

# Методы оценки неопределенности при использовании леса решающих деревьев в алгоритме глобальной оптимизации\*

Буянов А.Д.

Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского  
г. Нижний Новгород, Российская Федерация  
E-mail: [ar.buyanow@yandex.ru](mailto:ar.buyanow@yandex.ru)

**Аннотация.** В работе исследуется применение леса решающих деревьев в методах глобальной оптимизации в качестве альтернативы традиционному кригингу на основе гауссовских процессов. Рассматриваются преимущества RDF, включая линейную сложность обучения, устойчивость к шуму и возможность работы с категориальными признаками, а также его ограничения, такие как кусочно-постоянный характер предсказаний и отсутствие естественной оценки неопределенности. Предлагается четыре метода вычисления неопределенности для RDF: ковариационный, метод пересечений, метод ближайшего соседа и квадратичной аппроксимации, каждый из которых обеспечивает различный компромисс между точностью и вычислительной эффективностью. Для выбора новых точек в алгоритме SEGO разработан эвристический подход, учитывающий зоны минимального значения суррогата и максимальной неопределенности. Эксперименты на различных классах тестовых функций (уни-modalных, овражных, периодических, многоэкстремальных и зашумленных) показали, что предложенные методы превосходят кригинг по скорости оптимизации и качеству результатов, особенно на многомерных задачах.

**Ключевые слова:** глобальная оптимизация, лес решающих деревьев, суррогатная модель, оценка неопределенности, кригинг, SEGO.

## ВВЕДЕНИЕ

В задачах многомерной глобальной оптимизации, относящихся к классу NP-сложных проблем [1, 2], общем случае не существует алгоритмов, гарантированно находящих глобальный экстремум за полиномиальное время. Особую сложность представляют случаи, когда вычисление значения целевой функции в одной точке требует значительных вычислительных ресурсов (часы или дни расчетов в задачах проектирования сложных технических систем), и сама функция обладает сложным ландшафтом (мульти-modalностью, негладкостью, разрывами) в многомерном пространстве, где традиционные методы (например, градиентные и эволюционные) часто оказываются неэффективными. В таких условиях ограниченного бюджета вычислений, суррогатные модели играют ключевую роль, позволяя сократить количество дорогостоящих вычислений целевой функции. Традиционно для этой цели широко применяется кригинг [3] на основе гауссовских процессов, благодаря его способности давать не только точечные предсказания, но и оценку неопределенности значения в точке, что позволяет балансировать между исследованием и эксплуатацией пространства параметров. Однако вычислительная сложность этого метода, составляющая  $O(n^3)$  для одной итерации

(где  $n$  – количество уже вычисленных точек данных, а каждая итерация требует обращения матрицы ковариации размером  $n \times n$ ), наряду с чувствительностью к выбору ковариационной функции и ограниченной применимостью к негладким и разрывным функциям, делает его непрактичным для задач с  $n > 10^4$  точек, что стимулирует поиск более эффективных альтернатив.

В данной работе исследуется замена суррогатной модели на основе кригинга на модель ансамбля решающих деревьев [4] (Random Decision Forest, RDF) в его классической реализации в рамках алгоритма Super Efficient Global Optimization [5-9] (SEGO). В рамках работы исследуется способность RDF не только к сохранению ключевых преимуществ гауссовских процессов (например, оценку неопределенности через вычисление дисперсии), но и к улучшению эффективности поиска глобального оптимума многомерных функций.

## RDF КАК МЕТОД МАШИННОГО ОБУЧЕНИЯ

Классическая реализация RDF представляет собой ансамблевый метод, основанный на построении множества деревьев решений с последующим агрегированием их предсказаний. В основе метода лежат три ключевые техники. Во-первых, используется бутстрэп-агрегирование (bagging), при котором каждое дерево обучается на случайной подвыборке исходных данных, что позволяет снизить эффект переобучения. Во-вторых, применяется случайный выбор признаков – при каждом разбиении узла дерева рассматривается только подмножество доступных признаков, что уменьшает корреляцию между отдельными деревьями в ансамбле. В-третьих, окончательное предсказание модели формируется как среднее значение предсказаний всех деревьев.

Многочисленные исследования подтверждают эффективность RDF в задачах оптимизации: благодаря бутстрэп-агрегированию и случайному выбору признаков метод демонстрирует на 15-30% меньшую склонность к переобучению по сравнению с отдельными деревьями решений [10], сохраняя при этом устойчивость к шумам в данных (при 10% уровне шума точность снижается лишь на 2-3% против 7-9% у других методов [11]). Экспериментальные работы [12] подтверждают способность RDF эффективно обрабатывать разнородные данные, а его вычислительная сложность  $O(M \cdot n \cdot \log n)$  (где  $n$  – размер обучающей выборки,  $M$  – количество деревьев в ансамбле) [13] делает метод практичным для существенно многомерных задач.

