикторов

А.С. Нужный, ORCID: 0000-0003-3319-2523 <nuzhny@inbox.ru>
Институт проблем безопасного развития атомной энергетики РАН,
Россия, 115191, г. Москва, Большая Тульская ул., д. 52

Аннотация. В статье рассматривается задача обучения с учителем: требуется восстановить зависимость, отображающую векторное множество в скалярное по конечному набору примеров такого отображения – обучающей выборке. Данная задача относится к классу обратных задач, и, как и большинство обратных задач, является математически некорректной. Это выражается в том, что если строить решение методом наименьших квадратов по точкам обучающей выборки, то можно столкнуться с переобучением – ситуацией, когда модель хорошо описывает обучающее множество, но дает большую ошибку на тестовом. Нами применяется подход, когда решение ищется в виде ансамбля предиктивных моделей. Ансамбли строятся с использованием метода бэггинга. В качестве базовых обучаемых моделей в работе используются персептроны и деревья решений. Конечное решение получается путем взвешенного голосования предикторов. Весовые коэффициенты подбираются путем минимизации ошибки ансамбля на обучающей выборке. Для борьбы с переобучением при подборе весовых коэффициентов применяется байесовская регуляризация решения. Чтобы подобрать параметры регуляризации, в работе предложено использовать метод ортогонализованных базисных функций, который позволяет получить их оптимальные значения без использования ресурсоемких итерационных процедур.

Ключевые слова: обучение с учителем; бэггинг; некорректные задачи; Байесовская регуляризация обучения.

Для цитирования: Нужный А.С. Регуляризация Байеса при подборе весовых коэффициентов в ансамблях предикторов. Труды ИСП РАН, том 31, вып. 4, 2019 г., стр. 113-120. DOI: 10.15514/ISPRAS-2019-31(4)-7

Bayes regularization in the selection of weight coefficients in the predictor ensembles

A.S. Nuzhny, ORCID: 0000-0003-3319-2523 <nuzhny@inbox.ru>
Nuclear Safety Institute of the Russian Academy of Sciences,
52 Bolshaya Tulskaya st., Moscow 115191, Russia

Abstact. The supervised learning problem is discussed in the article: it is necessary to restore the dependence that maps a vector set into a scalar based on a finite set of examples of such a mapping - a training sample. This problem belongs to the class of inverse problems, and, like most inverse problems, is mathematically incorrect. This is expressed in the fact that if you construct the solution using the least squares method according to the points of the training sample, you may encounter retraining – a situation where the model describes the training set well, but gives a big error on the test one. We apply the approach when a solution is sought in the form of an ensemble of predictive models. Ensembles are built using the bagging method. Perceptrons and decision trees are considered as basic learning models. The final decision is obtained by weighted voting of predictors. Weights are selected by minimizing model errors in the training set. To avoid over-fitting in the selection of weights, Bayesian regularization of the solution is applied. In order to choose regularization parameters, it is proposed to use the method of orthogonalized basic functions, which allows obtaining their optimal values without using expensive iterative procedures.

Ключевые слова: supervised learning; bagging; ill-passed problem; Bayesian regularization of learning

Для цитирования: Nuzhny A.S. Bayes regularization in the selection of weight coefficients in the predictor ensembles. Trudy ISP RAN/Proc. ISP RAS, vol. 31, issue 4, 2019. pp. 113-120 (in Russian). DOI: 10.15514/ISPRAS-2019-31(4)-7

1. Введение

В работе рассматривается задача обучения с учителем, когда требуется восстановить зависимость, отображающую векторное множество X в скалярное Y по конечному набору примеров такого отображения – обучающей выборке: $D = \{y_i, \vec{x}_i\}_{i=1}^{L}$. Данная задача относится к классу обратных задач, и, как и большинство обратных задач, является математически некорректной. Это выражается в том, что если строить решение методом наименьших квадратов по точкам обучающей выборки, то можно столкнуться с переобучением – ситуацией, когда модель хорошо описывает обучающее множество, но дает большую ошибку на тестовом.

