Применение методов обработки естественного языка в задачах классификации радиологических отчётов

А.А. Слуднова[1], В.В. Шутько[1], А.В. Гайдель[1,2], А.В. Никоноров[1,2]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086
[2]Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 334001

Аннотация. В работе представлены результаты исследований по использованию методов обработки естественного языка в задаче классификации медицинских текстов. В качестве исходных данных использовался набор радиологических отчётов. Задача классификации сводилась к разделению отчётов с различными диагнозами (гидроцефалия, перелом, киста или отсутствие патологий). Используемые алгоритмы классификации (SGD и другие) сравнивались с поиском по подстроке, для каждого способа рассчитывались полнота и точность прогноза. На исследуемом ограниченном наборе данных наилучшие результаты классификации были достигнуты с использованием решающего дерева на наборе весов, полученных с помощью TF-IDF.

1. Введение

В настоящее время существует большое количество медицинских отчетов, записанных в свободном формате. Обработка таких медицинских отчетов и их классификация посредством применения методов обработки естественного языка и машинного обучения позволит получить структурированные выходные данные, которые потенциально могут пригодиться при контроле поставленных диагнозов [1], отслеживании состояния пациента, выявления невидимых для диагноста заболеваний [2], повышения качества лечения путем уточнения диагноза [3], а также использовать их в качестве замены первоначальных отчетов, написанных медицинскими специалистам в свободном формате, с целью повышения удобства хранения и доступа к ним для дальнейшей обработки.

2. Исходные данные

Данные были предоставлены Самарским государственным медицинским университетом. Набор данных содержал более 900 dicom-директорий с исследованиями различных органов. Для сортировки этих исследований был разработан алгоритм, проходящийся по всему набору данных и вычленяющий метаданные. В результате было получено, что среди предоставленных данных большую часть составляют исследования головного мозга и костей черепа (601 директория). Затем было произведено извлечение радиологических отчётов из каждой директории и составлена итоговая таблица с сопоставлением директории и поставленного диагноза, с которой затем сравнивались результаты классификации.

Все извлечённые тексты были предварительно обработаны: производилась токенизация, приведение слов к нижнему регистру и лемматизация, также было произведено удаление стоп-слов.

3. Классификация отчётов

В отобранном наборе отчётов встречались следующие диагнозы: гидроцефалия, переломы, гематомы, опухоли, различные кистозные образования, черепно-мозговые травмы и другие повреждения, а также отсутствие патологий.

В качестве диагнозов для классификации были выбраны следующие: без патологий, гидроцефалия, перелом, киста. Это решение было обусловлено тем, что перечисленные диагнозы составляли наибольшую долю из выборки.

3.1. Поиск по подстроке

В первую очередь была предпринята попытка осуществить классификацию отчётов на основании поиска по подстроке. Были подобраны шаблоны строк, соответствующих определённым диагнозам, после чего эти шаблоны искались в тексте отчётов и при совпадении отчёт относился в группу, соответствующей диагнозу.

При таком подходе удалось точно определить все отчёты, в которых встречались гидроцефалия, кисты и переломы, так как соответствующие им шаблоны были очень просты (в качестве шаблона можно было использовать первые несколько букв слов). Гораздо более трудной задачей было составление шаблонов для поиска отчётов, не содержащих патологий, так как в таких отчётах встречались различные формулировки: «патологии отсутствуют», «патологий нет», «без патологий», «патологии не выявлены» и другие.

Несмотря на то, что поиск по подстроке для определённых дефектов показал наивысшую точность для исследуемого набора данных, этот способ является далеко не самым оптимальным, потому что качество классификации сильно зависит от исходных данных, наличия в них опечаток, а также от наличия в данных различающихся диагнозов, подходящих под один шаблон, а само составление шаблонов может быть весьма трудоёмким процессом.

3.2. Классификация с использованием TF-IDF

TD-IDF (term frequency – inverse document frequency) – статистическая мера, которая используется для оценки важности слов в контексте отдельных документов, являющихся частью набора документов. Частота слова (TF) является отношением числа вхождений определённого слова к общему числу слов в документе и определяется формулой (1), где n_tявляется числом вхождения слова в конкретный документ, а в знаменателе находится общее число слов в данном документе.

$$tf(t,d) = \frac{n_t}{\sum_k n_k} \quad (1)$$

Обратная частота документа (IDF) — величина, обратная к частоте, с которой некоторое слово встречается в документах из набора. Эта величина задаётся формулой (2), где $|D|$– число документов в коллекции, а $|\{d_i \in D \mid t \in d_i\}|$– число документов из коллекции D, в которых встречается слово t.

$$idf(t,D) = \frac{|D|}{|\{d_i \in D | t \in d_i\}|} \quad (2)$$

Итоговая мера TF-IDF является произведением формул (1) и (2). При подобном подходе наибольший вес получают слова, которые редко встречаются в общем наборе документов, но при этом имеют высокую частоту в пределах конкретного документа [4].

Веса слов для исходного набора отчётов были получены с помощью TfidfVectorizer библиотеки sklearn. Затем полученная выборка делилась на обучающую и тестовую в соотношении 3:1. Для классификации были использованы несколько алгоритмов: стохастический градиентный спуск (SGD), решающее дерево (DT), случайный лес (RF) и K ближайших соседей (KNN).

