Анализ свойств ансамбля локально аппроксимирующих моделей

Р. И. Исламов[1]***, А. В. Грабовой***[1]***, В. В. Стрижов***[1]
islamov.ri@phystech.edu; grabovoy.av@phystech.edu; strijov@ccas.ru
[1]Московский физико-технический институт

Данная работа посвящена анализу свойств ансамбля локальных моделей. Для задачи регрессии предлагается использовать многоуровневый подход, согласно которому множество объектов разбивается на несколько подмножеств и каждому подмножеству соответствует одна локальная модель. Рассматривается задача построения универсального аппроксиматора — мультимодели, которая представлена в виде совокупности локальных моделей. В качестве решающей функции используется выпуклая комбинация локальных моделей. Коэффициенты выпуклой комбинации — шлюзовая функция — функция, значение которой зависит от объекта, для которого производится предсказание. Такой подход позволяет описывать те выборки, которые затруднительно описывать одной моделью. Для анализа свойств проводится вычислительный эксперимент. В качестве данных используются синтетические и реальные выборки. В данной работе реальные данные представлены выборками из boston house prices dataset, servo dataset.

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

1 Введение

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

Подход к мультимоделированию предполагает, что вклад каждой локальной модели в ответ зависит от рассматриваемого объекта. Мультимодель использует шлюзовую функцию, которая определяет значимость предсказания каждой локальной модели, входящей в ансамбль.

В данной работе каждая локальная модель является линейной. В качестве функционала качества рассматривается логарифм правдоподобия модели. Предлагается алгоритм нахождения оптимальных параметров ансамбля и локальных моделей.

Преимуществом данного подхода является его способность описывать те выборки, которые затруднительно описывать одной моделью, и разбивать выборку в соответствии с выбранными моделями.

Алгоритмы тестировались на синтетических и реальных данных. Реальные данные представляли собой boston house prices и servo datasets. Эксперименты показали преимущество использования многоуровневой модели и смеси моделей по сравнению с использованием одной модели.

В прикладных задачах данные порождены в результате использования нескольких источников, либо гипотеза порождения и вовсе не известна. В таких случаях качество предсказания можно повышать увеличивая количество моделей. Если моделей на самом деле меньше, чем предполагается, то веса лишних моделей будут малы и их вклад будет несущественен. Этим объясняется актуальность использования мультимоделирования.

1.1 Работы по теме

С момента своего появления мультимодельный подход стал предметом многих исследований. Были предложены различные типы архитектур локальных моделей, такие как SVM [1], Гауссовский процесс [2] и нейронные сети [3]. Другие работы была сосредоточены на различных конфигурациях, таких как иерархическая структура [4], бесконечное число экспертов [5] и последовательное добавление экспертов [6]. [7] предлагает модель ансамбля локальных моделей для машинного перевода. Стробирующая сеть обучается на предварительно обученной модели NMT ансамбля.

Ансамбль локальных моделей имеет множество приложений в прикладных задачах. Работы [8], [9] [10] посвящены применению смеси экспертов в задачах прогнозирования временных рядов. В работе [11] предложен метод распознавания рукописных цифр. Метод распознания текстов при помощи ансамбля локальных моделей исследуется в работах [12], а для распознавания речи — в [13], [14]. В работе [15] исследуется смесь экспертов для задачи распознавания трехмерных движений человека.

2 Постановка задачи построения ансамбля локальных моделей

Пусть задано множество объектов Ω. Будем считать, что множество объектов разбивается на k непересекающихся подмножеств Ω_k:

$$\Omega = \bigsqcup_{k=1}^{K} \Omega_k. \tag{3.1}$$

Также задано $\Omega', |\Omega'| = N$ — множество объектов для обучения, являющееся подмножеством Ω. Предполагается, что в Ω' представлены объекты из всех подмножеств Ω_k:

$$\Omega' = \bigsqcup_{k=1}^{K} \Omega'_k, \tag{3.2}$$

где $\Omega'_k \subset \Omega_k$. Пусть задан вектор $\mathbf{y} \in \mathbb{R}^N$ — вектор правильных ответов. Разбиение множества объектов Ω' на подмножества индуцирует разбиение вектора $\mathbf{y}$ на подвекторы $\mathbf{y}_k$.

Для каждого объекта из Ω задано признаковое описание в соответствии с подмножеством, в котором объект находится. Это отображение $\mathcal{K}_k$ из множества объектов в пространство признаков:

$$\mathcal{K}_k : \Omega_k \to \mathbb{R}^{n_k}, k \in \overline{1, K}. \tag{3.3}$$

В качестве общего пространства признаков будем рассматривать $\mathbb{R}^n = \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_K}$, полным признаковым описанием объекта является вектор $\mathbf{x} = [\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^K]$. Введем выборку данных $\mathcal{D}$:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i \in \overline{1, N}\}, \tag{3.4}$$

где $\mathbf{x}_i \in \mathbb{R}^n$ — полное признаковое описание объекта из Ω', а $y_i \in \mathbb{R}$ — значение целевой переменной, соответствующее этому объекту. Для каждого подмножества используется своя локальная модель.

