Простой общий алгоритм объяснения диагноза на выходе интеллектуальной системы диагностики в терминах примитивов естественного языка

Л. В. Уткин[1], А. А. Мелдо[2], М. С. Ковалев[3], Э. М. Касимов[4]
Высшая школа прикладной математики и вычислительной физики
Санкт-Петербургский политехнический университет Петра Великого
Санкт-Петербург, Россия
[1]lev.utkin@gmail.com, [2]anna.meldo@yandex.ru, [3]maxkovalev03@gmail.com, [4]kasimov.ernest@gmail.com

***Аннотация*. Предложен простой алгоритм объяснения решений интеллектуальных систем компьютерной диагностики рака. Алгоритм дает объяснения заболеваний в виде специальных предложений на естественном языке. Он состоит из двух частей. Первая часть реализована с помощью стандартной локальной объяснительной модели, например, известного метода LIME. Эта часть предназначена для выбора значимых признаков из сегментированного и детектированного подозрительного объекта или его представления признаков малой размерности. Вторая часть представляет собой набор простых классификаторов, которые преобразуют выбранные значимые признаки в один из классов, соответствующих простым фразам. Объясняющие предложения на естественном языке состоят из простых фраз (примитивов), так что каждое подмножество фраз описывает одну особенность подозрительного объекта, например, форму, структуру, включения, контуры. Примитивы рассматриваются как классы для множества классификаторов. Предложенный алгоритм является общим и может быть применен для реализации подсистемы объяснения различных заболеваний.**

Ключевые слова: объяснительный интеллект; интеллектуальные системы диагностики; машинное обучение; рак; классификация

I. Введение

Многие интеллектуальные системы диагностики (ИСД), основанные на использовании искусственного интеллекта (ИИ), были разработаны для того, чтобы помочь рентгенологам автоматически обнаруживать подозрительные объекты, например легочные новообразования в легких, на ранних стадиях [1–5]. Одним из наиболее чувствительных методов диагностики является компьютерная томография (КТ). У него есть много преимуществ, включая быстрое получение снимков и низкая стоимость использования [1]. Поэтому именно он очень часто используется в практике исследований онкологических заболеваний. Отсюда многие системы ИСД используют компьютерную томографию в качестве основы для анализа рака.

Технически, многие ИСД состоят из двух частей. Первая часть – это интеллектуальная подсистема обнаружения, сконцентрированная на сегментации и обнаружении подозрительных объектов и называемая системой или подсистемой CADe. Вторая часть – это интеллектуальная подсистема диагностики, которая фокусируется на классификации подозрительных объектов или и называется системой или подсистемой CADx.

Системы CADe стремятся к точному сегментированию объектов, представляющих интерес, и их отделению от других органов и тканей, например, от сосудов, жира, мышц и т.д. В ИСД рака легкого подсистема CADe направлена на отделении паренхимы легкого и новообразований в легком от других тканей. Важным элементом большинства систем CADe является процедура предварительной обработки, которая преобразует данные или значения пикселей в каждом изображении КТ в значения единиц Хаунсфилда (HU), которые рассматриваются как стандартная количественная шкала радиоплотности. Выходом системы CADe является множество сегментированных подозрительных объектов или небольших областей (bounding boxes), содержащих подозрительные объекты.

Система CADx направлена на решение нескольких задач. Прежде всего, она классифицирует каждый подозрительный объект, обнаруженный и сегментированный системой CADe, как злокачественный или доброкачественный. Для решения этой задачи многие системы CADx решают вторую задачу, которая реализует представление признаков подозрительных объектов малой размерности. Классификатор в системе CADx, использующий это представление, становится намного проще, поскольку он обучается на числовых векторах небольшого размера вместо трехмерных изображений КТ. Следует отметить, что существует множество подходов для упрощенного представления изображений КТ или

Исследование выполнено за счет гранта Российского научного фонда (проект № 18-11-00078)

объектов. Эти подходы могут быть рассмотрены в рамках радиомики, которая относится к интеллектуальному извлечению данных из радиологических изображений и предоставляет уникальный потенциал для решения медицинских задач с помощью алгоритмов машинного обучения [6]. Интересное представление сегментированных узелков легкого малой размерности было предложено в [7], где каждый узелок представлен набором гистограмм (гистограммы длин хорд, характеризующие форму новообразования, гистограммы радиоплотностей внутри и вокруг объекта). Гистограммы являются примерами представления сложных трехмерных объектов векторами признаков малой размерности, которые могут улучшить классификационные свойства, уменьшить количество ложноположительных случаев и упростить интерпретацию результатов.

