

# Анализ свойств ансамбля локально аппроксимирующих моделей

Р. И. Исламов<sup>1</sup>, А. В. Грабовой<sup>1</sup>, В. В. Стрижов<sup>1</sup>

islamov.ri@phystech.edu; grabovoy.av@phystech.edu; strijov@ccas.ru

<sup>1</sup>Московский физико-технический институт

Данная работа посвящена построению универсальной модели в виде ансамбля локальных моделей. Для решения задачи регрессии предлагается использовать многоуровневый подход. Множество объектов выборки разбивается на несколько подмножеств, каждому подмножеству ставится в соответствие одна локальная модель, оптимально аппроксимирующая данное подмножество. Для аппроксимации генеральной выборки строится универсальный аппроксиматор, который представлен в виде ансамбля локальных моделей. В качестве коэффициентов шлюзовой функции используется выпуклая комбинация локальных моделей, значение которой зависит от объекта, для которого производится предсказание. Ансамблевый подход позволяет описывать те выборки, которые затруднительно описать одной моделью. Для анализа свойств проводится вычислительный эксперимент. В качестве данных используются синтетические и реальные выборки.

**Ключевые слова:** *локальная модель; линейные модели; ансамбль моделей.*

## 1 Введение

В прикладных задачах данные порождены несколькими источниками. В таких случаях качество предсказания можно повышать увеличивая число моделей. Если моделей на самом деле меньше, чем предполагается, то веса лишних моделей будут малы и их вклад будет несущественен. Преимуществом ансамбля является его способность описывать те выборки, которые затруднительно описывать одной моделью.

В данной работе исследуется проблема построения ансамбля локальных моделей. *Локальная модель* — модель, которая аппроксимирует объекты, находящиеся в одной связной области в пространстве объектов. В качестве решающей функции используется выпуклая комбинация локальных моделей, при этом веса локальных моделей не постоянны, а зависят от положения объекта в пространстве объектов.

Ансамблевый подход предполагает, что вклад каждой локальной модели в целевую переменную зависит от рассматриваемого объекта. Ансамбль локальных моделей использует шлюзовую функцию, которая определяет значимость предсказания каждой локальной модели, входящей в ансамбль.

В данной работе каждая локальная модель является линейной. В качестве функции ошибки используется логарифм правдоподобия модели. Оптимальные параметры ансамбля и локальных моделей находятся при решении двухуровневой задачи оптимизации.

Алгоритмы тестировались на синтетических и реальных данных. Эксперименты показали преимущество использования многоуровневой модели и смеси моделей по сравнению с использованием одной модели.

С момента своего появления ансамблевый подход стал предметом многих исследований. Ансамбль моделей использовался в работах [1], [2], [3]. В работе [4] представлен обзор методов и моделей в задачах ансамбля моделей. В данной работе представлены виды шлюзовых функций. Приведен анализ разных моделей, которые могут выступать в качестве локальной модели.

Ансамблевый подход имеет множество приложений в прикладных задачах. В работе [5] предложен метод распознавания рукописных цифр. Метод распознавания текстов при

29 помощи ансамбля локальных моделей исследуется в работах [6], а для распознавания речи  
30 — в [7], [8]. В работе [9] исследуется смесь экспертов для задачи распознавания трехмерных  
31 движений человека.

32 Были предложены различные типы локальных моделей, такие как SVM [10], Гауссов-  
33 ский процесс [11] и нейронные сети [12]. Другие работы были сосредоточены на различных  
34 конфигурациях, таких как иерархическая структура [13], бесконечное число экспертов [14]  
35 и последовательное добавление экспертов [15]. [16] предлагает модель ансамбля локаль-  
36 ных моделей для машинного перевода. Стробирующая сеть обучается на предварительно  
37 обученной модели NMT ансамбля.

