Ж.В. Штадельманн, И.Н. Спиридонов

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ЛЕЙКОЦИТОВ НА ИЗОБРАЖЕНИЯХ МАЗКОВ КРОВИ

Аннотация
Определение лейкоцитарной формулы является важным этапом клинического анализа крови. В статье представлены классификатор, решающий задачу вычисления лейкоцитов по изображениям мазков крови, разработанный на основе варианта алгоритма AdaBoost, а также результаты его применения.

Клинический анализ крови является одним из видов анализов, наиболее часто востребованных врачами [1]. В процессе клинического анализа крови осуществляется подсчет форменных элементов, в том числе лейкоцитов. Так как лейкоциты являются главными агентами иммунной системы, состав лейкоцитарной формулы отражает состояние здоровья пациента, а количественные отклонения состава, так же как особенности клеточной морфологии, являются симптомами болезней [2], [3].

Проточная цитометрия, при которой учитываются электродинамические характеристики клеток, позволяет различить все пять типов лейкоцитов. В результате определяют общее число лейкоцитов и относительные концентрации лейкоцитов разных типов. Также большой объем крови (53 мкл [4]) позволяет анализировать большее число клеток и увеличивает статистическую достоверность анализа. Однако системы проточной цитометрии не позволяют анализировать особенности морфологии клеток (*табл. 1*), и в случаях, отличных от принятых в гематологии норм, необходим визуальный анализ [2], [5], [6].

Второй распространенный метод подсчета формулы кровы – визуальный метод. При применении визуального метода приготавливается мазок, который анализируется врачом-гематологом под микроскопом. Так как анализ проводится непосредственно врачом, он выделяет только такие морфологические особенности, которые знает. Стандартная процедура визуального анализа выполняется на ста клетках [2]. Доверительные интервалы позволяют оценить общую популяцию на базе образца с заданной вероятностью ошибки. В *табл. 2* представлены примеры результатов определения лейкоцитарной формулы, принятые в гематологии нормы [7], а также доверительный интервал, который с вероятностью 95 % содержит истинное число клеток, построенный по результатам исследования клеток с использованием распределения Стьюдента.

Второй распространенный метод подсчета формулы кровы – визуальный. При применении визуального метода приготавливается мазок, который анализируется врачом-гематологом под микроскопом. Так как анализ проводится непосредственно врачом, он выделяет только такие морфологические особенности, которые знает. Стандартная процедура визуального анализа выполняется на ста клетках [2]. Доверительные интервалы позволяют оценить общую популяцию на базе образца с заданной вероятностью ошибки. В *табл. 2* представлены примеры результатов определения лейкоцитарной формулы, принятые в гематологии нормы [7], а также доверительный интервал, который с вероятностью 95 % содержит истинное число клеток, построенный по результатам исследования клеток с использованием распределения Стьюдента.

Как видно из *табл. 2*, только в случае нейтрофилов интервал нормы шире доверительного интервала и соответственно вероятное число лейкоцитов попадает в норму только для этих клеток. Уменьшить ширину доверительного интервала и, следовательно, достоверно оценить число лейкоцитов возможно. Для этого необходимо увеличить количество исследуемых клеток. Поскольку визуальная классификация клеток под микроскопом является очень утомительной задачей, максимальное число клеток, анализируемых визуально, не превосходит 200 [2], [9]. Следовательно, необходима автоматическая система обработки изображений клеток, облегчающая работу лаборанта и способная оценить не только числа клеток, но и их морфологические особенности.

Метод

При определении формулы белой крови подсчитывают лейкоциты пяти классов (*рис. 1*): эозинофилы, лимфоциты, моноциты, нейтрофилы и базофилы. Так как базофилы редко встречаются в периферической крови, они исключены из классификации [6], [10].

Оценка качества препарата является первым этапом обработки мазков крови. Наличие области монослоя клеток является критерием качества препарата [11].

Этап обнаружения лейкоцитов, позволяющий ускорить обработку препарата, проводится после оценки его качества. Наличие ядра лежит в основе обнаружения лейкоцитов [12].

