

Обучение с экспертами для выборки со многими доменами

Н. А. Линдемманн, А. В. Грабовой

`lindemann.na@phystech.edu; andriy.graboviy@mail.ru`

Рассматривается задача аппроксимации multi-domain выборки единой мультимodelью – смесью экспертов. Каждый домен аппроксимируется локальной моделью. В работе рассматривается двухэтапная задача оптимизации на основе ЕМ-алгоритма. В качестве данных используется выборка отзывов сайта Amazon для разных типов товара, которая содержит в себе несколько доменов. В качестве локальной модели используется линейная модель, а в качестве признакового описания отзывов используются tf-idf вектора внутри каждого домена.

Ключевые слова: *Смесь экспертов, локальная модель, обучение с экспертом, классификация текстов.*

1 Введение

На текущий момент в машинном обучении появляется все больше задач связанных с данными, которые взяты с разных источников. Часто появляются выборки, которые состоят из большого количества доменов. Под *доменом* понимается подмножество объектов выборки, которые обладают некоторыми одинаковыми признаками. На текущий момент не существует полного теоретического обоснования построения смесей локальных моделей для аппроксимации такого рода выборок.

В работе рассматривается задача аппроксимации multi-domain выборки единой мультимodelью – смесью экспертов, и предлагается алгоритм бинарной классификации текстов, использующий дополнительную информацию о доменах. Использование этой информации позволяет использовать более простые и интерпретируемые модели.

Метод решения задачи состоит в построении мультимodelи, являющейся смесью локальных моделей. Каждый домен аппроксимируется локальной линейной моделью, смесь которых является итоговым алгоритмом классификации. Задача обучения модели сводится к двухэтапной задаче оптимизации на основе ЕМ-алгоритма. Схема работы алгоритма представлена на рис. 1.

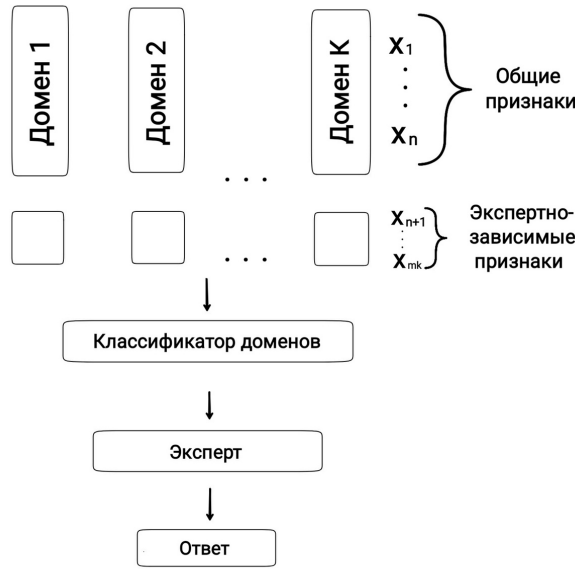


Рис. 1 Схема работы мультимодели.

В качестве данных используется выборка отзывов сайта Amazon для разных типов товара, которая содержит в себе несколько доменов. Каждый объект имеет экспертно-зависимое описание, которое определяется его принадлежностью к тому или иному домену. В качестве признакового описания отзывов используется tf-idf вектора внутри каждого домена.

2 Постановка задачи

2.1 Постановка задачи обучения одного эксперта

Задача бинарной классификации является задачей аппроксимации целевой функции

$$\mathbf{f} : \mathbb{R}^n \rightarrow \{-1, +1\},$$

где \mathbb{R}^n – пространство признакового описания объектов, а $\{-1, +1\}$ – метка класса объекта. Задачей локальной модели является аппроксимация функции \mathbf{f} на некотором домене. На основе общих признаков (x_1, \dots, x_n) эксперт генерирует экспертно-зависимые признаки $(x_{n+1}, \dots, x_{m_k})$, количество которых зависит от конкретного домена, и с помощью признаков $(x_1, \dots, x_n, x_{n+1}, \dots, x_{m_k})$ локальная модель делает предсказание о принадлежности объекта к одному из двух классов.

В качестве локальной модели будем использовать логистическую регрессию, которая будет предсказывать вероятность того, что объект с признаковым описанием \mathbf{x}_i принадлежит классу y_i :

$$p(y = y_i | \mathbf{x}_i, \mathbf{w}) = \sigma(y_i \mathbf{w} \cdot \mathbf{x}_i).$$

Рассмотрим правдоподобие выборки, а именно, вероятность наблюдать данный вектор \mathbf{y} у домена \mathbf{C} (выборка размера N). В предположении, что объекты выборки внутри одного домена независимы и из одного распределения, получим:

$$p(\mathbf{y} | \mathbf{C}, \mathbf{w}) = \prod_{i=1}^N p(y = y_i | \mathbf{x}_i, \mathbf{w}).$$

Далее рассмотрим логарифм правдоподобия:

$$\log p(\mathbf{y} \mid \mathbf{C}, \mathbf{w}) = \log \prod_{i=1}^N \sigma(y_i \mathbf{w} \cdot \mathbf{x}_i) = \sum_{i=1}^N \log \frac{1}{1 + e^{-y_i \mathbf{w} \cdot \mathbf{x}_i}} = - \sum_{i=1}^N \log(1 + e^{-y_i \mathbf{w} \cdot \mathbf{x}_i}).$$

