ОБРАБОТКА СИГНАЛОВ, ИЗОБРАЖЕНИЙ И РЕЧИ

УДК: 303.732.4

Е.В. Лисица[1], Н.Н. Яцков[1], В.В. Апанасович[1], Т.В. Апанасович[2]

ПРОГРАММНЫЙ ПАКЕТ CellDataMiner ДЛЯ АНАЛИЗА ЛЮМИНЕСЦЕНТНЫХ ИЗОБРАЖЕНИЙ РАКОВЫХ КЛЕТОК

Предлагается программный пакет CellDataMiner для анализа люминесцентных изображений раковых клеток. Проводится сравнительный анализ алгоритмов классификации и кластеризации данных с целью реализации в пакете наиболее эффективных из них. Работоспособность программного обеспечения проверяется на экспериментальных данных, представляющих результаты по исследованию опухоли молочной железы.

Введение

Метод люминесцентной микроскопии успешно используется для исследования срезов анатомических тканей в биомедицине [1]. В данном методе регистрируется интенсивность биомаркера объекта исследования (например, маркера онкологического заболевания) в ядрах и цитоплазмах клеток. Биомаркер – это специальное химическое вещество, отражающее протекающие в клетке процессы. В случае если биомаркер не обладает люминесцентными свойствами, то используются специальные красители, которые позволяют идентифицировать наличие биомаркера в образце. Поскольку в клетках одновременно протекает несколько процессов, для их анализа удобно использовать несколько маркеров, отражающих эти процессы. Все это привело к широкому внедрению на практике многоканальных экспериментов. Как правило, в многоканальных экспериментах используются отдельные красители для цитоплазм и ядер, информация о которых регистрируется в отдельном канале изображения. Результатом исследования является цветное RGB-изображение [2]. В большинстве случаев за один эксперимент получается около сотни изображений, которые содержат порядка тысячи объектов. Специалисты могут производить анализ изображений вручную, классифицировать объекты на них, считать их количество. Однако при обработке больших объемов данных такой подход становится неэффективным и приводит как к временным затратам, так и к ошибкам при проведении анализа.

На сегодняшний день существует большое количество программных средств, как коммерческих, так и находящихся в свободном доступе, которые предназначены для обработки и анализа изображений. Среди них, например, можно выделить проект по разработке открытого программного обеспечения (ПО) и форматов данных для обработки и хранения изображений, получаемых оптическим микроскопом в ходе анализа биологических объектов OME (Open Microscopy Environment). В рамках данного проекта реализованы программные пакеты Bio-Formats для чтения и записи данных в виде изображений с помощью стандартизированных открытых форматов и VisBio для визуализации и анализа многомерных данных [3]. Существуют и другие программные продукты для анализа многомерных данных, среди которых можно отметить пакет BioImageXD [4]. На рынке ПО имеется ряд средств для управления процессом получения изображений, одним из них является MicroManager [5]. Другое направление разработки ПО ориентировано на 4D- и 5D-визуализацию данных, получаемых при помощи оптических микроскопов, например FARSIGHT [6].

Одним из широко используемых программных пакетов для обработки изображений биологических объектов является ImageJ [7] и его расширение Fiji [8, 9]. Данное ПО представляет собой программу с открытым программным кодом, написанным на языке Java. Расширение функций осуществляется за счет подключения плагинов, реализующих алгоритмы сегментации и анализа; для устранения повторяющихся действий при анализе предусмотрен специальный язык макросов [10]. Разработаны специальные алгоритмы по обработке гистологических сним-

ков, получаемых с оптического микроскопа [11, 12], однако их применение к люминесцентным изображениями затруднительно. Эффективным программным средством для анализа люминесцентных изображений клеток является группа пакетов CellProfiler и CellProfiler Analyst [13, 14]. Пакет CellProfiler представляет собой набор модулей для сегментации и оценки характеристик сегментированных объектов. Он включает известные методы обработки, такие как пороговые, градиентные, и метод водораздела [15], где пользователю предлагается настраивать параметры методов вручную. CellProfiler Analyst – программный пакет, ориентированный на анализ многомерных данных. В нем интегрированы базовые методы классификации и кластерного анализа [14]. Однако использование стандартных методов, реализованных в программных пакетах, затруднено при сегментации ядер на трехканальных люминесцентных изображениях. Так, например, недостатком пороговых алгоритмов является низкое качество сегментации, характерное для обработки больших объемов данных. Группа градиентных методов имеет ограничения, связанные с чувствительностью к наличию артефактов и к перепадам интенсивности на изображении. Перекрытие цветовых каналов при регистрации изображений приводит к наличию большого количества артефактов на снимках. Широкое распространение для анализа биоданных получила библиотека C++ OpenCV, объединяющая набор алгоритмов компьютерного зрения, обработки изображений и численных алгоритмов [11]. Следует также отметить работы, посвященные разработке алгоритмов анализа данных об экспрессии генов, которые были получены с помощью изображений биоматриц ДНК [16] и гистологических изображений [17].

