ВОССТАНОВЛЕНИЕ ДИСПЕРСИИ ПРОЗРАЧНОЙ СРЕДЫ ПО ЭКСПЕРИМЕНТАЛЬНЫМ ДАННЫМ

Г. И. Рудой

Аннотация

Для восстановления нелинейной зависимости показателя преломления среды от длины волны рассматривается набор индуктивно порожденных моделей с целью выбора оптимальной. Применяется оригинальный алгоритм индуктивного порождения допустимых существенно нелинейных моделей. Предлагается метод типа Монте-Карло оценки стабильности полученного решения. Приводятся результаты вычислительного эксперимента на реальных данных.

Ключевые слова: *символьная регрессия, нелинейные модели, индуктивное порождение, стабильность решений, дисперсия прозрачной среды.*

1 Введение

В ряде прикладных областей возникает задача восстановления зависимости показателя преломления прозрачной для света среды от длины волны. При этом возможно его измерить с достаточно высокой точностью в широком диапазоне длин волн.

Физические соображения определяют некоторый общий вид зависимости [1, 2], которой должны описываться экспериментальные данные, и для эксперта представляют интерес конкретные параметры этой зависимости. Поэтому требуется возможность экспертной интерпретации полученной функции регрессии.

Кроме того, каждое измерение имеет свою допустимую погрешность, определяемую экспериментатором, поэтому необходимо оценить стабильность полученных результатов. Иными словами, требуется понять, насколько сильно меняется результирующая суперпозиция при небольшом изменении экспериментальных данных.

В настоящей работе исследуется возможность применения алгоритма, предложенного в [3] для восстановления искомой зависимости, и результаты его работы сравниваются с результатами применения SVM-регрессии. Кроме того, исследуется влияние штрафа за сложность на качество и структурную сложность получающихся суперпозиций, а также возможность оценки стабильности решения и вклада различных членов результирующей функции регрессии.

Во второй части данной работы формально поставлена задача восстановления дисперсии жидкости. В третьей части вкратце описывается алгоритм [3], используемый для порождения аналитической функции-суперпозиции, аппроксимирующей данные. В четвертой части описывается метод, позволяющий учитывать и анализировать вклад различных членов порожденной суперпозици и их зависимость от случайных шумов в данных. В пятой части приводятся результаты вычислительного эксперимента на реальных данных.

2 Постановка задачи

Дан набор из l результатов измерений коэффициента преломления для некоторого вещества s. То есть, для каждого измерения i даны:

- λ_i — длина волны.
- n_i — измеренный коэффициент преломления.

Требуется найти функцию $\hat{n} = \hat{n}(\lambda)$, минимизирующую функционал потерь в предположении о нормальности случайной ошибки эксперимента:

$$Q(n) = \sum_{i=1}^{l} \left|(n(\lambda_i) - n_i)^2\right| \to \min\,. \tag{1}$$

Далее, требуется оценить стабильность полученных результатов относительно небольшого изменения входных данных, изучив зависимость математического ожидания и дисперсии отклонения коэффициентов в полученной функции n от случайной нормально распределенной добавки к экспериментальным данным.

3 Алгоритм индуктивного порождения суперпозиций

Вкратце опишем предложенный в [3] алгоритм.

Пусть задано некоторое множество $G = \{g_1, \ldots, g_l\}$ элементарных порождающих функций. В начале работы набор суперпозиций $\mathcal{F} = \{f\}$ инициализируется случайными суперпозициями функций $g \in G$. Содержащиеся в $\mathcal{F}$ суперпозиции содержат как свободные переменные, соответствующие компонентам вектора-описания объектов из генеральной совокупности, так и численные константы, которые оптимизируются на каждом шаге алгоритмом Левенберга-Марквардта согласно введенному функционалу потерь (1). Также на каждой итерации над суперпозициями выполняется набор модифицирующих операций с целью улучшения качества суперпозиций.

Качество Q_f суперпозиции f вычисляется по совокупности точности приближения экспериментальных данных и структурной сложности суперпозиции по следующей формуле:

$$Q_f = \frac{1}{1 + Q(f)} \left(\alpha\hat{Q} + \frac{1 - \alpha\hat{Q}}{1 + \exp(C_f - \tau)} \right), \tag{2}$$

где:

- Q — среднеквадратичная ошибка на данной выборке;
- C_f — сложность суперпозиции, соответствующая количеству элементарных функций, свободных переменных и констант;
- $\hat{Q}$ — минимальная приспособленность суперпозиции из критерия останова;

- α — некоторый коэффициент, $0 \ll \alpha < 1$, характеризующий важность штрафа за сложность;
- τ — коэффициент, характеризующий желаемую сложность модели.