Таблица 1

Виды лейкоцитов и изменения нормального состава крови [1], [7], [8]

Вид лейкоцитов	Норма	Наименование	Изменения и аномалии
Лейкоциты	4...8,8 · 10^9/л	–	–
Эозинофил	Меньше 5 % лейкоцитарной популяции	Эозинофилоз Эозинопения Морфологические изменения	> 0,4 · 10^9/л < 0,05 · 10^9/л Гиперсегментация ядра Гипосегментация ядра Кольцеобразное ядро Наличие вакуолей Низкая зернистость
Лимфоцит	От 18 до 40 % лейкоцитарной популяции	Лимфоцитоз Лимфопения Морфологические изменения	> 4 · 10^9/л < 1 · 10^9/л Нерегулярность формы клеточной границы Не кольцеобразное ядро Двухядерная клетка Аномалии размера Полихроматизм цитоплазмы Наличие вакуолей Агрегаты лимфоцитов Превращение в плазматическую клетку Инклюзии
Моноцит	От 2 до 9 % лейкоцитарной популяции	Моноцитоз Монопения Морфологические изменения	> 0,8 · 10^9/л < 0,03 · 10^9/л Соотношение площадей ядра к цитоплазмы Наличие вакуолей Наличие инклюзии (малярийный плазмодий, эритроциты, ...) Кольцеобразное ядро Превращение в макрофаг
Нейтрофил	От 46 до 65 % лейкоцитарной популяции	Нейтрофилоз Нейтропения Морфологические изменения	> 8 · 10^9/л < 1,5 · 10^9/л Наличие ядерных проекций Гиперсегментация ядра Гипосегментация ядра Кольцеобразное ядро Ядро в форме грозди Фрагментация ядра Аномалии зернистости Наличие вакуолей Наличие тела Деле Наличие инклюзии, отходов фагоцитоза

Таблица 2

Результаты подсчета лейкоцитов, принятые в гематологии нормы и доверительный интервал

Вид лейкоцитов	Число исследованных клеток	Нормы согласно [7], % популяции	Доверительный интервал, % популяции
Нейтрофилы	56	От 45 до 70	Между 46 и 65
Лимфоциты	32	От 18 до 40	Между 22 и 41
Моноциты	7	От 2 до 9	Между 1 и 12
Эозинофилы	5	Менее 5	Между 0,7 и 9
Базофилы	0	Менее 1	–

Для получения характеристик, необходимых при классификации клеток, применяется сегментация ядра и цитоплазмы. Сегментация ядра обеспечивается введением порога в канале насыщенности цвета после преобразования изображения в цветовое пространство HSV [13] (*рис. 2*). Для сегментации цитоплазмы использовались критерии на базе энтропии Шэннона, алгоритмы обнаружения границ и операции морфологической обработки изображении для заполнения разрывов периметра [14], [15] (*рис. 2*).

Характеристики, использованные для классификации, были выбраны таким образом, чтобы они покрыли широкий спектр признаков лейкоцитов [15]: например, оконтуривание ядра отражает его топологию, а текстура цитоплазмы отражает ее содержание. Также выбор характеристик зависел от расстояний между максимумами их статистических распределений по типам лейкоцитов. Так как функции распределения характеристик лейкоцитов несимметричны, то они моделировались законом Вейбулла (*рис. 3*).

Для классификации лейкоцитов были использованы следующие характеристики:

- коэффициент корреляции текстуры ядра и цитоплазмы;
- средний квадрат текстуры ядра и цитоплазмы;
- контраст текстуры ядра и цитоплазмы;
- гомогенность текстуры цитоплазмы;
- среднее расстояние от центра до цитоплазмы;
- число бифуркаций и окончаний;
- эксцентриситет окружающего эллипса;
- площадь и периметр ядра;
- число Эйлера цитоплазмы;
- цвет цитоплазмы.

