Статистические модели для анализа категоризированных финансовых последовательностей

Самарин Игорь Александрович, группа 23.М03-мм

Санкт-Петербургский государственный университет
Математическое моделирование, программирование и искусственный интеллект

Научный руководитель: к.ф.-м.н., доцент Алексеева Н.П.
Рецензент: к.ф.-м.н., доцент Аль Джубури Ф.С.

Санкт-Петербург
2025 г.

Введение

Постановка задачи:

Оценить качество параметрических и непараметрических методов в анализе категоризованных финансовых последовательностей.

Используемые подходы:

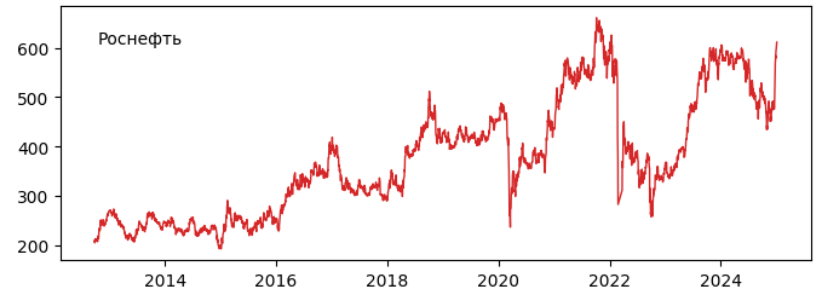
- Параметрический подход: выделение признаков через параметры отрицательного биномиального распределения;
- Непараметрический подход: обучение представлений моделью Word2Vec, архитектурами Continuous Bag Of Words и Skip-Gram.

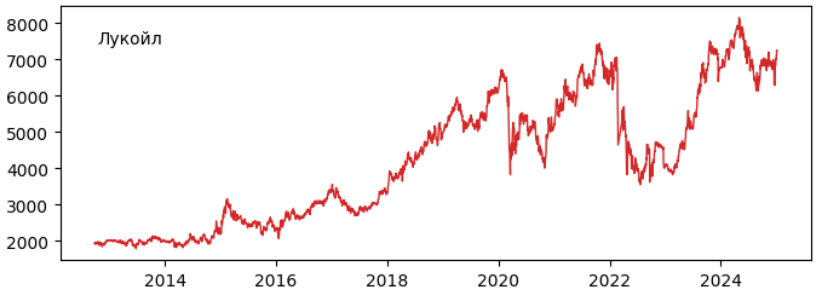
Качество подходов было оценено на задачах кластеризации и классификации.

Экспериментальные данные

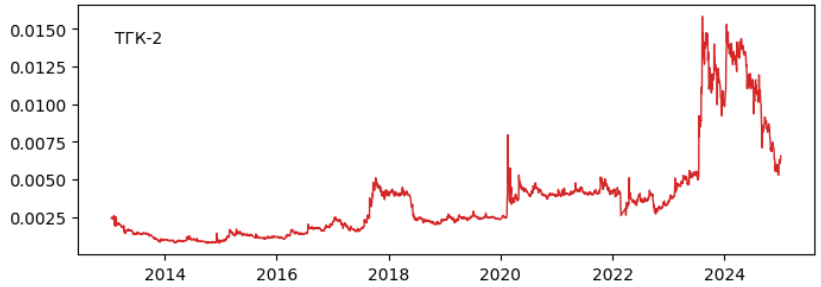
Исходные данные:

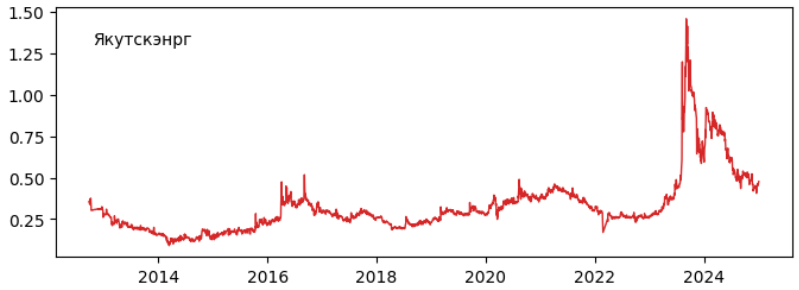
Дневные котировки популярных инструментов фондового рынка Московской биржи с 15.04.2013 по 03.09.2024.





Нефтегаз





Электроэнергетика

Экспериментальные данные. Первичная обработка

Преобразование данных:

Переход от численных значений к категориальным осуществлен по отношению текущего значения к предыдущему.

