МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Факультет Механико-математический факультет
Кафедра Математического моделирования

Направление подготовки Математическое моделирование

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

Можейко Евгений Александрович

Тема работы: Определение погрешностей характеристик объекта при решении обратной задачи светорассеяния с использованием кластеризации.

«К защите допущена»
Заведующий кафедрой,
д.ф.-м.н., проф.

Ковеня В.М. /_____
(фамилия , И., О.) / (подпись, МП)

«…»……20…г.

Научный руководитель
к.ф-м.н.,
ст.н.с. ИХКГ СО РАН

Юркин М.А. /_____
(фамилия , И., О.) / (подпись, МП)

«…»……20…г.

Дата защиты: «…»……20…г.

Новосибирск, 2016

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В лаборатории Цитометрии и Биокинетики института химической кинетики и горения разрабатывается прибор для измерения морфологических параметров (характеристик) клеток крови в потоке. Ядром прибора является сканирующий проточный цитометр. Его принцип работы основан на измерении индикатрис светорассеяния одиночных частиц в потоке со скоростью около ста частиц в секунду. В данном случае индикатриса светорассеяния представляет из себя вектор размерности 61, компонентами которого являются интенсивности рассеяния в направлении от 10° до 60° относительно направления падения электромагнитной волны [1-4].

Частицам крови сопоставляется приближенная физическая модель. Для определенного класса элементов крови существует своя модель. Например, для тромбоцитов это сплюснутый сфероид [1]. По индикатрисе светорассеяния частицы определяют четыре морфологических параметра модели. Однако вычисление даже при более простой модели частицы, например такой как сфера, с помощью метода нелинейной регрессии [3,5] занимает несколько секунд, что не сравнимо с потоком сто частиц в секунду.

По параметрам модели частицы так же можно определить индикатрису светорассеяния. Как вариант ускорения было предложено создать базу данных индикатрис [6]. Параметры моделей, соответствующих

индикатрисам из базы данных, выбирались как случайные величины, принадлежащие равномерному распределению по диапазонам параметров клеток, реально встречающихся в крови человека. Перебором ищется ближайшая к экспериментальной индикатриса по определенной мере с весами.

$$\|I\| = \sqrt{\sum_{i=10}^{i=60} w^2(i) \cdot I^2(i)} ,$$

I – индикатриса,

$w(\theta) = \frac{1^\circ}{\theta} \exp[-2\ln^2(\theta/54^\circ)]$ –весовая функция, которая примерно соответствует структуре экспериментальных погрешностей [3]. Далее везде используется эта норма. Заранее известные параметры ближайшей индикатрисы принимаются за параметры экспериментальной.

Кроме нахождения ближайшей индикатрисы, Баесовским подходом на пространстве параметров для экспериментальной индикатрисы считается плотность вероятности того, что ближайшая индикатриса имеет параметры элемента из базы данных. Далее методом Монте-Карло считается математическое ожидание и дисперсия. Это дает возможность оценить экспериментальную индикатрису, не только найдя ближайшую по базе данных, но и проанализировав математическое ожидание и дисперсию от каждого параметра, тем самым оценив интервал в пространстве параметров в котором вероятней всего находится реальные параметры

измеренной частицы. Данный интервал дает представление о погрешности определения параметров частицы по базе данных [1].

Вышеупомянутый способ нахождения параметров с использованием заранее насчитанной базы данных при программной реализации по времени занимает около секунды. В качестве ускорения было предложено предварительно обработать базу данных для нахождения ближайшей индикатрисы, чтобы было меньшее количество сравнений без потери точности. Для этой задачи был использован метод Кластеризации [7]. Но данный метод не учитывал возможность подсчета математического ожидания и дисперсии. Это является существенным упущением данного метода.