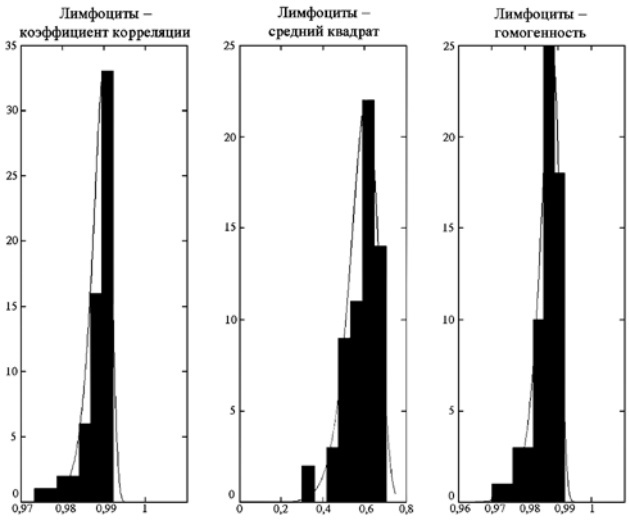

Рис. 3. Статистические распределения ядерных текстур лимфоцитов

Алгоритм классификации

База данных, использованная для классификации лейкоцитов, состояла из изображений мазков крови 12 пациентов. После предварительного этапа обнаружения были получены изображения, содержащие один лейкоцит, размером 17,6 х 17,6 мкм, соответствующим размеру моноцита – самого крупного лейкоцита [12].

База данных содержала 703 клетки и была разделена на две группы: для обучения и контроля классификатора, как описано в *табл. 3*.

Все лейкоциты, присутствующие в базе данных, были классифицированы визуально врачом-гематологом.

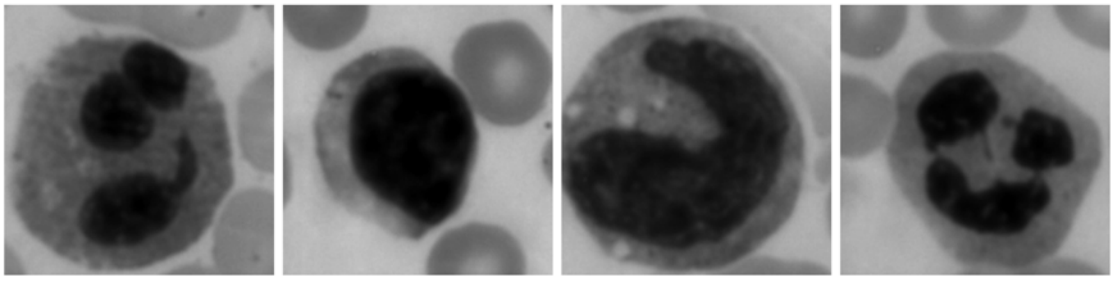

Рис. 1. Эозинофил, лимфоцит, моноцит и нейтрофил

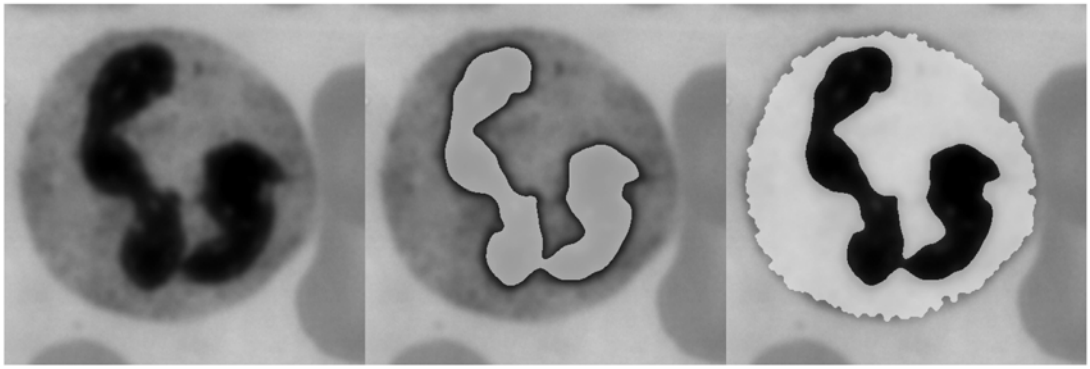

Рис. 2. Нейтрофил и соответствующие результаты сегментации ядра (в центре) и цитоплазмы (справа)

Таблица 3

Распределение лейкоцитов в базе данных

Группы:	Эозинофилы	Лимфоциты	Моноциты	Нейтрофилы
обучения	18	100	24	100
контроля	19	185	24	233
Итого:	37	285	48	333

Задачи обнаружения, распознавания и классификации целесообразно решать с использованием алгоритмов повышения эффективности, например AdaBoost [13]. Эти алгоритмы основываются на объединении большего количества так называемых слабых классификаторов в одном сильном, который проводит классификацию с малыми ошибками первого и второго рода и с низкими временными затратами [16], [17].

