Обучение с экспертами для выборки со многими доменами

Н. А. Линдеманн, А. В. Грабовой

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

Рассматривается задача аппроксимации multi-domain выборки единой мультимоделью – смесью экспертов. Каждый домен аппроксимируется локальной моделью. В работе рассматривается двухэтапная задача оптимизации на основе EM-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 мультимоделью – 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 мультимодели, 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 EM-алгоритма. Схема работы алгоритма представлена на рис. 1.

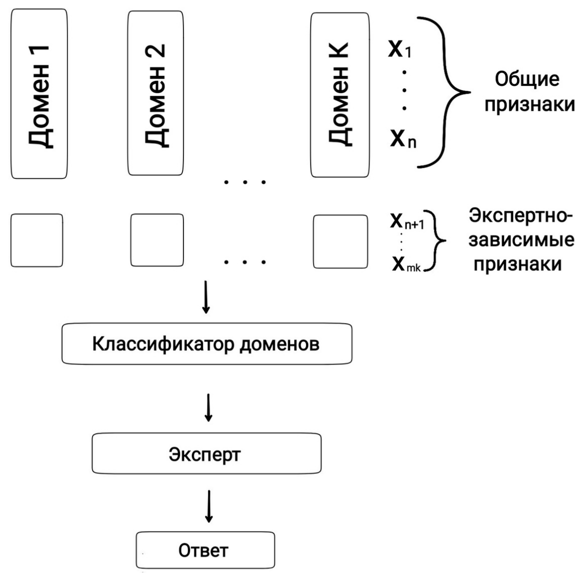

Рис. 1 Схема работы мультимодели.

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 пренадлежностью к тому или иному домену. В качестве признакового описания отзывов используется tf-idf вектора внутри каждого домена.

2 Постановка задачи

2.1 Постановка задачи обучнения одного эксперта

Задача бинарной классификации является задачей апроксимации целевой функции

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

где $\mathbb{R}^n$ – пространство признакового описания объектов, а $\{-1, +1\}$ – метка класса объекта. Задачей локальной модели является апроксимация функции $\mathbf{f}$ на некотором домене. На основе общих признаков $(x_1, \ldots, x_n)$ эксперт генерирует экспертно-зависимые признаки $(x_{n+1}, \ldots, x_{m_k})$, количество которых зависит от конкретного домена, и с помощью признаков $(x_1, \ldots, x_n x_{n+1}, \ldots, x_{m_k})$ 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 $\boldsymbol{x_i}$ принадлежит классу y_i:

$$p\left(y = y_i \mid \boldsymbol{x_i}, \mathbf{w}\right) = \sigma(y_i \mathbf{w} \cdot \boldsymbol{x_i}).$$

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

$$p\left(\boldsymbol{y} \mid \mathbf{C}, \mathbf{w}\right) = \prod_{i=1}^{N} p\left(y = y_i \mid \boldsymbol{x_i}, \mathbf{w}\right).$$

Далее рассмотрим логарифм правдоподобия:

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

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

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

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:

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

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 $p^k\bigl(\mathbf{w}_k\bigr)$. Получим совместное распределения параметров локальных моделей и вектора ответов:

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

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

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

3 Вычислительный эксперимент

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

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