Для борьбы с переобучением существует несколько подходов: метод валидационных выборок [1], метод регуляризации обучения [2], построение решения как суперпозиции ансамблей независимых предикторов [3]. В последнем случае работает эффект взаимной компенсации ошибок отдельных независимых предикторов при голосовании. Сами модели, участвующие в голосовании, строятся независимо друг от друга по обучающей выборке D.

Очевидно, что точность аппроксимации при таком подходе будет ограничена сверху тем обстоятельством, что конечная обучающая выборка позволит построить только конечное число независимых предикторов. После этого определенного улучшения точности предсказания можно добиться, например, введя дополнительно весовые коэффициенты для предикторов и проводя взвешенное голосование. Однако задача подбора весовых коэффициентов сама по себе также является некорректной, требующей регуляризации.

Для решения этой проблемы в статье предлагается искать весовые коэффициенты a_n перед предикторами $\Psi_n(\vec{x})$ путем минимизации квадратичной ошибки обучения в сумме со стабилизирующим функционалом в гауссовой форме:

$$\sum_i^L \left(y_i - \sum_{n=1}^N a_n \Psi_n(\vec{x}) \right)^2 + \lambda \sum_{n=1}^N a_n^2 , \quad (1)$$

где N – число базисных предикторов, а λ – регуляризационный множитель. Его значение будет находиться по методу Байеса [2,4].

В общем виде Байесовский метод поиска регуляризационного множителя сводится к дорогостоящей итерационной процедуре [2,4]. Однако в случае, когда решение ищется в виде ряда по набору базисных функций, а стабилизирующий функционал берется в гауссовой форме, можно применить метод ортогонализованных базисных функций (ОБФ) [5], который дает аналитическое выражение для λ, что существенно упрощает вычисления.

Ниже будет кратко изложена идея бэггинга на примере базовых обучаемых моделей двух типов – многослойных персептронов и деревьев решений; дано описание метода ортогонализованных базисных функций; предложен комбинированный алгоритм, использующий оба этих подхода; приведены результаты апробации предложенного алгоритма.

2. Ансамбли предиктивных моделей

В [3] был рассмотрен подход к задаче обучения с учителем, получивший название бэггинг. В нем решение строилось как суперпозиция различных (независимых) предиктивных моделей.

При этом сами модели получались обучением одного математического алгоритма на разных подмножествах обучающих данных. Было показано, что сложный предиктор, полученный голосованием ансамбля предиктивных моделей, дает в вероятностном смысле более точное решение по сравнению с простым предиктором.

Для построения ансамбля независимых предикторов с помощью какой-либо обучаемой модели необходимо обеспечить вариабельность процедуры обучения, внести в нее некоторый случайный фактор, чтобы различные эпохи обучения приводили в общем случае к разным конечным предикторам. Одним из способов обеспечения такой вариабельности является бэггинг, когда модель учится не на всем множестве данных, а только на случайной подвыборке. В результате предикторы, построенные на разных подмножествах данных, будут различны между собой и смогут создавать функциональный базис.

Кроме того, разнообразие предикторов может быть достигнуто благодаря особенностям самих обучаемых моделей. Например, многослойные персептроны [6] обучаются путем корректировки своих весов в сторону уменьшения ошибки обучения от некоторых начальных значений. В результате обучение сходится к состоянию, соответствующему некоторому минимуму ошибки (не обязательно глобальному). Какой конкретно минимум будет получен в результате обучения, зависит, в частности, от начальных значений весов персептрона. Таким образом, стартуя от разных начальных значений, мы будем получать разные предикторы.

В качестве еще одного примера можно привести деревья принятия решений, которые формируют набор разделяющих правил на основании обучающей выборки. При выборе разделяющего правила в узле дерева обычно используется так называемый «жадный» принцип, когда из всех возможных вариантов остается тот, который на данном шаге дает наилучшее значение разделяющего критерия [7]. Однако если при выборе разделяющего правила в каждом конкретном узле рассматривать только некоторое подпространство случайно выбранных входных признаков, а не все их множество, то можно построить ансамбль различных между собой деревьев решений. Данная идея реализуется в алгоритме, известном как *случайный лес* [8].