Алгоритмы повышения эффективности обычно выдают бинарный ответ, который не соответствует требованиям к полной лейкоцитарной классификации, однако классификация по большому числу классов возможна, например, с использованием алгоритмов SAMME или варианта Gentle AdaBoost [18], [19].

Итеративный процесс обучения классификатора состоит из описанных ниже этапов:

1. Определение весовых коэффициентов и целевой функции

$$w_i = \frac{1}{N} \mid i = 1, 2, ..., N;$$

$$\vec{F}(\vec{x}) = \vec{0}.$$

2. Для каждой реализации $m = 1$ до M:

a) выполняется согласование функции регрессии $\vec{y} = \vec{g}^{(m)}(\vec{x})$ и $\vec{z} = \vec{h}^{(m)}(\vec{x})$ весовым методом наименьших квадратов с использованием весовых коэффициентов w_i и $z_j = y_j^2 \mid j = 1, 2, ..., K$;

b) вычисляется выражение

$$r_j^{(m)}(\vec{x}) = K \cdot \frac{g_j^{(m)}(\vec{x})}{h_j^{(m)}(\vec{x})} \mid j = 1, 2, ..., K;$$

c) выполняется формирование слабого классификатора

$$f_j^{(m)}(\vec{x}) = r_j^{(m)}(\vec{x}) - \frac{1}{K} \sum_{k=1}^{K} r_k^{(m)}(\vec{x}) \mid j = 1, 2, ..., K;$$

d) выполняется уточнение целевой функции

$$\vec{F}(\vec{x}) \leftarrow \vec{F}(\vec{x}) + \vec{f}^{(m)}(\vec{x});$$

e) выполняется уточнение весовых коэффициентов w_i

$$w_i \leftarrow w_i \cdot e^{K^{-1} \cdot \vec{y}_i^T \cdot \vec{f}^{(m)}(\vec{x})} \mid i = 1, 2, ..., N$$

и их нормализация

3. Применение сильного классификатора

$$C(\vec{x}) = \arg \max_k F_k(\vec{x}),$$

где w_i – весовые коэффициенты данных; N – размер обучающей выборки; M – количество повторов процесса повышения эффективности; $\vec{y}$ – вектор классов, соответствующих векторам образцов $\vec{x}$; K – число классов; $\vec{f}^{(m)}(\vec{x})$ – слабый классификатор; $C(\vec{x})$ – результат сильной классификации с использованием целевой функции $\vec{F}(\vec{x})$.

Алгоритм AdaBoost и его варианты интересны тем, что их можно обучить, несмотря на число классов и признаков. В случае классификации лейкоцитов адаптивность алгоритмов AdaBoost позволяет переобучить классификатор и включить в него новый класс, что позволит в будущем классифицировать базофилы или бластные клетки при наличии достаточного количества изображений клеток в создании обучающей выборки.

Анализ полученных результатов

Вариант алгоритма Gentle AdaBoost для объединенной классификации был обучен и протестирован при классификации лейкоцитов. Результаты тестирования классификатора приведены в *табл. 4* и *5*.

Вероятность правильной классификации достигает 91,3 %.

Таблица 4

Результаты классификации лейкоцитов

Вид лейкоцитов	Итог	Эозинофилы	Лимфоциты	Моноциты	Нейтрофилы
Эозинофилы	19	14	1	3	1
Лимфоциты	185	0	177	4	4
Моноциты	24	1	4	16	3
Нейтрофилы	233	3	8	9	213

Таблица 5

Детализация родов ошибок каждого класса

Вид лейкоцитов	Ошибка первого рода, %	Ошибка второго рода, %
Эозинофилы	26,3	0,9
Лимфоциты	4,3	4,7
Моноциты	33,3	3,7
Нейтрофилы	8,6	3,5

Высокие значения ошибки первого рода обусловлены малыми размерами выборки лейкоцитов, поскольку соотношение содержащихся в формуле крови лейкоцитов очень низко в случае моноцитов и эозинофилов. Следовательно, минимизация погрешности, являющаяся основой применения алгоритмов повышения эффективности, не позволяет обучить их таким способом, чтобы их результаты для указанных видов лейкоцитов были удовлетворительными.

