Классификация товаров по ОКПД 2 кодам

Сергей Андреевич Фирсов

Московский физико-технический институт

Курс: Автоматизация научных исследований (практика, В. В. Стрижов)/Группа Б05-105

Эксперт: В. М. Старожилец

Консультант: А. Е. Вознюк

2024

Постановка задачи

Исследование направлено на решение задачи классификации товаров по кодам Общероссийского классификатора продукции по видам экономической деятельности (ОКПД 2) с использованием кратких текстовых описаний.
Выборка представлена парами "текстовое описание товара — код ОКПД2".

$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, $\mathbf{x}_i = \{t_i\}_{j=1}^{n}$ - т. описание, $y_i \in \mathbf{Y} = \{1, \ldots, k\}$.

Ограничения:

- Количество записей $m \approx 8$ миллионов, количество классов $k \approx 5000$.
- Структура классов несбалансирована: для некоторых классов доступно до 1000 записей, в то время как для других — более 200000.
- Текстовые описания часто содержат узкоспециализированную лексику, жаргонизмы, артикулы и числовые значения, что усложняет задачу классификации.

Определение модели

Используется модель логистической регрессии

$$P(y = 1|\mathbf{x}; \boldsymbol{\theta_k}) = \sigma(\mathbf{x}^\top \boldsymbol{\theta_k}),$$

где $\mathbf{x}$ обозначает вектор признаков наблюдения (с предварительно добавленной единицей для учета свободного члена), $\boldsymbol{\theta_k}$ — вектор параметров модели для класса k, а $\sigma(z) = \frac{1}{1+e^{-z}}$ — сигмоидная функция

Функция потерь и задача оптимизации:

$$\mathcal{L}(\boldsymbol{\theta_k}) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\sigma(\mathbf{x}_i^\top \boldsymbol{\theta_k})) + (1 - y_i) \log(1 - \sigma(\mathbf{x}_i^\top \boldsymbol{\theta_k})) \right],$$

$$\boldsymbol{\theta_k}^* = \arg\min_{\boldsymbol{\theta_k}} \mathcal{L}(\boldsymbol{\theta_k})$$

Алгоритм решения

Для преобразования текстов в векторное представление используются эмбеддинги, полученные с помощью библиотеки `spaCy`.
Для анализа текстовых данных и классификации используется модель логистической регрессии, реализованная в библиотеке `scikit-learn`.
Алгоритм решения:

1. Предварительная обработка данных: очистка текста от шума, нормализация и токенизация.
2. Фильтрация данных для улучшения свойств выборки.
3. Преобразование текстов в векторное представление.
4. Обучение модели логистической регрессии на обработанных данных.
5. Оценка качества модели с roc-auc и pr-auc.

Описание данных и их предобработка

Изначально данные — это пары значений: текстовое описание товара и его ОКПД2 код. Эти описания были составлены людьми и могут содержать орфографические ошибки, лишние символы, артикли, цифры и многое другое.