После того, как ансамбль предикторов $\Psi_n(\vec{x})$ построен, конечное решение может быть получено путем их простого голосования:

$$h(\vec{x}) = \frac{1}{N}\sum_{n=1}^{N}\Psi_n(\vec{x})$$

или взвешенного голосования:

$$h(\vec{x}) = \sum_{n=1}^{N} a_n\Psi_n(\vec{x}). \quad (2)$$

В данной работе рассматривается модель взвешенного голосования. Веса полагаются адаптивными параметрами ансамбля и подбираются путем минимизации ошибки обучения. Взвешенное голосование часто позволяет достичь лучшего или сравнимого с простым голосованием результата меньшим числом модулей, что делает модель более интерпретируемой и вычислительно менее затратной.

3. Метод ортогонализованных базисных функций

В методе ОБФ поиск решения $h(\vec{x})$ ведется в виде ряда по набору базисных функций (2). Предполагается, что векторы значений базисных функций в точках обучающей выборки $\vec{\Psi}_n = \{\Psi_n(\vec{x}_1), \Psi_n(\vec{x}_2)\}, \dots, \Psi_n(\vec{x}_L)$ ортогональны:

$$\sum_{i=1}^{L}\Psi_{m_i}\Psi_{n_i} = \delta_{m,n} \quad (3)$$

Если это условие не выполнено, то мы всегда можем построить линейное преобразование, приводящее к ортогональному набору векторов, который можно трактовать как значения

некоторых новых базисных функций в точках обучающей выборки. После того, как будут найдены коэффициенты разложения по ортогонализованным функциям, коэффициенты разложения решения по исходным функциям получаются обратным линейным преобразованием.

Делается стандартный для метода байесовской регуляризации набор предположений. Во-первых, предполагается, что данные зашумлены гауссовым шумом. В этом случае вероятность генерации решением $h(\vec{x})$ обучающего примера $y_i, \vec{x}_i$ можно оценить выражением:

$$P(y_i|h) = \frac{1}{Z_X}exp\left(-\beta\left(y_i - h(\vec{x}_i)\right)^2\right),$$

а генерации всей выборки $D = \{y_i, \vec{x}_i\}_{i=1}^{L}$, соответственно,

$$P(D|h) = \frac{1}{Z_X}exp\left(-\beta\sum_{i=1}^{L}\left(y_i - h(\vec{x}_i)\right)^2\right),$$

где β – параметр модели.

Во-вторых, делается предположение об априорной вероятности выбора того или иного решения. Априорная вероятность записывается в гауссовой форме:

$$P(h|\alpha,\beta) = \frac{1}{Z_A}exp\left(-\alpha\sum_{n=1}^{N}a_n^2\right).$$

Здесь также α – параметр модели.

По формуле Байеса

$$P(h|D,\alpha,\beta) = \frac{P(D|h)P(h|\alpha,\beta)}{P(D|\alpha,\beta)} \quad (4)$$

вероятность решения $h(\vec{x})$ будет равна [4]:

$$P(h|D,\alpha,\beta) = \frac{1}{Z_M}e^{-M}\ M = \beta, \text{где}\ \sum_{i=1}^{L}(y_i - h_i)^2 + \alpha\sum_{n=1}^{N}a_n^2$$

В приведенных формулах Z_X, Z_A, Z_M – нормировочные коэффициенты, которые получаются из условий нормировки соответствующих вероятностей на единицу. Если положить $\lambda = \frac{\alpha}{\beta}$, то функционал M будет полностью эквивалентен выражению (1).