Третья задача, которую должна решить система CADx, заключается в объяснении прогнозируемого диагноза. Объяснение диагнозов, выдаваемых ИСД, является важным условием эффективной эксплуатации ИСД, поскольку врач должен иметь возможность понять, как и почему было принято то или иное решение, предоставленное алгоритмом машинного обучения [8,9]. Под объяснением диагноза мы будем понимать причины, по которым соответствующая ИСД ставит диагноз, какие признаки подозрительного объекта ответственны за поставленный диагноз, какие внутренние и внешние особенности объекта соответствуют диагнозу. Эта задача сложна, потому что многие модели машинного обучения, используемые в ИСД, не объяснимы, они являются черными ящиками и не объясняют свои решения. Явными примерами таких моделей являются глубокие нейронные сети, которые являются основными компонентами CADe, а также систем CADx. Поэтому третья задача актуальна, и ее решение должно дополнять любые ИСД.

Для интерпретации и объяснения результатов классификации было разработано множество методов объяснения, например, известный метод локальной интерпретации LIME [10], метод SHAP [11]. Существует также важное семейство методов, основанных на гипотетических объяснениях, которые пытаются объяснить, что делать, чтобы достичь желаемого результата (другого диагноза), путем поиска оптимальных изменений признаков объясняемого входного примера, так что результирующая пример данных, называемая гипотетической, имеет иной диагноз, чем исходный пример [12, 13]. Многие из вышеупомянутых методов объяснения, начиная с LIME, основаны на методах возмущения [14, 15]. Эти методы предполагают, что вклад признака можно определить, оценив, как изменяется диагноз при изменении признака. Одним из преимуществ методов возмущений является то, что они могут применяться к модели черного ящика без необходимости доступа к внутренней структуре модели. Возможным недостатком метода возмущений является его вычислительная сложность, когда возмущенные входные примеры имеют высокую размерность. Это небольшая часть всех имеющихся подходов [16] в объяснительном ИИ. Большинство методов отбирают значимые признаки объекта, которые отвечают на вопрос, почему классификатор предсказал определенный диагноз исследуемого пациента. Фактически, эти значимые признаки можно рассматривать как объяснение результатов, полученных с помощью модели черного ящика (система CAD), когда ее входным данным является пациент, проходящий тестирование.

Подход, который отбирает значимые признаки объекта, интересен и полезен для объяснения во многих приложениях. Однако, если мы имеем дело с медицинскими изображениями и выбираем какую-либо область в сегментированном изображении 3D объектов КТ, это не поможет полностью объяснить диагноз. Представление признаков вектором малой размерности, например представление при помощи гистограмм, делает проблему объяснения еще более сложной с точки зрения понимания, потому что врач не знает, что означает каждая часть гистограммы или представления других признаков. Как указано в работе [17], в идеале, объяснение должно быть удобным для пользователя и представлено на естественном языке, например: «Это плоскоклеточный рак, потому что мы наблюдаем шаровидную форму, спикулы и воздушную полость».

Чтобы реализовать описанный выше подход для объяснения, было предложено несколько эффективных моделей глубокого обучения, которые генерируют описания изображений на естественном языке с помощью LSTM [18, 19]. Тем не менее, как указано в [17], на самом деле явные причины диагноза могут не отображаться на снимках. В результате авторы работы [17] предлагают модель на основе фраз для того, чтобы обойти это ограничение. Основная сложность рассматриваемых подходов состоит в том, что они требуют много обучающих примеров, чтобы избежать переобучения. К сожалению, большинство наборов данных, используемых для обучения медицинских систем ИСД, характеризуется небольшим количеством примеров. Эта проблема делает вышеуказанные подходы неприменимыми во многих медицинских приложениях.