Тем не менее малые значения ошибок первого и второго рода при классификации нейтрофилов и особенно лимфоцитов приводят к выводу, что использование варианта алгоритма AdaBoost, позволяющее провести классификацию по большему числу клеток, является перспективным выбором для

разработки системы вычисления формулы белой крови.

Заключение

Проведенные работы позволили построить автоматизированную систему определения формулы белой крови с вероятностью не ниже 90 %.

Несмотря на низкие значения вероятности классификации, полученные для моноцитов и эозинофилов, представленный классификатор обеспечивает достаточно высокую надежность классификации, что и можно применять в клинике.

Список литературы:

1. *Bain B.J.* Blood Cells: A Practical Guide. – Blackwell Publishing Ltd, Malden (Massachusetts), 2006.
2. *Bong H.H, Gulati G.L., Ashton J.K.* Differential Leukocyte Count: Manual or Automated, What Should It Be? // Yonsei Medical Journal. 1991. Vol. 32. № 4.
3. *Cornbleet J.* Clinical Utility of the Band Count // Clinics in Laboratory Medecine. 2002. Vol. 22. № 1.
4. *O'Neil P., Vital E., Betancourt-Loria N., Dinah M.* Performance Evaluation of the Complete Blood Count and White Blood Cell Differential Parameters on the AcT 5diff Hematology Analyzer // Laboratory Hematology. 2001. Vol. 7. PP. 116-124.
5. *Сафонова Л.П.* Пространственно-частотный анализ форменных элементов крови / Диссертация на соискание уч. степени канд. техн. наук. МГТУ им. Н.Э. Баумана, Москва, 1998.
6. *Webster J.G.* Medical Instrumentation Application and Design. – Fourth Edition. Wiley and Sons, 2010.
7. *Радченко В.Г.* Основы клинической гематологии. – Спб.: Диалект, 2003.
8. *Воробьев А.И.* Клинико-диагностическое значение лабораторных показателей в гематологии / РАМН ГНЦ, 2001.
9. *Pierre R.V.* Peripheral Blood Film Review // Clinics in Laboratory Medecine. 2002. Vol. 22. № 1.
10. *Козинец Г.И.* Атлас клеток крови и костного мозга. – М.: Триада – Х, 1998.
11. *Самородов А.В.* Оценка качества цитологических препаратов // Биомедицинская радиоэлектроника. 2008. № 10. С. 39-45.
12. *Штадельманн Ж.В., Самородов А.В., Спиридонов И.Н.* Метод автоматизированного обнаружения лейкоцитов на изображениях мазков крови на основе бустинга // Медицинская техника. В публикации.
13. *Stadelmann J.V., Spiridonov I.N., Samorodov A.V.* Leukocyte Classification Based on Nucleus Skeletization / Proceedings of the 6th Russian-Bavarian Conference on Bio-Engineering, Moscow, 2010.
14. *Штадельманн Ж.В., Самородов А.В., Спиридонов И.Н.* Классификация лейкоцитов с использованием текстурных характеристик их ядер // Тезисы доклада «Медико-технические технологии на страже здоровья», Кипр, Ларнака, 2010.
15. *Штадельманн Ж.В., Спиридонов И.Н.* Метод определения формулы белой крови // Биомедицинская радиоэлектроника. В публикации.
16. *Meir R, Rätsch G.* An Introduction to Boosting and Leveraging. – Advanced lectures on machine learning. Springer-Verlag, New York, 2003.
17. *Viola P., Jones M.J.* Robust Real-Time Face Detection // International Journal of Computer Vision. 2004. Vol. 57 (2). PP. 137-154.
18. *Ji Z., Rosset S., Zou H. Hastie T.* Multi-Class Adaboost. – Department of Statistics, Standford University, 2006.
19. *Huang J., Ertekin S., Song Y., Hongyuan Z., Giles C.L.* Efficient Multiclass Boosting Classification with Active Learning / 7th SIAM International Conference, Society for Industrial and Applied Mathematics, Minneapolis, 2007.

Жоэль Валентин Штадельманн,
аспирант,
Игорь Николаевич Спиридонов,
д-р техн. наук, профессор,
зав. кафедрой,
кафедра биомедицинских технических систем,
МГТУ им. Н.Э. Баумана,
г. Москва,
e-mail: joel.stadelmann@gmail.com