\* Статья публикуется по рекомендации программного комитета Всероссийской научно-технической конференции Автоматизация, <https://rusautocon.org>

Однако метод имеет и существенные ограничения. Так как в работе рассматривается классическая реализация RDF, его предсказания носят кусочно-постоянный характер, что может быть недостатком при работе с гладкими функциями. Кроме того, в отличие от кригинга, где оценка неопределенности имеет четкий вероятностный смысл, RDF не предоставляет естественной и интерпретируемой оценки неопределенности предсказаний, что может ограничивать его применение в некоторых задачах оптимизации.

Такие особенности классического RDF требуют разработки специальных подходов к оценке неопределенности при использовании этого метода в качестве суррогатной модели в алгоритмах оптимизации.

Чтобы адаптировать работу суррогатной модели на основе RDF, мы предлагаем модифицировать методы расчета значений и неопределенности в запрашиваемых точках.

#### КОРРЕКЦИЯ ПРЕДСКАЗАНИЙ RDF МЕТОДОМ ИНТЕРПОЛЯЦИИ НЕВЯЗОК

Основным недостатком модели RDF является ее неспособность точно воспроизводить известные значения в обучающих точках. Для устранения этого ограничения предлагается двухэтапная процедура коррекции предсказаний с использованием модели интерполяции обратных расстояний [14] (IDW).

На первом этапе строится вспомогательная IDW-модель, которая аппроксимирует разницу между фактическими значениями и предсказаниями RDF в обучающей выборке. На втором этапе итоговое предсказание для произвольной точки вычисляется как сумма исходного прогноза RDF и корректирующего слагаемого, полученного с помощью IDW-интерполяции.

Предлагаемый подход обеспечивает точное прохождение модели через опорные точки обучающей выборки, сохраняя при этом все преимущества RDF. Использование IDW в качестве корректирующего алгоритма обусловлено его вычислительной эффективностью и способностью к точной интерполяции, что делает процедуру коррекции ресурсоэффективной и простой в реализации.

#### ОПИСАНИЕ РЕАЛИЗАЦИЙ ВЫЧИСЛЕНИЯ НЕОПРЕДЕЛЕННОСТИ

В процессе усовершенствования алгоритма SEGO с использованием в качестве суррогата модели RDF были исследованы четыре альтернативных подхода к расчету неопределенности в точках:

1. Ковариационный метод – основан на аппроксимации гауссовского процесса с упрощенной оценкой ковариации между точками;
2. Метод пересечений – определяет неопределенность через оценку константы Липшица и вычисление пересечений между соседними точками;
3. Метод ближайшего соседа – использует метрику расстояния до ближайшей известной точки как меру неопределенности;
4. Метод квадратичной аппроксимации – строит локальные квадратичные аппроксимации между соседними точками для оценки неопределенности.

Каждый из этих подходов предлагает различный компромисс между вычислительной сложностью и точностью оценки неопределенности, что позволяет адаптировать алгоритм SEGO к конкретным условиям оптимизационных задач.

#### КОВАРИАЦИОННЫЙ МЕТОД

Первый подход оценки неопределенности в точке основан на модифицированной версии гауссовского процесса. В отличие от классического метода, требующего трудоемкого обращения полной ковариационной матрицы, предложенная реализация использует упрощенную схему вычисления.

Суть метода заключается в следующем: вместо точного вычисления обратной ковариационной матрицы применяется её диагональная аппроксимация, где учитываются только обратные значения диагональных элементов матрицы ковариации  $1 / \text{diag}(\text{cov}(A))$ . Для повышения устойчивости оценок полученные значения дополнительно обрабатываются сигмоидальной функцией, что обеспечивает их плавное изменение в допустимом диапазоне и гарантированная нулевая неопределенность в обучающих точках.

На рис. 1 приведено сравнение неопределенности, построенной с помощью ковариационного метода, с дисперсией кригинга с ядром RationalQuadratic [15] на одномерной тестовой функции Стыбинского-Танга.

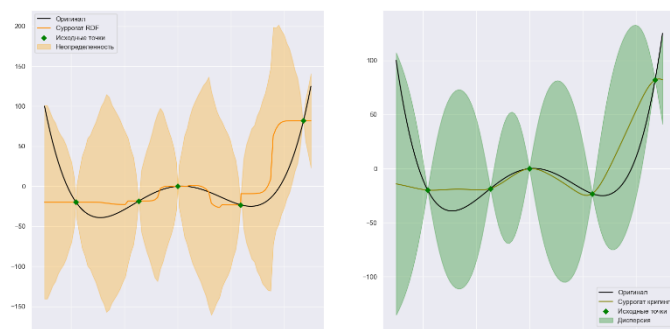


Рис. 1. Сравнение неопределенности RDF на основе ковариации (слева) и кригинга (справа)

Преимуществом данного подхода построения неопределенности является то, что за счет аппроксимации обращения матрицы ковариации удалось значительно сократить время вычисления, при этом сохранив паттерн поведения дисперсии кригинга.

