Случайные леса и метод хорд для интеллектуальной диагностики рака легких

Л. В. Уткин[1], М. А. Рябинин[2]
Санкт-Петербургский политехнический университет Петра Великого
[1]lev.utkin@gmail.com, [2]mihail-ryabinin@yandex.ru

А. А. Мелдо
Санкт-Петербургский клинический научно-практический центр специализированных видов медицинской помощи (онкологический)
anna.meldo@yandex.ru

***Аннотация*. В работе представлена новая архитектура интеллектуальной системы диагностики рака легкого по результатам анализа изображений, полученных методом компьютерной томографии. Обнаружение и классификация новообразований осуществляется в три основных этапа. Первый этап – сегментация изображения с использованием фильтрации в соответствии с заданным интервалом значений денситометрической плотности по шкале Хаунсфилда и выявление всех патологических объёмных образований. Второй этап – формирование вектора признаков, характеризующих каждое новообразование, с использованием метода хорд. На этом этапе формируется гистограмма длин хорд как универсальное представление поверхности объекта классификации. Третий этап − классификация полученного вектора признаков с использованием случайных лесов в целях определения, является ли новообразование злокачественным. Результаты тестирования системы диагностики на общедоступных наборах данных показали высокую точность классификации.**

Ключевые слова: рак легкого; случайный лес; метод хорд; классификация; компьютерная томография; система диагностики

I. Введение

Согласно [2], опухоль в легком может быть определена как патологическое объёмное образование, имеющее примерно сферическую структуру. Критериями доброкачественности являются ровный, чёткий контур, отсутствие в структуре признаков некроза, наличие обызвествлений, отсутствие изменений в окружающей лёгочной ткани и плевре. Критерии злокачественности опухоли, напротив, определяются как совокупность признаков, характеризующих экспансивный инвазивный рост: неровный нечёткий контур образования, признаки некроза к структуре, наличие радиарных тяжей, как проявление местного лимфангита, тракция прилежащей плевры. Было разработано множество систем диагностирования онкологических заболеваний (СДО) для обеспечения успешного обнаружения опухолей легких и для более обоснованного принятия решения о начале лечения на ранней стадии заболевания. Многие СДО основаны на применении методов фильтрации для обнаружения новообразований в легких на основе серий сканов компьютерной томографии (КТ). Детальный обзор соответствующих методов обнаружения и реализаций СДО можно найти в работе [5]. Серьезной проблемой этих СДО является относительно большое количество ложноположительных результатов, когда различные элементы легких распознаются как злокачественные новообразования, в то время как они таковыми не являются.

Чтобы решить эту проблему и «интеллектуализировать» процесс обнаружения злокачественных образований использовались многочисленные подходы на основе «неглубокого» обучения [8,11]. Многие предлагаемые в последние годы СДО используют также глубокие методы обучения, в том числе 2D и 3D сверточные нейронные сети (СНС) для решения задач классификации и сегментации [3,4,6,7,16].

Несмотря на большой интерес к методам глубокого обучения, существует много путей использования обычных методов машинного обучения, которые дают лучшие результаты по сравнению с СДО, использующими СНС. Так в [11] представлена методика, которая помогает сегментировать новообразования без применения методов глубокого обучения. Она использует деревья решений для классификации сегментированной области. В работе [9] отмечается, что информация о КТ-морфологии (размер, объем, форма, контур, структура) играет ключевую роль в скрининге, диагностике и классификации. Эта информация может быть эффективно использована при выявлении рака легкого. Геометрические параметры новообразований широко использовались для их обнаружения [13] дальнейшей классификации методами опорных векторов, k ближайших соседей, деревьями решений.

В представленной работе мы предлагаем СДО, процедура уменьшения числа ложноположительных случаев которой состоит из двух этапов. На первом этапе применяется метод хорд [12] для представления информации о поверхности и форме новообразований. Строится множество хорд между парами точек на поверхности новообразований, число которых определяется заранее. Вычисляются длины полученных хорд, и затем нормализуются в соответствии с самой длинной хордой. Строится гистограмма нормализованных длин хорд. Второй этап предполагает использовать гистограмму совместно с дополнительной информацией о

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

КТ-морфологии новообразования, включая плотность по шкале Хаунсфилда в анализируемой области. Полученный вектор признаков можно рассматривать как характеристическое представление каждого новообразования. Чтобы учесть тот факт, что объем данных для обучения может быть небольшим, мы предлагаем классифицировать полученное представление функции с использованием случайного леса [1]. Предполагается, что база данных содержит изображения КТ в формате DICOM.