Чтобы справиться с этой проблемой, мы предлагаем обобщенный простой алгоритм, который дает объяснения диагноза в форме предложений на естественном языке. Первая идея, лежащая в основе алгоритма, заключается в том, что существует ограниченное количество простых фраз (примитивов), которые описывают некоторый диагноз. Эти фразы обычно описывают форму, структуру, включения, контуры и другие особенности подозрительного объекта. Более того, каждая такая особенность объекта имеет небольшое количество фраз для описания. Это подразумевает, что полное объяснение диагноза на естественном языке может состоять из отдельных примитивов, касающихся каждой особенности. Другая идея алгоритма заключается в создании и обучении простых классификаторов, которые классифицируют подозрительные объекты или их представление малой размерности на классы, соответствующие примитивам. Третья идея заключается в реализации алгоритма в виде двух частей. Первая часть (Explainer 1) представляет собой стандартную модель объяснения, например, LIME [10],

которая используется для выбора значимых признаков из объекта или его представления. Вторая часть (Explainer 2) представляет собой набор классификаторов, цель которых – соединить выбранные значимые признаки, выбранные Explainer 1, с предложениями на естественном языке.

II. Предлагаемый алгоритм объяснения

Обобщенная схема подсистемы классификации с алгоритмами объяснения (Explainer 1, Explainer 2) показана на рис. 1. Предполагается, что классификатор представляет собой обученный черный ящик, например, глубокую нейронную сеть, которая обеспечивает некоторую диагностику. Существует набор данных, содержащий сегментированные подозрительные объекты. Мы не рассматриваем систему CADe, потому что она не влияет на алгоритм объяснения. Мы используем ее выход только в виде сегментированных объектов. В соответствии со способом представления признаков каждый сегментированный объект сохраняется в форме представления признаков малой размерности. Следует отметить, что представление признаков малой размерности не является необходимым в этой схеме. Оно упрощает классификатор и Explainer 1, но эти части могут быть реализованы без использования этого представления.

Explainer 1 – это локальная модель объяснения, например LIME [10] или SHAP [11], которая использует объекты из обучающей выборки или их представление малой размерности, а также результаты классификации (диагностики) для отбора значимых признаков, соответствующих каждому сегментированному объекту. Кроме того, рассматривается подход, который не зависит от модели черного ящика. Это означает, что детали модели черного ящика неизвестны, за исключением ее входных и выходных данных. Объяснения получены путем подбора интерпретируемой модели (классификатора) локально вокруг каждого примера обучения.

LIME [10] является одним из простых и эффективных методов реализации Explainer 1. В соответствии с методом, предлагается аппроксимировать модель черного ящика, обозначенную как f, простой функцией g в окрестности интересующей точки $\mathbf{x}$, прогноз которой с помощью f необходимо объяснить при условии, что аппроксимирующая функция g принадлежит множеству объяснительных моделей, например, линейных моделей или деревьев решений. Чтобы построить функцию g в соответствии с LIME, генерируется новый набор данных, состоящий из возмущенных примеров, и предсказания, соответствующие возмущенным примерам, получают с помощью модели, требующей объяснения. Новым примерам присваиваются веса в соответствии с их близостью к точке $\mathbf{x}$ с использованием метрики расстояния, например, евклидова расстояния. Объясняющая локальная модель обучается на новых сгенерированных примерах путем решения задачи оптимизации, целевая функция которой оценивает, насколько объяснение близко к предсказанию модели черного ящика. Локальная линейная модель и является результатом LIME. Прогноз объясняется анализом коэффициентов локальной линейной модели.

Результатом объяснения для каждого объекта является множество значимых признаков (наибольшие коэффициенты локальной линейной модели). Количество значимых признаков для каждого сегментированного объекта ограничено значением K, которое можно рассматривать как параметр настройки модели. Если Explainer 1 предоставляет линейную комбинацию признаков с некоторыми коэффициентами, тогда признаки с наибольшими значениями коэффициентов выбираются для обработки Explainer 2. Следует отметить, что результат объяснения может быть представлен как простой «разреженный» вектор признаков так, что его ненулевые элементы соответствуют значимым признакам. Классификатор, а также Explainer 1 являются стандартными и зависят от ИСД и соответствующего представления признаков малой размерности.

Рассмотрим Explainer 2, который является ключевым компонентом алгоритма объяснения, создающего объяснения на естественном языке.