## 38 2 Постановка задачи построения ансамбля локальных моделей

39 Пусть задано множество объектов  $\Omega$ , признаки которых описываются матрицей

$$40 \quad \mathbf{X} \in \mathbb{R}^{N \times n}, \quad (2.1)$$

41 где  $N$  — число объектов во множестве, а  $n$  — размерность признакового пространства.  
42 Каждому объекту  $\omega_i$  из  $\Omega$  соответствует признаковое описание  $\mathbf{x}_i \in \mathbb{R}^n$ , которое является  
43  $i$ -ой строкой матрицы  $\mathbf{X}$ , и значение целевой переменной  $y_i \in \mathbb{R}$ . Введем выборку данных  
44  $\mathfrak{D}$ :

$$45 \quad \mathfrak{D} = \{(\mathbf{x}_i, y_i) \mid i \in \overline{1, N}\}. \quad (2.2)$$

46 Будем считать, что множество индексов  $I = \{1, 2, \dots, N\}$  разбивается на  $K$  непересе-  
47 кающихся подмножеств  $I_k$ :

$$48 \quad I = \bigcup_{k=1}^K I_k. \quad (2.3)$$

49 Разбиение индексного множества  $I$  индуцирует разбиение множества объектов  $\Omega$  на под-  
50 множества  $\Omega_k$

$$51 \quad \Omega = \bigcup_{k=1}^K \Omega_k, \quad \Omega_k = \{\omega_i \in \Omega \mid i \in I_k\} \quad (2.4)$$

52 и выборку  $\mathfrak{D}$  на подвыборки  $\mathfrak{D}_k$

$$53 \quad \mathfrak{D} = \bigcup_{k=1}^K \mathfrak{D}_k, \quad \mathfrak{D}_k = \{(\mathbf{x}_i, y_i) \in \mathfrak{D} \mid i \in I_k\} \quad (2.5)$$

54 Для каждого подмножества объектов  $\Omega_k$  используется своя локальная модель.

55  
56 **Определение 1.** Модель  $\mathbf{g}_k$  называется локальной, если она аппроксимирует аппрокси-  
57 мирует подвыборку  $\mathfrak{D}_k = \{(\mathbf{x}_i, y_i) \in \mathfrak{D} \mid i \in I_k\}$ .

58 В данной работе локальные модели объединены в ансамбль локальных моделей.

59  
60 **Определение 2.** Ансамбль локальных моделей — мультимодель, определяющая прав-  
61 доподобие веса  $\pi_k$  каждой локальной модели  $\mathbf{g}_k$  на признаковом описании объекта  $\mathbf{x}$ .

$$62 \quad \mathbf{f} = \sum_{k=1}^K \pi_k \mathbf{g}_k(\mathbf{x}, \mathbf{w}_k), \quad \pi_k(\mathbf{x}, \mathbf{V}) : \mathbb{R}^{n \times |\mathbf{V}|} \rightarrow [0, 1], \quad \sum_{k=1}^K \pi_k(\mathbf{x}, \mathbf{V}) = 1, \quad (2.6)$$

где  $\mathbf{f}$  — мультимодель,  $\mathbf{g}_k$  — локальная модель,  $\pi_k$  — шлюзовая функция,  $\mathbf{V}$  — параметры шлюзовой функции.

В работе в качестве локальной модели  $\mathbf{g}_k$  и шлюзовой функции  $\pi$  рассматриваются следующие функции:

$$\mathbf{g}_k(\mathbf{x}, \mathbf{w}_k) = \mathbf{w}_k^\top \mathbf{x}, \quad \pi(\mathbf{x}, \mathbf{V}) = \text{softmax}\left(\mathbf{V}_1^\top \sigma(\mathbf{V}_2^\top \mathbf{x})\right), \quad (2.7)$$

где  $\mathbf{V} = \{\mathbf{V}_1, \mathbf{V}_2\}$  — параметры шлюзовой функции,  $\sigma(x)$  — сигмоидная функция. Введем понятие расстояние между двумя объектами.

