УДК 519.237.8 + 576.08 DOI: 10.17587/it.26.56-61

А. И. Лепский, alexlep97@gmail.com,
Санкт-Петербургский государственный университет

Сравнительный анализ алгоритмов кластеризации лейкоцитов по FS и SS параметрам при цитофлуориметрическом исследовании крови

При проведении лабораторных исследований в биологии и медицине большое практическое значение имеет получение формальных правил для оценки численных значений эмпирических данных. Одной из нерешенных задач при цитомертическом исследовании крови является автоматическая типологизация лейкоцитов по размеру и сложности внутриклеточной структуры. Возможным подходом в этом случае может быть применение методов кластерного анализа. Однако при кластеризации белых клеток крови по указанным выше параметрам остается много нерешенных вопросов. В статье исследованы различные алгоритмы кластеризации. При проведении численных экспериментов было показано, что иерархические методы и метод K-средних не дают положительных результатов. Для дальнейшего изучения вопросов, связанных с автоматической типологизацией лейкоцитов крови, наиболее перспективным является метод DBSCAN. Для проведения численных экспериментов был создан программный код, написанный на языке Python.

Ключевые слова: *кластерный анализ, проточная цитометрия, машинное обучение*

За иммунитет организма, как способ его защиты от различных патогенных воздействий, отвечают лейкоциты (белые клетки крови). Их можно разделить на три основные группы: гранулоциты (содержащие крупные сегментированные ядра и имеющие специфическую зернистость цитоплазмы), лимфоциты и моноциты. Две последние группы клеток имеют простое несегментированное ядро и небольшую зернистость цитоплазмы [1].

Математическая иммунология как научная дисциплина начала формироваться во второй половине XX века. В 1964 г. Ф. Барнет и Н. Йерне сформулировали клонально-селекционную теорию иммунитета [2]. В начале 70-х гг. прошлого века были предложены первые математические модели иммунного ответа, построенные согласно этой теории. Они представляли собой системы четырех обыкновенных дифференциальных уравнений и описывали эволюцию В-лимфоцитов в ходе иммунного ответа [3,4]. Большую роль в развитии отечественной математической иммунологии сыграли работы академика Г. И. Марчука. Его модели включали до 15 обыкновенных дифференциальных уравнений [5, 6].

Из последних работ в данном направлении можно отметить систему, построенную С. Р. Кузнецовым, которая содержит не только обыкновенные дифференциальные уравнения, но и уравнения в частных производных первого порядка [7]. Модификации этой модели позволяют изучать динамику пролиферации и дифференцировки неоднородной клеточной популяции [8, 9], а также процессы формирования иммунной памяти Т-лимфоцитов [10].

Экспериментальной основой для построения математических моделей иммунного ответа являются результаты исследования крови. Самым эффективным подходом для получения таких данных сейчас является проточная цитофлуориметрия [11].

Проточный цитофлуориметр — прибор для измерения оптических свойств клеток. Детектор прямого светорассеяния (*forward scatter, FS)* определяет их размер. Детектор бокового светорассеяния (*side scatter, SS*) позволяет судить о сложности внутреннего строения клет-

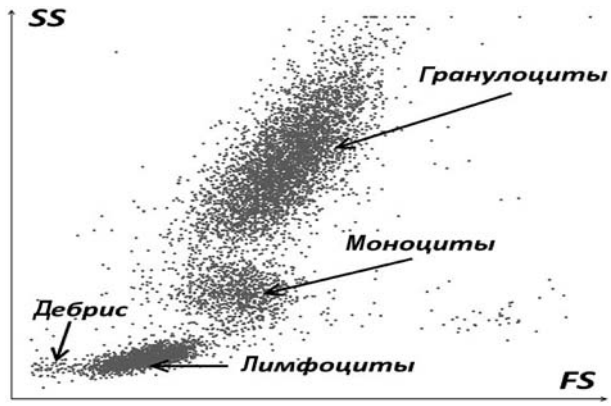

Рис. 1. Распределение лейкоцитов по размеру и сложности строения клеток

ки (соотношение между ядром и цитоплазмой, наличие гранул и других внутриклеточных включений). Используя *FS* и *SS* параметры, можно провести первичный анализ популяций лейкоцитов по размеру и сложности внутреннего строения клеток.

