Разбор статьи "Trade the tweet: Social media text mining and sparse matrix

factorization for stock market prediction" Andrew Suna, Michael Lachanskib, Frank J. Fabozzic

Борисов Александр, МФТИ

**Постановка задачи**

Исследование потенциальной использование текстовой информации от созданных пользователем микроблогов для прогнозирования уровня цен на бирже. Используется модель скрытого пространства, сопоставляется движение цены на акции и контента в социальных сетях.

**Данные**

Используются Текстовые данные принадлещие компании StockTwits.com. Было использовано около 45 миллионов сообщений от StockTwits с 1 января 2011 года по 31 августа 2015 года. В каждой сущности было около 40 различных признаков, включая контент, счетчик подписчиков, время отправления и теги.

**Предложенный метод решения**

Сначала исходные тексты твитов подвергаются первичной обработке, используя библиотеку R tm, а именно: все записи за один день объединяются в один большой текст. Для внутридневных экспериментов данные дополнительно разделены на AM, PM периоды. Далее текст очищается от слов которые могут вызвать дополнительный шум: URL-адреса веб-сайтов и смайлики. Весь текст сохраняется в нижнем регистре для удобства сравнения. Затем создается словарь терминов, которые записываются в т.н. “матрицу терминов”, где строки соответствуют слагаемым в словаре, а столбцы соответствуют документам. Далее текст за каждый день нормализуется по методу косинусов путем деления на евклидову норму. Таким образом, данные преобразуются для наилучшего использования в весовой схеме, предложенной Лоугрэмом и МакДоналдом в 2011 году вместе с моделью SFM

**Актуальность / обоснование метода**

Выбор в пользу модели “разреженной матрицы факторизации” (SMF) был сделан по причине того, что она имеет два благоприятных качества для прогнозирования цен: методология производит матрицы низкого ранга, когда контролируются определенные значения параметров и разреженная матрица выбирает только наиболее важные параметры для прогнозирования фондового рынка, сводя к минимуму риск переобучения.

**С чем сравнивают**

Данные, полученные при помощи модели SMF сравниваются со следующими величинами:

• Предыдущий доход / цена: ожидается, что доходы и цены будут

как и предыдущий день

•Данными, полученными при помощи Модели авторегрессии (AR), прогнозирующими доходность для каждой акции в отдельности

• Случайные данные: прогнозы возврата на основе таймере рынка и случайных догадках

**Результаты**

Используемая SMF-модель и данные с StockTwits показали, что работают лучше

чем большинство базовых моделей. Точность прогнозирования составляла 51,37% для ежедневного прогнозирования.

**Доступность данных**

Данные находятся в открытом доступе на сайте <https://stocktwits.com/> и распространяются без ограничений.

**Доступность реализации**

В статье не опубликованы источники на открытый код проекта. Сложность самого проекта оценивается как сложная.

**Плюсы / минусы статьи по вашему мнению**

Точность предсказания, по моему мнению, была недостаточной. Большинство инвесторов

не считайте точность прогнозирования 51% значимой

**Идеи для дальнейших исследований (указанные в статье / предложенные вами)**

Использование данной модели и алгоритма на других данных , полученных из проверенных источников, т.к. в статье указано, что в отличие от источников высококачественных новостей, таких как Wall Street Journal, информации StockTwits может потребоваться всего несколько часов чтобы повлиять на рынок, а не целый день, как у журнала с Wall Street.