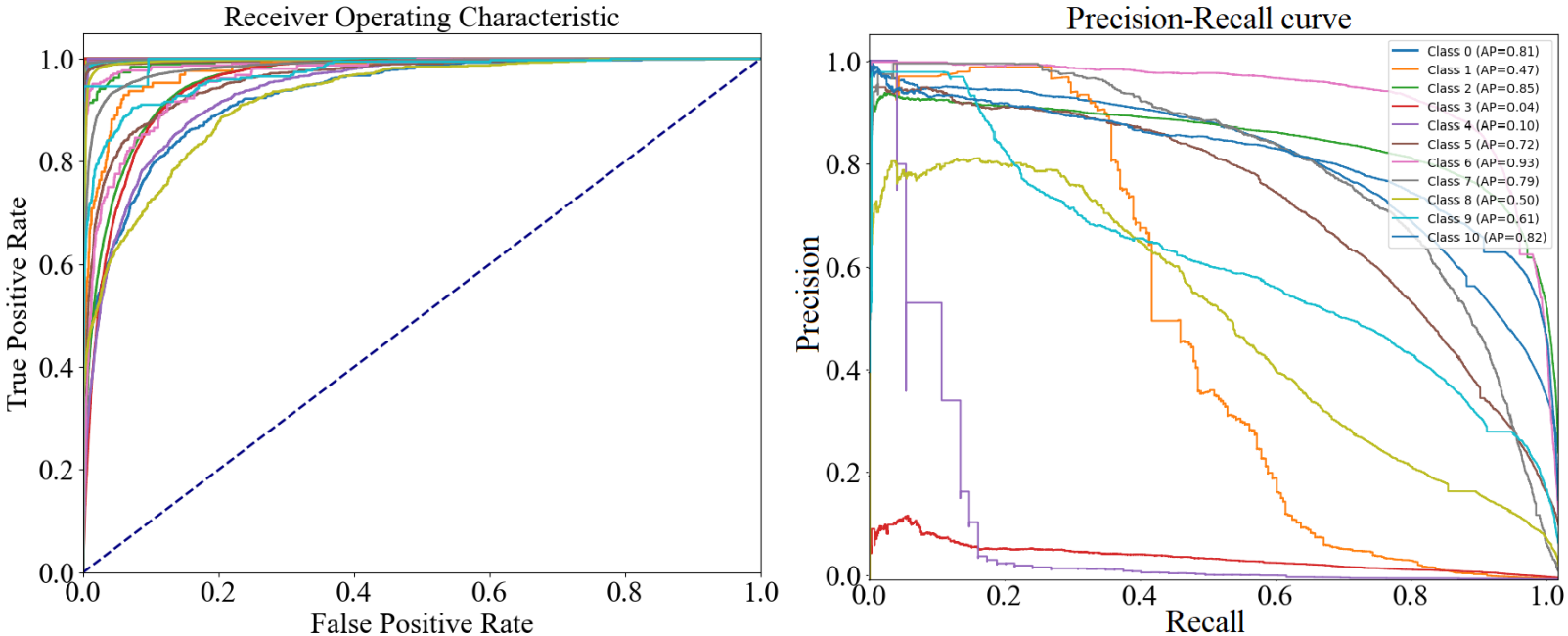
- Данные очистили и типизировали.
- В данных введено обозначение: префикс кода длины N (далее префикс N) — первые N цифр кода.
- Из-за вычислительной сложности эксперимента было принято выделить лишь несколько больших классов для анализа.
- Фильтрация для избавления от слишком мелких подклассов, чтобы избегать дизбаланса классов и улучшать качество классификации.

Условия эксперимента

- Модель "sklearn.linear_model.LogisticRegression"
- Предобученная модель NLP "ru_core_news_lg"
- Проводились отдельные эксперименты с библиотекой "gensim", для получения другой вариации эмбеддингов. Но результаты классификации оказались значительно хуже.
- Размерность векторов эмбеддингов сильно влияет на классификацию. В экспериментах выбрана размерность 300, как оптимальная для использования "large" модели из "spaCy".
  Выбор был обоснован экспериментом со сравнением качества эмбеддингов полученных с помощью "large" и "small" моделей.

