ОПТИМИЗАЦИЯ ПАРАМЕТРОВ SVM-РЕГРЕССИИ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА ГЛОБАЛЬНОГО ПОИСКА

К.А. Баркалов, И.Б. Мееров, А.Н. Половинкин, С.В. Сидоров, Н.Ю. Золотых

Нижегородский госуниверситет им. Н.И. Лобачевского

Рассматривается задача выбора оптимальных параметров для метода построения регрессии с использованием алгоритма опорных векторов. Предлагается подход, основанный на оптимизации функции ошибки перекрестного контроля с использованием алгоритма глобального поиска для решения задачи глобальной оптимизации. Приводятся результаты вычислительного эксперимента.

Введение

Одним из распространенных методов для решения задачи восстановления регресии является алгоритм SVM-регрессии [1]. Среди преимуществ данного алгоритма можно выделить возможность эффективного моделирования нелинейных зависимостей, а также независимость его обобщающей способности от размерности пространства признаков. Однако в отдельных случаях практическое применение алгоритма ограничено вследствие того, что точность метода сильно зависит от выбора его параметров [3]. Наиболее часто применяемые на практике подходы для выбора оптимальных параметров в итоге сводятся к решению задачи глобальной оптимизации [2, 3]. В данной работе предлагается новый метод, основанный на оптимизации функции ошибки перекрестного контроля с использованием основанного на информационно-статистическом подходе алгоритма глобального поиска.

1. Оптимизация параметров алгоритма SVM-регресии

В работе рассматривается задача восстановления регрессии в следующей постановке. Пусть задана обучающая выборка $\{(x_i, y_i), i = 1,..., N\}$, где $x_i \in R^d$ – вектор признаков, $y_i \in R$ – ответ. Требуется найти функцию $f(x)$, принадлежащую некоторому определенному классу K, минимизирующую значение эмпирического риска (ошибка предсказания на обучающей выборке). Для алгоритма построения SVM-регрессии функцию $f(x)$ в общем виде можно записать как

$$f(x) = w^T \phi(x) + b,$$

где $\phi(x)$ – нелинейное (в общем случае) отображение $R^d \to R^m$, $w \in R^m$ – вектор коэффициентов линейной функции в новом пространстве признаков R^m. В качестве функции потерь используется кусочно-линейная функция ε-чувствительности:

$$L_\varepsilon(y, f(x)) = \max(0, |y - f(x)| - \varepsilon),$$

где ε – заранее заданный порог (в случае если предсказанное значение отличается от истинного на величину, меньшую данного порога, ошибка считается равной нулю). Функция эмпирического риска записывается в виде:

$$R_{emp}(w) = \frac{1}{N} \sum_{i=1}^{N} L_\varepsilon(y_i, f(x_i)).$$

Рассматриваемая задача минимизации эмпирического риска сводится к задаче квадратичной оптимизации

$$\min_w \frac{\|w\|^2}{2} + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right), \qquad (1)$$
$$f(x_i) - y_i \le \varepsilon + \xi_i,\ y_i - f(x_i) \le \varepsilon + \xi_i^*,\ \xi_i, \xi_i^* \ge 0,$$

где C – параметр регуляризации, который соответствует отношению сложности модели и эмпирической ошибки в минимизируемой функции. Используя метод множителей Лагранжа, задачу (1) можно свести к двойственной форме:

$$\max \sum_{i=1}^{N} y_i\left(\alpha_i^* - \alpha_i\right) - \varepsilon\sum_{i=1}^{N}\left(\alpha_i^* + \alpha_i\right) - \frac{1}{2}\sum_{i,j=1}^{N}\left(\alpha_i^* - \alpha_i\right)\left(\alpha_j^* - \alpha_j\right)K(x_i, x_j), \qquad (2)$$
$$\sum_{i=1}^{N}\left(\alpha_i^* - \alpha_i\right) = 0,\ 0 \le \alpha_i, \alpha_i^* \le C,$$

где α_i, α_i^* – множители Лагранжа, $K(x_i, x_j)$ – функция ядра, которая соответствует скалярному произведению в новом пространстве признаков R^m. Одними из наиболее часто применяемых на практике ядер являются радиальные базисные функции:

$$K\left(x_i, x_j\right) = \exp\left(-\left\|x_i - x_j\right\|^2 \Big/ 2\sigma^2\right).$$

Как показано в [3], обобщающая способность алгоритма SVM-регрессии существенным образом зависит от выбора параметров C, ε и σ. В работе [2] предлагается метод, основанный на минимизации ошибки перекрестного контроля. Идея метода заключается в разделении тренировочной выборки случайным образом на S частей $\{G_s, s = 1,...,S\}$, обучении модели на $(S-1)$ части и использовании оставшейся части для вычисления тестовой ошибки. Усредненная ошибка по всем тестовым множествам используется в качестве оценки обобщающей способности алгоритма

$$MSE_{CV} = \frac{1}{N}\sum_{s=1}^{S}\sum_{i \in G_s}\left(y_i - f(x_i \mid \theta_s\right)^2,$$