На рис. 1 изображена характерная картина распределения белых клеток крови в осях *FS* и *SS*. Лимфоциты являются самыми маленькими клетками с простым внутренним строением, моноциты крупнее и сложнее по внутреннему строению, гранулоциты являются самыми сложными и самыми большими лейкоцитами. Важно отметить еще одну группу точек, расположенную на рис. 1 ниже и левее всех остальных клеток, это так называемый дебрис (осколки клеток).

Кластерный анализ субпопуляций лейкоцитов

Нерешенной проблемой остается задача автоматической типологизации белых клеток крови, так как гейтирование различных групп (субпопуляций) лейкоцитов при цитометрическом исследовании проводится вручную [11, 12]. Согласно общепринятому соглашению: "Гейт (от англ. *gate* — ворота) — это инструмент для выделения из всего массива полученных данных отдельных популяций клеток, удовлетворяющих определенным условиям" [12, с. 31]. В настоящие время для создания гейтов используются программные средства, позволяющие оператору графически выделять группы клеток на экране консоли (выделение прямоугольного региона, выделение полигонального региона, выделение эллиптического региона) [12, с. 32], что в свою очередь обусловливает субъективность эксперимента и большие значения погрешностей при вычислении числа клеток в различных субпопуляциях.

В связи с этим возникает еще одна проблема. Доказательная медицина — это процесс систематического пересмотра, оценки и использования результатов клинических исследований в целях оказания оптимальной медицинской помощи пациентам, который сочетает в себе определенные принципы и методы. Благодаря их действию инструкции и стратегии в медицине основываются на текущих подтверждающих данных об эффективности разных форм лечения и медицинских услуг в целом. В числе прочего эти принципы предъявляют повышенные требования к качеству лабораторных исследований [13].

Неоднократно в различных медицинских учреждениях и у нас в стране и за рубежом проводился один и тот же эксперимент: нескольким исследователям предлагалось выполнить гейтирование клеток из одной и той же пробы, и всегда получались разные результаты, максимальная погрешность достигала 30 % [14—17].

Свой вклад в погрешность результатов обработки анализов крови вносят два вида ошибок: случайные и системные.

Случайные ошибки сопутствуют любому измерению, как бы тщательно оно не проводилось, и проявляются в некотором различии результатов измерения одного и того же элемента, выполненного данным методом. Они обусловлены в том числе точностью работы персонала лаборатории (неточное считывание результатов, ошибка утомления, неверный подбор класса точности инструментов, психологическая ошибка, например, оказание предпочтения каким-либо цифрам и т. д.)

Системные ошибки зависят от применяемых приборов и реактивов, определяются точностью приборов, происходят от неправильного или неточного выполнения операции, зависят от личных способностей оператора, его органов чувств, привычек и т. д.

Субъективный фактор имеет большое значение и при случайных, и при системных ошибках, поэтому для улучшения качества некоторых видов медицинских услуг необходима разработка машинных (формальных) методов гейтирования клеток крови. Возможным подходом к решению этой задачи может быть кластерный анализ результатов цитометрического исследования.

Произвольный алгоритм кластеризации является отображением

$$A : \begin{cases} X \to \mathbb{N}, \\ \bar{x}_i \mapsto k, \end{cases}$$

которое ставит в соответствие любому элементу $\bar{x}_i$ из некоторого множества X единственное натуральное число k, являющееся номером кластера, которому принадлежит $\bar{x}_i$. Процесс кластеризации разбивает X на попарно дизъюнктные подмножества X_h, которые называются кластерами:

$$X = \bigcup_{h=1}^{m} X_h,$$

где для $\forall h,\ l \mid 1 \leqslant h,\ l \leqslant m$: $X_h \cap X_l = \varnothing$.