Первый этап эксперимента

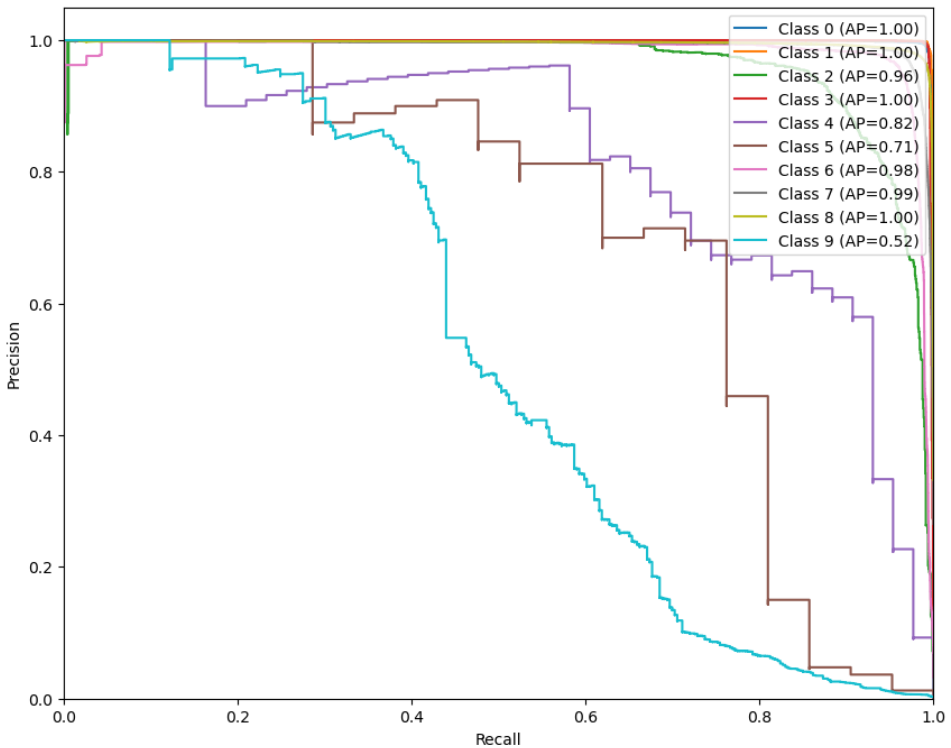
На графиках изображены ROC и PR кривые после классификации по префиксу 3.



Наблюдаем отличные показатели ROC и PR, для всех кроме классов 3 и 4 (красный и фиолетовый). После анализа выявлена проблема с построением эмбеддингов, так как описания этих классов содержат множество узкоспециализированной лексики.

Второй этап эксперимента

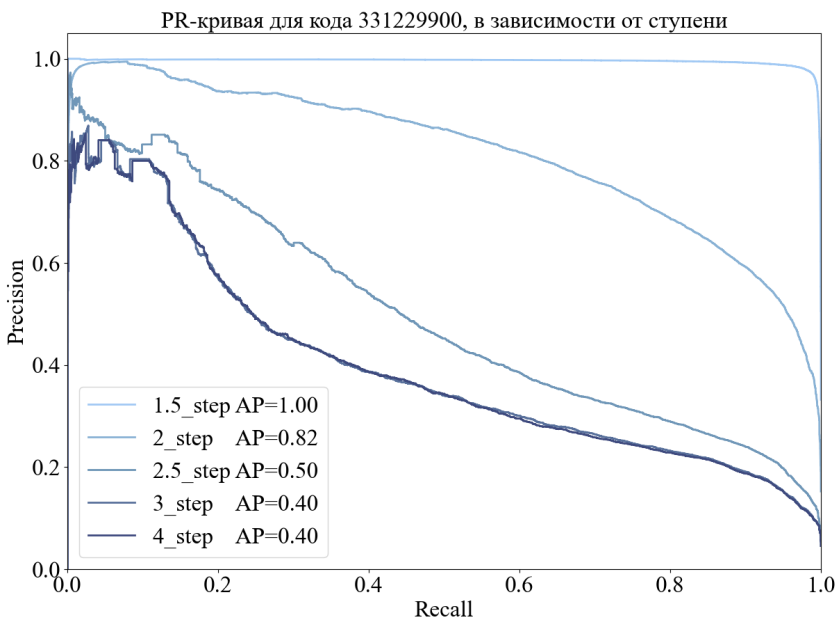
На графике изображена PR кривая после классификации по префиксу 3. Выборка изменилась — убрали классы с некачественными эмбеддингами.