Решение (2) ищется как наиболее вероятное. Максимизация вероятности решения $P(h|D,\alpha,\beta)$ (или минимизация M) по коэффициентам разложения a_n в случае выполнения условия (3) приводит к следующему выражению для их значений [5]:

$$a_n = \frac{\beta}{\beta+\alpha}\sum_{i=1}^{L}y_i\Psi_{n_i} \quad (5)$$

Параметры α и β находятся, как наиболее правдоподобные, путем максимизации логарифма знаменателя в формуле Байеса (4) [2, 4]. Последний выражается через нормировочные коэффициенты при экспонентах:

$$lnP(D|H) = lnZ_M - lnZ_A - lnZ_X.$$

В результате, как показано в [5], получается следующее выражение для логарифма знаменателя:

$$lnP(D|H) = \frac{\beta^2 S}{\beta+\alpha} + \frac{N}{2}ln\frac{\alpha}{\beta+\alpha} - \beta\vec{y}^2 + \frac{L}{2}ln\frac{\beta}{\pi},$$

где

$$S = \sum_{n=1}^{N}\left(\vec{y}\vec{\Psi}_n\right)^2, \quad (6)$$

$\vec{y}$ – вектор значений искомой функции в точках обучающей выборки.

Максимизация данного выражения по α и β сводится к системе двух нелинейных уравнений, которая имеет единственное аналитическое решение для параметров модели. В результате α и β выражаются через число точек в обучающем множестве L, число базисных функций N, квадрат вектора значений функции в точках обучающего множества $\vec{y}$ и величину (6):

$$\alpha = \frac{1}{2}\frac{(L-N)}{\left(\frac{L}{N}S - y^2\right)}, \qquad \beta = \frac{1}{2}\frac{(L-N)}{(\vec{y}^2 - S)} \tag{7}$$

Подробные выкладки приведены в работе [5].

Вычислив регуляризационные параметры α и β, можно получить коэффициенты разложения по ортогонализованным функциям (5) и, применив к ним преобразование, обратное ортогонализующему, получить коэффициенты разложения по исходным функциям. Таким образом, метод ортогонализованных базисных функций приводит к единственному решению для коэффициентов a_n в выражении (2), соответствующему максимуму правдоподобия.

4. Алгоритм построения взвешенного ансамбля

В качестве базовых обучаемых моделей в работе рассматривались многослойные персептроны и деревья принятия решений. Архитектура персептронов подбиралась заведомо простой, чтобы переобучение отдельного персептрона на обучающей выборке было невозможно. Их разнообразие обеспечивалось тем, что в каждом случае обучение проводилось от разных начальных значений синоптических коэффициентов и на разных подвыборках обучающих данных.

В случае деревьев принятия решений, разнообразие предикторов обеспечивалось тем, что каждое дерево строилось на случайно выбранной подвыборке обучающего множества (бэггинг).

Алгоритм построения решения выглядит следующим образом:

1) на обучающем множестве строится серия различных между собой предикторов;
2) вычисляются векторы значений полученных базисных предикторов в точках обучающей выборки;
3) векторы значений проверяются на линейную независимость, если условие независимости не выполнено, то число предикторов уменьшается;
4) строится ортогонализующее линейное преобразование этих векторов;
5) вычисляются параметры α и β по формулам (7);
6) находятся коэффициенты разложения по ортогонализованным базисным функциям (5);
7) рассчитываются коэффициенты разложения решения по исходным предикторам, для чего к коэффициентам разложения по ортогонализованным функциям применяется линейное преобразование, обратное ортогонализующему.

В работе [4] рассматривалась задача аппроксимации функции по точкам. Решение искалось в виде ряда по набору базисных функций путем минимизации функционала (1). Для поиска регуляризационного множителя применялся байесовский подход к регуляризации решения. Предложенный в [4] итерационный алгоритм выбора параметров модели дает значения весовых коэффициентов a_n, близкие к коэффициентам, полученным алгоритмом, рассматриваемым в данной статье. Однако алгоритм, рассмотренный в [4], обладает большими вычислительными затратами.

