ПОДГОТОВКА ДАННЫХ С ВЫСОКИМ УРОВНЕМ ОШИБОК И ШУМА К ОБРАБОТКЕ АЛГОРИТМАМИ МАШИННОГО ОБУЧЕНИЯ

Р.Э. Мнацаканян
Научный руководитель: А.В. Кудинов
Томский политехнический университет
rafaelmnatsakanyan@yandex.ru

Введение

Алгоритмы машинного обучения применяются там, где строгий алгоритмический подход решения задачи либо очень труден и дорог, либо пока не придуман. Одной из таких задач стало предугадывание высоковероятных случаев повторного вызова скорой помощи. Задача имеет высокую потенциальную социальную полезность, так как позволит снизить издержки на повторные выезды к одному и тому же больному, настаивая на превентивной госпитализации в предсказанных случаях.

Задача машинного обучения

В данном проекте требуется решить задачу классификации, распределив вызовы по двум категориям – предшествующие повторному вызову, или нет. Для данных целей можно применить, как нелинейные классификаторы, так и нейронные сети различных топологий. Однако при любом подходе для обучения классификатора необходимы хорошо представимые математически, чистые и хорошо структурированные данные, которые несут в себе много смысловой нагрузки [1].

Обрабатываемые данные

Медицинские данные разбиты на три связанные таблицы. Первая таблица (270000 записей) содержит информацию обо всех вызовах скорой помощи в Томске за полуторагодичный период. Данные заполняются диспетчером и врачом, который вел прием. Из-за специфики работы СМП данные очень лаконичны и содержат только самое необходимое для коммуникации диспетчера и врача, а именно:

- адрес вызова;
- описание пациента и его проблемы;
- комментарий от диспетчера;
- временные метки каждого этапа обработки вызова;
- результат вызова с поставленным диагнозом.

Вторые две таблицы (около 1000000 записей), связанные с первой при помощи уникальных, персонифицированных ID пациентов, описывают различные врачебные приемы, каждая строка в которых содержит дату приема, профиль врача, анамнез, результат осмотра и рекомендацию специалиста.

Фактически задача подготовки данных разбилась на две подзадачи. Подготовить таблицу вызовов СМП и подготовить таблицы медицинских осмотров.

Таблица вызовов СМП

Полученные от медучреждений данные являлись очень зашумленными и фактически малопригодными для обработки в сыром виде. Фундаментальная проблема в том, что данные в информационную систему люди вносят вручную, что приводит к огромному количеству опечаток, пропусков в данных и орфографических ошибок. Слабо унифицированные данные, в которых название одной улицы может быть записано десятком разных форм, были подвергнуты множеством различных методов улучшения текстовой информации [3].

Для первого приближения необходимо найти все вызовы к каждому уникальному человеку. Для этого необходимо сгруппировать вызовы самостоятельно, потому что ID пациентов оказался не заполнен для многих вызовов. Для данного приближения было предположено, что объединение таких признаков, как “пол”, “дата рождения”, “улица” и “номер дома” будет достаточно для указания на конкретного человека.

Пол и дата рождения были проверены на грубые ошибки методом подсчета уникальных обозначений для пола и построением гистограмм для даты рождения [1]. Таким образом, все найденные варианты записи пола были каталогизированы и заменены на ‘Ж’ и ‘М’, для женского и мужского пола соответственно. Анализ гистограммы, в свою очередь, показал сильно выбивающиеся элементы в ряду, которые не поддавались логическому объяснению. Каждый такой случай был рассмотрен отдельно, как например, человек с указанным 2070 годом рождения. Все эти данные были исправлены и впоследствии представляли собой распределения, объяснимые знаниями из реальной жизни. Так, например, женщин, вызывающих скорую помощь больше на 10%, а на возрастной гистограмме заметна просадка количества вызовов от пожилых людей, которые жили во время Великой отечественной войны.

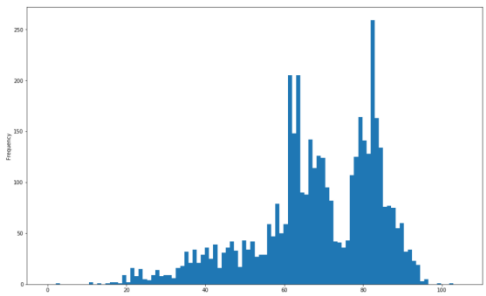


Рис. 1. Отчищенная возрастная гистограмма людей с кардиологическими проблемами

Поля улица и номер дома были унифицированы еще хуже. Например, слово микрорайон было обнаружено в 8 разных вариантах написания. Для того, чтобы уменьшить количество различных названий семантически одинаковых объектов, были составлены списки различных написаний одних и тех же объектов. Программно каждый найденный объект из этого списка заменялся одним выбранным наименованием. Из-за невозможности проверить все данные вручную, проблема с разными окончаниями улиц была решена при помощи укорачивания всех слов до определенной длины. Таким образом определить, является ли конкретный вызов – вызовом с одного адреса, стало возможно в автоматизированном режиме.