II. Сегментация легких на основе фильтрации

В соответствии с анализом большинства СДО [3, 5], процедура обнаружения новообразований в легких на основе серий сканов КТ обычно состоит из следующих этапов: предварительная обработка изображения (обнаружение «кандидатов» для новообразований фильтрацией и сегментацией тканей); сокращение числа ложноположительных случаев (исключение ложных новообразований, которые неверно идентифицированы на этапе фильтрации); классификация новообразований.

Цель процедуры предварительной обработки состоит в том, чтобы отделить область исследования (лёгочную ткань) от других органов и тканей (органы средостения, мягкие ткани грудной стенки, костные структуры) и уменьшить вычислительную сложность следующих этапов [5]. Она также включает этап сегментации легких. В соответствии с этой процедурой данные или значения пикселей в каждом изображении преобразуются в значения плотности по шкале Хаунсфилда, обозначаемые как HU (Hounsfield Unit), в которой за 0 принята плотность воды. Воздух обычно имеет значения около -1000 HU, средние значения денситометрической плотности лёгочной ткани -600 – -400. Используя разницу плотностного диапазона между лёгочной тканью, обладающей т.н. естественной контрастностью, и мягкими тканями, имеющими положительные значения по шкале Хаунсфилда от +40 до +80, метод сегментации представляется, несомненно, эффективным. Маскируя пиксели, которые находятся за пределами интересующей области плотностного диапазона, мы пытаемся оставить для анализа только легочную ткань. Мы используем интервал от -60 до 100 значений плотности для реализации сегментации. Примеры фильтрации показаны на рис. 1 и 2.

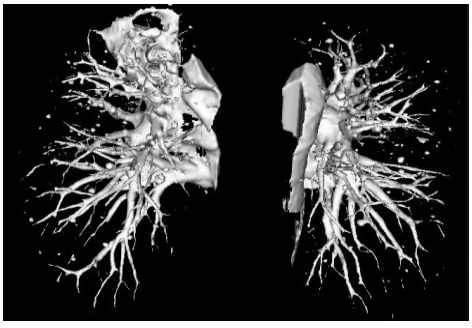

Рис. 1. 3D изображение тканей легких перед фильтрацией

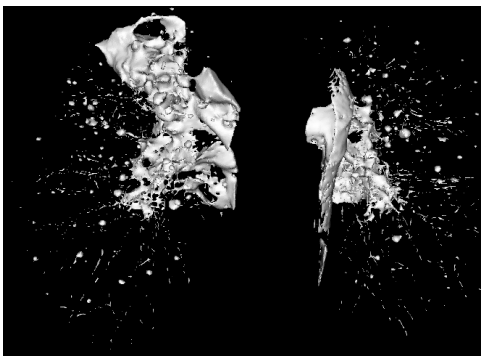

Рис. 2. 3D изображение тканей легких после фильтрации

III. Метод хорд

Прежде чем исследовать, как уменьшить ложноположительные случаи, рассмотрим один из интересных методов описания формы, метод хорд, предложенный в [12]. Метод основан на построении большого числа отрезков, соединяющих случайные пары точек на поверхности фигуры, которые называются хордами. Длины хорд нормируются в соответствии с длиной самой длинной хордой. Множество длин хорд теперь можно рассматривать как распределение вероятностей или гистограмму. Метод хорд инвариантен к размеру объектов, их перемещению и повороту. Он устойчив по отношению к «шумам» или искажениям поверхности объекта.

Насколько нам известно, нет методов, которые используют метод хорд в СДО для уменьшения числа ложноположительных случаев в легких и их классификации. Поэтому его использование в задаче обнаружения новообразований представляет большой интерес.

IV. Сокращение ложноположительных случаев

Процедура обнаружения опухоли направлена на то, чтобы идентифицировать различные элементы в изображениях грудной клетки и отделить новообразования от нормальных анатомических структур. Согласно [5], точная сегментация имеет решающее значение для реализации различных диагностических и лечебных процедур.