ПОСТАНОВКА ЗАДАЧИ

1. КЛАСТЕРИЗАЦИЯ

1.1. ВИД КЛАСТЕРИЗОВАНОЙ БАЗЫ ДАННЫХ

Пусть $I = \{I_{\beta_1}, \dots, I_{\beta_N}\}$ – множество индикатрис базы данных. Кластер W_i имеет следующий вид: $\{T_i, C_i, R_i\}$. Где T_i либо множество кластеров , либо множество элементов базы данных. Обозначим через $P(W_i)$ множество всех элементов базы данных принадлежащих подкластерам W_i. Кластерная структура базы данных $W_0 = \{T_0, C_0, R_0\}$, $P(T_0) = I$. Обозначим через $V_{i_k} \neq \emptyset \;\; k = 1, \dots, M$ кластеры из кластерной структуры, у которых T_{i_k} состоит из элементов базы данных. Тогда:

$$C_{\mathrm{i}} = \frac{\sum_{I_{\beta_k} \in P(W_i)} I_{\beta_k}}{n_i}$$

$$\mathrm{R_i} = \max_{I_{\beta_k} \in V_i} \left\| C_i - I_{\beta_k} \right\|$$

C_{i} называется центром кластера, $\mathrm{R_i}$ радиусом кластера. n_i количество элементов в $\mathrm{P}(W_i)$

Введем $J = \{i_1, \dots, i_M\}$. Кластерная структура построена таким образом, что для них выполнятся следующее:

1. $\bigcup_{k=1}^{M} V_{i_k} = \mathrm{I}$
2. $V_{i_k} \bigcap V_{i_l} = \emptyset \quad k \neq l$.

Получается своеобразное дерево, где каждому кластеру соответствует центр, радиус и либо только кластеры, если это ветвь, либо только элементы, если это лист.

1.2. ПОИСК БЛИЖАЙШЕЙ

1. В кластере ищется минимальный шар D (в пространстве индикатрис) с центром в I_{exp} такой , что шар соответствующий одному из подкластеров, полностью лежит в нем.

2. Все подкластеры лежащие дальше радиуса шара D, исключаются из рассмотрения.

3. Остальные подкластеры рассматриваются. То есть либо перебор по всем элементам подкластера, если элементы подкластера индикатрисы, либо, для каждого такого подкластера, возвращаемся к пункту 1 [7].

Приближенное представление (Кластер 3 отбрасывается):

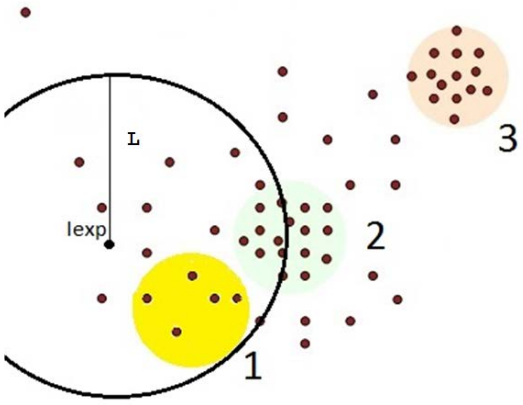

Рис 1.

Таким образом гарантированно находится ближайшая без потери точности относительно полного перебора. Само разбиение на кластеры производится по алгоритму FOREL [8]. При использовании кластеризации количество сравнений с элементами базы данных существенно уменьшается, и добавляется незначительное количество сравнений с центрами кластеров.

2. ПОДСЧЕТ МАТЕМАТИЧЕСКОГО ОЖИДАНИЯ ДЛЯ ПОЛНОГО ПЕРЕБОРА

1. Баесовским подходом определяется плотность вероятности P(β):

$$S(\beta_i) = \left\| I_{exp} - I_{\beta_i} \right\|^2$$

$$P(\beta) = \frac{1}{k} S(\beta_i)^{-k_{eff}/2}, \qquad k = [\int_{B} S(\beta_i)^{-k_{eff}/2} d\beta]$$

B – пространство параметров.

k_{eff} - эффективное число степеней свободы [1].

I_{β_i} - i-ая индикатриса из базы данных.

I_{exp} - экспериментальная индикатриса.

$\beta_i = (\beta_i^1, \beta_i^2, \beta_i^3, \beta_i^4)$ четырехмерный вектор параметров i-ого элемента базы данных.

2. Методом Монте-Карло считается математическое ожидание:

$$\langle f(\beta)\rangle = \int_B f(\beta)P(\beta)d\beta$$

$$\langle f(\beta)\rangle \approx \frac{\overline{\boldsymbol{g}}}{\overline{\boldsymbol{h}}}, \qquad \bar{\boldsymbol{x}} \equiv \frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{x_i}, \qquad \boldsymbol{g_i} = f\left(\beta_i\right)\boldsymbol{h_i}, \qquad \boldsymbol{h_i} = Np_i S(\beta_i)^{-\boldsymbol{k_{eff}/2}}$$

$f(\beta_i)$ – произвольная функция параметров.

