МЕТОДИКА ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ДАННЫХ И ПОДГОТОВКИ ДАТАСЕТА ДЛЯ АНАЛИЗА ПРЕПАРАТОВ ЖИДКОСТНОЙ ЦИТОЛОГИИ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

Курочка К. С., Зайцева Л. П., Гуменников Е. Д.
Кафедра информационных технологий, Гомельский государственный университет имени П. О. Сухого;
Учреждение Гомельский областной клинический онкологический диспансер,
централизованная цитологическая лаборатория
Гомель, Республика Беларусь
E-mail: guma178@gstu.by

В данной работе рассматриваются методы и подходы к подготовке датасета для обучения моделей машинного обучения в задачах анализа препаратов жидкостной цитологии. Особое внимание уделяется процессам сбора, очистки, аннотирования и форматирования данных, необходимым для обеспечения высокой точности и надежности автоматизированных систем диагностики. Представлены рекомендации по оптимизации этапов предварительной обработки данных, а также обсуждаются особенности работы с медицинскими изображениями и цитологическими образцами.

Введение

Исследование клеточных структур методом автоматизированного анализа изображений цифровой микроскопии играет важную роль в биомедицинских исследованиях, особенно в области онкологии. Одним из ключевых элементов, участвующих в клеточном метаболизме и пролиферации, являются ядрышковые организаторы. Эти участки хромосом содержат гены, отвечающие за синтез рибосомальной РНК (рРНК), что делает их активность важным показателем уровня рибосомного синтеза и общей активности клетки. Анализ изменений в структуре и количестве ядрышковых организаторов может служить ценным диагностическим критерием при оценке злокачественных новообразований и прогнозировании их агрессивности.

В публикации рассматриваются методы подготовки датасета для обучения ИИ в анализе жидкостной цитологии, включая сбор, аннотирование и обработку данных, что способствует качественному обучению моделей искусственного интеллекта и повышает эффективность автоматизированных решений.

I. Обзор исходных данных, постановка задачи

На рисунке 1 приведен пример изображения препарата полученного методом цифровой микроскопии. Подобные изображения являются исходными данными для искомой модели машинного обучения. Перед такой компьютерной программой стоит задача локализации клеточных структур и их компонентов, а именно ядрышковых организаторов.

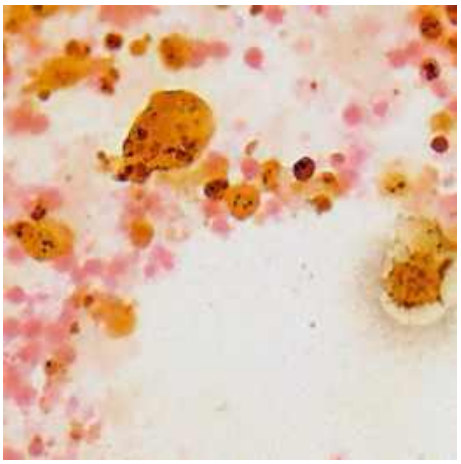

Рис. 1 – Цитологический препарат

На данном изображении клеточные структуры представлены эллиптическими областями повышенной контрастности. Однако формы и размеры этих фрагментов в значительной мере отличаются относительно друг друга. Так же присутствуют элементы окрашенные в аналогичные цвета и обладающие нечеткой границей, которые не являются искомыми клеточными структурами. Таким образом задача локализации таких объектов является трудно формализуемой. Похожее описание применимо к другому типу искомых объектов – ядрышковых организаторов. Лишь с тем отличием что их контрастность куда выше контрастности клеточных структур и их размер значительно меньше сравнительно с первым описанным классом. На изображении также встречаются фрагменты похожего цвета и контрастности не являющиеся искомыми объектами. Ядрышковые организаторы всегда содержатся внутри клеточных структур. Традиционные алгоритмы основаны на методах компьютерного зрения: пороговая сегментация, морфологические операции, фильтрация и анализ контуров. Эти

подходы позволяют выделить объекты по цвету, форме и текстуре. Однако они чувствительны к шуму, вариативности окрашивания и артефактам. При незначительных изменениях входных данных такие методы требуют ручной настройки параметров, что снижает их универсальность.

Более продвинутые методы используют машинное обучение, где объекты распознаются на основе заранее извлечённых признаков – например, гистограмм градиентов, локальных бинарных шаблонов или др. Классификаторы, такие как SVM или случайные леса, обучаются на размеченных данных. Хотя эти методы демонстрируют лучшую устойчивость, они не всегда справляются с сложными случаями перекрытия или деформации клеток.