Отображение A задает на X отношение эквивалентности. В качестве независимых представителей классов эквивалентности выбирают элементы, называемые центроидами. В n-мерном евклидовом пространстве E^n координаты центроидов равны среднему арифметическому соответствующих координат всех элементов (векторов), входящих в кластер (класс эквивалентности). Если отождествить каждый вектор из E^n с материальной точкой единичной массы, то центроиды можно рассматривать как центры масс кластеров [18].

Число кластеров в некоторых случаях известно заранее, однако чаще всего ставится задача определить оптимальное число кластеров с точки зрения того или иного критерия качества кластеризации.

Для оценки качества кластеризации будем использовать *Silhouette Coefficient* [19, 20]. Коэффициент силуэта не предполагает знания истинных меток объектов и позволяет оценить качество кластеризации, используя только саму (неразмеченную) выборку и результат кластеризации. Сначала силуэт определяется отдельно для каждого объекта. Обозначим a — среднее расстояние от данного объекта до объектов из того же кластера, b — среднее расстояние от данного объекта до объектов из ближайшего кластера (отличного от того, в котором лежит сам объект). Тогда силуэтом данного объекта называется величина

$$s = \frac{b - a}{\max(a, b)}.$$

Силуэтом выборки называется среднее значение силуэта объектов данной выборки.

С помощью силуэта можно выбирать оптимальное число кластеров (если оно заранее неизвестно) — выбирается число кластеров, максимизирующее значение силуэта. Силуэт зависит от формы кластеров и достигает больших значений на более выпуклых кластерах, получаемых с помощью алгоритмов, основанных на восстановлении плотности распределения.

В статье сравниваются результаты численных экспериментов кластеризации точек на евклидовой плоскости с помощью четырех различных методов, с учетом специфики распределения параметров лейкоцитов крови в осях *FS* и *SS*.

Численное моделирование проводили с помощью программного кода, написанного на языке программирования *Python* 3.7.2. В качестве интегрированной среды разработки использовали оболочку *PyCharm*, разработанную для языка программирования *Python* компанией *JetBrains* на основе *IntelliJ IDEA*. *PyCharm* предоставляет средства для анализа кода, графический отладчик, инструмент для запуска юнит-тестов. Для выполнения процедур, реализующих алгоритмы кластеризации, подключались библиотеки *sklearn* и *scipy*. Предварительную стандартизацию данных не проводили, так как оба признака расположены примерно в одном диапазоне. Подбор оптимальных параметров осуществлялся по сетке $k = [1, ..., n]$ (где n — общее число точек) для метода одиночной связи, K-средних и *EM*, и по сетке $n = [1, ..., 10]$; $e = [1, ..., 100]$ (число соседей и радиус) для метода *DBSCAN*. Оптимальным результатом считался тот, который был получен благодаря параметрам, максимизирующим значение коэффициента силуэта.

Метод одиночной связи

Под иерархическими методами кластеризации понимается группа алгоритмов, направленных на построение дерева вложенных кластеров, новые классы эквивалентности создаются путем объединения более мелких кластеров и, таким образом, их дерево формируется от листьев к стволу.

Сначала представим результаты для метода одиночной связи [21—23], где за расстояние между кластерами принимается дистанция между центроидами. Численное моделирование процесса кластеризации лейкоцитов проводили с помощью процедуры *sklearn.cluster.AgglomerativeClustering*. На рис. 2 (см. четвертую сторону обложки) представлена типичная картина формирования кластеров этим методом.

Среднее время работы программы составляет 8,64 с. Этот алгоритм удовлетворительно показал себя на самых разнообразных данных. Из недостатков стоит отметить, что приемлемый результат достигается при достаточно большом числе кластеров. Поэтому помимо трех классов, выделяющих лимфоциты, моноциты и гранулоциты, возникает огромное число мелких кластеров, что существенно ухудшает восприятие результатов типологизации. Возможно, что именно по этой причине не всегда удается достоверно отделять дебрис от лимфоцитов. Еще одной проблемой является определение оптимального критерия остановки при построении иерархического дерева или, иными словами, выбор оптимального числа кластеров.