где θ_s – решение задачи (2), полученное при использовании в качестве тренировочной выборки множества $D \setminus \theta_s$. В случае когда число объектов в обучающей выборке невелико, можно использовать LOO (leave-one-out) ошибку

$$MSE_{LOO} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f(x_i \mid \theta_i\right)^2,$$

где θ_i – решение задачи (2), полученное при использовании в качестве тренировочной выборки множества $D \setminus \{(x_i, y_i)\}$. В силу того, что решение задачи квадратичного программирования (2) для каждого набора параметров (C, ε, σ) существует и единственно, для заданной тренировочной выборки можно рассмотреть leave-one-out-ошибку MSE_{LOO} как фунцию, зависящую от C, ε и σ:

$$MSE_{LOO} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f\left(x_i \mid \theta_i(C, \varepsilon, \sigma)\right)\right)^2 = F(C, \varepsilon, \sigma). \qquad (3)$$

Таким образом, задача выбора оптимальных параметров алгоритма построения SVM-регрессии свелась к задаче минимизации функции $F(C, \varepsilon, \sigma)$. В общем случае функция является многоэкстремальной, что требует применения алгоритмов глобальной оптимизации для нахождения её оптимума.

2. Параллельный алгоритм глобального поиска

Рассматриваемый алгоритм основан на подходе к решению задач условной глобальной оптимизации, использующем информационно-статистическую теорию глобального поиска [4, 5, 7].

Рассмотрим задачу безусловной глобальной оптимизации вида

$$\varphi = \varphi(y) = \min\{\varphi(y) : y \in D\},$$
$$D = \{y \in R^N : a_i \le y_i \le b_i,\ 1 \le i \le N\}, \quad (4)$$

где целевая функция $\varphi(y)$ удовлетворяет условию Липшица с соответствующей константой L (которая в общем случае может быть не задана). В обсуждаемом подходе (см. [4, 6]) решение многомерных задач сводится к решению эквивалентных им одномерных задач (редукция размерности). Идея метода заключается в том, чтобы, используя кривые типа развертки Пеано $y(x)$, однозначно отображающие отрезок $[0,1]$ на N-мерный гиперкуб D,

$$D = \{y \in R^N : -2^{-1} \le y_i \le 2^{-1},\ 1 \le i \le N\} = \{y(x) :\ 0 \le x \le 1\},$$

свести исходную задачу к следующей одномерной задаче:

$$\varphi(y(x^*)) = \min\{\varphi(y(x)) : x \in [0,1]\}. \quad (5)$$

Таким образом, рассматриваемая схема редукции размерности сопоставляет многомерной задаче с липшицевой минимизируемой функцией и липшицевыми ограничениями одномерную задачу, в которой соответствующие функции удовлетворяют равномерному условию Гельдера.

Недостатком подобного подхода является потеря части информации о близости точек в многомерном пространстве. Это объясняется тем, что точка $x \in [0,1]$ имеет лишь левых и правых соседей, а соответствующая ей точка $y(x) \in R^N$ имеет соседей по 2^N направлениям. Сохранить часть информации о близости точек позволяет использование множества отображений

$$Y_L(x) = \{y^1(x), ..., y^L(x)\} \quad (6)$$

вместо применения единственной кривой Пеано $y(x)$. Каждая кривая Пеано $y^i(x)$ из $Y_L(x)$ может быть получена в результате некоторого сдвига вдоль главной диагонали гиперинтервала D [5] или поворота [8].

Использование множественных отображений (6) позволяет решать задачу (4) путем параллельного решения L задач вида (5) на наборе отрезков $[0,1]$. Каждая одномерная задача решается на отдельном процессоре с использованием развертки $y^s, 1 \le s \le L$. Результаты испытания в точке x^k, полученные конкретным процессором для решаемой им задачи, интерпретируются как результаты испытаний во всех остальных задачах (в соответствующих точках $x^{k1}, ..., x^{kL}$) и рассылаются другим процессорам. При таком подходе испытание в точке $x^k \in [0,1]$, осуществляемое в s-й задаче, состоит в последовательности действий:

1. Определить образ $y^k = y^s(x^k)$ при соответствии $y^s(x)$;
2. Проинформировать остальные процессоры о начале проведения испытания в точке y^k (*блокирование* точки y^k);
3. Вычислить величину $\varphi(y)$. Пара $\{y^s(x^k), z^k = \varphi(y^s(x^k))\}$ является результатом испытания в точке x^k;

4. Определить прообразы $x^{kl} \in [0,1]$, $1 \le l \le L$, точки y^k и интерпретировать испытание, проведенное в точке $y^k \in D$, как проведение испытаний в L точках $x^{k1},...,x^{kL}$ с одинаковыми результатами $\varphi(y^1(x^{k1}) = ... = \varphi(y^L(x^{kL})) = z^k$.
5. Проинформировать остальные процессоры о результатах испытания в точке y^k, разослав им пары (y^k, z^k).