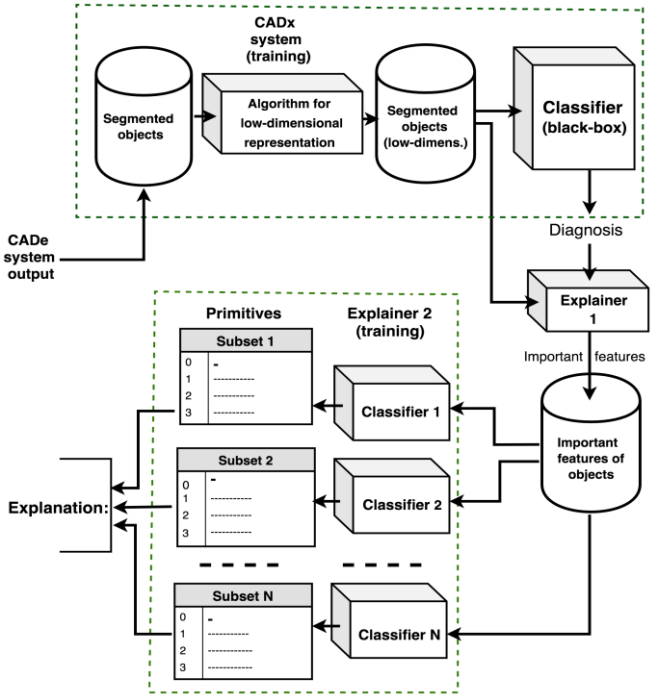

Рис. 1. Общая структура алгоритма объяснения

Разобьем набор простых фраз, описывающих особенности подозрительного объекта с точки зрения диагностики, на n непересекающихся подмножеств $s_1,\ldots,s_n$. Каждое подмножество содержит простые фразы (примитивы), относящиеся к одному понятию, например, структуре формы объекта. Подмножество, соответствующее структуре формы объекта, может, например, включать фразы: «форма сферическая», «форма треугольная» и т.д. Важно, чтобы только один элемент (фраза) можно выбрать из каждого подмножества для объяснения поставленного диагноза. Среди всех объектов или их объяснений при помощи Explainer 1 есть подмножество объектов, которое соответствует одной из

простых фраз из подмножества фраз, например, фразе «форма сферическая». Это подразумевает, что это подмножество объектов принадлежит одному классу под названием «форма сферическая». Другое подмножество объектов принадлежит другому классу, например, под названием «форма треугольная». Из этого следует, что каждую фразу можно рассматривать как класс или метку класса. Следовательно, можно построить классификатор, который будет определять классы всех объектов, соответствующих фразам из одного подмножества. Важно добавить случай, когда объект не принадлежит ни одному классу (фразе) из подмножества фраз. В этом случае мы введем пустой или фиктивный класс.

Таким же образом, мы можем построить и обучить N классификаторов для каждого подмножества так, чтобы входными данными всех классификаторов были исходные объекты или их значимые признаки, а выходом были бы классы, которые определяются одной из фраз из подмножества. Другими словами, каждый вектор значимых признаков классифицируется в один из классов (фраз) из подмножества, включая пустые классы. Если мы обозначим множество всех векторов значимых признаков как $\{h_1, \ldots, h_N\}$, то каждому вектору соответствует какое-то значение класса p_k. Классы обозначаются как $\{0,1,2, \ldots, c_k\}$, где c_k – количество фраз в подмножестве s_k. Пустой класс обозначается как 0. В итоге имеем N классификаторов как функции

$$f_k : \{h_1, \ldots, h_N\} \rightarrow \{0,1,2, \ldots, c_k\},\ k=1,..,N,$$

которые могут быть реализованы, например, с помощью нейронных сетей или случайных лесов, в зависимости от представления вектора значимых признаков.

Чтобы реализовать Explainer 2, каждый объект в обучающем множестве должен быть размечен набором примитивов, описывающих объект.

На этапе тестирования полученный вектор значимых признаков из Explainer 1, классифицируется с использованием всех классификаторов. Определенное количество фраз в соответствии с результатами классификации объединяются в предложение, которое можно рассматривать как объяснение. Если результат классификации классификатора равен 0, тогда предложение не содержит элементов из соответствующего подмножества. Кроме того, если известны распределения вероятностей классов для каждого классификатора, то можно ввести порог η для выбора значимых фраз. Порог – параметр настройки всей системы объяснения.