#### МЕТОД ПЕРЕСЕЧЕНИЙ

В качестве другого подхода предлагается рассмотреть оценку неопределенности на основе вычисления обобщенной константы Липшица  $\tilde{L}$  [16, 17] и поиска пересечения поверхностей наклона между  $k$  ближайшими соседями исходных точек.

Между всеми исходными точками попарно проводятся линии и вычисляется пара точек с максимальным углом наклона  $\alpha$ . После этого вычисляется модуль коэффициента наклона  $k_{max} = \tan(\alpha)$ . Оценкой константы Липшица будет  $\tilde{L} = r \cdot k_{max}$  где  $r$  – коэффициент надежности для того, чтобы оценка константы Липшица не была занижена, а также увеличена зона охвата огибающей, чтобы гарантированно захватить минимум на функциях со сложным ландшафтом. В данной работе рассматривается коэффициент надежности  $r = 2$ .

Далее для каждой целевой точки вычисляются  $k$  ближайших соседей среди известных точек. Алгоритм поиска соседей старается подобрать кандидатов «равномерно» по

всем координатам вокруг целевой точки для лучшего исследования пространства. Если точка лежит вне многогранника соседей, добавляется симметричная точка на границе области.

Для каждой исходной точки составляются пары с их соседями. Из обеих точек каждой пары  $(p_1, p_2)$  строится поверхность наклона  $\tilde{L}$ . Для одномерного случая это будут линии, для многомерного – многомерные конусы. Для нахождения точки пересечения поверхностей наклона необходимо решить систему уравнений (1).

$$\begin{cases} y = y_1 - \tilde{L} \cdot \|x - p_1\| \\ y = y_2 - \tilde{L} \cdot \|x - p_2\| \end{cases} \quad (1)$$

Приравняв эти уравнения и упрощая выражение, получаем уравнение пересечения (2).

$$\|x - p_2\| - \|x - p_1\| = (y_2 - y_1) / \tilde{L} = h \quad (2)$$

Введем параметр  $t$  вдоль оси между точками (3).

$$x = p_1 + t \cdot (p_2 - p_1), \text{ где } t \in [0, 1] \quad (3)$$

Тогда, подставляя эту замену в уравнение, получаем (4).

$$\begin{aligned} \|x - p_1\| &= t \cdot d \\ \|x - p_2\| &= (1-t) \cdot d \\ \text{где } d &= \|p_2 - p_1\| \end{aligned} \quad (4)$$

Подставляем в исходное уравнение и решаем относительно  $t$  (5).

$$\begin{aligned} (1-t) \cdot d - t \cdot d &= h \\ d - 2 \cdot t \cdot d &= h \\ t &= (d - h) / (2d) \end{aligned} \quad (5)$$

Точка пересечения вычисляется по формуле (6).

$$\begin{aligned} x_{intersect} &= p_1 + t \cdot (p_2 - p_1) \\ y_{intersect} &= y_1 - \tilde{L} \cdot t \cdot d \end{aligned} \quad (6)$$

Данная система уравнений имеет решение при  $|h| \leq d$ .

С точки зрения геометрической интерпретации:

- при  $h = 0$  ( $y_1 = y_2$ ) точка пересечения – середина отрезка;
- при  $h > 0$  точка смещается ближе к  $p_2$ ;
- при  $h < 0$  точка смещается ближе к  $p_1$ ;
- при  $|h| = d$  пересечение совпадает с одной из точек;
- при  $|h| > d$  разница высот между  $y_1$  и  $y_2$  настолько велика, что под таким углом пересечения не будет.

Пересечение с минимальным значением  $y_{intersect}$  будем считать точкой с наибольшей неопределенностью и записывать значение исходной функции в ней.

На рис. 2 приведено сравнение неопределенности, построенной с помощью метода пересечений, с дисперсией кригинга с ядром RationalQuadratic на одномерной тестовой функции Стыбинского-Танга.

Основное достоинство подхода – устойчивость к застреванию в локальных минимумах и выполнение всех запланированных итераций оптимизации. Недостатками же являются сложность реализации, трудности с геометрической интерпретацией, а также сильная зависимость от начального распределения точек. Например, для периодической функции, если стартовые точки лежат в плоскости, ортогональной оси значений, алгоритм мгновенно завершит работу, возвращая исходные точки без исследования пространства. Однако данная проблема можно решать следующими способами: использованием квази-случайных последовательностей, комбинирование равномерного и

кластерного отбора начальных точек, предварительное скачивание пространства параметров с выбором точек в конкретных зонах.