Метод K-средних

Алгоритм *K-means* — популярный метод кластеризации, изобретенный во второй половине прошлого века [24, 25]. Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге (на первом шаге нужно знать число кластеров и выбрать расположение их центров масс). Затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния. Моделирование проводили с помощью процедуры *sklearn.cluster.KMeans.* Пример результатов работы процедуры, реализующей метод K-средних, приведен на рис. 3 (см. четвертую сторону обложки).

Среднее время работы программы 3,71 с. Как видно из рис. 3, метод K-средних плохо показал себя на тестовых данных. Ни один из результатов численного моделирования не является удовлетворительным из-за того, что алгоритм разделяет гранулоциты пополам. Возможно, это связано с тем, что метод сходится к локальному минимуму, игнорируя глобальный. Также одним из недостатков алгоритма является то, что итоговые метки классов зависят от первоначального выбора центроидов.

EM-алгоритм

Если *a priori* известно что данные представляют собой смесь нормальных распределений с неизвестными параметрами, то их кластеризация может проводиться с помощью *EM*-алгоритма [26]. Каждая итерация алгоритма состоит из двух шагов. На *E*-шаге для каждого объекта вычисляется вероятность его принадлежности к кластеру. На *M*-шаге вычисляются параметры нормального распределения и решается задача максимального правдоподобия. Затем определяется кластерная принадлежность для каждого объекта. Численное моделирование проводили с помощью процедуры *sklearn.mixture.GaussianMixture.* Результаты работы программного кода, реализующего эту процедуру, приведены на рис. 4 (см. четвертую сторону обложки).

Среднее время работы программы составляет 3,32 с. Как видно из рисунка, этот алгоритм показал себя чуть лучше, чем метод K-средних, однако результат все же не является удовлетворительным. *EM*-алгоритм имеет схожие недостатки с методом K-средних (чувствительность к начальным данным и необходимость задавать число кластеров до начала процесса).

DBSCAN

Последним рассмотрим метод кластеризации, основанный на оценке плотности распределения экспериментальных данных, *DBSCAN (Density-based spatial clustering of applications with noise*) [27]. Этот алгоритм группирует плотно расположенные элементы, помечая как выбросы те точки, которые находятся в областях с малой плотностью. В отличие от метода K-средних и *EM*-алгоритма, *DBSCAN* не требует указывать число кластеров заранее. Предварительно нужно лишь задать параметры, определяющие радиус окрестности и число соседних точек, попадающих в окрестность указанного радиуса. Анализ проводили с помощью процедуры *sklearn.cluster.DBSCAN.* Пример результатов работы алгоритма *DBSCAN* показан на рис. 5 (см. четвертую сторону обложки).

Среднее время работы программы 4,05 с. В целом алгоритм хорошо кластеризовал тестовые данные. Однако у него тоже есть свои недостатки. Во-первых, *DBSCAN* не может хорошо кластеризовать наборы данных с большой разницей в плотности, поскольку не удается выбрать приемлемую для всех кластеров комбинацию "число соседей" и "радиус". Во-вторых, нет объективного критерия выбора оптимальных параметров, и они подбирались вручную.

Заключение

При численном моделировании процесса кластеризации лейкоцитов по *FS* и *SS* параметрам хуже всего проявил себя метод K-средних. Разделение гранулоцитов на два кластера является ошибочным в принципе. Несущественно лучше результаты кластерного анализа с помощью *EM*-алгоритма. К тому же оба метода предполагают априорное знание числа кластеров.

В целом иерархический алгоритм одиночной связи позволяет получать удовлетворительные результаты, но не более того. Остается открытой проблема завершения процесса кластеризации и определения предпочтительного числа кластеров. Не удается достоверно отделять лимфоциты от мелкого дебриса.

