ПРИМЕНЕНИЕ МЕТОДОВ СОКРАЩЕНИЯ ПРИЗНАКОВОГО ПРОСТРАНСТВА ДЛЯ ТЕКСТУРНОЙ КЛАССИФИКАЦИИ ПАТОЛОГИЙ ДИССЕМИНИРОВАННОГО ТУБЕРКУЛЁЗА ЛЁГКИХ

К.А. Костин
С.В. Аксёнов
Томский политехнический университет
kak@tpu.ru

На сегодняшний день, согласно статистике Всемирной организации здравоохранения, туберкулёз лёгких наряду с тремя другими опасными заболеваниями органов дыхания входит в рейтинг десяти самых распространённых причин смерти человека в мире. По состоянию на 2016 год, туберкулёз занимает в нём десятую позицию – в год он забирает жизни около 1,4 миллионов человек по всему миру [1]. В России, по состоянию на 2016 год, на 100 000 человек приходится 11,5 % смертей от этого заболевания – это около 16 500 тысяч в год [2].

Одной из главных проблем борьбы с туберкулёзом на сегодняшний день, для которой до сих пор не найдено эффективного решения, является сложность его ранней диагностики. Наиболее информативными методами исследования пациентов на сегодняшний день являются лучевые методы, в частности компьютерная томография (КТ). Но диагностика с помощью такого подхода затруднена из-за высоких требований к квалификации и опыту врача, что и является частой причиной трудностей с постановкой диагноза на ранних стадиях развития заболевания [3]. Таким образом, разработка системы медицинской диагностики для автоматического детектирования и классификации различных форм туберкулёза лёгких является одной из актуальных задач на сегодняшний день.

Одной из наиболее распространённых форм туберкулёза лёгких, которая часто на ранних стадиях развития может быть расценена врачом как пневмония, является диссеминированная форма [4]. Существующие исследования по данному направлению в основном направлены на выделение некоторых визуальных характеристик патологий диссеминированного туберкулёза на снимках КТ или рентгене [5] и построение методов по их детектированию на основе алгоритмов компьютерного зрения [6].

Для построения системы диагностики диссеминированной формы туберкулёза лёгких авторами данной работы уже был реализован метод классификации патологий по данным КТ с помощью текстурных характеристик *GLCM* (*Gray Level Co-Occurrence Matrix*) [7]. Задачами текущего этапа работы над системой являются:

1. исследование пространства текстурных признаков *GLCM*;
2. применение методов сокращения этого пространства с целью улучшения результатов классификации;
3. оценка информативности признаков *GLCM* для выделения патологий диссеминированной формы туберкулёза.

Исследование проводилось на неперсонифицированных КТ-данных пациентов с подтверждённым диагнозом диссеминированного туберкулёза. Каждый снимок КТ был размечен: участки лёгких, содержащие патологию, выделены как отдельные регионы – рис. 1.

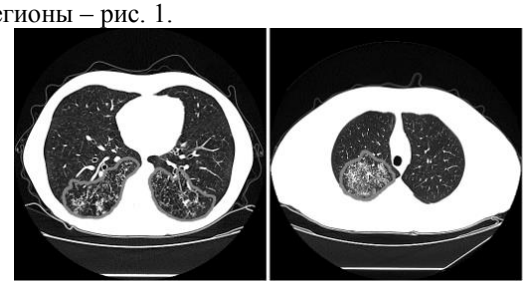

Рис. 1. Пример КТ-снимков, на которых границами выделены регионы, содержащие патологии диссеминированной формы

Существующая методология детектирования и классификации патологий диссеминации включает в себя следующие этапы:

1. сегментация участков снимка, включающих в себя только лёгкие, с помощью фильтрации изображения и кластеризации *k-means*;
2. расчёт текстурных признаков *GLCM* для каждого пикселя лёгких на снимке с экспериментально подобранными параметрами;
3. классификация векторов признаков с помощью алгоритма случайного леса и отделение участков класса патологии от участков здоровой ткани.

Для того, чтобы исследовать информативность используемых текстурных характеристик для описания патологий диссеминированной формы, были построены графики классифицируемых объектов в пространстве наиболее важных для классификатора признаков. Также были применены методы сокращения признакового пространства (*PCA – Principal Component Analysis* и *t-SNE – t-distributed Stochastic Neighbor Embedding*), позволившие преобразовать исходные признаки и так же отобразить наиболее важные из них на графиках. Это позволило лучше понять распределение значений текстурных характеристик, а также определить степень разделимости классов патологии и здоровой ткани – рис. 2.