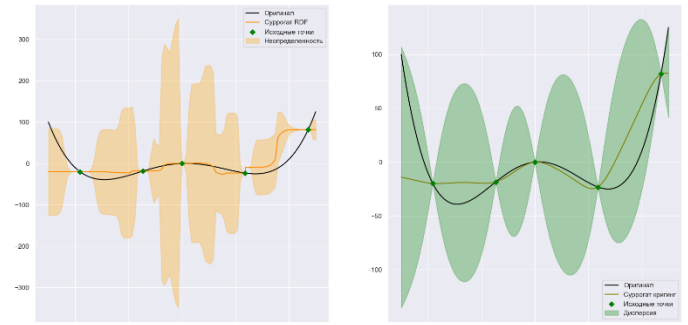


Рис. 2. Сравнение неопределенности RDF на основе пересечений (слева) и кригинга (справа)

### МЕТОД БЛИЖАЙШЕГО СОСЕДА

Самым простым с алгоритмической точки зрения является метод оценки неопределенности, использующий расстояние от целевой точки до ближайшего известного соседа. Метод вычисляет неопределенность как евклидово расстояние до ближайшей точки обучающих данных, найденное с помощью k-d-дерева для эффективного поиска. Полученные расстояния нормируются относительно диапазона значений целевой функции.

Этот подход обеспечивает простую, но эффективную оценку локальной неопределенности, где большие расстояния до ближайших соседей соответствуют областям с высокой неопределенностью, а малые – хорошо изученным регионам пространства параметров. Нормировка позволяет сопоставить оценку неопределенности с масштабом изменений целевой функции.

На рис. 3 приведено сравнение неопределенности, построенной с помощью метода ближайшего соседа, с дисперсией кригинга с ядром RationalQuadratic на одномерной тестовой функции Стыбинского-Танга.

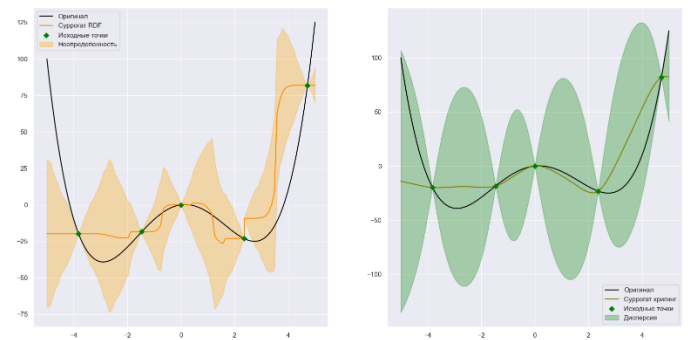


Рис. 3. Сравнение неопределенности RDF на основе метода ближайшего соседа (слева) и кригинга (справа)

Данный метод обладает ключевыми преимуществами в виде простоты реализации и наглядной геометрической интерпретацией, а также высокой скорости оценки неопределенности в точке, что обеспечивает более быстрый процесс оптимизации в сравнении с другими подходами. Однако он имеет существенный недостаток – склонность к преждевременной сходимости к локальному минимуму, что приводит к досрочному завершению оптимизационного процесса до исчерпания выделенного вычислительного бюджета.

МЕТОД КВАДРАТИЧНОЙ АППРОКСИМАЦИИ

В качестве еще одного подхода вычисления неопределенности опробован метод аппроксимации квадратичной функции, или же построения парабол между известными точками на основе расстояния между ними.

Для каждой целевой точки вычисляются  $k$  ближайших соседей среди известных точек. Алгоритм поиска соседей старается подобрать кандидатов «равномерно» по всем координатам вокруг целевой точки для повышения качества исследования пространства и итоговой аппроксимации. В случае, если точка находится вне выпуклой оболочки соседей, наиболее удаленная точка заменяется проекцией на ближайшую границу области определения.

Далее вычисляются центр масс  $x_{mid}$  соседей  $x_{neib}$  (7), среднее евклидово расстояния  $d_{aver}$  от целевой точки  $x^*$  до них (8) и покоординатные смещения целевой точки  $d_x$  относительно центра масс (9).

$$x_{mid} = \frac{\sum_{i=1}^k x_{neib_i}}{k} \tag{7}$$

$$d_{aver} = \frac{\sum_{i=1}^k \sqrt{\sum_{j=1}^n (x_{neib_{ij}} - x_j^*)^2}}{k} \tag{8}$$

$$d_{x_j} = x_{mid_j} - x_j^*, j = \overline{1, n} \tag{9}$$

На основе этих параметров строится квадратичная аппроксимация  $var$  (10), масштабированная относительно диапазона значений целевой функции  $y_{diff}$ , которая и будет оценкой неопределенности в рамках данного подхода.

$$var = 0,5 \cdot y_{diff} \cdot \left( - \left( \sum_{j=1}^n x_j^{*2} \right) + d_{aver}^2 \right) \tag{10}$$

На рис. 4 приведено сравнение неопределенности, построенной с помощью метода квадратичной интерполяции, с дисперсией кригинга с ядром RationalQuadratic на одномерной тестовой функции Стыбинского-Танга.