III. Заключение

В статье рассмотрен алгоритм, обеспечивающий объяснение на естественном языке. Он был разработан для объяснения произвольного диагноза рака. Алгоритм является общим, но может быть изменен при реализации для определенного заболевания. Каждый его шаг может иметь разные реализации, которые в конечном итоге определяют его эффективность и точность. Разработка и исследование различных модификаций алгоритма и его адаптация к определенному заболеванию являются важными направлениями для дальнейших исследований.

Список литературы

[1] Zhang G., Jiang S., Yang Z., Gong L., Ma X., Zhou Z., Bao C., and Liu Q. Automatic nodule detection for lung cancer in ct images: A review // Computers in Biology and Medicine. 2018. vol. 103, pp. 287-300.

[2] Afshar P., Mohammadi A., Plataniotis K.N., Oikonomou A., and Benali H. From hand-crafted to deep learning-based cancer radiomics: Challenges and opportunities // arXiv:1808.07954v1, Aug 2018.

[3] Cheplygina V., de Bruijne M., Pluim J.P.W. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis // arXiv:1804.06353, Apr 2018.

[4] Litjens G., Kooi T., Bejnordi B.E., Setio A.A.A., Ciompi F., Ghafoorian M., van der Laak J.A.W.M., van Ginneken B., Sanchez C.I. A survey on deep learning in medical image analysis // Medical Image Analysis. 2017. vol. 42, pp. 60-88.

[5] Zhang J., Xia Y., Cui H., Zhang Y. Pulmonary nodule detection in medical images: A survey // Biomedical Signal Processing and Control. 2018. vol. 43, pp. 138-147.

[6] Thawani R., McLane M., Beig N., Ghose S., Prasanna P., Velcheti V., Madabhushi A. Radiomics and radiogenomics in lung cancer: A review for the clinician // Lung Cancer. 2018. vol. 115, pp. 34-41.

[7] Meldo, A., Utkin, L. A new approach to differential lung diagnosis with ct scans based on the siamese neural network, in IOP Conference Series: Journal of Physics: Conference Series. 2019. Vol. 1236. pp. 012058-5.

[8] Holzinger A., Biemann C., Pattichis C., Kell D. What do we need to build explainable AI systems for the medical domain? // arXiv:1712.09923, 2017.

[9] Holzinger A., Langs G., Denk H., Zatloukal K., Muller H. Causability and explainability of artificial intelligence in medicine // WIREs Data Mining and Knowledge Discovery. 2019. vol. 9(4), p. e1312.

[10] Ribeiro M., Singh S., Guestrin C. Why should I trust you? Explaining the predictions of any classifier // arXiv:1602.04938v3, 2016.

[11] Lundberg S., Lee S.I. A unified approach to interpreting model predictions, in Advances in Neural Information Processing Systems. 2017. pp. 4765-4774.

[12] Wachter S., Mittelstadt B., Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GPDR // Harvard Journal of Law & Technology. 2017. vol. 31, pp. 841-887.

[13] Looveren A.V., Klaise J. Interpretable counterfactual explanations guided by prototypes // arXiv:1907.02584, 2019.

[14] Fong R., Vedaldi A. Explanations for attributing deep neural network predictions, in Explainable AI. Volume 11700 of LNCS. Springer, Cham 2019. 149-167.

[15] Vu M., Nguyen T., Phan N., Gera M.T. Evaluating explainers via perturbation // arXiv:1906.02032v1, 2019.

[16] Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D. A survey of methods for explaining black box models // ACM computing surveys. 2019. vol. 51, pp. 93.

[17] Hendricks L., Hu R., Darrell T., Akata Z. Grounding visual explanations, in Proceedings of the European Conference on Computer Vision (ECCV). 2018. pp. 264-279.

[18] Hendricks L., Hu R., Darrell T., Akata Z. Generating visual explanation, in Proceedings of the European Conference on Computer Vision (ECCV). 2016. pp. 3-19.

[19] Karpathy A., Fei-Fei L. Deep visual-semantic alignments for generating image descriptions // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. vol. 39, pp. 664-676.