Прогнозирование направления движения цены биржевых инструментов по новостному потоку.

Ахияров В., Борисов А., Говоров И., Дробин М., Мухитдинова С., Родионов В.

akhiarov.va@phystech.edu, borisov.as@phystech.edu, govorov.is@phystech.edu, drobin.me@phystech.edu, muhitdinova.sm@phystech.edu, rodionov.vo@phystech.edu

МФТИ (ГУ)

Аннотация: В работе рассматривается задача классификации направления движения временных рядов. Классификация производится с помощью анализа признаков из отчётов 8-K, которые компании обязаны заполнять при значительных событиях, таких как банкротство, выбор совета директоров и пр. Рассматривается несколько моделей классификации. В одних используются только признаки из отчётов, 1-граммы которых встречающиеся более 10 раз. В других к предыдущему этапу применяется неотрицательная матричная факторизация (NMF). И в последней, ансамбле, объединяются предыдущие подходы путём голосования большинства. В качестве прикладной задачи рассматривается задача распознавания направления движения акций по новостям, выраженных отчётами 8-K. Модели классификации, исследованные в этой работе, сравнивается в точности и статистической значимости с простыми моделями, использующими только прогнозируемый показатель доход на акцию или использующую только финансовые показатели.

Ключевые слова: *метрическая классификация, анализ текстов, классификация временных рядов, новостной поток*

1 Введение

1. **Прогнозирование направления движения цены биржевых инструментов по новостному потоку.** Мотивируемое тем, что флуктуации цен на бирже, сильно зависящие от политической, географической и т.д. обстановок, интересные не только при скальпинге. Для среднесрочных торгов и инвестиций такие данные так же имеют большую роль, позволяя корректировать вложения. Как правило, крупные изменения в политике, природные катаклизмы и все события которые именяют распределение цен котировок, освещаются в прессе.
2. Исследование строится вокруг постоянных изменений цен биржевых котировок, новостей, и алгоритма NMF вектора.
3. Требуется на основе большого количество новой информации (предоставляемой в разрозненном текстовом виде) касающейся компаний, перечисленных на фондовом рынке, предсказать повышение, понижение либо стабилизацию цен на акции, ценные бумаги и т.д. Необходимо разработать модель, которая также учитывает недавнее движение акций, и так называемую “неожиданную прибыль”(отчет о прибылях и убытках компании, значительно отличающийся (в положительном или отрицательном направлении) от ожиданий аналитиков (согласованного прогноза)
4. **Методы исследования.** В работе приведены другие, которые как улучшают уже существующие, так и вводят новые методы обработки естественного языка. Так в Xie et al. (2013) вводится дерево представлений об информации в новостях, в Bollen et al. (2010) использованы данные из Twitter’a. Bar-Haim et al. (2011) распознают лучших экспертов-инвесторов, a Leinweber and Sisk (2011) исследуют влияние новостей и времени усвоения новостей в событийной торговле. В Kogan et al. (2009) приводится

предсказание риска по финансовым отчётам и в Engelberg (2008) - закономерность о том, что лингвистическая информация (возможно из-за когнитивной нагрузки при обработке) имеет более долгосрочную предсказуемость цен, нежели количественная информация.

5. **Решаемая в данной работе задача.** Построить и исследовать модель прогнозирования направления движения цены. Задано множество новостей S и множество временных меток T, соответствующих времени публикации новостей из S. 2. Временной ряд P, соответствующий значению цены биржевого инструмента, и временной ряд V, соответствующий объему продаж по данному инструменту, за период времени T'. 3. Множество T является подмножеством периода времени T'. 4. Временные отрезки w=[w0, w1], l=[l0, l1], d=[d0, d1], где w0 < w1=l0 < l1=d0 < d1. Требуется спрогнозировать направление движения цены биржевого инструмента в момент времени t=d0 по новостям, вышедшим в период w.
6. **Предлагаемое решение.** 8K - отчеты компаний об их внутренних событиях. Данная отчетность выходит строго в период между закрытием торгов в один день и их открытием на следующий день. Из отчета 8K убираются все HTML-теги, таблицы и прочее. Используется метод NMF вектора. Вычитается из цен сегодняшнего открытия торгов вчерашние цены закрытия торгов с поправкой на индекс. Берется текст отчета 8K и на выходе нейронной сети функция, принимающая три значения : *UP-цена открытия следующего дня больше на 1+*DOWN- цена открытия следующего дня меньше на 1+*STAY - цена открытия следующего дня в пределах +/-1
7. Работа, описывающая наиболее близкое решение
8. **Плюсы метода:** Большой объем данных Он более доступен небольшим инвесторам, чем real-time trading tools, которыми пользуются большие трейдинговые компании Он показывает accuracy на 10смотрят "изменение цены"-"изменение индекса" => чистое влияние все дивидендные гэпы убирали
   Минусы: Исследование проведено на рынке США, где отчеты выходят не в торговое время => вся информация отражается мгновенно в цене акции от открытии результаты не имеют значения на практике => невозможно извлечь финансовую прибыль Метод не улавливает такие эффекты, как: slippage, transaction costs, borrowing costs
9. -
10. Эксперимент будет проведен на финансовых данных: данные о котировках (с интервалом в один тик) нескольких финансовых инструментов (GAZP, SBER, VTBR, LKOH) за 2 квартал 2017 года с сайта Finam.ru; для каждой точки ряда известны дата, время, цена и объем. И на текстовых данных: экономические новости за 2 квартал 2017 года от компании Форексис; каждая новость является отдельным html файлом.