N – размер базы данных.

p_i=1/N в случае если сгенерированные элементы базы данных принадлежат равномерному распределению.

Очевидно, в этих формулах требуются расстояния от экспериментальной индикатрисы до всех элементов из базы данных. Однако при использовании кластеризации сравнений получается примерно в шесть раз меньше. Следовательно возникает недостаток данных для вычисления.

АЛГОРИТМ РЕШЕНИЯ

1. **АЛГОРИМ РЕКУСИВНОГО ВЫЧИСЛЕНИЯ БЛИЖАЙШЕЙ И ТРЕБУЕМЫХ РАССТОЯНИЙ** (*):

   L_p- текущий показатель принятия кластера.

   L_{min} - текущее минимальное расстояние от элемента из базы данных до I_{exp} из подсчитанных.

   I_0 – индикатриса расстояние которой до I_{exp} равно L_{min}.

   Начинаем с W_0.

1. Для кластера W_i: если $i \in J$ переходим к пункту 1.1 иначе к 1.2

1.1. Если $\min_{I_{\beta_s} \in W_i}\{\|I_{exp} - I_{\beta_s}\|\} < L_{min}$ обновляем значение L_{min} и I_0 , иначе ничего.

1.2. Если $L_{\text{p}} > \min\limits_i\{\|C_i - I_{exp}\| + R_i\}$, то $L_{\text{p}} := \min\limits_i\{\|C_i - I_{exp}\| + R_i\}$. В любом случае, переходим к пункту 1.2.1.

1.2.1. Для всех кластеров W_{i_j} $j = 1, \dots, M_i$ принадлежащих T_i если $\max\{\|C_{i_j} - I_{exp}\| - R_{i_j}, 0\} < L_p$ то кластер W_{i_j} переходит на рассмотрение в пункт 1 , иначе ничего.

2. **ПОДСЧЕТ МАТЕМАТИЧЕСКОГО ОЖИДАНИЯ С УСЛОВИЕМ КЛАСТЕРИЗАЦИИ**

2.1. Теоретическое предположение.

В алгоритме кластеризации FOREL используется фиксированный радиус для всех кластеров находящихся на одном иерархическом уровне. Отброшенные кластеры, как правило, находятся далеко от экспериментальной индикатрисы. Было сделано предположение, что если принять расстояния до всех элементов отброшенного кластера равными расстоянию до центра кластера, то значения математического ожидания при полном переборе и с использованием кластеризации будут незначительно различаться. Предположение подкреплялось тем, что радиус отброшенного кластера небольшой относительно расстояния от его центра до экспериментальной индикатрисы. Формулы для математического ожидания, согласованные с этим предположением, принимают следующий вид:

$$\langle f(\beta)\rangle \approx \frac{\overline{\boldsymbol{g}}}{\overline{\boldsymbol{h}}}, \qquad \bar{\boldsymbol{x}} \equiv \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{x_i}, \qquad \boldsymbol{g_i} = f\left(\beta_i\right)\boldsymbol{h_i},$$

$$\boldsymbol{h_i} = \begin{cases} \boldsymbol{N p_i} S(\beta_i)^{-\boldsymbol{k_{eff}/2}}, & \textit{если } i \in \boldsymbol{Q} \\ \boldsymbol{N p_i}\left\|C_j - I_{exp}\right\|^{-\boldsymbol{k_{eff}/2}}, & \textit{если } \beta_i \in \boldsymbol{P}(\boldsymbol{T_j}), \textit{где } \boldsymbol{j} \in \boldsymbol{J} \end{cases}$$

$Q -$ множество индексов i, для которых для I_{β_i} подсчитано расстояние до I_{exp}.

$J -$ множество индексов j таких, что W_j отброшенный кластер.

2.2. Алгоритм подсчета математического ожидания (**):

1. В алгоритме (*) Считаются $N_1 < N$ расстояний. $N_1 \approx \frac{N}{6}$.

