Литобзор

7 ноября 2016 г.

Аннотация

Испокон веков люди хотели знать будщее. Некоторые по религиозным соображениям, некоторые с корыстными целями, а некоторые просто так. Развитие статистических методов во второй половинеXX века и развитие вычислительной техники начала XXI века позволили использовать все более сложные статистические модели, которые смогли показать очень высокие результаты. В данной работе будут рассмотрены несколько моделей нацеленных на предсказание значений акций компаний на бирже.

Часть I
Формулировка проблемы

Решение проблемы прогнозирования поведения акций компаний является востребованной проблемой, так как с помощью купли-продажи акций в нужные моменты времени может быть получен значительный доход. Сама задача прогнозирования может быть сформулирована либо как задача регрессии(то есть предсказания вещественной переменной в один или несколько моментов времени например: цена акций в конце дня), либо как задача классификации(то есть задача отнесения к одному из нескольких классов например: уменьшение или повышение стоимости акций. Среди профессиональных трейдеров существует так называемая гипотеза эффективного рынка [1], утверждающая что в любой момент времени значения стоимостей акций полностью отражают всю имеющуюся информацию, то есть что нельзя лучше чем случайно предсказать значение в будущие моменты времени. Вопреки этому утверждению, выходят все новые статьи, утверждающие возможность предсказания с помощью современных статистических методов. Здесь рассматриваются методы предсказания с помощью обработки текстовой информации.

Часть II

Возможные подходы к решению проблемы

Разделяют несколько подходов к прогнозированию поведения цен акций. Один рассматривает значения цен как временной ряд и использует различные существующие инструменты для моделирования временных рядов. Это назвывается technical analysis. Другой использует такую дополнительную информацию как, например, многочисленные финансовые обзоры и статьи или просто комментарии из социальных сетей или блогов. Это называется fundamental analysis. И тут уже применяются очень много различных методов: начиная от эврестических и заканчивая сложными рекуррентными сетями с миллионами настраевымых параметров.

Часть III

Подходы, использованные в рассмотренных статьях

В рассмотренных статьях так или иначе используется обучение с учителем для решения поставленных проблем. Также делается и проверяется предположение, что публикации в различных изданиях и блогах хорошо кореллируют с изменениями в ценах акций. Причем обычно считается что больший объем публикаций соответствует большему изменению в ценах, а также что большое количество негативных мнений в публикациях соответствует уменьшению цен акций, в то время как позитивные мнения способствуют увеличению цен акций. Эффективное извленчение настроений и представление документов и является основной сложностью в данном подходе. Во всех рассмотренных статьях использовался подход bag-of-words, со стеммингом и извлением различных слов, которые, по мнению авторов, не несут никакого положительного или отрицательного оттенка. Для того чтобы уменьшить размерность пространства признаков некоторые авторы ранжируют документы с помощью $tf - idf$ и критерия χ^2. Для определения того негативная новость или нет для каждого документа считается полярность(polarity). Polarity - эвристическая метрика, вычисляемая по размеченным оттенкам для каждого слова. Она обычно определяется как сумма всех оттенков слов. Авторы отмечают, что исключение документов с полярностью близкой к нулю помогает улучшить точность предсказаний. Также, один из авторов фильтровал документы по частоте встречаемых слов с оттенками чтобы исключить документы с очень большим количеством слов с оттенками. Такие документы могут быть сгенерированы автомати-

ческими системами для различных нужд плохих людей. Далее, матрица признаков и лейблов подается на вход регрессору/классификатору в зависимости от формулировки задачи из I. Самые лучшие результаты показал SVM с ядром-смесью нескольких полиномильных ядер также известный как MKL(Multi Kernel Learning). Сравнимые результаты показывает также Random Forest. К сожалению, не получилось найти подходы с применением бустинга.

Часть IV
Проблемы подходов, использованных в рассмотренных статьях

1 Проблемы, выделяемые авторами статей

При анализе решений, авторами отмечается ряд проблем связанных с применением текстовой информации.

- Во-первых, может быть такое, что не вышедшие статьи влияют на цены акций, а цены акций на вышедшие статьи. Если статьи выходят позже, делать что-то на бирже уже поздно и остается только ждать выхода новых статей. В этом свете авторами отмечается, что различные источники по-разному влияют на изменения курсов акций. Ежедневные финансовые публикации кореллируют с изменением курсов только в тот же самый день, в то время как блоги вроде twitter или lifejournal кореллируют дольше, но корелляция монотонно убывает. С чем это связано, вообще говоря, не очень понятно. Может быть трейдеры считают авторитетными только ежедневные издания.

- Освещение в прессе, очевидно, зависит от медийности компании и популярности того чем она занимается для обычного человека. Конечно, многие будут писать в твиттер что у них взорвался телефон, но вот про какие-нибудь компании-добытчики нефти будут писать только узкоспециализированные издания. Этим примером авторы хотят показать, что количестно публикаций как статистика для какой-либо количественной оценки изменения курса акций - это не самая удачная идея.

- Использование оффлайн подхода может привести к тому, что не все актуальные данные будут использованы при оценке параметров модели и приведут к худшим результатам по сравнению с онлайн моделью

2 Примечания автора обзора

- Ни одна из рассмотренных статей не использует word embeddings для sentiment analysis. Модель и использованием word embeddings сейчас является state of the art в sentiment analysis.
- Предпроцессинг далеко не идеален, и удаляет огромную часть информации из текста. Также, использование словарей оттенков слов предполагает наличие подобных словарей для всех слов и языков, что конечно, не так. С помощью подхода авторов нельзя распознать такие частые конструкции как ирония и сарказм, которые появляются очень часто в обзорных статьях и отзывах.

Часть V
Предлагаемые способы решения проблем.

Использование word embeddings позволяет отказаться от словарей сделанных людьми и перейти к пространствам более низких размерностей. Это значит что предсказательная система сможет работать для статей на любом языке, а также что матрица признаков не будет разреженной. Это очень хорошо для некоторых алгоритмов машинного обучения.

Использование рекуррентных нейронных сетей с модификациями памяти вроде LSTM или GRU может лучшим результатам благодаря «сохранению» информации о новостях на протяжении нескольких итераций.

В недавней статье[8] описывается применение моделей со скрытыми переменными и em алгоритма для разрешения неоднозначностей в модели skip-gram. Добавление дополнительных скрытых переменных типа оттенок может позволить получить еще более качественную модель для определения настроений отзыва или статьи и, как следствие, получения лучшего предсказания цен акций.

Список литературы

[1] https://en.wikipedia.org/wiki/Efficient-market_hypothesis

[2] http://michael.hahsler.net/research/misc/ICCTS_2012_NewsSentimentAnalysis.pdf

[3] http://uir.ulster.ac.uk/33264/1/File07376681.pdf

[4] https://arxiv.org/pdf/1607.01958.pdf

[5] http://airccse.org/journal/ijsc/papers/3211ijsc03.pdf

[6] http://www.icwsm.org/papers/3--Godbole-Srinivasaiah-Skiena.pdf

[7] https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1529/1904

[8] https://arxiv.org/abs/1502.07257