Лучшие результаты были получены при использовании метода *DBSCAN*. Однако при наличии большого объема шумов или большой разницы в плотности распределения лейкоцитов подбор параметров, необходимых для выполнения этого алгоритма, вызывает большие затруднения. Поэтому возникают задачи, связанные с оценкой устойчивости и робастности результатов кластеризации, полученных этим методом.

Основной вывод из проведенных вычислительных экспериментов состоит в том, что необходимо модифицировать существующие алгоритмы так, чтобы они давали хорошие результаты вне зависимости от структуры данных и уровня шумов. Возможным подходом к решению перечисленных выше задач может быть доработка методов *DBSCAN* и одиночной связи с учетом специфики распределения лейкоцитов в осях *FS-SS*.

Список литературы

1. **Хаитов Р. М., Игнатьева Г. А., Сидорович И. Г.** Иммунология: Учебник. М.: Медицина, 2000. 432 с.

2. **Аронова Е. А.** Иммунитет. Теория, философия и эксперимент: Очерки из истории иммунологии XX века. М.: КомКнига, 2006. 160 с.

3. **Смирнова О. А., Степанова Н. В.** Математическая модель колебаний при инфекционном иммунитете // Колебательные процессы в биологических и химических системах: Тр. Второго Всесоюзного симпозиума по колебательным процессам в биологических и химических системах. Пущино-на-Оке: НЦБИ АН СССР, 1971. Т. 2. С. 247—251.

4. **Bell G. I.** Mathematical model of clonal selection and antibody production // J. Theor. Biol. 1970. Vol. 29, N. 2. P. 191—232.

5. **Марчук Г. И.** Математические модели в иммунологии: вычислительные методы и эксперименты. М.: Наука, 1991. 299 с.

6. **Бочаров Г. А., Марчук Г. И.** Прикладные проблемы математического моделирования в иммунологии // Журн. вычисл. математики и матем. физики. 2000. Т. 40, № 12. С. 1905—1920.

7. **Кузнецов С. Р.** Математическая модель иммунного ответа // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2015. № 4. С. 72—87.

8. **Kuznetsov S. R., Kudryavtsev I. V., Orekhov A. V., Polevshchikov A. V., Serebriakova M. K., Shishkin V. I.** A mathematical model for predicting of IGD—CD27 + B lymphocytes levels in donors' blood // Биоинформатика регуляции и структуры геномов и системной биологии. Новосибирск. 2016. С. 166.

9. **Кузнецов С. Р., Лыкосов В. М., Орехов А. В., Шишкин В. И.** Модель пролиферации и дифференцировки неоднородной клеточной популяции // Математическое и компьютерное моделирование в биологии и химии III Международная научная Интернет-конференция. 2014. С. 95—103.

10. **Кузнецов С. Р., Лыкосов В. М., Орехов А. В., Шишкин В. И.** Процесс формирования иммунной памяти Т-лимфоцитов и его зависимость от числа пройденных клетками делений // Устойчивость и процессы управления. Материалы III международной конференции. 2015. С. 487—488.

11. **Зурочка А. В., Хайдуков С. В., Кудрявцев И. В., Черешнев В. А.** Проточная цитометрия в медицине и биологии. 2-е изд. Екатеринбург: УрО РАН, 2014. 574 с.

12. **Балалаева И. В.** Проточная цитофлуориметрия: Учебно-методическое пособие. Нижний Новгород: Нижегородский госуниверситет, 2014. 75 с.

13. **Основы** доказательной медицины: Учебное пособие для системы послевузовского и дополнительного профессионального образования врачей / Под общей редакцией академика РАМН, профессора Р. Г. Оганова. М.: Силицея-Полиграф, 2010. 136 с.

14. **Pedersen N. W.** Automated Analysis of Flow Cytometry Data to Reduce Inter-Lab Variation in the Detection of Major Histocompatibility Complex Multimer-Binding T Cells // Front Immunol 8: 2017. p. 858.

15. **Daneau G.** CD4 results with a bias larger than hundred cells per microliter can have a significant impact on the clinical decision during treatment initiation of HIV patients // Cytometry B Clin Cytom 92(6): 2017. P. 476—484.