Значит, в данном случае принцип максимального правдоподобия приводит к минимизации логистической функции потерь по всем объектам из данного домена:

$$\mathcal{L}(\mathbf{C}, \mathbf{y}, \mathbf{w}) = \sum_{i=1}^N \log(1 + e^{-y_i \mathbf{w} \cdot \mathbf{x}_i}) \rightarrow \min_{\mathbf{w}}.$$

2.2 Постановка задачи построения смеси экспертов

Обобщим подход аппроксимации одного домена на случай, когда в данных присутствует несколько доменов. Пусть всего имеется K доменов в выборке, тогда всю выборку \mathbf{C} можно представить в виде:

$$\mathbf{C} = \bigsqcup_{k=1}^K \mathbf{C}'_k,$$

где \mathbf{C}'_k множество объектов, принадлежащих k -му домену. Множеству объектов из домена $\mathbf{C}'_k \subset \mathbf{C}$ соответствует задача линейной регрессии для выборки $\mathbf{X}'_k \subset \mathbf{X}, \mathbf{y}'_k \subset \mathbf{y}$. Модель \mathbf{g}_k аппроксимирующая выборку $\mathbf{X}'_k, \mathbf{y}'_k$ является локальной моделью для выборки \mathbf{X}, \mathbf{y} .

Определение 1. Модель \mathbf{g} называется локальной моделью для выборки \mathbf{U} , если \mathbf{g} аппроксимирует некоторое не пустое подмножество $\mathbf{U}' \subset \mathbf{U}$.

Определение 2. Мультимодель \mathbf{f} называется смесью экспертов, если:

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

где \mathbf{g}_k является k -й локальной моделью, π_k — шлюзовая функция, вектор \mathbf{w}_k является параметрами k -й локальной моделью, а \mathbf{V} — параметры шлюзовой функции.

Пусть \mathbf{w}_k является случайным вектором, который задается плотностью распределения $p^k(\mathbf{w}_k)$. Получим совместное распределения параметров локальных моделей и вектора ответов:

$$p(\mathbf{y}, \mathbf{W} \mid \mathbf{X}, \mathbf{V}) = \prod_{i=1}^N \left(\sum_{k=1}^K \pi_k p_k(y_i \mid \mathbf{w}_k, \mathbf{x}_i) \right),$$

где $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_K\}$. Оптимальные параметры находятся при помощи максимизации правдоподобия:

$$\hat{\mathbf{V}}, \hat{\mathbf{W}} = \arg \max_{\mathbf{V}, \mathbf{W}} p(\mathbf{y}, \mathbf{W} \mid \mathbf{X}, \mathbf{V}).$$

3 Вычислительный эксперимент

В качестве первого вычислительного эксперимента использовался метод из работы [5].

Сперва выборка отзывов с сайта Amazon приводится к признаковому описанию tf-idf вектора. Далее с помощью ЕМ-алгоритма модель обучается в течение 15 эпох. График

71 зависимости точности классификации на тестовом наборе данных в зависимости от номера
72 эпохи обучения представлен на рис. 2.

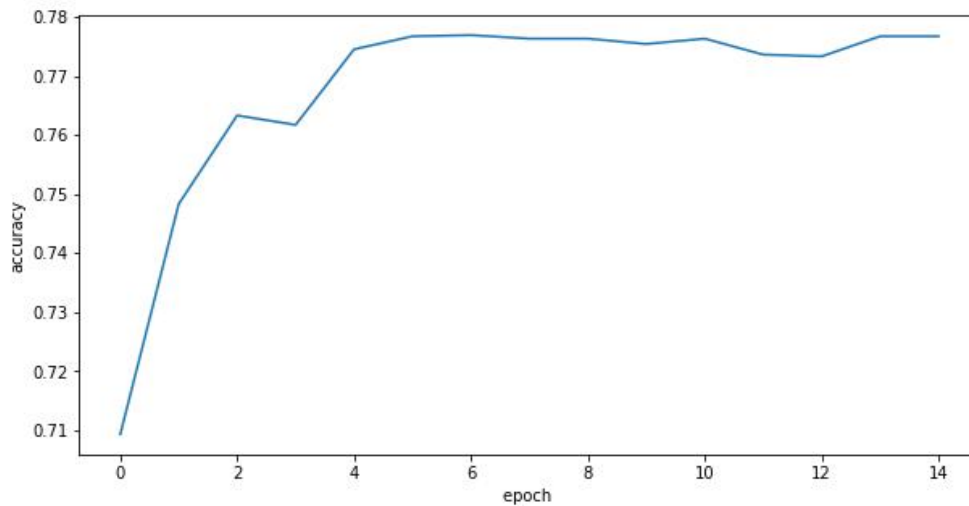


Рис. 2 Зависимость точности классификации в зависимости от номера эпохи.

Литература

- [1] J. Jiang. A Literature Survey on Domain Adaptation of Statistical Classifiers // ?????, 2007
- [2] А.В. Грабовой, В.В. Стрижов. Анализ выбора априорного распределения для смеси экспертов // ?????, 2018
- [3] G. Wilson, D.J. Cook. A Survey of Unsupervised Deep Domain Adaptation // ACM Transactions on Intelligent Systems and Technology, 2020
- [4] M. Wang, W. Deng. Deep Visual Domain Adaptation: A Survey // Manuscript accepted by Neurocomputing, 2018
- [5] J. Guo, D.J. Shah, R. Barzilay. Multi-Source Domain Adaptation with Mixture of Experts // Conference on Empirical Methods in Natural Language Processing, 2018

Поступила в редакцию