**Определение 3.** Расстоянием между двумя объектами  $\omega_1$  и  $\omega_2$  из  $\Omega$  называется число, равное расстоянию между векторами  $\mathbf{x}_1, \mathbf{x}_2$  признаковых описаний этих объектов:

$$\rho(\omega_1, \omega_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_2. \quad (2.8)$$

Будем считать, что каждый объект имеет признаковое описание  $\mathbf{x}$ , взятый из некоторого вероятностного распределения. Пусть этому распределению соответствует вероятностная мера  $F(x)$  в пространстве признаков. Тогда пространство локальных моделей является гильбертовым пространством, в котором введено скалярное произведение.

**Определение 4.** Скалярным произведением между двумя локальными моделями  $\mathbf{g}_i$  и  $\mathbf{g}_j$  называется число, вычисляемое по формуле

$$\langle \mathbf{g}_i, \mathbf{g}_j \rangle = E\left(\mathbf{g}_i(\mathbf{x} \cdot \mathbf{w}_i) \cdot \mathbf{g}_j(\mathbf{x}, \mathbf{w}_j)\right), \quad (2.9)$$

где  $E(\xi \cdot \eta) = \int_{\mathbb{R}^n} \xi(\mathbf{x})\eta(\mathbf{x})dF(\mathbf{x})$  — математическое ожидание произведения случайных величин.

Для нахождения оптимальных параметров мультимодели используется функция ошибки следующего вида:

$$\mathcal{L}(\mathbf{V}, \mathbf{W}) = \sum_{(\mathbf{x}, y) \in \mathcal{D}} \sum_{k=1}^K \pi_k(\mathbf{x}, \mathbf{V}) (y - \mathbf{w}_k^\top \mathbf{x}^k)^2 + R(\mathbf{V}, \mathbf{W}), \quad (2.10)$$

где  $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k]$  — параметры локальных моделей,  $R(\mathbf{V}, \mathbf{W})$  — регуляризация параметров. Оптимальные параметры определяются из выражения

$$\hat{\mathbf{V}}, \hat{\mathbf{W}} = \arg \min_{\mathbf{V}, \mathbf{W}} \mathcal{L}(\mathbf{V}, \mathbf{W}). \quad (2.11)$$

## 3 Вычислительный эксперимент

### 3.1 Постановка проблемы

Данный эксперимент ставится для того, чтобы показать, что одна линейная модель плохо аппроксимирует выборку, объекты которой порождены несколькими источниками.

В качестве данных используются синтетическая выборка. Создается две подвыборки объектов, имеющих по одному признаку  $x^k$ , описываемых линейной моделью с нормальным шумом:

$$\mathbf{y}_k = \alpha_k \mathbf{x}^k + \varepsilon, \quad k \in \{1, 2\}, \quad x_k, y_k \in \mathbb{R}, \quad \varepsilon \in \mathcal{N}(0, 1). \quad (3.1)$$

Эти две подвыборки сливаются в одну общую выборку, описываемую вектором целевой переменной  $\mathbf{y}$  и матрицей признаков  $\mathbf{X}$ :