16. **Qian Yu.** FlowGate: towards extensible and scalable web-based flow cytometry data analysis. XSEDE, 2015.

17. **Omana-Zapata I.** Accurate and reproducible enumeration of T-, B-, and NK lymphocytes using the BD FACSLyric 10-color system: A multisite clinical evaluation // PLoS One 14(1): 2019. e0211207.

18. **Орехов А. В.** Марковский момент остановки агломеративного процесса кластеризации в евклидовом пространстве // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2019. Т. 15, Вып. 1. С. 76—92.

19. **Rousseeuw P. J.** "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics.

20. **Amorim R. C., Hennig C. (2015).** "Recovering the number of clusters in data sets with noise features using feature rescaling factors" // Information Sciences. 324: 126—145.

21. **Everitt B. S.** Cluster Analysis. Chichester, West Sussex, UK: John Wiley & Sons Ltd, 2011. 330 p.

22. **Hartigan J. A.** Clustering algorithms. New York, London, Sydney, Toronto: John Wiley & Sons Inc. Press, 1975. 351 p.

23. **Aldenderfer M. S., Blashfield R. K.** Cluster analysis. Newburg Park, Sage Publications Inc. Press, 1984. 88 p.

24. **Steinhaus H.** Sur la division des corps materiels en parties // Bull. Acad. Polon. Sci. C1. III. 1956. Vol. IV. P. 801—804.

25. **Lloyd S.** Least squares quantization in PCM // IEEE Transactions on Information Theory. 1982. Vol. 28, Iss. 2. P. 129—137. doi:10.1109/TIT.1982.1056489.

26. **Dempster A. P., Laird N. M., Rubin D. B.** Maximum Likelihood from Incomplete Data via the EM Algorithm // Journal of the Royal Statistical Society, Series B. 1977. Vol. 39 (1). P. 1—38.

27. **Ester M., Kriegel H.-P., Sander J., Xu. X.** A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. AAAI Press, 1996. P. 226—231.

A. I. Lepsky, alexlep97@gmail.com,
Saint Petersburg State University, Saint Petersburg, ..., Russian Federation

Comparative Analysis of Leukocyte Clustering Algorithms According to FS and SS Parameters in a Cytofluorimetric Blood Test

In laboratory studies in biology and medicine, obtaining formal rules for evaluating the numerical values of empirical data is of great practical importance. One of the unsolved problems in the cytometric blood test is the automatic typology of leukocytes. A possible approach, in this case, could be the use of cluster analysis methods. However, with the clustering of white blood cells, many unresolved issues remain. The article explores various clustering algorithms. When conducting numerical experiments, it was shown that hierarchical methods and the K-means method do not give positive results. The DBSCAN method is the most promising for further study. A program code written in Python was created to conduct numerical experiments.

Keywords: *cluster analysis, flow cytometry, machine learning*

DOI: 10.17587/it.26.56-61

References

1. **Haitov R. M., Ignat'eva G. A., Sidorovich I. G.** Immunologiya: Uchebnik, Moscow, Medicina, 2000, 432 p. (in Russian).

2. **Aronova E. A.** Immunity. Theory, Philosophy, and Experiment: Essays on the History of Immunology of the 20th Century, Moscow, KomKniga, 2006, 160 p. (in Russian).

3. **Smirnova O. A., Stepanova N. V.** *Kolebatel'nye processy v biologicheskih i himicheskih sistemah: Trudy Vtorogo Vsesoyuznogo simpoziuma po kolebatel'nym processam v biologicheskih i himicheskih sistemah*. Pushchino-na-Oke: NCBI AN SSSR, 1971, vol. 2, pp. 247—251 (in Russian).

4. **Bell G. I.** *J. Theor. Biol*., 1970, vol. 29, no. 2, pp. 191—232.

5. **Marchuk G. I.** Mathematical models in immunology: computational methods and experiments, Moscow, Nauka, 1991, 299 p. (in Russian).