Стохастический градиентный спуск, для которого использовалась L1-регуляризация, а в качестве функции потерь была выбрана функция потерь Хьюбера (эти параметры были получены как оптимальные при поиске по сетке из всех параметров), показал общую взвешенную точность 0.91 и полноту 0.90, при этом общие потери полноты были связаны с классификацией отчётов, где встречались кистозные образования (для таких отчётов точность составила 1, а полнота только 0.33).

Для решающего дерева с оптимальными параметрами на этих же данных взвешенная точность составила 0.92, а полнота 0.92. Случайный лес показал более низкое качество классификации со взвешенной точность. 0.79 и полнотой 0.85.

Классификация по методу K ближайших соседей показала результаты 0.80 по взвешенной точности и 0.82 по полноте.

Для большей наглядности все полученные с помощью подхода TF-IDF результаты приведены в таблице 1.

Таблица 1. Результаты классификации.

Классификатор	Взвешенная точность	Взвешенная полнота
SGD	0.91	0.90
DT	0.93	0.92
RF	0.79	0.85
KNN	0.80	0.82

3.3. Классификация с использованием Word2Vec

Модель Word2Vec основана на гипотезе локальности (то есть слова, которые встречаются в одинаковых окружениях, имеют близкие значения) [5]. Подобный подход, в отличии от предыдущего, только выигрывает от больших объёмов данных. Word2Vec предсказывает вероятность слова по его окружению (контексту) согласно формуле (3), где w_o обозначает вектор целевого слова, w_c – вектор контекста, вычисленный из векторов других слов, окружающих целевое слово, а $s(w_1, w_2)$ – функцию, сопоставляющую двум векторам одно число (например, косинусное расстояние).

$$P(w_o|w_c) = \frac{e^{s(w_o,w_c)}}{\sum_{w_i \in V} e^{s(w_i,w_c)}} \quad (3)$$

Для качественного обучения модели Word2Vec количество данных в исходной выборке было недостаточным, поэтому с этим подходом были получены наихудшие результаты (средняя взвешенная точность 0.48 и средняя взвешенная полнота 0.69 для классификации с использованием логистической регрессии). В дальнейшем имеет смысл использовать предварительно обученные на медицинских данных сторонние модели.

4. Заключение

В рамках данной работы была предпринята попытка классификации радиологических отчётов, содержащих информацию об исследованиях головного мозга и костей черепа. Были применены и сравнены три различных подхода: поиск по подстроке, расчёт весов слов с помощью td-idf и применение Word2Vec.

Несмотря на то, что на предоставленных данных наилучшие результаты при поиске отчётов с патологиями были достигнуты с использованием поиска по подстроке, подобный подход не является общим и качество результатов, полученных с его применением, сильно зависит от структуры исходных данных и способов написания.

Подход с использованием TF-IDF также показал хорошие результаты, при этом наилучшие значения точности и полноты были получены с использованием решающего дерева.

Подход с использованием Word2Vec показал наихудшие результаты, однако это не является показательным, так как исходная выборка была невелика, а Word2Vec является подходом,

который выигрывает при большом разнообразии данных. Для небольших наборов данных целесообразно в дальнейшем использовать предварительно обученные модели.

5. Литература

[1] Wang, Ya. Natural language processing of radiology reports for identification of skeletal site-specific fractures / Ya. Wang, S. Mehrabi, S. Sohn, E.J. Atkinson, Sh. Amin, H. Liu // DMC Medical Informatics and Decision Making. – 2019. – Vol. 19(3).

[2] Fu, S. Natural language processing for the identification of silent brain infarcts from neuroimaging reports / S. Fu, L.Y. Leung, Ya. Wang, A.-O. Raulli, D.F. Kallmes, K.A. Kinsman, K.B. Nelson, M.S. Clark, P.H. Luetmer, P.R. Kingsbury, D.M. Kent, H. Liu // JMIR Medical Informatics. – 2019. – Vol. 7(2).

[3] Kaggal, V.C. Toward a learning health-care system – knowledge delivery at the point of care empowered by Big Data and NLP / V.C. Kaggal, R. K. Elayavilli, S.M. Mehrabi, J.J. Pankratz, S. Sohn, Ya. Wang, D. Li, M.M. Rastegar, S.P. Murphy, J.L. Ross, R. Chaundry, J.D. Buntrock, H. Liu // Innovations in Clinical Informatics. – 2016. – Vol. 8(1). – P. 13-22.

[4] Jones, K.S. A statistical interpretation of term specificity and its application in retrieval / K.S. Jones // Journal of Documentation. – 1972. – Vol. 28(1). – P.11-21.

[5] Mikolov, T. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, G.S. Corrado, K. Chen, J. Dean // International Conference on Learning Representations, 2013.

Natural language processing for radiological reports classification

A.A. Sludnova[1], V.V. Shutko[1], A.V. Gaidel[1,2], A.V. Nikonorov[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

Abstract. The paper presents the results of studies on the use of natural language processing methods in the classification of medical texts. A set of radiological reports was used as initial data. The classification task was reduced to the separation of reports with various diagnoses (hydrocephalus, fracture, cyst or no pathologies). The classification algorithms (SGD and others) were compared with a substring search, the accuracy and recall were calculated. On the limited data set, the best classification results were achieved by using a decision tree with TF-IDF weights.