Проблема с домами имела иную природу и специфику. Из-за отсутствия договоренностей в нумерации домов, номер дома часто не просто число, а указание номера и строения, номера и дроби и так далее.

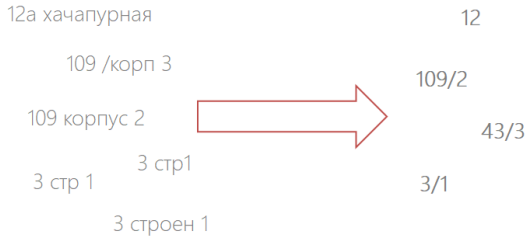
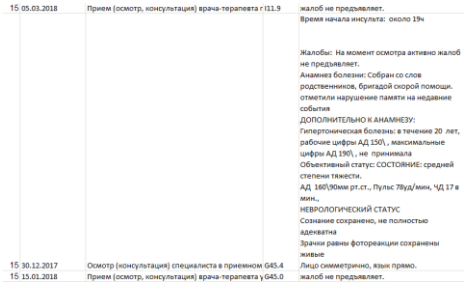


Рис. 2. Преобразование номеров домов.

Проблема, ухудшается человеческими ошибками и опечатками. Было решено удалить все буквенные символы, оставив лишь числа, разделив их слешем, если они располагались не подряд. Данное решение позволило уменьшить количество различных вариантов больше чем на 60%.

Таблица приемов СМП

Данная таблица содержит запись врачебного приема в свободной форме.



15	05.03.2018	Прием (осмотр, консультация) врача-терапевта г I11.9	жалоб не предъявляет. Время начала инсульта: около 19ч Жалобы: На момент осмотра активно жалоб не предъявляет. Анамнез болезни: Собран со слов родственников, бригадой скорой помощи. отметили нарушение памяти на недавние события ДОПОЛНИТЕЛЬНО К АНАМНЕЗУ: Гипертоническая болезнь: в течение 20 лет, рабочие цифры АД 150\ , максимальные цифры АД 190\ , не принимала Объективный статус: СОСТОЯНИЕ: средней степени тяжести. АД 160\90мм рт.ст., Пульс 78уд/мин, ЧД 17 в мин., НЕВРОЛОГИЧЕСКИЙ СТАТУС Сознание сохранено, не полностью адекватна Зрачки равны фотореакции сохранены живые
15	30.12.2017	Осмотр (консультация) специалиста в приемном G45.4	Лицо симметрично, язык прямо.
15	15.01.2018	Прием (осмотр, консультация) врача-терапевта у G45.0	жалоб не предъявляет.

Рис. 3. Наполнение таблицы приемов

Большие массивы текста невозможно использовать напрямую для машинного обучения, поэтому существует несколько подходов их преобразования. Было решено использовать частотный анализ слов и word embedding (превращение слов в векторы, где семантически близкие слова являются математически близкими векторами).

Используемые технологии

Для анализа данных и последующего обучения был использован стандартный набор инструментов для анализа и машинного обучения:

- Язык программирования Python.
- Библиотеки для машинного обучения Tensorflow, Keras и scikit-learn [2].
- Библиотеки для обработки данных: Pandas, matplotlib и другие [2].

Заключение

Классификаторы, оперирующие лишь данными скорой помощи, могут предсказать до **23,5%** повторных вызовов. До обработки данных, успешное предсказание находилось в пределах статистической погрешности.

Подготовка данных позволила значительно улучшить точность предсказания классификаторов, однако все еще ведется работа по интегрированию информации о врачебных приемах в систему, для улучшения точности предсказывания повторных вызовов.

Список использованных источников

1. Jeffrey Stanton, An introduction to datascience – Syracuse University, 2013.
2. Рашка С. Python и машинное обучение / пер. с англ. А. В. Логунова. М.: ДМК Пресс, 2017. – 418 с.: ил.
3. Davy Cielen, Arno D.B. Meysman, Mohamed Ali Introducing Data Science – Manning Publications Co., 2016.