6. **Bocharov G. A., Marchuk G. I.** *ZHurn. vychisl. matematiki i matem. Fiziki*, 2000, vol. 40, no. 12, pp. 1905—1920 (in Russian).

7. **Kuznecov S. R.** *Vestnik Sankt-Peterburgskogo universiteta. Prikladnaya\linebreak matematika. Informatika. Processy upravleniya*, 2015, no. 4, pp. 72—87 (in Russian).

8. **Kuznetsov S. R., Kudryavtsev I. V., Orekhov A. V., Polevshchikov A. V., Serebriakova M. K., Shishkin V. I.** *Bioinformatika regulyacii i struktury genomov i sistemnoj biologii*, Novosibirsk, 2016, p. 166.

9. **Kuznecov S. R., Lykosov V. M., Orekhov A. V., Shishkin V. I.** *Matematicheskoe i komp'yuternoe modelirovanie v biologii i himii III Mezhdunarodnaya nauchnaya Internet-konferenciya*, 2014, pp. 95—103 (in Russian).

10. **Kuznecov S. R., Lykosov V. M., Orekhov A. V., Shishkin V. I.** *Ustojchivost' i processy upravleniya. Materialy III mezhdunarodnoj konferencii,* 2015, pp. 487—488 (in Russian).

11. **Zurochka A. V., Hajdukov S. V., Kudryavcev I. V., Chereshnev V. A.** Flow cytometry in medicine and biology, Ekaterinburg: UrO RAN, 2014, 574 p. (in Russian).

12. **Balalaeva I.V** Flow cytofluorimetry: a teaching aid, Nizhnij Novgorod, Nizhegorodskij gosuniversitet, 2014, 75 p. (in Russian).

13. **Oganov R. G.** ed. The basics of evidence-based medicine. Textbook for postgraduate and continuing professional education of doctors, Moscow, Siliceya-Poligraf, 2010, 136 p. (in Russian).

14. **Pedersen N. W.** Automated Analysis of Flow Cytometry Data to Reduce Inter-Lab Variation in the Detection of Major Histocompatibility Complex Multimer-Binding T Cells, *Front Immunol* 8: 2017, p. 858.

15. **Daneau G.** CD4 results with a bias larger than hundred cells per microliter can have a significant impact on the clinical decision during treatment initiation of HIV patients, *Cytometry B Clin Cytom* 92(6): 2017, pp. 476—484.

16. **Qian Yu.** FlowGate: towards extensible and scalable web-based flow cytometry data analysis, XSEDE, 2015.

17. **Omana-Zapata I.** Accurate and reproducible enumeration of T-, B-, and NK lymphocytes using the BD FACSLyric 10-color system: A multisite clinical evaluation, *PLoS One* 14(1): 2019, e0211207.

18. **Orekhov A. V.** *Vestnik Sankt-Peterburgskogo universiteta. Prikladnaya matematika. Informatika. Processy upravleniya*, 2019, vol. 15, iss. 1, pp. 76—92 (in Russian).

19. **Rousseeuw P. J.** Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis, Computational and Applied Mathematics.

20. **Amorim R. C., Hennig C.** Recovering the number of clusters in data sets with noise features using feature rescaling factors". Information Sciences, 2015, 324: 126—145.

21. **Everitt B. S.** Cluster Analysis. Chichester, West Sussex, UK, John Wiley & Sons Ltd, 2011, 330 p.

22. **Hartigan J. A.** Clustering algorithms. New York, London, Sydney, Toronto, John Wiley & Sons Inc. Press, 1975, 351 p.

23. **Aldenderfer M. S., Blashfield R. K.** Cluster analysis. Newburg Park, Sage Publications Inc. Press, 1984, 88 p.

24. **Steinhaus H.** Sur la division des corps materiels en parties, *Bull. Acad. Polon. Sci.* C1. III. 1956, vol. IV, pp. 801-804.

25. **Lloyd S.** Least squares quantization in PCM, *IEEE Transactions on Information Theory*, 1982, vol. 28, iss. 2, pp. 129—137, doi:10.1109/TIT.1982.1056489

26. **Dempster A. P., Laird N. M., Rubin D. B.** Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B,* 1977, vol. 39 (1), pp. 1—38.