2. Далее эти расстояния используются в формуле (**).

Время подсчета самой формулы (**) не вносит значимый вклад в длительность исполнения всего алгоритма подсчета погрешностей параметров.

В данном подходе не требуется больше информации, чем мы получили при поиске ближайшей с использованием кластеризации. В алгоритме поиска ближайшей параллельно вычисляются математические ожидания для $f(\beta_i) = \beta_i^j$ f(β_i)=1 и f(β_i)= $\beta_i^j \beta_i^j$ $j = 1, \dots, 4.$, $\beta_i = (\beta_i^1, \beta_i^2, \beta_i^3, \beta_i^4)$ [1].

Программа, реализующая данный алгоритм, была написана на языке LabView.

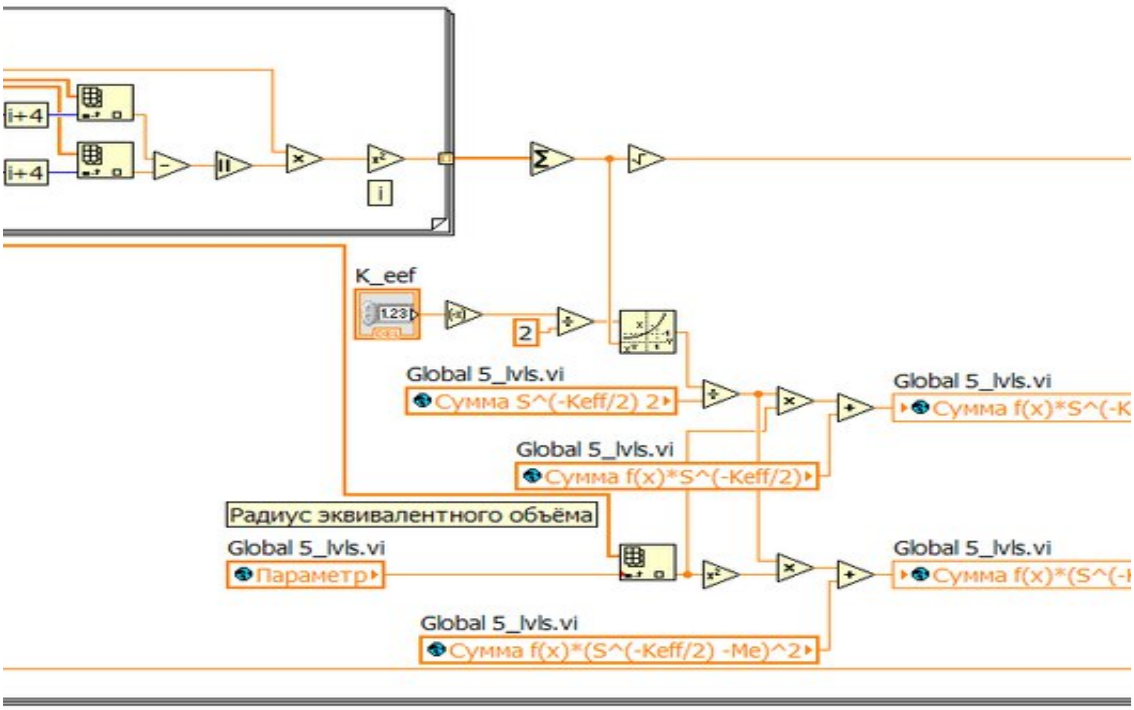

Рис 2.

ТЕСТИРОВАНИЕ

Задача решалась для тромбоцитов с моделью сплюснутый сфероид. База данных размера 65909. База данных была кластеризована иерархическим образом. В итоге получилась глубина кластерной структуры пять уровней. Были проведены успешные тесты работоспособности программы поиска ближайшей с использованием кластеризации. Программа была апробирована на тысяче тестовых индикатрисах. Результат поиска ближайшей всегда совпадал с полным перебором. Проведены тесты среднего количества сравнений для 1000 экспериментальных индикатрис.

