# Обнаружение острых состояний и хронических заболеваний в тексте анамнеза

В медицинских картах, в том числе медицинских электронных картах (МЭК), информацию о хронических заболеваниях и острых состояниях в прошлом записывают в анамнез. Однако, часто текст анамнеза не структурирован, повествование ведётся на естественном языке с большим числом аббревиатур и специальных медицинских терминов. Для прогнозирования осложнений и острых состояний, постановки диагнозов, назначения некоторых препаратов важно знать об острых состояниях пациента в прошлом. С помощью методов машинного обучения и интеллектуального анализа текстов можно обучить модель для автоматического обнаружения упоминания в тексте болезней или острых состояний.

Для пациентов с сердечно-сосудистыми заболеваниями высок риск острого инфаркта миокарда (ОИМ) и острого нарушения мозгового кровообращения (ОНМК) – инсульта. Наличие таких состояний у пациента в прошлом может повлиять на выбор тактики лечения, поэтому важно всегда знать о таких состояниях. Также, для построения любых предсказательных моделей необходима информация об ОИМ и ОНМК. Поэтому была построена модель для автоматического анализа текста анамнеза болезни, которая способна обнаружить упоминание об остром состоянии в прошлом, отрицание пациентом таких состояний или отсутствие какого-либо упоминания о таких состояниях.

Для обучения модели был использован набор деперсонализированных данных пациентов с острым коронарным синдромом (ОКС), поступивших в ФГБУ «НМИЦ им. В. А. Алмазова» Минздрава России с 2010 по 2015 год. В 2682 МЭК есть анамнез болезни. 897 анамнезов были размечены вручную и с помощью ключевых фраз для дальнейшего обучения модели. Примеры ключевых фраз: «перенес инфаркт миокарда», «признаки ишемии миокарда», «выставлен диагноз не Q-ИМ» – указывают на наличие ОИМ, «инфаркт миокарда: отрицает», «ОИМ в анамнезе отрицает» - указывают на отсутствие ОИМ. Итак, в 701 анамнезе есть ОИМ, в 163 анамнезах пациент отрицает, и в 33 анамнезах информация об ОИМ отсутствует. В 1787 анамнезах построенная модель должна автоматически обнаружить упоминание об ОИМ.

Перед обучением модели текст анамнеза был обработан: приведен к нижнему регистру, удалены все небуквенные символы, нормализован – все слова приведены к начальной форме: существительные в именительный падеж, глаголы в форму инфинитива и т. д. Далее на основе этих слов была составлена модель теска Bag-of-words, которая каждому тексту ставит в соответствие вектор, размерность которого определяется числом уникальных слов и словосочетаний во всех текстах, а значение компонент вектора – наличием или отсутствием уникальных слова или фразы в каждом конкретном тексте. Далее к набору векторов применяется преобразование TF-IDF, которое позволяет увеличить вес тех слов, которые встречаются в большинстве документов, но редко (скорее всего, это термины), и уменьшить веса слов, которые встречаются часто и в каждом документе (скорее всего, это слова-пустышки – союзы, местоимения и т.п.) [1]. TF-IDF преобразование выполняется по следующим формулам:

где *t* – набор слов, *d* – один документ, *nt* – число вхождений конкретного слова t в документ, – общее число слов в документе, *|D|* – число документов в коллекции, – число документов из коллекции D, в которых встречается конкретное слово t. В контексте данной задачи документы – это анамнезы. Полученный набор векторов после преобразования делится на тренировочное, валидационное и тестовое множества.

Необходимо распознать три класса: «-1» – пациент отрицает ОИМ, «0» – информация об ОИМ отсутствует, «1» – есть ОИМ в анамнезе. Классы сильно неуравновешенные, поэтому перед обучением модели используется алгоритм SMOTE для генерации дополнительных объектов, чтобы увеличить меньшие по размеру классы [2]. Для решения задачи многоклассовой классификации используется бустинг-модель над решающими деревьями. С помощью перебора по сетке и кросс-валидации определяются значения внешних параметров модели. На тренировочном множестве модель показала точность 99%, на тестовом множестве – 93 %.

В Таблице 1 представлены слова и словосочетания, которые в наибольшей степени повлияли на определение класса классификатором. Так как примеров в пользу ОИМ больше всего, характерных слов и фраз алгоритм больше всего нашёл для наличия ОИМ. По значимым словам и фразам легко понять, как врачи чаще всего описывают острое состояние, какие есть сопутствующие патологии (ИБС), какие типы ОИМ бывают (передний, задний), и даже как лечат (стентирование и анальгин, трамадол).