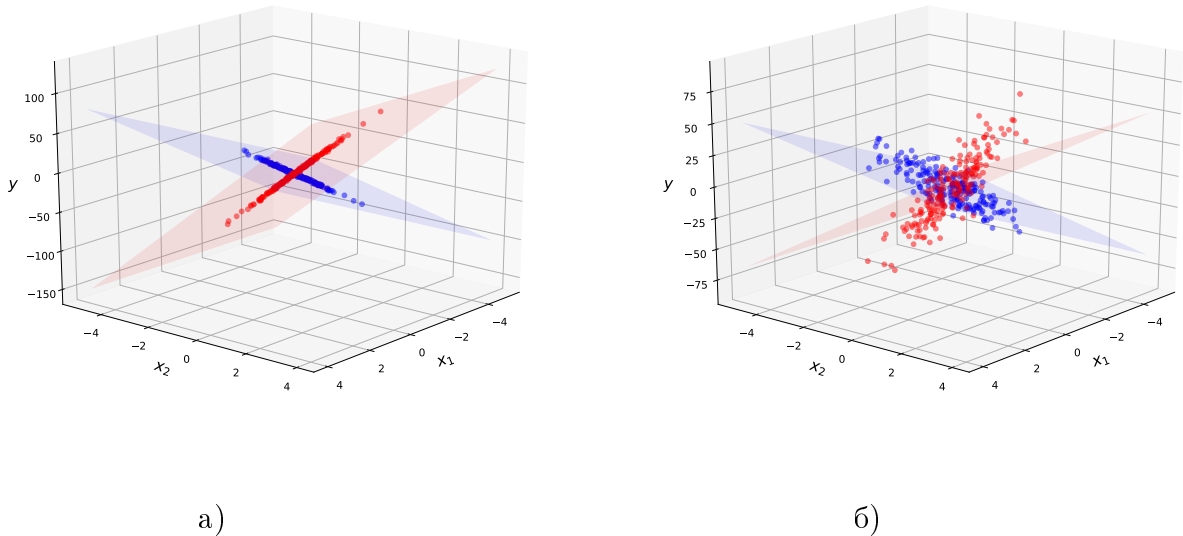
$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}^1 & \mathbf{0} \\ \mathbf{0} & \mathbf{x}^2 \end{pmatrix}. \quad (3.2)$$

На общей выборке обучается линейная модель. Линейная модель хорошо аппроксимирует данную выборку (см. рис. 1а).

Во втором эксперименте две подвыборки сливаются в одну общую выборку, описываемая целевой переменной  $\mathbf{y}$  и матрицей признаков  $\hat{\mathbf{X}}$ :

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad \hat{\mathbf{X}} = \begin{pmatrix} \mathbf{x}^1 & \varepsilon_1 \\ \varepsilon_2 & \mathbf{x}^2 \end{pmatrix}, \quad (3.2)$$

где  $\varepsilon_1, \varepsilon_2 \in \mathcal{N}(0, 1)$ . На построенной модели также обучается линейная модель. В этом случае линейная модель плохо аппроксимирует данную выборку (см. рис. 1б).



**Рис. 1** а) Признаки, соответствующие другому подмножеству, заполнялись нулями, б) признаки, соответствующие другому подмножеству, заполнялись случайными числами. Точки соответствуют правильным ответам, плоскости задают предсказание линейной модели для каждого из подмножеств.

Данный эксперимент показывает, что для аппроксимации выборки, порожденной несколькими источниками, одна модель не подходит. Таким образом, при построение модели для обучения нужно учитывать гипотезу порождения данных. Нередко оказывается,

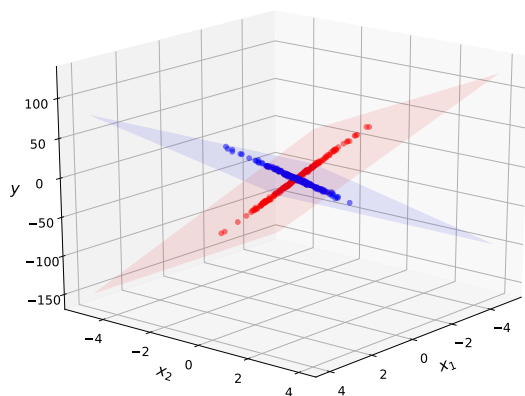
что данные порождены несколькими источниками. В этом случае для лучшей аппроксимации можно использовать ансамбль локальных моделей, где каждая локальная модель обрабатывает свою область признакового пространства (в одной области объекты имеют схожие признаки, объекты из разных областей имеют разные признаковые описания).