Второй множитель в (2) выполняет роль штрафа за слишком большую сложность суперпозиции, что подавляет эффект переобучения и позволяет получать более простые суперпозиции ценой большей ошибки на обучающих данных при большей обобщающей способности.

Отметим, что параметры α и τ нельзя подбирать методом скользящего контроля или любыми другими внутренними относительно метода обучения методами.

4 Метод исследования стабильности решения

Для оценки стабильности решения предлагается следующий подход. Фиксируется структурный вид суперпозиции f, являющейся ответом описанного выше алгоритма, и исследуется зависимость стандартного отклонения ее коэффициентов как функция стандартного отклонения нормально распределенного случайного шума в исходных данных.

Иными словами, выбираются значения σ_λ и σ_n. Затем для этих значений генерируется шум $\xi^\lambda \in \mathcal{N}(0, \sigma_\lambda$ и $\xi^n \in \mathcal{N}(0, \sigma_n)$, который прибавляется к данным выборки. То есть, исходная выборка (λ_i, n_i) заменяется на

$$(\lambda_i + \xi_i^\lambda, n_i + \xi_i^n) \mid \xi_i^n \in \mathcal{N}(0, \sigma_n), \xi_i^\lambda \in \mathcal{N}(0, \sigma_\lambda).$$

Для полученной зашумленной выборки вычисляются значения коэффициентов фиксированной суперпозиции путем оптимизации методом Левенберга-Марквардта.

Данная процедура для фиксированной пары σ_λ и σ_n повторяется до достижения некоторого критерия останова (например, по количеству итераций), после которого для каждого коэффициента суперпозиции f считается эмпирическое матожидание и эмпирическое стандартное отклонение, основываясь на полученных на различных итерациях значениях.

Повторяя данную процедуру для различных σ_λ и σ_n, можно оценить зависимость стандартного отклонения коэффициентов суперпозиции от стандартного отклонения шума.

Из физических соображений ясно, что линейная зависимость означает стабильное решение ввиду отсутствия явных особых точек, тогда как отклонения от линейности символизируют ту или иную ошибку в структурной формуле и могут являться свидетельством переобучения.

5 Вычислительный эксперимент

В вычислительном эксперименте используются данные, полученные в ходе физического эксперимента по изучению возможности определения состава смеси прозрачных

материалов по суммарной дисперсионной зависимости, если известна экспериментальная зависимость дисперсии для каждого конкретного материала. Рассматривается один материал, для которого имеется 18 экспериментальных точек, соответствующих коэффициенту преломления при разных значениях длины волны. Характерное значение $n \approx 1.33$, $\lambda \in (400 - 950)$ нм.

В силу замечания о невозможности подбора структурных параметров α и τ, разделение на обучающую и контрольную выборку не производилось, однако переобучения удается избежать и без такого разделения, опираясь целиком на штраф за сложность.

Из физических соображений следует [2], что зависимость коэффициента преломления от длины волны должна выражаться суммой отрицательных четных степеней дисперсии, поэтому множество элементарных функций состоит из стандартных операций сложения и умножения:

$$g_1(x_1, x_2) = x_1 + x_2,$$

$$g_2(x_1, x_2) = x_1 x_2,$$

а также из функции

$$g_3(\lambda, p) = \frac{1}{\lambda^{2p}}.$$

В ходе вычислительного эксперимента константы, меньшие 10^{-7}, принудительно заменялись на 0.

В результате применения описанного выше алгоритма со значениями $\hat{Q} = 0.95$, $\alpha = 0.05$, $\tau = 10$ получена следующая формула (константы округлены до третьей значащей цифры для удобства чтения):

$$n(\lambda) = 1.35 + \frac{5.82}{\lambda^2} + \frac{3.58 \cdot 10^{-5}}{\lambda^4}, \tag{3}$$

со сложностью 13, среднеквадратичной ошибкой $2.2 \cdot 10^{-5}$ и значением $Q_f \approx 0.0475$.

Отметим, что обычно в приложениях учитывают только квадратичный член, а более высокими степенями пренебрегают. Коэффициент при $\frac{1}{\lambda^4}$ указывает на полное согласие полученных результатов с принятой практикой.

