МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Механико-математический факультет
Кафедра математических методов геофизики

Выпускная квалификационная работа бакалавра

БОРОВКОВА Софья Всеволодовна

Использование кластеризации при решении параметрической обратной задачи светорассеяния

Научные руководители:
к.ф-м.н,
М.А.Юркин

к.ф-м.н,
Г.В.Дятлов

Новосибирск 2014

СОДЕРЖАНИЕ

1 ВВЕДЕНИЕ

Клетки крови несут в себе большое количество информации о состоянии организма, поэтому важно уметь точно и быстро проводить её анализ, определяя те или иные параметры кровяных клеток, такие как форма, объём, показатель преломления и т.д. В настоящее время эта задача решается с достаточно хорошей точностью в лаборатории Цитометрии и Биокинетики ИХКГ СО РАН с помощью сканирующего проточного цитометра [1-4].

Метод основан на измерении индикатрисы светорассеяния одиночных частиц в потоке со скоростью около 100 частиц в секунду (время измерения одной частицы ~1 мс) и решении параметрической обратной задачи светорассеяния с целью определения их морфологических параметров. В данном случае результатом измерения является вектор, составленный из значений интенсивности рассеянного света, полученных под углом рассеяния от 10° до 60°, с интервалом в 1°. При этом предполагается определенная модель исследуемых объектов, описываемая несколькими параметрами. Стандартный метод нелинейной регрессии для решения такой обратной задачи реально применим лишь в случае сферических частиц, для которых решение прямой задачи, т.е. моделирование индикатрисы светорассеяния для заданной частицы занимает порядка 1 мс [3,5]. Но и в этом случае решение обратной задачи занимает секунды из-за того, что необходимо использовать

метод глобальной оптимизации. Поэтому были предложены способы оптимизации [6,7].

В общем случае (для несферических частиц) решение прямой задачи для одной частицы занимает большое количество времени (порядка 1 минуты), например, с помощью метода дискретных диполей [8], что делает регрессию абсолютно нереалистичной. В качестве решения было предложено использовать интерполяцию методом ближайших соседей по предварительно насчитанной базе данных индикатрис [1,2]. Для четырех параметров модели данная база данных состоит ~10^5 теоретических индикатрис в диапазоне параметров, соответствующих литературным данным о морфологии конкретного класса клеток, для каждой и которых известны параметры используемой клетки. Таким образом, экспериментально полученная индикатриса сравнивается поочерёдно с каждым элементом базы данных. Параметры, соответствующие ближайшей теоретической индикатрисе, дают информацию о параметрах экспериментальной. Создание базы данных производится один раз для одного класса клеток, например, тромбоцитов крови человека, с использованием суперкомпьютера, и соответствующее время вычислений оправдано в связи с дальнейшей характеризацией большого количества клеток (только в одной пробе исследуется обычно 10^3–10^4 клеток).

2 ПОСТАНОВКА ЗАДАЧИ

Стандартный метод решения обратной задачи светорассеяния занимает порядка 1 с, что велико по сравнению со временем измерения и не позволяет проводить обработку экспериментальных данных в реальном времени. Более того, это станет ещё большей проблемой при увеличении объёма базы данных. С учетом типичной скорости измерений, существует потребность тратить на решение обратной задачи не более 10 мс. Считаем, что информация об индикатрисе светорассеяния хорошо отражается значениями при сканировании под углами от 10° до 60°, с шагом в 1°. Таким образом, можно считать, что индикатриса светорассеяния – это вектор. Также можно ввести норму в этом пространстве, в нашем случае это

$$\|I\| = \sqrt{\sum_{i=10}^{i=60} w(i) \cdot I^2(i)}\,, \qquad (1)$$

где I – индикатриса, а $w(\theta) = \frac{1^\circ}{\theta} \exp[-2\ln^2(\theta / 54^\circ)]$ – весовая функция, которая примерно соответствует структуре экспериментальных погрешностей [3]

База данных представляет массив из порядка 10^5 индикатрис, параметры клеток, соответствующих каждой из них, нам известны. Понятно, что поиск ближайшей индикатрисы при помощи простого перебора требует количества сравнений, равного количеству элементов в базе данных. В данной работе предлагается метод, предполагающий дополнительную предварительную обработку базы данных, а именно, её кластеризацию, с целью уменьшить время поиска индикатрисы, ближайшей к экспериментальной.