2 Постановка задачи

Задача ставится с целью классификации текстов(финансовых отчетов), поэтому необходимо построить соответствующее признаковое пространство.В нашем распоряжении имеются отчеты и время, когда данный отчет появился. В нашем случае $\mathfrak{D}$ — наша выборка, представляющая собой описание текста при помощи двух векторов. Вектора текста $\mathbf{x}(t) = [x_1, \ldots, x_m]^\top$ И вектора времени $\mathbf{t} = [t_1, \ldots, t_m]^\top$. Вектор текста описывает наличие или отсутствие в тексте отобранных признаков: слов, обладающих наибольшей релевантностью. Поэтому этот вектор будет бинарным. Задана выборка $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, где $y \in \{0, 1\}$, 0 — stay, 1 — move. Будут рассматриваются модели-претенденты $\mathfrak{F} = \{f(\mathbf{w}, \mathbf{x})\}$ а именно: логистическая регрессия, линейный вектор опорных векторов, случайный лес

и градиентный бустинг. Моделью будет являться параметрическое семейство функций $f(\mathbf{w}, \mathbf{x}) = \mu(\mathbf{w}^\top \mathbf{x})$, где в общем случае задач классификации $\mu = \frac{1}{1+exp(-\mathbf{w}^\top \mathbf{x})}$.
Для нашей задачи мы будем использовать 2 функции ошибки. Первую выбирем на основе F_1-меры:

$$S_1 = 1 - F_1 = 1 - 2 \cdot \frac{Precision \times Recall}{Precision + Recall}$$

где точность $Precision = \frac{TP}{TP+FP}$, полнота $Recall = \frac{TP}{TP+FN}$ с обозначениями TP — истино-положительное решение, TN — истино-отрицательное решение, FP — ложно-положительное решение, FN — ложно-отрицательное решение (из задачи бинарной классификации с классами $\{-1\}$ и $\{1\}$). Вторая функция ошибки будет основана на AUC — ROC:

$$S_2 = 1 - \int_{\infty}^{-\infty} TPR(T) \cdot FPR(T) dT$$

где $FPR = \frac{FP}{FP+TN}$ — доля неверно принятых объектов, а $TPR = \frac{TP}{TP+FN}$ — доля верно принятых объектов.
Выборку $\mathfrak{D}$ разобъем на 3 части: Тестовая $\mathfrak{D}_t$ (в ней будут лишь данные с 2002 по 2009 год). На этой выборке будет происходить обучение. Дополнительная $\mathfrak{D}_a$ (данные с 2009 по 2011 год). На ней будут донастраиваются параметры. Контрольная $\mathfrak{D}_c$ (данные с 2011 по 2013 год). На ней будем тестировать качество построенной модели.
Бинарной классификацией будет отображение f признакового описания текста $\mathbf{x} \in \mathbf{x}(t)$ в метку класса $\{0, 1\}$:

$$f : (\mathbf{w}, \mathbf{x}) \mapsto y$$

Где вектор $\mathbf{w}$ — вектор параметров классификатора.
Таким образом основной задачей становится поиск оптимального классификатора $f(\mathbf{x}_i)$ при $\mathbf{x}_i \in \mathfrak{D}_t$ из условия:

$$\hat{f} = \text{argmin}_f S$$

где $S = \{S_1, S_2\}$.

Литература

[1] Hongping Hu, Li Tang, Shuhua Zhang, Haiyan Wang (2018) *Predicting the direction of stock markets using optimized neural networks with Google Trends*, Neurocomputing.

[2] Mikhail Kuznetsov, Anastasia Motrenko, Rita Kuznetsova, Vadim Strijov (2016) *Methods for Intrinsic Plagiarism Detection and Author Diarization*, CLEF (Working Notes).

[3] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, Dan Jurafsky (2014) *On the Importance of Text Analysis for Stock Price Prediction*, Proceedings of the Ninth International Conference on Language Resources and Evaluation.

[4] Anna Potapenko, Artem Popov, Konstantin Vorontsov (2017) *Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks*, CoRR.

[5] Andrew Sun, Michael Lachanski, Frank J. Fabozzi (2016) *Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction*, International Review of Financial Analysis.

[6] Усманова К. Р., Кудияров С. П., Мартышкин Р. В., Замковои А. А., Стрижов В. В. (2018) *Анализ зависимостей между показателями при прогнозировании объема грузоперевозок*, Системы и средства информатики.