Таблица 1 – 15 наиболее значимых слов и словосочетаний для определения ОИМ в тексте анамнеза и классы, на которые эти слова указывают

|  |  |  |  |
| --- | --- | --- | --- |
| **Слово или фраза** | **Значимость для классификатора** | **Класс** | **Комментарий** |
| оим | 0,052 | 1 | Аббревиатура: острый инфаркт миокарда |
| ибс оим | 0,042 | 1 | Аббревиатуры: ишемическая болезнь сердца. ИБС включает в себя стенокардию, инфаркт миокарда и атеросклеротический кардиосклероз. |
| представление ибс | 0,041 | 1 | Указывает на наличие ИБС |
| миокард отрицать | 0,041 | -1 | Пациент отрицает ОИМ |
| инфаркт | 0,039 | 1 |  |
| им | 0,036 | 1 | Аббревиатура: инфаркт миокарда |
| отрицать | 0,035 | -1 | Пациент отрицает наличие острого состояния |
| миокард ранее | 0,027 | 1 | Указывает на проблемы с миокардом ранее |
| нижний | 0,022 | 1 | Встречается в словосочетаниях: нижний ИМ и нижняя стенка левого желудочка – указывает на локализацию ОИМ |
| миокард | 0,018 | 1 | Указывает на проблемы с миокардом |
| ранее | 0,018 | 1 | Указывает на наличие патологий ранее |
| провести стентирование | 0,016 | 1 | Стентирование – это операция, которую часто проводят при ОИМ |
| ознакомить | 0,015 | 0 | Указывает на анамнезы следующего вида: «c и/б, анамнезом ознакомлен, дополнений нет.» – анамнез без истории болезни |
| передний | 0,014 | 1 | Встречается в словосочетаниях: передняя стенка, передняя локализация, передний ИМ – указывает на локализацию ОИМ |
| анальгин трамадол | 0,013 | 1 | Препараты, которые вводят вместе с тромботическими препаратами при ОИМ, чаще всего во время транспортировки до медицинского центра |

Аналогичный алгоритм был применен для поиска упоминания ОНМК и артериальной гипертензии (АГ). Наиболее значимые слова для ОНМК и АГ указаны в Таблице 2 и Таблице 3 соответственно. При этом точность классификации для ОНМК составляет 94 %, а для АГ 95 % на тестовом множестве.

Таблица 2 – 10 наиболее значимых слов и словосочетаний для определения ОНМК в тексте анамнеза и классы, на которые эти слова указывают

|  |  |  |
| --- | --- | --- |
| **Слово или фраза** | **Значимость для классификатора** | **Класс** |
| онмк | 0,111 | 1 |
| онмк отрицать | 0,068 | -1 |
| отрицать | 0,057 | -1 |
| инсульт | 0,051 | 1 |
| гемипарез | 0,023 | 1 |
| онмк по | 0,023 | 1 |
| отмечать | 0,021 | -1 |
| онмк не | 0,019 | -1 |
| отрицать онмк | 0,017 | -1 |
| без клиника | 0,016 | -1 |

Таблица 3 – 10 наиболее значимых слов и словосочетаний для определения АГ в тексте анамнеза и классы, на которые эти слова указывают

|  |  |  |  |
| --- | --- | --- | --- |
| **Слово или фраза** | **Значимость для классификатора** | **Класс** | |
| ад отрицать | 0,036 | -1 |
| без анамнез | 0,036 | -1 |
| аг нет | 0,029 | -1 |
| гб отрицать | 0,027 | -1 |
| длительный | 0,024 | 1 |
| ад | 0,023 | 1 |
| не повышаться | 0,021 | -1 |
| отрицать | 0,021 | -1 |
| аг отрицать | 0,02 | -1 |
| гипертензия | 0,017 | 1 |

Предложенный подход для определения острых состояний и хронических заболеваний в тексте позволяет быстро и с высокой точностью находить нужные заболевания. Точность при поиске ОИМ, ОНМК и АГ колебалась в промежутке от 91 % до 97 % на тестовых множествах. При этом время выполнения всех преобразований текста, обучение модели и выполнение предсказаний занимает около 1,5 минут. Данный подход можно применять для текстов на естественном языке, если нет возможности, найти необходимое состояние с помощью ключевых слов, так как может упоминаться не только наличие этого состояния, но и его отсутствие. Однако, для обучения модели необходимо разметить вручную достаточное количество текстов.