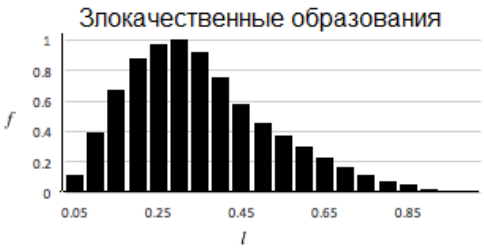

Рис. 3. Гистограмма для злокачественных образований

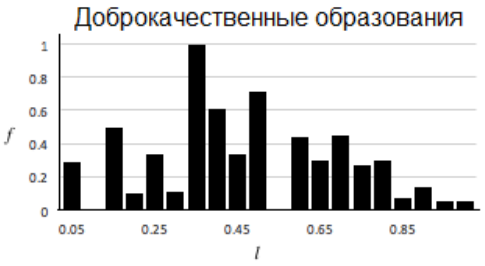

Рис. 4. Гистограмма для доброкачественных образований

Мы используем стандартные библиотеки Python для обнаружения и сегментации новообразований. Объекты на КТ-изображениях разделяются анализом каждого вокселя, применяя библиотеку Python SimpleITK (метод Connected Treshold).

Другой процедурой, используемой после идентификации образований, является сокращение ложноположительных случаев, целью которого является устранение неправильно идентифицированных элементов. Предлагаемый в работе подход заключается в применении метода хорд для представления информации о форме элементов. Оказывается, что гистограммы, соответствующие злокачественным и доброкачественным образованиям совершенно различны (рис. 3 и 4). Для злокачественных образований гистограмма имеет более гладкую форму. Следовательно, их можно использовать для классификации образований. Мы предлагаем также расширить вектор признаков дополнительными признаками, характеризующими морфологическую информацию о каждом образовании. Таким образом, полный вектор признаков состоит из следующих элементов:

1. гистограмма длин хорд, которая характеризуется числом b интервалов;

2. среднее значение плотности по шкале Хаунсфилда вокселей, составляющих анализируемый узел;

3. относительные размеры: l/s, w/s, h/s, где $s=l+w+h$, l, w, h – длина, ширина и высота, соответственно;

4. абсолютные размеры в мм: l, w, h;

5. объем v охватывающего параллелепипеда в мм^3;

6. общий объем q вокселей в образовании в мм^3;

7. относительная плотность: q/v.

В результате мы имеем вектор $b+10$ признаков, характеризующих каждый узел.

Второй шаг обработки изображений в СДО предполагает использование полученного вектора для обучения случайного леса [1], который является одной из наиболее эффективных композиционных моделей классификации, состоящей из большого числа деревьев решений, построенных на основе случайного выбора подмножеств примеров обучающей выборки и подмножеств признаков.

V. Численные эксперименты

Чтобы исследовать предлагаемую СДО, мы используем набор данных, содержащий снимки КТ-сканирования N=228 пациентов в формате DICM из Minisite Harvard Tunor Hunt Challenge Minisite http://www.topcoder.com.

F-мера (F_1-мера) используется для оценки точности классификации системы САПР. В отличие от стандартной точности классификации (доля правильно классифицированных случаев на выборке данных), F-мера учитывает несбалансированность классов.

Разработанное программное обеспечение реализовано на Python. Чтобы оценить точность, мы выполняем кросс-валидацию со 100 повторениями, где в каждом цикле произвольно выбираем данные для обучения (N_{train}=3N/4) and для тестирования (N_{test}=N/4).

Первый вопрос заключается в том, как F-мера зависит от количества интервалов b в гистограмме. Это число определяет соответствующее количество признаков, характеризующих форму образований. На рис. 5 показана эта зависимость. Наибольшее значение F-меры для данного набора данных достигается при b=20.

Рис. 5. Зависимость F-меры от числа интервалов в гистограмме

Мы также исследуем, как F-мера зависит от количества хорд c, используемых для представления каждого узла. Зависимость показана на рис. 6, где видно, что наилучшие результаты достигаются при c=105. Из результатов можно сделать вывод, что существуют некоторые оптимальные значения параметров настройки b и c, которые должны быть получены для каждого набора данных.