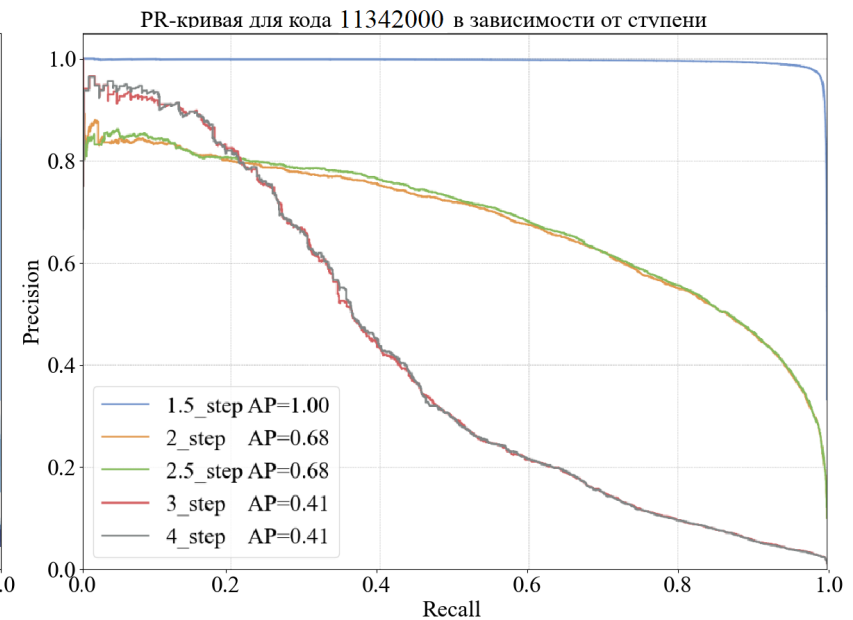


Наблюдаем хорошие показатели классификации для префикса длины 3.

Вычислительный эксперимент

Text	1 st	1.5 st	2 st	2.5 st	3 st	4 st	Embedding
укроп свежий	1	11	113	1131	11319	11319000	[0.3,0.2..]
яблоки	1	12	124	1241	12410	12410000	[-0.2,0.3..]





Иерархическая классификация

Мы достаточно хорошо умеем прогнозировать префиксы 3-4, а так как данные имеют иерархическую структуру, хотелось бы использовать результаты предсказания первых цифр при предсказании дальнейших.

- Эксперимент 1: Будем добавлять предсказанные цифры в вектор признаков.
- Эксперимент 2: Будем разделять выборку на подклассы, по результатам классификации по первым цифрам. Далее для каждого класса - обучаем свою модель.

Идеи похожи друг на друга. Вторая по сути представляет собой более строгое использование результата предыдущей классификации, по сравнению с первой.

Сравнение способов

Выборка одна и та же, всего 736 тысяч записей.
Так как всё же эксперименты чуть отличаются — не можем сравнивать их графики roc и pr. Будем анализировать качество модели, в зависимости от ступени:

Method	**Standart**	**Class split**	**Hierarchical**
prefix 4	87.6%	88.4%	88.2%
prefix 5	85.6%	86.4%	86.1%
prefix 6	83.7%	84.9%	84.7%

В таблице представлено количество правльных ответов модели (в процентах) в зависимости от метода. Standart - обычный способ, Class split - с разделением на классы, Hierarchical — с добавлением нового признака.
Вывод: Относительное количество ошибок уменьшилось примерно на 5%, абсолютная точность эксперимента - на 1%. При этом время и вычислительная сложность увеличились в разы.

Заключение

- Предложен алгоритм для решения поставленной задачи классификации.
- Реализована модель выполняющая этот алгоритм
- Исследовано влияние гиперпараметров на результаты модели
- Исследовано качество модели в зависимости от глубины классификатора
- Исследованы методы улучшенния модели и их качество

Пути улучшения:

- Улучшение качества эмбеддингов
- Объединение и усовершенствование предложенных улучшенных методов классификации
- Борьба с несбалансированностью классов

Список литературы

- Lane, H., Howard, C., & Hapke, H. (2019). *Natural Language Processing in Action*. Manning Publications.
- Marra de Artiñano, I., Riottini Depetris, F., & Volpe Martincus, C. (2021). *Automatic Product Classification in International Trade: Machine Learning and Large Language Models*.
- Lewis, D. D., et al. (2004). *RCV1: A New Benchmark Collection for Text Categorization Research.* Journal of Machine Learning Research, 5.
- Haav, H.-M. (2021). *Assessment of HS Code Correctness*.
- Muñoz, E. (2020). *Introduction to Natural Language Processing: Word Embeddings & Sentiment Analysis with Python*.