Среднее число сравнений для 1000 индикатрис: 11573

Показателем близости математического ожидания и дисперсии для поиска полным перебором и поиска с использованием кластеризации выступали следующие формулы:

$$P_{ME} = \frac{\sum_{i=1}^{100} \frac{\left|ME_i^0 - ME_i\right|}{\sqrt{{ME_i^0}^2 + {ME_i}^2}}}{1000} \qquad P_{SD} = \frac{\sum_{i=1}^{100} \frac{\left|SD_i^0 - SD_i\right|}{\sqrt{{SD_i^0}^2 + {SD_i}^2}}}{1000}$$

$ME_i -$ математическое ожидание для i-ого эксперимента с использованием кластеризации.

$ME_i^0 -$ математическое для i-ого эксперимента при полном переборе.

$SD_i^{0^2}$ – дисперсия для i-ого эксперимента при полном переборе.

$SD_i{}^2$ – дисперсия для i-ого эксперимента с использованием кластеризации.

P_{ME} и P_{SD} считались для каждого из четырех параметров, т.е. $f(\beta_i) = \beta_i^j$.

j $= 1, \ldots, 4$, $\beta_i = (\beta_i^1, \beta_i^2, \beta_i^3, \beta_i^4)$.

	P_{ME}	P_{SD}
Радиус шара эквивалентного объема.	0,08	0,87
Соотношение полуосей.	0,10	0,50
Показатель преломления.	0,03	0,77
Угол ориентации.	0,11	0,62

ЗАКЛЮЧЕНИЕ

Был разработан алгоритм вычисления погрешностей параметров при решении обратной задачи с помощью кластеризации. Написана программа на языке LabView, предназначающаяся как модуль для использования программой, работающей непосредственно со сканирующим проточным цитометром. Она выполняет поиск ближайшего элемента базы данных за меньшее число сравнений. Число сравнений выступает как главный показатель длительности работы программы. С уменьшением числа сравнений в среднем в шесть раз было получено ускорение в два раза в реальной продолжительности работы программы. Близость математического ожидания и стандартных отклонений определяемых параметров для полного перебора и перебора с использованием кластеризации достаточна для использования алгоритма в определенных задачах.

Дальнейшее развитие данной работы состоит в следующем:

- Исследование зависимости точности вычисления математического ожидания заданной функции параметров частицы от увеличения объема базы данных и использования других способов кластеризации.
- Оптимизация данной рекурсивной программы для получения улучшения быстродействия более близкого к ускорению в шесть раз.

ЛИТЕРАТУРА

1. Accurate measurement of volume and shape of resting and activated blood platelets from light scattering / Moskalensky A.E., Yurkin M.A., Konokhova A.I. *et al.* // J. Biomed. Opt.– 2013.–V. 18.– № 1.–P. 17001.
2. High-precision characterization of individual E. coli cell morphology by scanning flow cytometry / Konokhova A.I., Gelash A.A., Yurkin M.A. *et al*// Cytometry A. –2013. –V. 83A. – № 6. P. 568–575.
3. Is there a difference between T- and B-lymphocyte morphology? / Strokotov D.I., Yurkin M.A., Gilev K.V.*et al*. // J. Biomed. Opt.– 2009.– V. 14. – № 6.– P. 064036.
4. Optics of white blood cells: optical models, simulations, and experiments / Maltsev V.P., Hoekstra A.G., Yurkin M.A. // Advanced Optical Flow Cytometry: Methods and Disease Diagnoses / ed. Tuchin V.V. Weinheim: WileyWCH. – 2011.– P. 63–93.
5. Light-scattering flow cytometry for identification and characterization of blood microparticles / Konokhova A.I., Yurkin M.A., Moskalensky A.E. *et al*// J. Biomed. Opt. –2012.– V. 17.– P. 057006.
6. Yurkin M.A.,The discrete dipole approximation: an overview and recent developments / Yurkin M.A., Hoekstra A.G. // J. Quant. Spectrosc. Radiat. Transfer.– 2007.–V. 106. – № 1–3.– P. 558–589.
7. Боровкова С.В. *Использование кластеризации при решении параметрической обратной задачи светорассеяния.* Квалиф. работа

на соиск. степ. бакалавра, Новосибирск: Новосиб. гос. университет (2014).

8. ЗагоруйкоН.Г. Прикладные методы анализа данных и знаний / ЗагоруйкоН.Г. // Новосибирск: Изд-во Ин-та математики, 1999. – 270 с.