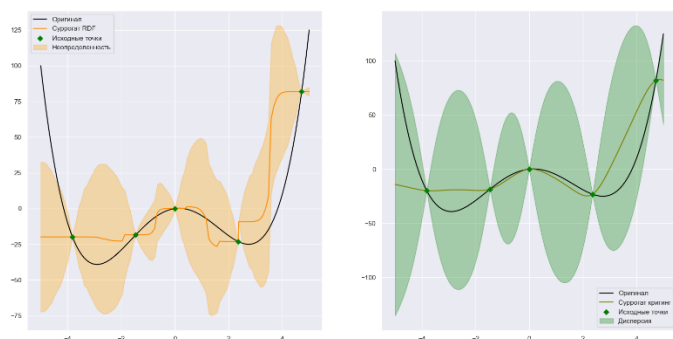


Рис. 4. Сравнение неопределенности RDF на основе квадратичной аппроксимации (слева) и кригинга (справа)

Данный метод демонстрирует значительные преимущества в виде способности учитывать локальную кривизну целевой функции за счет квадратичной аппроксимации, что обеспечивает более точную оценку неопределенности по сравнению с линейными методами. Равномерный подбор соседей по всем координатам и автоматическая коррекция для точек вне выпуклой оболочки повышают устойчи-

вость алгоритма. Однако подход имеет существенные ограничения: вычислительная сложность возрастает с увеличением размерности задачи из-за необходимости анализа многомерных расстояний и координатных смещений, а качество аппроксимации сильно зависит от начального распределения точек. Кроме того, метод может давать завышенные оценки неопределенности в областях с резкими изменениями градиента функции, где квадратичная аппроксимация становится недостаточно точной.

ЭВРИСТИЧЕСКИЙ МЕТОД ВЫБОРА ТОЧЕК В АЛГОРИТМЕ SEGO

Для использования представленных подходов в оптимизации SEGO, для начала необходимо построить суррогат неопределенности.

Суррогатная модель неопределенности строится на основе точек, сгенерированных с использованием квази-случайной последовательности на основе «золотого» сечения [18]. Значения неопределенности в этих точках вычисляются одним из представленных методов, после чего методом обратного взвешенного расстояния (IDW) формируется суррогатная модель неопределенности.

В классическом подходе SEGO с кригингом новые точки выбираются путем минимизации дисперсии, однако для моделей на основе решающих деревьев это затруднительно из-за негладкой природы суррогатов неопределенностей. В методе пересечений точка SEGO выбирается в области самого нижнего пересечения, тогда как для других подходов применяется эвристическая стратегия.

Эвристический выбор следующей точки рассматривает две ключевые области: окрестность текущего минимума RDF-суррогата и зону максимальной неопределенности. Первая потенциальная точка отбирается как середина отрезка между точкой текущего минимума и ее ближайшим соседом, либо как ближайшая граничная точка, если она расположена ближе. Вторая точка выбирается из обучающего набора как точка с максимальным значением неопределенности. Окончательный выбор новой точки для оценки осуществляется путем сравнения разности значений суррогатной модели и неопределенности в этих двух точках, предпочтение отдается точке с минимальной такой разностью.

ТЕСТИРОВАНИЕ ПОДХОДОВ НА КЛАССАХ ФУНКЦИЙ

Для тестирования представленного функционала предлагается сравнить их время работы и точность оптимизации SEGO с использованием кригинга в качестве суррогата и с использованием RDF в качестве суррогата с представленными выше подходами вычисления неопределенности и выбором точки на очередном этапе оптимизации. Качество работы алгоритмов будет сравниваться на пяти классах функций: унимодальные, овражные, периодические, многоэкстремальные и зашумленные. Каждый класс содержит по 10 функций. Все функции и их минимумы представлены в источниках [19-28].

В качестве входных данных эксперимента выступают:

- исходные тестовые функции размерностью с размерностью  $n$  от 2 до 10 и областью определения  $x_i \in [-5, 5], i = \overline{1, n}$ ;
- количество соседей  $k$  для вычисления неопределенности вычисляется по формуле (11);
- количество точек, необходимых для построения суррогата неопределенности, взято  $100 \cdot n$ ;
- дебет по количеству заказов точек установлен  $100 \cdot n$ .