Кластеризация, или кластерный анализ множества – это процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы по какому-либо заранее заданному признаку. Такая группировка зачастую упрощает решение многих практических задач анализа данных. Вместо того, чтобы хранить в памяти или перебирать все объекты, достаточно использовать типичный представитель каждого кластера (назовём его центром кластера), перечислить номера объектов, входящих в данный кластер, и указать максимальное отклонение элементов кластера от центра (радиус) [9]. Этой информации будет достаточно для решения нашей задачи.

Спектр применений кластерного анализа очень широк. Однако универсальность применения привела к появлению большого количества несовместимых терминов, методов и подходов, затрудняющих однозначное использование и непротиворечивую интерпретацию кластерного анализа. Так, например, использование кластеризации, предлагаемое в данной работе, насколько известно автору, ранее не встречалось.

Таким образом, для каждого из кластеров вычисляется положение центра и радиус, что зачастую позволяет сравнивать экспериментальную индикатрису лишь с центрами некоторых кластеров, чтобы понять, что ближайшая индикатриса находится в другом кластере, тем самым избегая сравнений со всеми элементами «отброшенного» кластера и экономя время.

3 АЛГОРИТМ РЕШЕНИЯ ОБРАТНОЙ ЗАДАЧИ ПРИ ПОМОЩИ КЛАСТЕРНОЙ СТРУКТУРЫ

Пусть $I = \{I_1,\ldots,I_N\}$ – исходное множество векторов. Предположим, что у нас уже есть некоторая кластерная структура на I, полученная с помощью того или иного алгоритма кластеризации:

$X = \{X_1,\ldots,X_n\}$,

где X_i содержит номера индикатрис, отнесенных к i-му кластеру, $I = 1,..,n$. Для каждого из кластеров ищем центр, используя

$$C_i = \frac{\sum_{k \in X_i} I_k}{n_i}, \qquad (2)$$

где n_i – количество элементов в кластере X_i, и радиус

$$R_i = \max_{k \in X_i} \left\| I_k - C_i \right\|. \qquad (3)$$

Пусть $C = \{C_1, \ldots, C_n\}$ – множество центров, $R = \{R_1, \ldots, R_n\}$ – множество радиусов, I_{exp} – экспериментальная индикатриса, I_0 – ближайшая к ней из базы данных.

В процессе поиска I_0, каждый раз, когда нам встречаются соответствующие выражения, обновляем значение

$$M = \min\{\left\| I_{\text{exp}} - I_i \right\|, \left\| I_{\text{exp}} - C_j \right\| + R_j\}, \qquad (4)$$

где i, j такие, для которых уже вычислялись значения выражений в фигурных скобках.

В качестве начального значения берем $M = \left\| I_{\text{exp}} - I_1 \right\|$. Начинаем перебирать все кластеры из X. Понятно, что $\left\| I_{\text{exp}} - I_0 \right\| \le M$. Таким образом, кластеры, для которых выполнено

$$\left\| I_{\text{exp}} - C_i \right\| - R_i > M , \qquad (5)$$

можно исключить из дальнейшего рассмотрения целиком, тем самым избегая сравнения со всеми их элементами. Если же

$$\left\| I_{\text{exp}} - C_i \right\| - R_i \le M , \qquad (6)$$

рассматриваем кластер более подробно.

Здесь заметим, что кластер X_i можно так же предварительно разбить на подкластеры, для каждого из них посчитать центр и радиус, и, в случае выполнения неравенства (6), проделать аналогичные действия, исключив некоторые кластеры целиком. Структура, устроенная таким образом, называется иерархической. В случае иерархической структуры аналогичную процедуру проделываем до тех пор, пока дальнейшее разбиение на подкластеры не прекратится. Тогда пользуемся простым перебором. Таким образом, мы гарантированно найдем I_0, но количество сравнений с индикатрисами из базы данных будет меньше N.

3.1 Проблемы кластерного анализа

Существует ряд проблем, связанных с кластерным анализом и возникающих при решении данной задачи.