27. **Ester M., Kriegel H.-P., Sander J., Xu. X.** A density-based algorithm for discovering clusters in large spatial databases with noise, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* / Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. AAAI Press, 1996, pp. 226—231.

Рисунки к статье А. И. Лепского

«СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ЛЕЙКОЦИТОВ ПО FS И SS ПАРАМЕТРАМ ПРИ ЦИТОФЛУОРИМЕТРИЧЕСКОМ ИССЛЕДОВАНИИ КРОВИ»

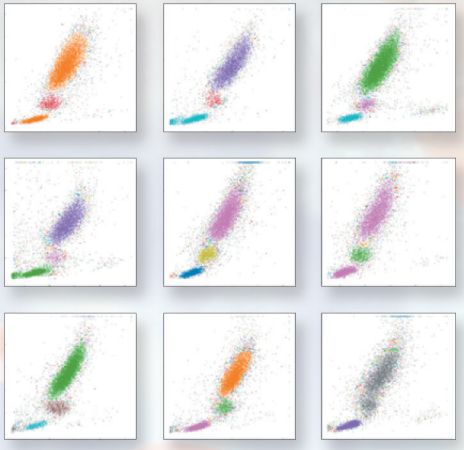

Рис. 2. Результаты численного моделирования процесса кластеризации лейкоцитов методом одиночной связи

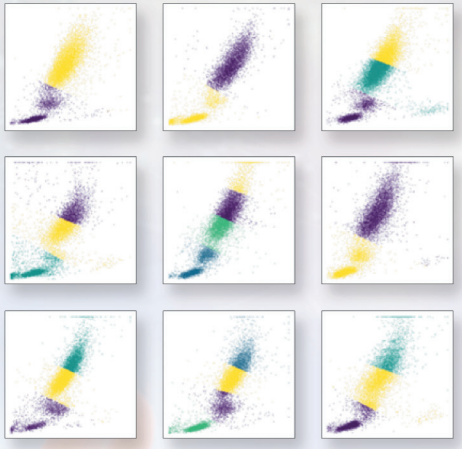

Рис. 3. Результаты численного моделирования методом К-средних

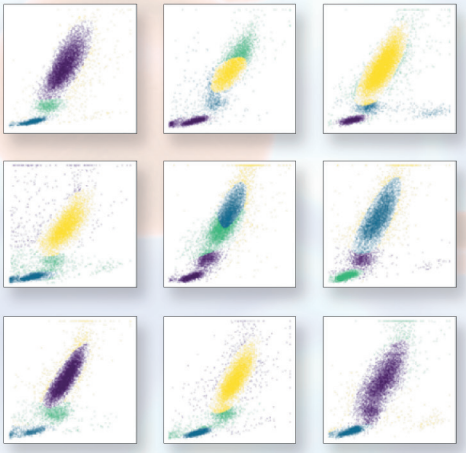

Рис. 4. Результаты кластеризации с помощью *EM*-алгоритма

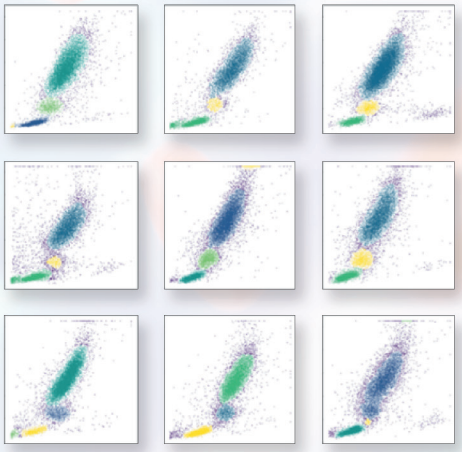

Рис. 5. Результаты кластеризации лейкоцитов методом *DBSCAN*

ISSN 1684-6400. Информационные технологии. 2020. Том 26. № 1. 1–64. Индекс: 72656