УДК 004.912

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ НЕСТРУКТУРИРОВАННОГО МЕДИЦИНСКОГО ТЕКСТА

Ленивцева Ю.Д. (Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к.т.н., Копаница Г.Д.
(Университет ИТМО, г. Санкт-Петербург)

В данной работе представлен сравнительный анализ методов машинного обучения в задаче классификации неструктурированного медицинского текста на примере аллергологических анамнезов. Данный подход является основой для разработки метода извлечения данных об аллергии из медицинских текстов.

Введение. Актуальность структурирования медицинских данных в России обуславливается потребностью связать большое количество разнородных агентов и источников медицинских данных, распределенных территориально. В связи с этим остро стоит проблема интеграции и обмена медицинскими данными. Обмен данными между медицинскими организациями представляет сложность в силу наличия на российском рынке большого количества перекрывающихся форматов хранения данных и необходимости их синхронизации. На сегодняшний день существуют такие международные стандарты обмена данными как openEHR, ISO13606, FHIR и терминологические стандарты SNOMED CT и LOINC, обладающие высоким потенциалом обмена данными. Большинство провайдеров медицинских услуг готовы работать с международными стандартами, в частности с FHIR. Ресурсы FHIR структурируют полезные медицинские данные, большая часть из которых хранится в неструктурированном виде в электронных медицинских картах.

Основная часть. Цель работы: сравнить методы машинного обучения для классификации неструктурированных медицинских записей на примере аллергологических анамнезов. Данные для исследования были предоставлены ФГБУ «НМИЦ им. В. А. Алмазова». В результате поиска с помощью регулярных выражений по корням слов «аллергия» и «непереносимость» было выделено более 269 тысяч записей, которые могут содержать информацию об аллергиях и непереносимостях. Для корректного извлечения аллергенов и реакций 11670 медицинских записей были вручную размечены на 4 класса: AL – записи, содержащие информацию об аллергене или группе аллергенов, а также упоминание о том, что аллергия у пациента присутствует; R – записи, содержащие информацию о реакциях, при этом аллерген может быть не уточнен; NN – записи, отрицающие наличие аллергии; N – записи, не имеющие отношение к аллергиям. Для классификации были выбраны следующие модели: логистическая регрессия, метод k ближайших соседей, наивный байесовский классификатор, линейный метод опорных векторов, а также два ансамблевых классификатора на основе логистической регрессии и линейного метода опорных векторов. В качестве метрик качества классификации были выбраны точность, полнота и F-мера. Лучшие результаты показал ансамблевый классификатор на основе линейного метода опорных векторов. Также в работе были проанализированы слова, которые оказывают наибольшее влияние на результаты классификации для каждого класса. Для визуализации классов был использован метод t-SNE.

Выводы. В дальнейшем полученные результаты классификации будут использованы для разработки метода извлечения аллергенов и реакций из медицинских записей. Интеллектуальные методы структурирования медицинских текстов позволят повысить качество предсказательного моделирования и обеспечить интероперабельность медицинских систем.

Ленивцева Ю.Д. (автор) Подпись
Копаница Г.Д. (научный руководитель) Подпись