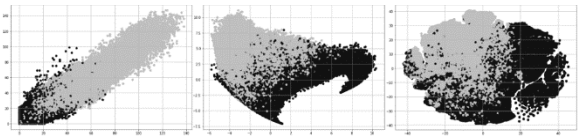

Рис. 2. Графики распределения объектов классификации в пространстве наиболее важных признаков. Слева-направо: исходные признаки, признаки *PCA*, признаки *t-SNE*

Как видно на рис. 2, классы в пространстве наиболее важных признаков (для примера приведено двухмерное пространство) трудно разделимы, что вызывает сложности при создании модели, которая могла бы достаточно эффективно разделить классифицируемые объекты в этих пространствах. Видно, что методы *PCA* и *t-SNE* значительно видоизменяют исходное пространство, однако даже они не позволяют выделить такое подпространство, в котором объекты класса патологии (светлый цвет на графиках) могли бы быть эффективно отделимы от объектов, относящихся к здоровой ткани лёгких (тёмный цвет на графиках).

Для сравнения эффективности классификации при применении методов сокращения пространства признаков, были проведены эксперименты для нескольких моделей на неизменённом признаковом пространстве и на сокращённом с помощью *PCA* – количество признаков было уменьшено с 40 до 14. Метод *t-SNE* в данных экспериментах не рассматривался из-за его низкой вычислительной эффективности. Данные эксперимента включали в себя набор признаков, полученных с 99 КТ-срезов для обучения и 417 КТ-срезов для тестирования. Сравнительные результаты представлены в таблице 1.

Таблица 1. Результаты экспериментов

Алгоритм	Точность	Полнота	*F*-мера
Лог. регр.	0,21	0,65	0,31
CatBoost	0,23	0,64	0,33
Случ. лес	0,35	0,58	0,43
Лог. регр. (*PCA*)	0,18	0,68	0,28
CatBoost (*PCA*)	0,21	0,64	0,32
Случ. лес (*PCA*)	0,29	0,60	0,39

Полученные результаты демонстрируют, что применение метода *PCA* для линейного преобразования признакового пространства позволило улучшить результаты по метрике полноты (*Recall*), пожертвовав тем самым точностью (*Precision*). Таким образом, преобразовав исходное признаковое пространство, удалось уменьшить количество пропусков патологий диссеминированного туберкулёза на КТ-снимке в счёт увеличения ложных срабатываний. Также, работа в сокращённом пространстве увеличивает вычислительную эффективность при обучении модели – в среднем скорость обучения возросла в 2,5 раза.

Результаты проведённого исследования показывают, что методы сокращения размерности признакового пространства являются достаточно эффективными и позволяют не только увеличить производительность и точность решения, но и получить некоторую проекцию признакового пространства, в которой классы могут быть разделены лучше. Основной вывод данной работы формулируется из результатов исследования текстурных признаков *GLCM* – построение эффективного классификатора на этих данных затруднено и необходимо искать другие более информативные характеристики для описания патологий диссеминированного туберкулёза. Поэтому в будущем планируется изменить методологию классификации, отдельно выделив этапы детектирования области патологий и извлечения локальных текстурных дескрипторов для их описания.

Данные для исследования предоставлены Национальной академией наук Белоруссии. Исследование выполнено при поддержке гранта РФФИ №16-47-700289.

Список использованных источников

1. The top 10 causes of death [Электронный ресурс] / World Health Organization. 2018. URL: http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (дата обращения: 18.11.2018).
2. Russian Federation – Tuberculosis Country brief, 2016 / World Health Organization. 2018. URL: http://www.euro.who.int/__data/assets/pdf_file/0010/335539/RUS_TB_Brief_0223-AM-edits-D1-20-03-17.pdf?ua=1 (дата обращения: 18.11.2018).
3. Berlin L. Accuracy of Diagnostic Procedures: Has It Improved Over the Past Five Decades? / L. Berlin // AJR. – 2007. – Vol. 188. – P. 1173–1178.
4. Дейкина О.Н. Дифференциальная диагностика пневмонии и туберкулёза лёгких // автореф. дис. канд. мед. наук: 14.00.05, 14.00.26. Моск. гос. мед.-стом. ун-т. – 2005. – 25 с.
5. Feng F. Radiological characterization of disseminated tuberculosis in patients with AIDS / F. Feng, G. Xia, Y. Shi, Z. Zhang // Radiology of Infectious Diseases – 2016 – Vol. 3, Issue 1 – Pp. 1-8.
6. Ramya R. Automatic tuberculosis screening using canny Edge detection method / R. Ramya, P.S. Babu // ICECS – 2015 – Pp. 282-285.
7. Костин К.А., Ламонова Т.С., Аксёнов С.В. Классификация патологий диссеминированного туберкулёза лёгких с помощью методов машинного обучения // Научная сессия ТУСУР – 2018: Материалы международной научно технической конференции студентов, аспирантов и молодых учёных – избранные статьи: в трёх частях., Томск, 16-18 Мая, 2018. – Томск: ТУСУР, 2018 – часть 3, С. 129-132.