5.1 Влияние штрафа за сложность

Исследуем, как влияет добавление нечетных степеней на получающийся результат, заменив функцию g_3 в порождающем наборе на

$$g_3(\lambda, p) = \frac{1}{\lambda^p}.$$

Следует отметить, что при тех же $\alpha = 0.05$ и $\tau = 10$ результирующей функцией остается (3).

Увеличим τ до 25. Получим следующую формулу:

$$n(\lambda) = 1.34 + \frac{11.6}{\lambda} + \frac{17.37}{\lambda^2} + \frac{0.0866}{\lambda^3} + \frac{2.95 \cdot 10^{-4}}{\lambda^4} + \frac{8.54 \cdot 10^{-7}}{\lambda^5},$$

сложность которой составляет 31, и для которой среднеквадратичная ошибка на выборке составляет $\approx 3.9 \cdot 10^{-7}$, а значение $Q_f \approx 0.0498$.

Иными словами, при большей желаемой сложности, регулируемой параметром τ, выигрывает более сложная (а в данном случае и физически некорректная) модель, которая лучше описывает экспериментальные данные.

То есть, как и следовало ожидать, чрезмерное увеличение τ ведет к переобучению.

5.2 SVM

В качестве базового алгоритма используется SVM-регрессия с RBF-ядром [4]. Параметр γ ядра подбирался по методу скользящего контроля, наилучшим результатом является комбинация из 15 опорных векторов с $\gamma \approx 2 \cdot 10^{-6}$, при этом среднеквадратичная ошибка при кросс-валидации с тестовой выборкой, содержащей по 2 объекта, составляет $8.96 \cdot 10^{-8}$. Однако, проинтерпретировать полученную решающую функцию не представляется возможным.

5.3 Исследование стабильности решения

Для оценки стабильности решения фиксировалась формула (3) и исследовалась зависимость стандартного отклонения ее коэффициентов от стандартного отклонения нормально распределенного случайного шума в исходных данных описанным выше методом. Критерием останова в нем являлось достижение 10000 итераций для каждой пары $(\sigma_\lambda, \sigma_n)$.

Численные значения эмпирического матожидания и дисперсии для первого, второго и третьего коэффициентов формулы (3) для некоторых значений $(\sigma_\lambda, \sigma_n)$ приведены в таблицах 1, 2 и 3 соответственно. По строкам указаны значения для σ_n, по столбцам — для σ_λ.

	0	10^{-5}	10^{-4}	0.001	0.01
0	$(1.36; 5.97 \cdot 10^{-7})$	$(1.36; 2.34 \cdot 10^{-6})$	$(1.36; 2.39 \cdot 10^{-5})$	$(1.36; 2.34 \cdot 10^{-4})$	$(1.36; 0.00238)$
0.01	$(1.36; 0)$	$(1.36; 2.37 \cdot 10^{-6})$	$(1.36; 2.38 \cdot 10^{-5})$	$(1.36; 2.37 \cdot 10^{-4})$	$(1.36; 0.00231)$
0.1	$(1.36; 0)$	$(1.36; 2.37 \cdot 10^{-6})$	$(1.36; 2.35 \cdot 10^{-5})$	$(1.36; 2.34 \cdot 10^{-4})$	$(1.36; 0.00237)$
1	$(1.36; 1.03 \cdot 10^{-7})$	$(1.36; 2.36 \cdot 10^{-6})$	$(1.36; 2.34 \cdot 10^{-5})$	$(1.36; 2.39 \cdot 10^{-4})$	$(1.36; 0.00235)$
10	$(1.36; 2.74 \cdot 10^{-7})$	$(1.36; 2.37 \cdot 10^{-6})$	$(1.36; 2.35 \cdot 10^{-5})$	$(1.36; 2.35 \cdot 10^{-4})$	$(1.36; 0.00236)$

Таблица 1: Значения матожидания и стандартного отклонения для первого коэффициента

На графиках 1, 2 и 3 представлены поверхности уровня дисперсии для первого, второго и третьего коэффициентов соответственно.