Каждый процессор имеет свою копию программных средств, реализующих вычисление функций задачи, и решающее правило алгоритма. Для организации взаимодействия на каждом процессоре создается очередь, в которую процессоры помещают информацию о выполненных итерациях в виде пар: точка очередной итерации и значение из (4).

Различные варианты алгоритма глобального поиска для решения одномерных и многомерных задач и соответствующая теория сходимости представлены в работах [4-8].

3. Результаты вычислительного эксперимента

Рассмотрим функцию $y(x) = 0.5 + 0.4\sin(2\pi x)\cos(6\pi x)$. Пусть $x_i = 0.05 \cdot (i-1)$, $i = 1,...,21$, $y_i = y(x_i) + N(0,0.05)$, где $N(0,0.05)$ – гауссово распределение с математическим ожиданием, равным 0, и стандартным отклонением, равным 0.05. На рис. 1 приведен график зависимости ошибки перекрестного контроля MSE_{LOO} от ε и σ при фиксированном значении параметра C=1. Как видно из графика, функция, описывающая данную зависимость, содержит несколько локальных минимумов в области поиска.

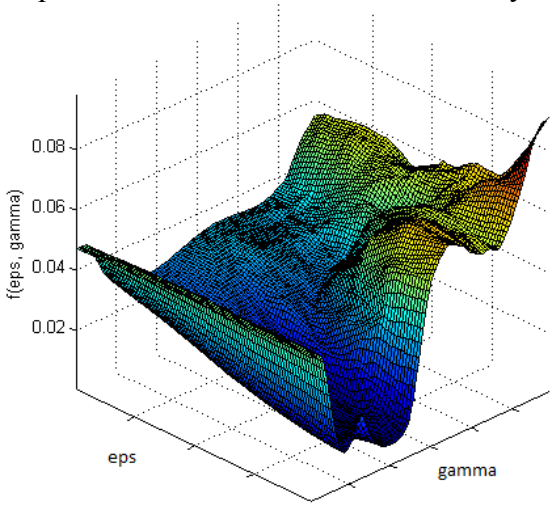

Рис. 1. График зависимости ошибки перекрестного контроля от параметров

Для каждой задачи использовалось правило остановки $\|y - y^*\| \le \rho$, т.е. достижение известного глобального оптимума с заданной точностью ρ в евклидовой метрике.

При решении использовались следующие параметры алгоритма: точность поиска $\rho = 0.01$, параметр надежности $r = 2.5$, точность развертки $M = 12$, количество разверток $L = 2$. Общее число итераций, выполненных параллельной версией алгоритма до достижения критерия останова, равно 1550 (для достижения той же точности методом полного перебора потребовалось бы 10000 итераций), найденное значение оптимума равно 0.006728.

4. Заключение

В работе предложен новый метод для нахождения оптимальных параметров SVM-регрессии, основанный на оптимизации функции ошибки перекрестного контроля с ис-

пользованием основанного на информационно-статистическом подходе алгоритма глобального поиска. Проведенный вычислительный эксперимент показал преимущество подхода над методом полного перебора, обычно используемым на практике. Среди возможных продолжений работы наиболее перспективным представляется расширение метода для нахождения оптимальных параметров алгоритмов обучения с учителем, предназначенных для решения задач классификации.

Работа выполнена в рамках программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы», государственный контракт № 11.519.11.4015.

Литература

1. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. – Springer, 2008.
2. Ito K., Nakano R. Optimization Support Vector Regression Hyperparameters Based on Cross-Validation // Proceedings of the International Joint Conference on Neural Networks. 2003. Vol. 3. P. 2077–2083.
3. Ren Yu., Bai G. Determination of Optimal SVM Parameters by Using GA/PSO // Journal of Computers. 2010. Vol. 5. No 8, P. 1160-1168.
4. Стронгин Р.Г. Численные методы в многоэкстремальных задачах (Информационно-статистические алгоритмы). М.: Наука, 1978.
5. Стронгин Р.Г. Параллельная многоэкстремальная оптимизация с использовани-ем множества разверток // Ж. вычисл. матем. и матем. физ. 1991. Т.31, №8. С. 1173 – 1185.
6. Strongin R.G., Sergeyev Ya.D. Global optimization with non-convex constraints. Sequential and parallel algorithms. Kluwer Academic Publishers, Dordrecht, 2000.
7. Gergel V.P., Strongin R.G. Parallel computing for globally optimal decision making on cluster systems // Future Generation Computer Systems. 2005. Vol. 21. № 5. P. 673-678.
8. Баркалов К.А., Сидоров С.В., Рябов В.В. Параллельные вычисления в задачах многоэкстремальной оптимизации // Вестник ННГУ. 2009. №6(1). С. 171-177.