Наиболее эффективным решением задачи распознавания клеток и ядрошковых организаторов являются современные архитектуры глубоких нейронных сетей, относящиеся к классу одноэтапных детекторов. Эти модели объединяют этапы локализации и классификации объектов в единую структуру, что позволяет обрабатывать изображения с высокой точностью. Они обучаются напрямую на изображениях, автоматически извлекая релевантные признаки, и демонстрируют устойчивость к шуму, вариациям формы и плотности объектов.

Благодаря своей архитектуре, одноэтапные детекторы особенно хорошо подходят для задач, где требуется быстрое и точное обнаружение множества мелких объектов — как в случае с клетками и ядрошковыми организаторами.

В задачах анализа цитологических изображений локализация объектов оказывается более практичным и эффективным подходом по сравнению с сегментацией. Метод локализации позволяет быстро и надёжно обнаруживать клетки и ядрошковые организаторы с помощью ограничивающих рамок, обеспечивая высокую точность при значительно меньших затратах на аннотирование и обучение модели. В отличие от сегментации локализация демонстрирует лучшую масштабируемость, устойчивость к вариативности изображений и высокую скорость обработки.

Для обучении модели искусственного интеллекта необходим датасет, содержащий изображения препаратов и аннатирующий файл, содержащий N строк. Кажадая строка содержит информацию о классе объекта и его габаритной рамки в виде «N X1 Y1 X2 Y2», где N код класса, X1 и Y1 координаты нижнего левого угла габарита объекта, а X2 и Y2 верхнего правого.

II. Конвертация существующего датасета предназначенного для задачи сегментации

Для обучения модели локализации клеточных структур и ядрышковых организаторов возможно использование датасета предназначенного для инстанс-сегментации аналогичных объектов. В существующем наборе данных каждое изображение аннотируется текстовым файлом содержащим количество строк равное количеству размеченных объектов искомых классов на соответствующем изображении. Каждая строка содержит код класса и пары вещественных чисел кодирующих координаты вершин полигонов ограничивающих сегмент изображения содержащий объект.

Путем простых математических манипуляций возможно рассчитать границы рамок для задачи локализации и сформировать датасет для решения описанной задачи. Расчет координат границ рамок возможно вычислить по следующим формулам:

$$x_{ul} = min\,(X)\,; y_{ul} = min\,(Y)$$

$$x_{br} = max\,(X)\,; y_{br} = max\,(Y)$$

где X – координаты X вершин полигона;

Y – координаты Y вершин полигона;

x_{ul} – координата x верхнего левого угла рамки;

y_{ul} – координата y верхнего левого угла рамки;

x_{br} – координата x нижнего правого угла рамки;

y_{br} – координата y нижнего правого угла рамки;

Таким образом конвертируется датасет ориентированный на задачу сегментации для задачи локализации.

III. Контекстно-ориентированная фильтрация артефактов изображений

Для повышения качества обучающего датасета при распознавании клеток и ядрошковых организаторов в цитологических препаратах разумно применение метод предварительной обработки изображений, заключающийся в размытии ложных или обрезанных объектов. Такие структуры, часто встречающиеся на краях изображений или в виде артефактов, могут негативно влиять на обучение модели, создавая шум и снижая точность распознавания. Размытие позволяет нейтрализовать их влияние, сохраняя при этом естественную текстуру фона, что делает изображения более приближенными к реальным условиям микроскопии. Этот подход способствует улучшению устойчивости модели и снижению количества ложных срабатываний.

1. Курочка К. С., Ковалев В. А. Организация распределенных вычислений при обработке цифровых биомедицинских изображений // Информатика. – 2018. – №. 4 (24). – С. 66-73.
2. Weeks, S. E. The nucleolus: a central response hub for the stressors that drive cancer progression / S. E. Weeks, B. J. Metge, R. S. Samant // Cellular and Molecular Life Sciences. – 2019. – Vol. 76. – С. 4511-4524.
3. Богуш, Р. П. Обнаружение объектов на изображениях с большим разрешением на основе их пирамидально-блочной обработки / Р. П. Богуш, И. Ю. Захарова, С. В. Абламейко // Информатика. – 2020. – Т. 17, № 2. – С. 7-16.