	0	10^{-5}	10^{-4}	0.001	0.01
0	$(5.82; 0)$	$(5.82; 9.12 \cdot 10^{-4})$	$(5.82; 0.00913)$	$(5.82; 0.0871)$	$(5.55; 1.87)$
0.01	$(5.82; 5.41 \cdot 10^{-5})$	$(5.82; 9.17 \cdot 10^{-4})$	$(5.82; 0.00904)$	$(5.82; 0.0867)$	$(5.56; 1.81)$
0.1	$(5.82; 5.37 \cdot 10^{-4})$	$(5.82; 0.00105)$	$(5.82; 0.00907)$	$(5.82; 0.0873)$	$(5.56; 1.81)$
1	$(5.82; 0.00538)$	$(5.82; 0.00549)$	$(5.82; 0.0106)$	$(5.82; 0.0866)$	$(5.56; 1.82)$
10	$(5.82; 0.0511)$	$(5.82; 0.0516)$	$(5.82; 0.0525)$	$(5.82; 0.103)$	$(5.59; 1.75)$

Таблица 2: Значения матожидания и стандартного отклонения для второго коэффициента

	0	10^{-5}	10^{-4}	0.001	0.01
0	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 6.21 \cdot 10^{-7})$	$(0; 0.0657)$
0.01	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 6.23 \cdot 10^{-7})$	$(5.7 \cdot 10^{-4}; 0.0536)$
0.1	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 6.21 \cdot 10^{-7})$	$(0; 0.106)$
1	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 0)$	$(3.58 \cdot 10^{-5}; 1.16 \cdot 10^{-7})$	$(3.58 \cdot 10^{-5}; 6.32 \cdot 10^{-7})$	$(0; 0.185)$
10	$(3.59 \cdot 10^{-5}; 9.71 \cdot 10^{-7})$	$(3.59 \cdot 10^{-5}; 9.83 \cdot 10^{-7})$	$(3.59 \cdot 10^{-5}; 9.89 \cdot 10^{-7})$	$(3.59 \cdot 10^{-5}; 1.18 \cdot 10^{-6})$	$(2.4 \cdot 10^{-4}; 0.0882)$

Таблица 3: Значения матожидания и стандартного отклонения для третьего коэффициента

Из графиков видно, что от шума, накладываемого на значения длины волны, дисперсия значений первого и второго коэффициентов практически не зависит. В то же время дисперсия значений первого коэффициента зависит от дисперсии шума коэффициента преломления практически линейно, тогда как для второго коэффициента после некоторого характерного значения зависимость теряется. Кроме того, дисперсия третьего коэффициента формулы не имеет явной гладкой зависимости от дисперсии шума длины волны или коэффициента преломления, что еще раз подтверждает его достаточно случайную природу.

Физическая интерпретация этих результатов — при построении прибора для измерения дисперсии сред значительное внимание следует уделять точности измерения коэффициента преломления, тогда как измерения длины волны могут быть неточны вплоть до нескольких процентов. Кроме того, предложенный метод прямо указывает, на каких интервалах шума какой будет выигрыш в точности предсказания от небольшого увеличения точности.