- Проблема обоснования качества результатов анализа. Понятно, что кластеризацию множества можно осуществлять разными способами, в результате чего будут получены соответствующие кластерные структуры. Таким образом, встает вопрос о критерии качества конкретной структуры, а также о сравнении двух таких структур.
- Многие области исследований характеризуются недостаточностью знаний об изучаемых объектах. В частности, параметры индикатрис рассеяния из нашей базы данных имеют равномерное распределение в пространстве параметров, но какое распределение имеют сами индикатрисы в своем пространстве – неизвестно, что в некоторой степени затрудняет выбор метода кластеризации.
- При увеличении размерности объектов сложно представить себе, как всё устроено на самом деле, достаточно легко представлять себе процесс в двумерном случае и даже рисовать картинки. Но что происходит в пространствах больших размерностей представить невозможно.

4 МЕТОДЫ КЛАСТЕРИЗАЦИИ

Кластеризация, как уже было сказано ранее, – это группировка N объектов по похожести их свойств. Одно и то же множество из N объектов можно разбить на n кластеров ($n<N$) по-разному. Поэтому необходимо

определить термин «похожесть», ввести некоторый критерий похожести. Этим критерием, по большому счету, и определяется метод кластеризации.

Самый известный критерий состоит в том, что в один кластер должны собираться объекты, максимально похожие на центр. Мы рассматриваем объекты метрического пространства, поэтому в качестве меры похожести можно использовать расстояние между элементами базы данных.

При проведении эксперимента использовались два нижеописанных алгоритма кластеризации, имеющие относительно низкую временную сложность, что весьма актуально при таком объёме базы данных.

4.1 Метод случайных индикаторов

1) Задается заранее количество кластеров n.
2) Случайным образом выбираем n элементов множества $\{x_1,\ldots,x_n\}$. Каждой точке соответствует будущий кластер, т.е. это их «индикаторы».
3) Для каждой из точек множества считаем расстояние от неё до каждого из индикаторов: $\|I_i - x_j\|$, $i = 1,\ldots,N$.
4) Точку относим к тому кластеру, который соответствует ближайшему из индикаторов, т.е. $\|I_i - x_k\| = \min_j \|I_i - x_j\| \Rightarrow I_i \in X_k$, $I = 1,\ldots,N$.

Понятно, что при различном выборе «индикаторов» получим разные варианты кластерных структур. Предлагается выбрать лучшую из них при помощи эксперимента. А именно, ту структуру, которая даст наибольшее

ускорение поиска ближайшей индикатрисы, т.е. наименьшее (в среднем) количество сравнений в процессе поиска.

4.2 FOREL

1) Вычисляем минимальный радиус R_0, такой что шар радиуса R_0 охватывает все N точек. Понятно, что разбиение множества на один кластер не имеет смысла, поэтому постепенно будем уменьшать радиус шара.
2) Берем радиус $R_1 < R_0$ и помещаем центр шара в произвольный элемент множества.
3) Для точек, лежащих внутри этого шара, считаем координаты центра масс.
4) Переносим центр в этот центр тяжести и проделываем аналогичные действия, до тех пор, пока центр не останется на месте, т.е. на очередном шаге состав внутренних точек, а следовательно, и их центр масс не изменится. То есть мы попадем в локальный максимум плотности сгущения точек.
5) Точки, попадающие внутрь сферы после её остановки, мы приписываем отдельному кластеру и далее исключаем из рассмотрения. Продолжаем процесс для оставшихся точек, пока все элементы множества не будут распределены по кластерам.

Доказана сходимость алгоритма за конечное число шагов [9].Как и в предыдущем методе, при различном выборе радиуса R_1 и начальных точек на

каждом шаге получим разные варианты кластерных структур. Предлагается выбрать лучшую из них при помощи эксперимента. А именно, ту структуру, которая даст наибольшее ускорение поиска ближайшей индикатрисы, т.е. наименьшее (в среднем) количество сравнений в процессе поиска.

5 РЕЗУЛЬТАТЫ

Была разработана программа в среде Wolfram Mathematica 8.0 осуществляющая кластеризацию базы данных каждым из выше описанных способов, а также поиск ближайшей индикатрисы по экспериментальной с использованием кластерной структуры. Были проведены эксперименты для тестовых индикатрис, в ходе чего были получены следующие результаты.