### 3.2 Базовый алгоритм. Построение ансамбля локальных моделей

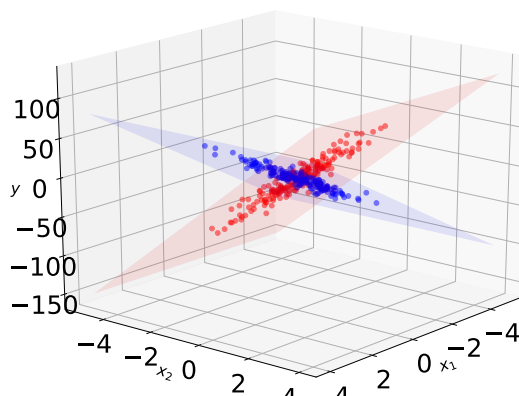
Данный эксперимент ставится для того, чтобы показать, что ансамбль локальных моделей может быть использован в случае, когда выборка порождена несколькими источниками, и лучше аппроксимирует выборку, чем одна модель.

В данном эксперименте для аппроксимации используется ансамбль двух линейных локальных моделей.

В первом эксперименте ансамбль обучается на выборке  $(y, X)$ , а во втором — на выборке  $(y, \hat{X})$ . Ансамбль двух линейных локальных моделей хорошо аппроксимирует выборку (см. рис. 2 а, б).



а)



б)

**Рис. 2** а) Признаки, соответствующие другому подмножеству, заполнялись нулями, б) признаки, соответствующие другому подмножеству, заполнялись случайными числами. Точки соответствуют правильным ответам, плоскости задают предсказание мультимодели для каждого из подмножеств.

Данный эксперимент показывает, что для аппроксимации выборки, порожденной несколькими источниками, подходит ансамбль локальных моделей. Качество аппроксимации ансамбля моделей выше, чем при использовании лишь одной модели.

## Литература

- [1] M. Serdar Yumlu, Fikret S. Gorgen, and Nesrin Okay. Financial time series prediction using mixture of experts. In *Computer and Information Sciences - ISCIS 2003*, pages 553–560. Springer Berlin Heidelberg, 2003.
- [2] Yiu-ming Cheung, Wai Leung, and Lei Xu. Application of mixture of experts model to financial time series forecasting. October 1995.

- [3] Andreas S. Weigend and Shanming Shi. Predicting daily probability distributions of s&p500 returns. *Journal of Forecasting*, 19(4):375–392, 2000.
- [4] S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, aug 2012.
- [5] Reza Ebrahimpour, Mohammad Moradian, Alireza Esmkhani, and Farzad Jafarlou. Recognition of persian handwritten digits using characterization loci and mixture of experts. *JDCTA*, 3:42–46, January 2009.
- [6] Andrew Estabrooks and Nathalie Japkowicz. A mixture-of-experts framework for text classification. In *Proceedings of the 2001 workshop on Computational Natural Language Learning*. Association for Computational Linguistics, 2001.
- [7] S. Iman Mossavat, Oliver Amft, Bert de Vries, Petko N. Petkov, and W. Bastiaan Kleijn. A bayesian hierarchical mixture of experts approach to estimate speech quality. In *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, jun 2010.
- [8] Fengchun Peng, Robert A. Jacobs, and Martin A. Tanner. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91(435):953–960, sep 1996.
- [9] Cristian Sminchisescu, Atul Kanaujia, and Dimitris N. Metaxas.  $BM^3e$ : Discriminative density propagation for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):2030–2044, nov 2007.
- [10] Ronan Collobert, Samy Bengio, and Yoshua Bengio. A parallel mixture of SVMs for very large scale problems. *Neural Computation*, 14(5):1105–1114, may 2002.
- [11] Volker Tresp. Mixtures of gaussian processes. In *Advances in Neural Information Processing Systems 13*, pages 654–660. MIT Press, 2001.
- [12] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarsz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.
- [13] Michael I. Jordan and Robert A. Jacobs. Hierarchies of adaptive experts. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 985–992. Morgan-Kaufmann, 1992.
- [14] Carl Edward Rasmussen and Zoubin Ghahramani. Infinite mixtures of gaussian process experts. In *In Advances in Neural Information Processing Systems 14*, pages 881–888. MIT Press, 2001.
- [15] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Lifelong learning with a network of experts.
- [16] Ekaterina Garmash and Christof Monz. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

Поступила в редакцию