

0.1. Найдено А.В. Классификация молочных желез по классам температурных аномалий при диагностике злокачественных образований на основе текстовой информации

В основе заботы о собственном здоровье лежат регулярные медицинские обследования. При этом важно знать о предрасположенности организма к различным заболеваниям, поскольку на эти аспекты необходимо обратить внимание в первую очередь. Особенно, если речь идет о таком серьезном заболевании, как рак молочной железы. Данные, с которыми проводится работа в представленной статье, были получены от врачей словенской клиники, обследующих пациентов на предмет наличия рака молочных желез. Обследование выполнялось методами микроволновой радиотермометрии, после чего специалистами для каждой молочной железы была определена группа риска, указывающая уровень температурных аномалий. Подобная оценка основывается на реальной картине здоровья пациента, будем считать ее достоверной. Также данные содержали такую информацию, как анамнез, примечания врачей и описание результатов осмотра пациентов. Такие данными будем считать предварительными, содержащими предположения. Цель статьи — понять, есть ли в контексте имеющихся данных согласованность между этими двумя типами информации: достоверной и предварительной.

Пусть есть n молочных желез, где $n = 5782$. Для каждой i -й молочной железы имеем набор текстовых характеристик на словенском языке d_{1i}, d_{2i}, d_{3i} , $i = \overline{1, n}$, где d_1 — анамнез, d_2 — примечания врача, d_3 — данные осмотра. Объединив фрагменты текста, получим $d = \{d_i = \{d_{1i}, d_{2i}, d_{3i}\}\}$, где d_i — текстовое описание i -й молочной железы. Для набора текстовых описаний d выполнялась предобработка: из всего текста оставлены только слова, удалены стоп-слова и выполнена лемматизация. После чего с помощью предобученной на различных текстовых данных модели [1] получили векторное представление [2] $V = \{v_{ij}, i = \overline{1, n}, j = \overline{1, m}\}$, где m — заданный объем получаемого векторного представления. Был построен классификатор $h : X \rightarrow Y$, где $X = V$, $Y = \{0, 1, 2, 3, 4, 5\}$ — множество классов температурных аномалий, полученных врачами-специалистами словенской клиники с помощью метода микроволновой радиотермометрии, где 0 — температурные аномалии отсутствуют или в пределах нормы, 5 — температурные аномалии характерны для рака молочной железы.

В качестве метода машинного обучения был использован случайный лес. Для оценивания качества классификации применялась метрика out-of-bag, которая на тестовой выборке показала результат в районе 0.5.

В итоге имеем, что полученная модель машинно-

го обучения имеет высокую погрешность. Планируется уточнить модель, расширив признаковое пространство для повышения информативности. В задаче определения класса температурных аномалий текстовые данные можно рассматривать, как дополнительное обоснование полученных результатов.

Научный руководитель — д.т.н. Гермашев И. В.

Список литературы

- [1] EXPLOSION Slovenian pipeline optimized for CPU: sl_core_news_lg. [Электронный ресурс]. URL: https://spacy.io/models/sl#sl_core_news_lg (дата обращения 13.09.2024).
- [2] BOJANOWSKI P., GRAVE E., JOULIN A., MIKOLOV T. Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. 2017. Vol. 5.. P. 135–146.