Приведём результаты численного эксперимента для тромбоцитов. Тромбоциты – это небольшие (2–4 мкм диаметром) дискообразные безъядерные клеточные фрагменты, циркулирующие в кровотоке, чутко реагирующие на повреждения сосуда и играющие критически важную роль в гемостазе и тромбозе [10]. Считается, что такую частицу можно хорошо приблизить сплюснутым шаром. Таким образом, тромбоцит описывается четырьмя параметрами: отношение полуосей, радиус шара эквивалентного объёма, угол ориентации и показатель преломления [1].

Для исследования алгоритмов кластеризации была использована база данных индикатрис тромбоцитов, содержащая 10 000 элементов. Параметры тромбоцитов случайно выбирались из диапазонов: радиус шара эквивалентного объема от 0.5 мкм до 2 мкм, отношение полуосей от 1 до 8,

показатель преломления от 1.35 до 1.5, угол ориентации от 0° до 90° [1]. Полученная кластерная структура применялась к экспериментальным индикатрисам тромбоцитов, в результате определялось среднее количество сравнений необходимое для нахождения ближайшей индикатрисы из базы данных.

Таблица 1. Среднее количество сравнений при поиске ближайшей индикатрисы при помощи кластерной структуры, полученной методом случайных индикаторов.

n	1	2	3	4	5
5	1770	1820	1556	1540	1882
8	1591	1619	1476	1728	1489
10	1421	1573	1516	1412	1456
20	1401	1510	1473	1398	1404
50	1801	1756	1844	1754	1789

где *n* – количество кластеров, в верхней строке указан номер эксперимента.

Как видно из данных таблицы, использование кластерной структуры, полученной «методом случайных индикаторов», позволяет получить ускорение примерно в 7 раз.

Таблица 2. Среднее количество сравнений при поиске ближайшей индикатрисы при помощи кластерной структуры, полученной алгоритмом FOREL.

R_1/R_0	1	2	3	4	5
0.9	1049	1191	1147	1202	1122
0.7	866	928	903	899	912
0.5	960	1171	1094	1154	1140
0.3	1564	1493	1598	1475	1562

где R_0 – половина максимального расстояние между точками множества, R_1 – радиус, которым ограничиваем кластер, в верхней строке указан номер эксперимента.

Как видно из данных таблицы, использование кластерной структуры, полученной с помощью алгоритма FOREL позволяет получить ускорение более, чем в 10 раз.

Также во всех экспериментах было проверено, что использование кластерного подхода не изменяет итогового ответа, т.е. ближайшие индикатрисы определяются ровно те же, что и в случае прямого перебора.

6 ЗАКЛЮЧЕНИЕ

Разработан общий подход использования кластеризации для ускорения нахождения ближайшего к заданному элементу из базы данных. Это позволяет ускорить решение параметрической обратной задачи светорассеяния для несферических биологических клеток. Выполнена программная реализация двух алгоритмов кластеризации: метода ближайших индикаторов и алгоритма FOREL, а также алгоритма использования кластерной структуры для нахождения ближайшей индикатрисы.

Были проведены численные эксперименты с использованием базы данных индикатрис тромбоцитов человека, и экспериментальных индикатрис тромбоцитов, измеренных на сканирующем проточном цитометре. Предварительная кластеризация базы данных позволила ускорить решение обратной задачи светорассеяния в 7 и 10 раз при использовании метода ближайших индикаторов и алгоритма FOREL соответственно. Это ускорение соответствует наиболее оптимальным параметрам алгоритма – $n = 20$ для первого метода и $R_1/R_0 = 0.7$ для второго.

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. G.V.Dyatlov, An optimization methodwith precomputed starting points for solving the inverse Mie problem, / G.V.Dyatlov, K.V. Gilev, M.A. Yurkin, V.P. Maltsev, // Inv. Probl. 28, 045012 (2012).
7. G.V.Dyatlov, An optimization method for solving the inverse Mie problem based on adaptive algorithm for construction of interpolating

database,/G.V.Dyatlov, K.V. Gilev, M.A. Yurkin, V.P. Maltsev, // J. Quant. Spectrosc. Radiat. Transfer 131, 202–214 (2013).

8. 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.
9. Загоруйко Н.Г. Прикладные методы анализа данных и знаний / Загоруйко Н.Г. // Новосибирск: Изд-во Ин-та математики, 1999. – 270 с.
10. A. D. Michelson, Platelets, Second Edition, 2nd ed., A. D. Michelson, Ed., Academic Press - 2007.