Установленное ограничение на параметр  $k$  обусловлено необходимостью предотвращения экспоненциального роста вычислительной сложности при увеличении размерности пространства ( $n \geq 5$ ). Эмпирические исследования на тестовых функциях различной размерности демонстрируют, что превышение указанного порогового значения  $k$  не приводит к статистически значимому улучшению точности оптимизации, однако вызывает существенное увеличение временных затрат на построение суррогатной модели и выполнение оптимизационного процесса, таким образом, установленное ограничение на количество соседей обеспечивает оптимальный компромисс между точностью и временем работы алгоритма.

$$k = \begin{cases} 2^n, & \text{если } n < 5 \\ 16, & \text{иначе} \end{cases} \quad (11)$$

В качестве выходных данных эксперимента выступает абсолютная ошибка оптимизации по каждой функции относительно ее минимума и время оптимизации каждой функции.

Ниже на рис. 5-9 представлены графики сравнения оптимизации SEGO на классах тестовых по результирующему среднему проценту абсолютной ошибки найденного минимума относительно диапазона данных функции и по среднему времени оптимизации по классу функций. Цветам на графиках соответствуют следующие подходы: красный – кригинг, желтый – ковариационный метод, фиолетовый – метод пересечений, зеленый – метод ближайшего соседа, голубой – метод квадратичной аппроксимации. Примечание: на всех изображениях на графиках точности (левые изображения) линии, представляющие методы ковариации и квадратичной аппроксимации, частично или полностью перекрываются.