- $\mathrm{x_k}$: изменение от x_{k-1} до $2 \cdot x_{k-1}$ процентов, где $k \geq 2$;
- $\mathrm{x_1}$: изменение от 0.25 до 0.50 процентов;
- $\mathrm{n_0}$: изменение от -0.25 до 0.25 процентов;
- $\mathrm{y_1}$: изменение от -0.50 до -0.25 процентов;
- $\mathrm{y_k}$: изменение от $2 \cdot y_{k-1}$ до y_{k-1} процентов, где $k \geq 2$;

где x_k — правая граница $\mathrm{x_k}$, а y_k — левая граница $\mathrm{y_k}$ интервала.

Пример преобразованных данных:

Последовательность t_1 представима последовательностью букв t_2 и последовательностью n-грамм t_3 при $n = 3$.

$$t_1 : (10.5\ 10.6\ 10.6\ 10.4) \mapsto t_2 : (n_0\ x_2\ n_0\ y_3) \mapsto t_3 : ([n_0\ x_2\ n_0]\ [x_2\ n_0\ y_3]).$$

Параметрический подход

Предположение:

Наиболее информативные категориальные подпоследовательности подчиняются отрицательному биномиальному распределению.

$$X \sim \mathrm{NB}(r, p), \quad \mathbb{P}(X = k) = \frac{\Gamma(r+k)}{k!\Gamma(r)} p^r (1-p)^k.$$

Значения параметров в лингвистике [Alexeyeva et al., 2013]:

- r — количество неупотребления, контекстных замен слова;
- p — вероятность неупотребления слова.

Параметрический подход. Вторичная обработка

Разделим последовательность на m временных интервалов. Найдем для всех n-грамм значения $X_1^{(j)}, \ldots, X_m^{(j)}$, где $X_i^{(j)}$ — количество вхождений j-ой n-граммы в i-ый интервал.

Оценка параметров:

Параметры по методу максимального правдоподобия:

$$\hat{p} = \frac{\hat{r}}{\hat{r} + \bar{x}}, \quad \hat{r} : m \ln(\hat{p}) - m\psi(\hat{r}) + \sum_{i=1}^{m} \psi(\hat{r} + X_i) = 0,$$

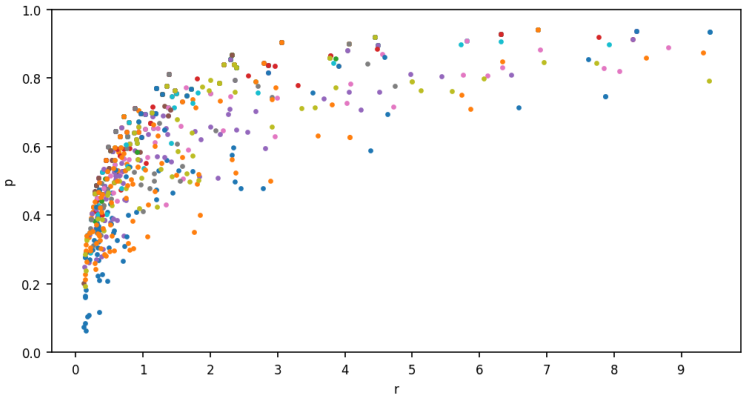
где $\psi(x) = (\ln \Gamma(x))'$, m — количество интервалов.

Соответствие распределений:

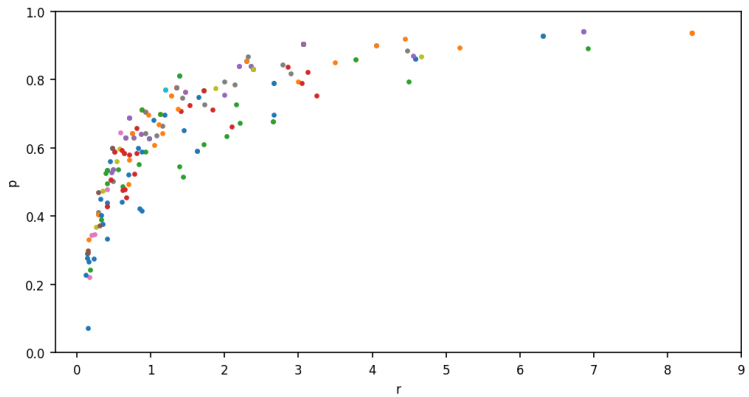
Гипотеза соответствия эмпирического закона распределения с теоретическим по критерию хи-квадрат.

Параметрический подход. Параметры распределения

Точечные диграммы значений параметров n-грамм.