Определение 1. *Модель $\mathbf{g}_k$ называется локальной, если она аппроксимирует некоторую пару $(\Omega'_k, \mathbf{y}_k)$.*

В приведенном определении подразумевается, что локальная модель $\mathbf{g}_k$ использует только соответствующее признаковое описание $\mathbf{x}_i^k \in \mathbb{R}^{n_k}$ объекта — подвектор вектора $\mathbf{x}_i$, соответствующий отображению $\mathcal{K}_k$. В данной работе локальный модели объединены в ансамбль локальных моделей.

Определение 2. *Ансамбль локальных моделей — мультимодель, определяющая правдоподобие веса π_k каждой локальной модели $\mathbf{f}_k$ на признаковом описании объекта $\mathbf{x}$.*

$$\mathbf{f} = \sum_{k=1}^{K} \pi_k \mathbf{g}_k(\mathbf{x}^k, \mathbf{w}_k), \qquad \pi_k\left(\mathbf{x}, \mathbf{V}\right) : \mathbb{R}^{n \times |\mathbf{V}|} \to [0,1], \qquad \sum_{k=1}^{K} \pi_k\left(\mathbf{x}, \mathbf{V}\right) = 1, \tag{3.5}$$

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

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

$$\mathbf{g}_k(\mathbf{x}^k, \mathbf{w}_k) = \mathbf{w}_k^\mathsf{T}\mathbf{x}^k, \qquad \boldsymbol{\pi}\left(\mathbf{x}, \mathbf{V}\right) = \operatorname{softmax}\left(\mathbf{V}_1^\mathsf{T}\sigma\left(\mathbf{V}_2^\mathsf{T}\mathbf{x}\right)\right), \tag{3.6}$$

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 ω_1 и ω_2 из Ω называется число, равное расстоянию между векторами признаковых описаний этих объектов, и вычисляемое по формуле*

$$\rho(\omega_1, \omega_2) = ||\mathcal{K}_1(\omega_1) - \mathcal{K}_1(\omega_2), \mathcal{K}_2(\omega_1) - \mathcal{K}_2(\omega_2), \ldots, \mathcal{K}_k(\omega_1) - \mathcal{K}_k(\omega_2)||_2\,. \tag{3.7}$$

Будем считать, что каждый объект имеет векторное описание $\mathbf{x}$, взятый из некоторого вероятностного распределения. Пусть этому распределению соответствует вероятностная мера $\mathsf{F}(x)$ в пространстве признаковых описаний объектов. 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*

$$\langle \mathbf{g}_i, \mathbf{g}_j \rangle = \mathbb{E}\left[\mathbf{g}_i(\mathbf{x}^i, \mathbf{w}_i), \mathbf{g}_j(\mathbf{x}^j, \mathbf{w}_j)\right], \tag{3.8}$$

где $\mathbb{E}\left[\xi, \eta\right] = \int\limits_{\mathbb{R}^n} \xi(\mathbf{x})\eta(\mathbf{x})\mathsf{dF}(\mathbf{x})$ — *математическое ожидание произведения случайных величин.*

Для нахождения оптимальных параметров мультимодели используется функция ошибки следующего вида:

$$\mathcal{L}\left(\mathbf{V}, \mathbf{W}\right) = \sum_{(\mathbf{x}, y) \in \mathcal{D}} \sum_{k=1}^{K} \pi_k\left(\mathbf{x}, \mathbf{V}\right)\left(y - \mathbf{w}_k^\mathsf{T}\mathbf{x}^k\right)^2 + R\left(\mathbf{V}, \mathbf{W}\right), \tag{3.9}$$

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

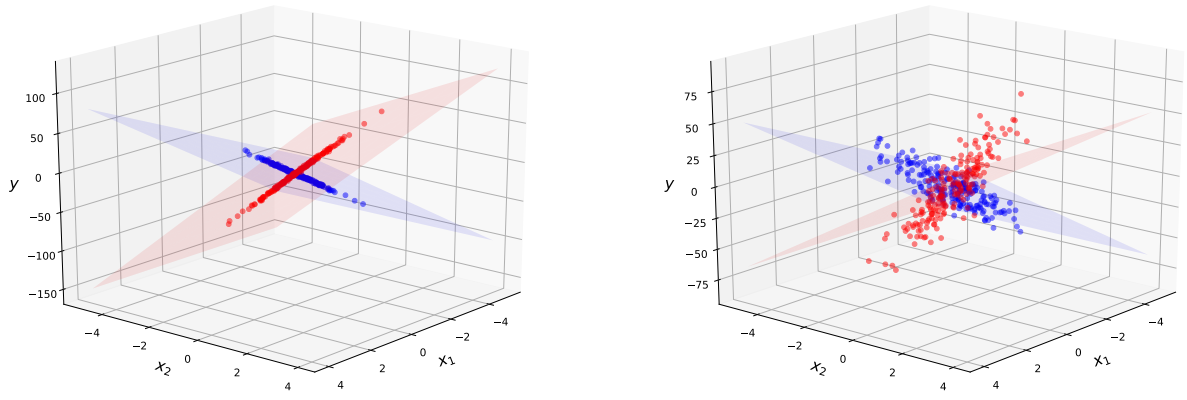
3 Вычислительный эксперимент

3.1 Постановка проблемы

В данном эксперименте изучается проблема, когда выборка имеет не одну порождающую модель, а несколько. В качестве данных используются синтетическая выборка. Создается две подвыборки объектов, описываемых линейной моделью с нормальным шумом. В каждом подвыборке объекты имеют один признак. Будем обозначать признак объектов из первого подмножества как x_1, а признак объектов из второго подмножества как x_2. Эти две подвыборки сливаются в одну общую выборку.