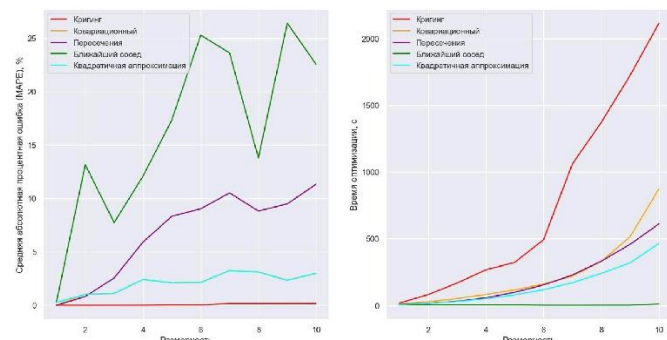


Рис. 5. Результаты для унимодальных функций

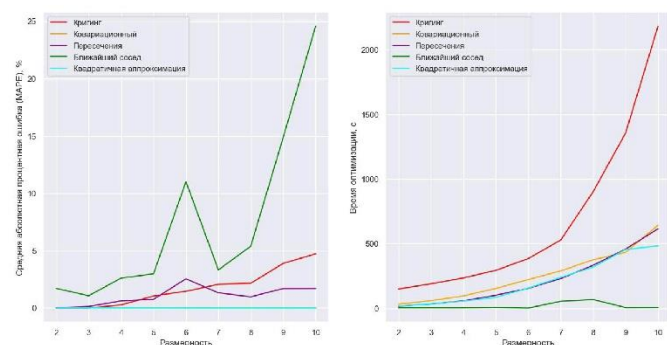


Рис. 6. Результаты для обранных функций

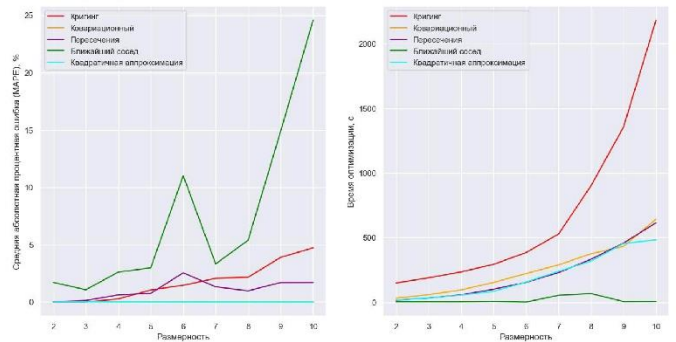


Рис. 7. Результаты для периодических функций

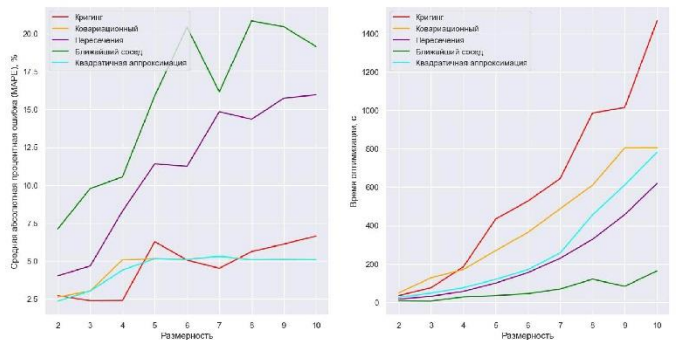


Рис. 8. Результаты для многоэкстремальных функций

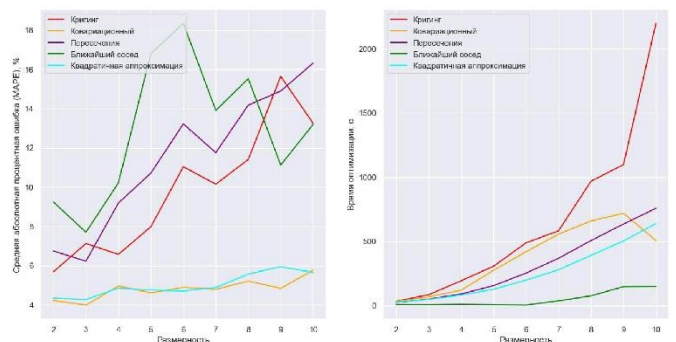


Рис. 9. Результаты для зашумленных функций

Проведенный анализ динамики оптимизации на различных классах тестовых функций демонстрирует, что реализация алгоритма SEGO с использованием кригинга в качестве суррогатной модели характеризуется наименьшей вычислительной эффективностью и значительной вариабельностью результатов, зависимой от размерности пространства параметров и специфики целевой функции. Напротив, метод ближайшего соседа показывает максимальное быстрое действие, что при малой точности делает его предпочтительным для задач начального приближения. Наиболее сбалансированные показатели эффективности по критериям временных затрат и качества оптимизации демонстрируют методы упрощенной ковариации и квадратичной аппроксимации, которые на большинстве тестовых случаев превосходят кригинг по обоим параметрам.

ЗАКЛЮЧЕНИЕ

Проведенное исследование демонстрирует значительное преимущество использования Random Decision Forest (RDF) в качестве суррогатной модели в алгоритме SEGO по сравнению с традиционным подходом на основе кригинга. Реализованные методы оценки неопределенности для RDF позволили достичь не только более высокой скорости оптимизации, но и улучшения качества конечных решений. Экспериментальные результаты на тестовых функциях различных классов показали, что эффективность того или иного метода вычисления дисперсии существенно зависит от характера и размерности оптимизируемой функции.

Перспективные направления дальнейших исследований включают:

– модификация представленных методов с целью сокращения вычислительных затрат;

– создание адаптивных механизмов автоматического выбора оптимального метода оценки неопределенности в зависимости от характеристик целевой функции (ее вида и размерности).

Полученные результаты открывают новые возможности для совершенствования методов глобальной оптимизации, особенно в задачах, требующих баланса между вычислительной эффективностью и точностью получаемых решений.

ЛИТЕРАТУРА

1. Pardalos P.M. Quadratic Programming with One Negative Eigenvalue is NP-Hard / P.M. Pardalos, S.A. Vavasis // Journal of Global Optimization. – 1991. – Vol. 1, no 1. – P. 15-22. DOI: 10.1007/BF00120662.
2. Nesterov Y. Global Quadratic Optimization via Conic Relaxation // Mathematical Programming. – 2000. – Vol. 89, no 3. – P. 469–486. DOI: 10.1007/s101070000240.
3. Williams C.K.I. Gaussian Processes for Machine Learning / C.K.I. Williams, C.E. Rasmussen. – MA: MIT Press, 2006. – 248 p.
4. Breiman L. Random Forests // Machine Learning. – 2001. – Vol. 45, no.1. – P. 5--32.
5. Jones D.R. Efficient global optimization of expensive black-box functions / D.R. Jones, M. Schonlau, W.J. Welch // Journal of Global optimization. – 1998. – Vol. 13, no. 4. – P. 455-492.
6. Бурнаев Е. Сравнительный анализ процедур оптимизации на основе гауссовских процессов / Е. Бурнаев, М. Панов, Д. Кононенко, И. Коноваленко // Сборник: Информационные технологии и системы. – 2012. – С. 167-172.
7. Sasena M.J. Exploration of Metamodeling Sampling Criteria for Constrained Global Optimization / M.J. Sasena, P. Papalambros, P. Goovaerts // Engineering Optimization. – 2002. – Vol. 34, no, 3. – P. 263-278.
8. Watson G. Infill sampling criteria to locate extremes / G. Watson, R.J. Barnes // Mathematical Geology. – 1995. – Vol. 27, no. 5. – P. 589-608.
9. Sasena M. Flexibility and efficiency enhancements for constrained global design optimization with Kriging approximations: Ph.D. thesis. University of Michigan, 2002.
10. Fernández-Delgado M. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? / M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim // Journal of Machine Learning Research. – 2014. – Vol. 15. – P. 3133-3181.

11. Prasad A.M., Iverson L.R., Liaw A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction // Ecosystems. 2006. Vol. 9, № 2. P. 181-199. DOI: 10.1007/s10021-005-0054-1.

12. Goldstein B.A. Random Forests for Genetic Association Studies / B.A. Goldstein, E.C. Polley, F.B.S. Briggs // Statistical Applications in Genetics and Molecular Biology. – 2011. – Vol. 10, no. 1. – Article 32. DOI: 10.2202/1544-6115.1691.