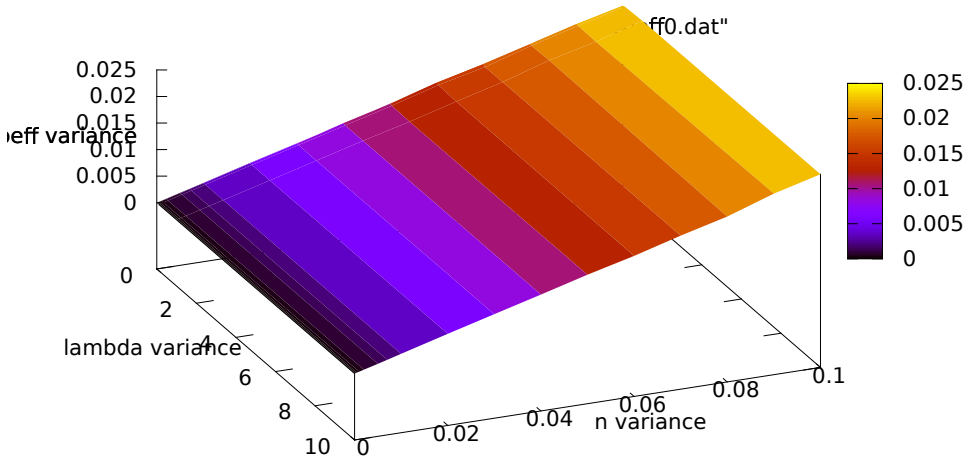

Рис. 1: Поверхность уровней дисперсии для первого коэффициента

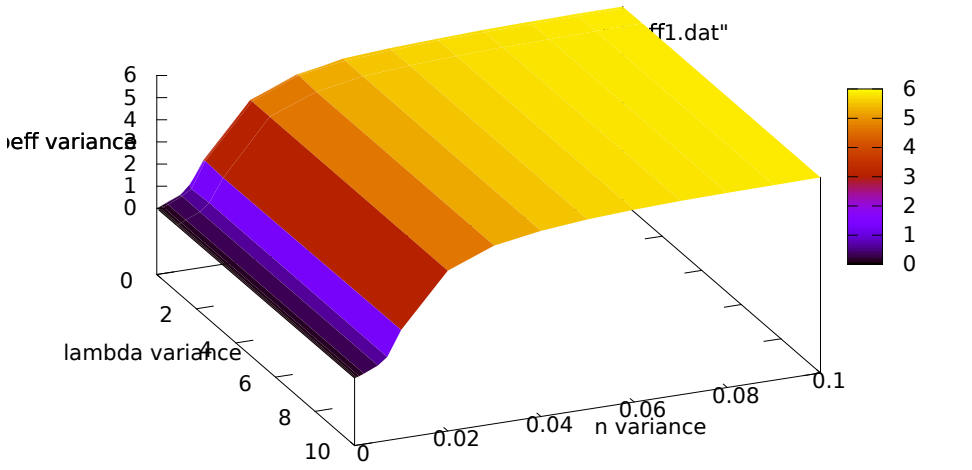

Рис. 2: Поверхность уровней дисперсии для второго коэффициента

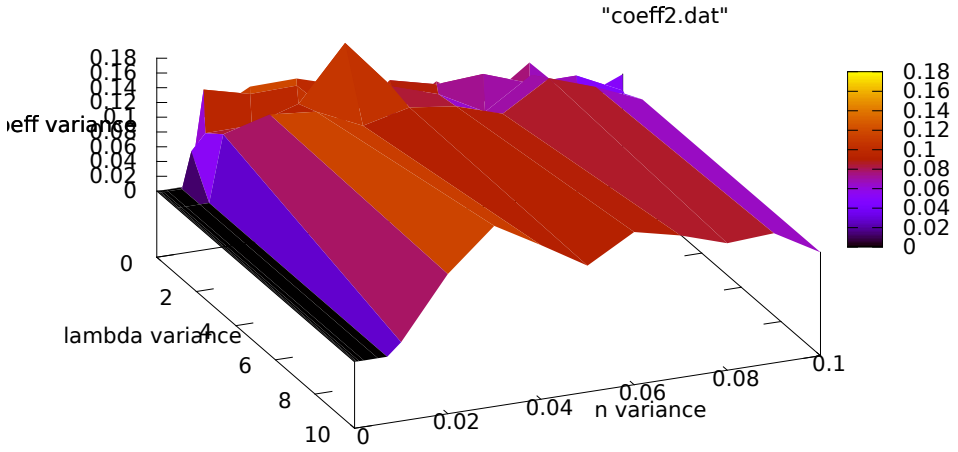

Рис. 3: Поверхность уровней дисперсии для третьего коэффициента

6 Заключение

Предложенный в [3] алгоритм позволяет получить интерпретируемую аналитическую формулу, описывающую зависимость коэффициента преломления среды от длины волны. Введенный штраф за сложность позволяет избежать переобучения без прибегания к методам вроде скользящего контроля, и таким образом отпадает необходимость в контрольной выборке.

Хотя другие алгоритмы, такие как SVM-регрессия, могут демонстрировать более высокое качество приближения данных, их результаты неинтерпретируемы и не защищены от переобучения «по построению», поэтому требуют разделения выборки на обучающую и контрольную. Кроме того, их структурные параметры так же требуют оценки по методам вроде кросс-валидации.

Предложенный в настоящей работе метод оценки стабильности решения позволяет исследовать вклад различных членов результирующей суперпозиции в решение, и зависимость изменения этих членов от случайных шумов во входных данных. В частности, в прикладных областях данный метод позволяет выявить, какие именно элементы признакового описания объектов в генеральной совокупности наиболее чувствительны к шуму.

Список литературы

[1] Сивухин, Д. В.: *Оптика.* ФИЗМАТЛИТ, 3 редакция, 2005.

[2] Н., Серова В.: *Полимерные оптические материалы.* Научные основы и технологии, 2011.

[3] Рудой, Г. И. и Стрижов, В. В.: *Алгоритмы индуктивного порождения суперпозиций для аппроксимации измеряемых данных.* Информатика и ее применения, 7(1):44–53, 2013.

[4] Вапник, В. Н.: *Восстановление зависимостей по эмпирическим данным.* М.: Наука, 1979.