В первом эксперименте для объектов из одного подмножества признаки, соответствующие другому подмножеству, берутся нулевыми. На общей выборке обучается линейная модель. Так как признаки объектов, соответствующие другой подвыборке, взяты нулевыми, то это никак не мешает модели обучаться на общих данных. Точность предсказания при таком построении модели высока.

Во втором эксперименте признаки объектов, не соответствующие подмножеству, берутся из нормального распределения $\mathcal{N}(0, 1)$, то есть признаки становятся зашумленными, при этом все также используется одна общая модель. Такое построение общей выборки усложняет обучение линейной модели, так как разделение объектов на принадлежность подмножеству становится труднее. Следствием этого является снижение точности предсказания.



а) б)

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.

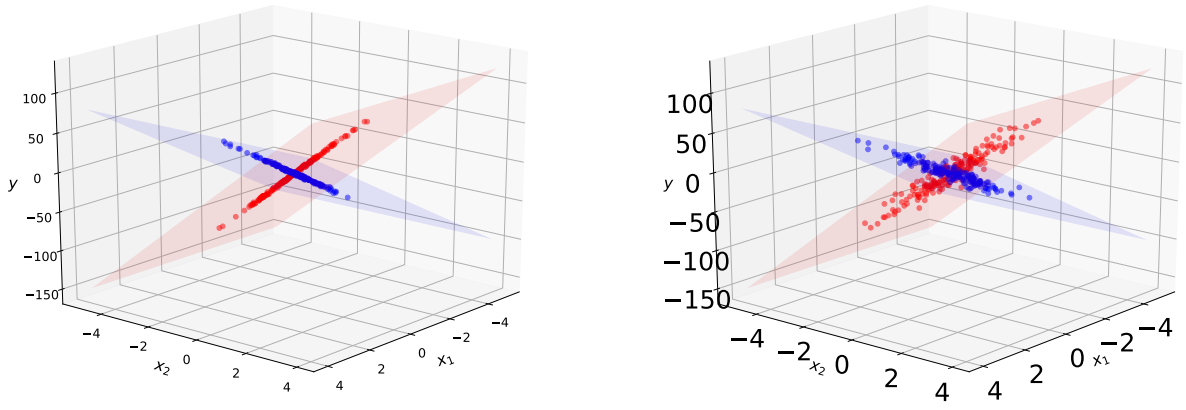
Таким образом, при построение модели для обучения нужно учитывать гипотезу порождения данных. Нередко оказывается, что данные порождены не одним источником, а несколькими. В этом случае лучше использовать мультимодель — совокупность локальных моделей, где каждая локальная модель обрабатывает свою область 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

В данном эксперименте используется мультимодель с двумя линейными локальными моделями. Обучение этого ансамбля локальных моделей происходит в два этапа: на первом этапе оптимизируются параметры локальных моделей, на втором — параметры шлюзовой функции. Используется те же два подхода построения общей выборки, как и в предыдущем пункте.

В первом эксперименте для объектов из одного подмножества признаки, соответствующие другому подмножеству, берутся нулевыми. При обучении мультимодели на такой выборке параметры локальных моделей становятся близкими друг к другу. Это объясняется тем, что для аппроксимации данной выборки достаточно одной локальной модели.

Во втором эксперименте признаки объектов, не соответствующие подмножеству, берутся из нормального распределения $\mathcal{N}(0, 1)$. В данном эксперименте мультимодель обучается так, что каждая локальная модель аппроксимирует одно из подмножеств. Параметры локальных моделей получаются близкими к нужным: один из коэффициентов близок к коэффициенту, при помощи которого порождалось подмножество, а второй — к нулю. В данном случае двух локальных моделей достаточно для аппроксимации общей выборки.



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.

Литература

[1] 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.

[2] Volker Tresp. Mixtures of gaussian processes. In *Advances in Neural Information Processing Systems 13*, pages 654–660. MIT Press, 2001.

[3] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.

[4] 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.

[5] 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.

[6] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Lifelong learning with a network of experts.

[7] 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.

[8] M. Serdar Yumlu, Fikret S. Gurgen, 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.

[9] Yiu-ming Cheung, Wai Leung, and Lei Xu. Application of mixture of experts model to financial time series forecasting. October 1995.

[10] Andreas S. Weigend and Shanming Shi. Predicting daily probability distributions of s&p500 returns. *Journal of Forecasting*, 19(4):375–392, 2000.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

Поступила в редакцию