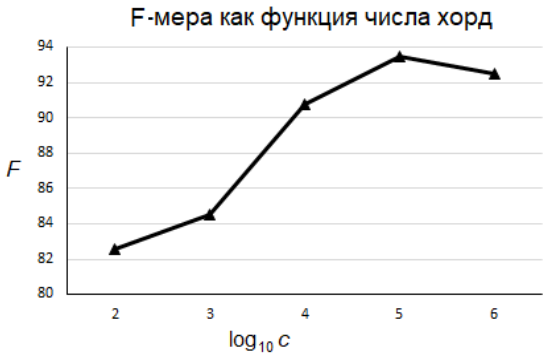

Рис. 6. Зависимость F-меры от числа хорд

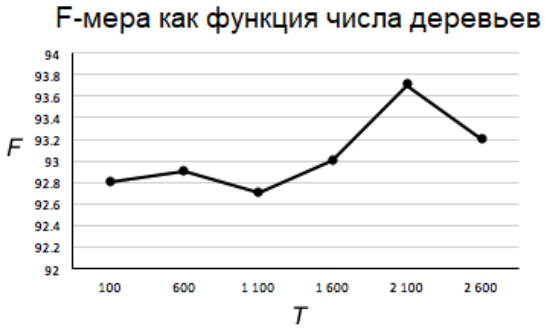

Рис. 7. Зависимость F-меры от числа деревьев

Мы также исследуем, как F-оценка зависит от числа деревьев решений деревьев в случайном лесу. Соответствующая зависимость показана на рис. 7, где наибольшее значение F-меры равна 93.7. Это достигается при количестве деревьев равным 2100.

VI. Заключение

Одна из реализаций СДО была представлена в работе. Главная особенность системы заключается в том, что она использует методы «неглубокого» обучения, которые значительно упрощают процедуру обучения. Важно отметить, что метод хорд в сочетании с сегментацией изображений легких на основе пороговых значений плотностей позволяет получить интересные результаты, которые сравнимы с другими существующими современными подходами, применяемыми для построения СДО.

Предложенную реализацию СДО можно рассматривать как первую попытку использования метода хорд. Более того, мы рассмотрели только использование случайных лесов для классификации новообразований. Однако мы предполагаем, что использование более сложных классификаторов, например глубокого леса [15], может значительно повысить эффективность системы. Это является интересным направлением для дальнейших исследований. Другим направлением для дальнейших исследований является совместное использование предлагаемого представления признаков со стандартным представлением изображения, где классификация изображений выполняется путем применения СНС или снова глубокого леса с элементами обработки исходных изображений большой размерности.

Список литературы

[1] Breiman L. Random forests // *Machine learning*, 45(1):5–32, 2001.

[2] Choi W.J. and Choi T.S. Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor // *Computer Methods and Programs in Biomedicine*, 113:37–54, 2014.

[3] Chon A., Balachandar N., and Lu P. Deep convolutional neural networks for lung cancer detection // Technical report, Stanford University, 2017.

[4] Dey R., Lu Z., and Hong Y. Diagnostic classification of lung nodules using 3D neural networks // arXiv:1803.07192v1, March 2018.

[5] Firmino M., Morais A.H., Mendoca R.M., Dantas M.R., Hekis H.R., and Valentim R. Computer-aided detection system for lung cancer in computed tomography scans: review and future prospects // *Biomedical engineering online*, 13(1):41, 2014.

[6] Hamidian S., Sahiner B., Petrick N., and Pezeshk A. 3D convolutional neural network for automatic detection of lung nodules in chest CT // *Proc SPIE Int Soc Opt Eng*, 10134:1013409, Mar 2017.

[7] Huang X., Shan J., and Vaidya V. Lung nodule detection in CT using 3D convolutional neural networks // *14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 379–383. IEEE, April 2017.

[8] John J. and Mini M.G. Multilevel thresholding based segmentation and feature extraction for pulmonary nodule detection // *Procedia Technology*, 24:957–963, 2016.

[9] Khosravan N. and Bagci U. Semi-supervised multi-task learning for lung cancer diagnosis // arXiv:1802.06181v1, Feb 2018.

[10] Mobiny A., Moulik S., Gurcan I., Shah T., and Van Nguyen H. Lung cancer screening using adaptive memory-augmented recurrent networks // arXiv:1710.05719v1, Oct 2017.

[11] Nithila E.E. and Kumar S.S. Automatic detection of solitary pulmonary nodules using swarm intelligence optimized neural networks on CT images // *Engineering Science and Technology, an International Journal*, 20(3):1192–1202, 2017.

[12] Smith S.P. and Jain A.K. Chord distribution for shape matching // *Computer vision, graphics, and image processing*, 20(3):259–271, 1982.

[13] Tan M., Deklerck R., Jansen B., Bister M., and Cornelis J. A novel computer-aided lung nodule detection system for CT images // *Medical physics*, 38(10):5630–5645, 2011.

[14] Walawalkar D. A fully automated framework for lung tumour detection, segmentation and analysis // arXiv:1801.01402, Jan 2018.

[15] Zhou Z.-H. and Feng J. Deep forest: Towards an alternative to deep neural networks // *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*, pages 3553–3559, Melbourne, Australia, 2017.

[16] Zhu W., Liu C., Fan W., and Xie X. DeepLung: Deep 3D dual path nets for automated pulmonary nodule detection and classification // arXiv: 1801.09555v1, Jan 2018.