13. Louppe G. Understanding Random Forests: From Theory to Practice. – URL: <https://arxiv.org/abs/1407.7502> (дата обращения: 14.05.2025).

14. Shepard D. A two-dimensional interpolation function for irregularly-spaced data // Proceedings of the 1968 23rd ACM national conference. – 1968. – P. 517-524. DOI: 10.1145/800186.810616.

15. Rasmussen C.E., Williams C.K.I. Gaussian Processes for Machine Learning. – URL: <http://www.gaussianprocess.org/gpml/chapters/> (дата обращения: 14.05.2025).

16. Hansen P. Global optimization of Lipschitz functions / P. Hansen, B. Jaumard // Journal of Optimization Theory and Applications. – 1992. – Vol. 75, no, 1. – P. 193-211.

17. Strongin R.G. Global Optimization with Non-Convex Constraints / R.G Strongin, Ya.D. Sergeyev. – Springer, 2000.

18. Swinbank R. Fibonacci grids: A novel approach to global modelling / R. Swinbank, James R. Purser // Quarterly Journal of the Royal Meteorological Society. – 2006. – Vol. 132, no, 619. – P. 1769-1793.

19. Yang X.-S. Nature-Inspired Optimization Algorithms. – London: Elsevier, 2014. – 300 p.

20. Arnold D.V. Noisy Optimization: Theory and Practice. – Berlin: Springer, 2002. – 210 p.

21. Jamil M., Yang X.-S. A literature survey of benchmark functions for global optimization problems. – URL: <https://arxiv.org/abs/1308.4008> (дата обращения: 09.07.2025).

22. Jin Y. Evolutionary Optimization in Noisy Environments / Y. Jin, J. Branke // Journal of Global Optimization. – 2005. – Vol. 32, no. 1. – P. 1-25. DOI: 10.1007/s10898-004-3174-0.

23. Suganthan P.N. Benchmarking Optimization Algorithms / P.N. Suganthan et al. – Singapore: Nanyang Technological University, 2005. – 45 p.

24. Virtual Library of Simulation Experiments. – URL: <https://www.sfu.ca/~ssurjano/optimization.html> (дата обращения: 09.07.2025).

25. Global Optimization Benchmarks. – URL: [http://www.optima.amp.i.kyoto-u.ac.jp/member/student/hedar/Hedar\\_files/TestGO.htm](http://www.optima.amp.i.kyoto-u.ac.jp/member/student/hedar/Hedar_files/TestGO.htm) (дата обращения: 09.07.2025).

26. Pasillas R. Test-Functions-for-Optimization. – URL: <https://github.com/raulppasillas/test-functions-for-optimization> (дата обращения: 09.07.2025).

27. Rend S. Optimization-Test-Functions. – URL: <https://github.com/sorend/functions> (дата обращения: 09.07.2025).

28. CEC Benchmark Functions. – URL: <https://www.ntu.edu.sg/home/EPNSugan/> (дата обращения: 09.07.2025).

# Approaches to Estimating Uncertainty when Using a Random Decision Forest in Global Optimization Algorithm

Buyanov A.D.

National Research Nizhny Novgorod State University named after N.I. Lobachevsky  
Nizhny Novgorod, Russian Federation

E-mail: [ar.buyanow@yandex.ru](mailto:ar.buyanow@yandex.ru)

**Abstract.** This paper explores the use of Random Decision Forest as a surrogate model for global optimization, providing an alternative to Gaussian process-based kriging. Although RDF has advantages such as linear training complexity, robustness to noise, and support for categorical features, its piecewise-constant predictions and lack of inherent uncertainty quantification present challenges. To overcome these limitations, we present four approaches for estimating uncertainty in RDF: covariance-based, intersection-based, nearest-neighbor and quadratic approximation, each of which balances accuracy and computational efficiency. Addition-

ally, we introduce a heuristic for the SEGO algorithm that dynamically selects new evaluation points by considering both surrogate minima and high-uncertainty regions. Empirical tests on diverse function classes show that our approach surpasses kriging. This is in terms of optimization speed and solution quality. It is particularly effective in high-dimensional settings. The function classes include unimodal, ravine-like, periodic, multimodal and noisy.

**Keywords:** global optimization, random decision forest, surrogate modeling, uncertainty estimation, kriging, SEGO.

## Библиографическое описание статьи

Буянов А.Д. Методы оценки неопределенности при использовании леса решающих деревьев в алгоритме глобальной оптимизации // *Машиностроение: сетевой электронный научный журнал*. – 2026. – Т.13, №1. – С. 21-27. DOI: 10.24892/RIJIE/20260104

## Reference to article

Buyanov A.D. Approaches to estimating uncertainty when using a random decision forest in global optimization algorithm, *Russian Internet Journal of Industrial Engineering*, 2026, vol.13, no.1, pp. 21-27. DOI: 10.24892/RIJIE/20260104