$n = 3$



$n = 4$

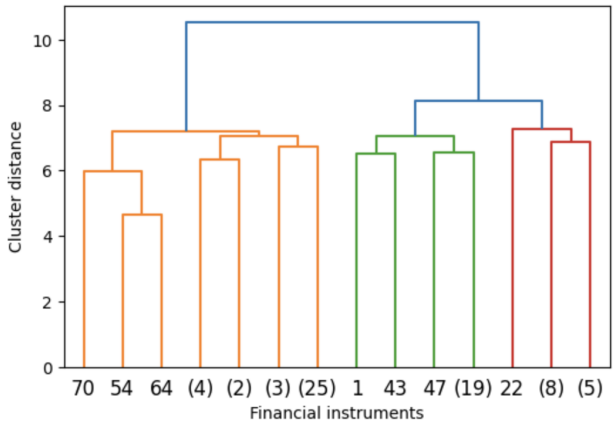
Поведение параметров категоризированных последовательностей похоже на поведение параметров слов в текстах [Samarin I., 2023].

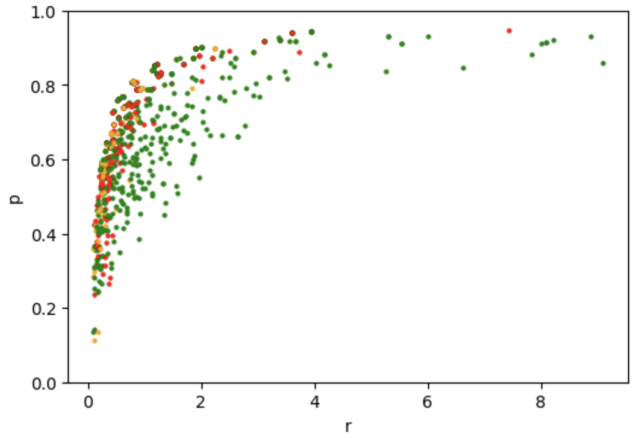
Параметрический подход. Кластеризация

Последовательности были кластеризированы по агрегированным векторам параметров распределений.

Результаты:

Последовательности делятся на три кластера.





В зеленом кластере электроэнергетические последовательности, в красном — нефтегазовые, в оранжевом — банковского сектора и торговли.

Параметрический подход. Классификация

Последовательности были классифицированы по двум секторам экономики — электроэнергетической и нефтегазовой отрасли.

Результаты:

В таблице представлены значения критерия $f1$-score.

	2-gram	3-gram	4-gram
Decision Tree	0.5142	0.5428	0.5714
Random Forest	0.5714	0.7428	0.7456
Logistic Regression	0.6285	0.7714	0.7804

Увеличение длины n-граммы приводит к росту значений качества классификации.

Непараметрический подход

Предположение:

Категориальные подпоследовательности представимы численными последовательностями, сохраняющими семантическую связь.

$$X_i \equiv X_j \Leftrightarrow \rho(\mathcal{F}(X_i), \mathcal{F}(X_j)) \approx 0,$$

где $\mathcal{F}(\cdot)$ — полносвязная или рекуррентная нейронная сеть, а $\rho(\cdot)$ — произвольная функция расстояния (например, евклидова).

Непараметрический подход. Вторичная обработка

Сопоставим контекстно близким по значению n-граммам близко расположенные векторные представления.

Обучение представлений:

Назовем контекстом окно ширины $2k+1$, а слово на $k+1$ позиции центральным. Будем перемещать окно слева направо и обучать для каждой n-граммы векторы v_u и v_w.

Варианты архитектур:

- Continuous Bag of Words: центральное слово по контексту;
- Skip-Gram: слова из контекста по центральному слову.

Вектором последовательности является Mean Pooling вектор представлений n-грамм, входящих в последовательность.

Непараметрический подход. Классификация

Последовательности были классифицированы по двум секторам экономики — электроэнергетической и нефтегазовой отрасли.

Результаты:

В таблице представлены значения критерия $f1$-score.

	25-dim	50-dim
Continuous Bag of Words	0.7351	0.7351
Skip-Gram	0.7559	0.7559

Увеличение размерности не приводит к улучшению классификации.

Итоги

- Описаны параметрические и непараметрические методы анализа категоризированных финансовых последовательностей;
- Показана схожесть поведения параметров отрицательного биномиального распределения в данных разного происхождения;
- Показана целесообразность использования параметрического метода для экспериментальных данных малого объема.

Полученные результаты могут быть использованы в анализе категориальных переменных.