Итерационный алгоритм использует процедуру, где на каждом шаге решается система линейных алгебраических уравнений (СЛАУ), число уравнений которой равно числу базисных функций N. Таким образом, количество операций, необходимое для одной итерации, пропорционально N^3, а сложность всего алгоритма $\sim N^3 T$, где T – число итераций. Для надежной сходимости алгоритму обычно требуется не менее 10 итераций.

Наиболее ресурсоемким этапом рассматриваемого в данной статье алгоритма является шаг 4 – переход к ортогональному представлению. Вычислительная стоимость такой процедуры аналогична стоимости решения СЛАУ $\sim N^3$ операций. Таким образом, вычислительная стоимость приведенного алгоритма – порядка N^3 операций, что кратно меньше стоимости итерационного алгоритма $\sim N^3 T$.

Приведенная в работе модель сравнивалась с решением, полученным простым голосованием предикторов. Тестирование проводилось на стандартном дата сете для задачи регрессии, содержащем информацию о ценах на жилые объекты в Бостоне и параметрах, характеризующих эти объекты [9]. Данные были взяты из библиотеки scikit-learn (http://scikit-learn.org). Всего в выборке содержится 506 точек. В экспериментах 400 точек были использованы для обучения, на оставшихся 106-и проводилось тестирование.

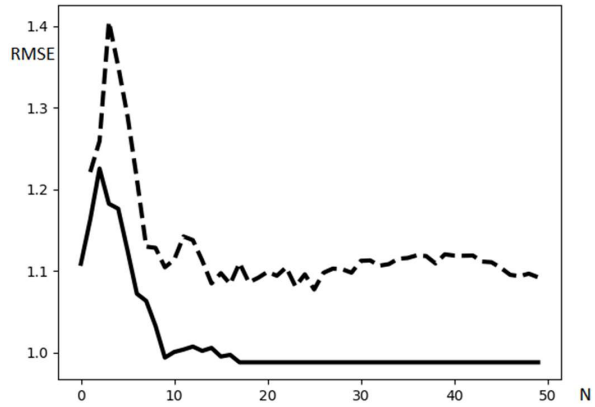

Рис. 1. Графики зависимости значения RMSE от числа персептронов: пунктирная линия – решение ищется методом простого голосования, сплошная – решение представляется суммой модулей с весами, рассчитанными методом ОБФ

Fig. 1. Plots of dependence of RMSE value on the number of perceptrons: dotted line – solution is determined by simple voting, solid line – solution is determined by the voting with weights, calculated by OBF method

В первом эксперименте в качестве базовой модели использовался многослойный персептрон.

На рис. 1 приведены графики зависимости квадратного корня из среднеквадратичных ошибок (RMSE) ансамблей предикторов на тестовом множестве от числа персептронов в них.

Пунктирная линия демонстрирует изменение RMSE модели, в которой решение выбирается простым усреднением модулей, сплошная линия – изменение RMSE модели, в которой веса перед модулями выбирались методом ортогонализованных базисных функций.

На рис. 2 приведены аналогичные графики для ансамблей, в которых в качестве базовой модели использовались деревья решений. Пунктирный график – зависимость RMSE от числа предикторов ансамбля, в котором решение получалось путем простого голосования, сплошной – та же зависимость для метода, в котором веса подбирались с помощью ОБФ.

В большинстве экспериментов ошибка взвешенного голосования оказывалась меньше, чем простого. Как правило, метод взвешенного голосования начинал выигрывать уже при малом количестве предикторов. На графиках видно, что в какой-то момент ошибка взвешенного голосования перестает меняться. Это связано с тем, что добавление очередного предиктора приводит к их линейной зависимости и последний добавленный модуль автоматически исключается алгоритмом. Таким образом, подбор коэффициентов методом ортогонализованных базисных функций позволяет не только добиваться более высокой точности ансамбля, но и ограничиваться при этом меньшим количеством модулей.

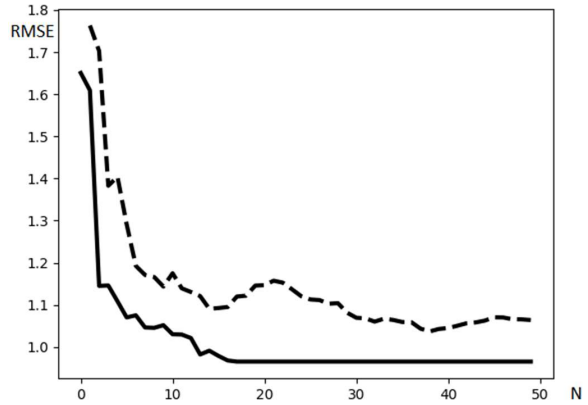

Рис. 2. Графики зависимости значения RMSE от числа решающих деревьев: пунктирная линия – решение ищется методом простого голосования, сплошная – решение представляется суммой модулей с весами, рассчитанными методом ОБФ
Fig. 2. Plots of dependence of RMSE value on the number of decision trees: dotted line – solution is determined by simple voting, solid line – solution is determined by the voting with weights, calculated by OBF method

5. Заключение

В работе рассматривалась задача подбора весовых коэффициентов в ансамбле голосующих предикторов. Задача решалась методом минимизации регуляризационного функционала, состоящего из суммы ошибки обучения и стабилизирующего функционала, взятого в гауссовой форме. Регуляризационный множитель подбирался по методу Байеса. Применение метода ортогонализованных базисных функций позволило уйти от итерационной процедуры подбора регуляризационного множителя. В этом подходе регуляризационный множитель имеет аналитическое выражение через исходные данные.

Предложенный алгоритм сравнивался с методом простого голосования предикторов. Проведенные численные эксперименты показали, что применение метода ортогонализованных базисных функций для подбора весовых коэффициентов, позволяет, во-первых, получать решение, приводящее, как правило, к меньшей ошибке на тестовой выборке, во-вторых, включающее меньшее число исходных модулей.

Список литературы / References

[1]. H.Zhu, R.Rohwer. No free lunch for cross-validation. Neural Computation, vol. 8, issue 7, 1996, pp. 1421-1426.
[2]. David J. C. MacKay. Bayesian Interpolation. Neural Computation, vol. 4, issue 3, 1992, pp. 415-447.
[3]. Breiman L. Bagging predictors. Machine Learning, vol. 24, no. 2, 1996, pp. 123–140.
[4]. А.С. Нужный, С.А. Шумский. Регуляризация Байеса в задаче аппроксимации функции многих переменных. Математическое моделирование, 2003, том 15, № 9, стр. 55-63 / A.S. Nuzhny, S.A. Shumsky. The Bayes regularization in the problem of function of many variables approximation. Matematical Modeling, vol. 15, no. 9, 2003, pp. 55–63 (in Russian).
[5]. A.S. Nuzhny. Bayesian regularization in the problem of point-by-point function approximation using orthogonalized basis. Mathematical Models and Computer Simulations, vol. 4, issue 2, 2012, pp. 203-209.
[6]. Simon Haykin. Neural Networks: A Comprehensive Foundation (2nd Edition). Prentice Hall, 1998, 842 p.
[7]. Breiman L., Friedman J.H., Stone C.J. , Olshen R.A. Classification and regression trees. Chapman and Hall/CRC, 1984, 368 p.
[8]. Breiman Leo. Random Forests. Machine Learning, vol. 45, no. 1, 2001, pp. 5—32.

[9]. Harrison D. and Rubinfeld D.L. Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management, vol. 5, no. 1, 1978, pp. 81-102.

Информация об авторе / Information about the author

Антон Сергеевич НУЖНЫЙ – кандидат физико-математических наук, научный сотрудник ИБРАЭ РАН. Его научные интересы включают машинное обучение, распознавание образов, интеллектуальный анализ данных, информационный поиск, некорректные задачи.

Anton Sergeevich NUZHNY – Candidate of Physics and Mathematics, Research Fellow, IBRAE RAS. His research interests include machine learning, pattern recognition, data mining, information retrieval, and incorrect tasks.