К основным ограничениям существующего ПО для анализа люминесцентных трехканальных биоизображений можно отнести:

– отсутствие автоматических алгоритмов сегментации;

– невысокую точность локализации биообъектов изображений, связанную с использованием традиционных методов, не адаптированных для анализа многоканальных изображений;

– отсутствие учета статистической взаимосвязи между каналами регистрации многоканальных изображений;

– трудности обработки больших объемов данных;

– анализ изображений без учета характеристик биообъектов (ядра, клетки, мембраны) на пиксельном уровне изображения, что значительно ограничивает точность классификации данных;

– требование достаточно глубокого знания пользователем не только применяемых методов анализа данных, но и определенных навыков программирования для более точной настройки параметров алгоритмов или осуществления взаимодействия между алгоритмами;

– ограниченное представление методов классификации сегментированных объектов изображений.

Возможными методами улучшения автоматической обработки трехканальных биомедицинских изображений являются применение объектно-ориентированной методологии анализа изображений на пиксельном уровне, учет статистической взаимосвязи между различными каналами изображений и последующее использование методов анализа данных для выявления скрытых закономерностей в оценке параметров сегментированных объектов изображений, не поддающихся обнаружению традиционными методами или экспертным путем [18, 19].

В данной работе представлен программный пакет для локализации, сегментации и классификации ядер клеток трехканальных люминесцентных изображений, интегрирующий наиболее эффективные алгоритмы сегментации, объектно-ориентированный анализ биообъектов на пиксельном уровне и методы анализа данных. Для всестороннего исследования разработанного программного пакета используется набор экспериментальных данных, полученных в ходе анализа тканей опухоли молочной железы [20, 21].

1. Экспериментальные данные

В настоящем исследовании рассматриваются микрочипы срезов тканей опухолей молочной железы. Изображения представляют собой популяции клеток, окрашенные в зеленые, синие и красные цвета (трехканальные люминесцентные сигналы в системе RGB). В цитоплазмах раковых клеток регистрируются процессы с участием белка цитокератина. Белок маркируется циа-

ниновым красителем Cy3 и регистрируется в зеленом цветовом канале изображения. Красный канал изображения зарезервирован для индикации ядер раковых клеток. В ядрах раковых клеток находится белок эстроген-рецептор, для маркировки которого применяется краситель [20]. Для маркировки ядер используется краситель 4,6-диамидино-2-фенилиндол дигидрохлорид (DAPI, 4',6-diamidino-2-phenylindole) и зарезервирован синий канал. Размер изображений – 2048×2048 пикселов в каждом из трех каналов, разрешающая способность – 0,2 мкм на пиксел [20–22].

2. Программный пакет CellDataMiner

Разработан и реализован программный пакет CellDataMiner, интегрирующий объектно-ориентированную методологию анализа биообъектов люминесцентных изображений на пиксельном уровне с последующим использованием методов анализа данных для выделения групп раковых клеток или исследования стадий развития онкологического заболевания. На данном этапе оптимальной средой разработки является Matlab. Пакет содержит проверенные и опубликованные библиотеки математических алгоритмов, включая алгоритмы цифровой обработки данных и имитационного моделирования, а также предоставляет возможности для создания графических интерактивных интерфейсов приложений и некоммерческой установки разрабатываемых пакетов.

Созданное ПО CellDataMiner реализует следующий набор принципиально новых функций, обеспечение которых не поддерживается или ограничено в других программных продуктах, находящихся в открытом доступе [23]:

– полностью автоматическую сегментацию ядер на изображениях биологических объектов;

– попиксельный анализ сегментированных объектов (результаты анализа отображаются в виде дополнительной таблицы);

– анализ распределений характеристик сегментированных объектов на пиксельном и интегральном уровнях;

– снижение размерности данных, т. е. выделение основных групп признаков сегментированных объектов для последующей интерпретации и визуализации данных;

– классификацию и кластеризацию раковых клеток;

– табличную и графическую визуализацию промежуточных и итоговых результатов анализа;

– интерактивное взаимодействие пользователя с объектами изображения.

Каждый из функциональных блоков ПО является независимым, что позволяет пользователю пропустить часть из них с целью оптимального выбора схемы анализа данных. В разработанном пакете наборы люминесцентных изображений могут анализироваться последовательно автоматически, без участия оператора, что позволяет значительно упростить анализ данных и вместе с тем сократить затраты на проведение исследований.

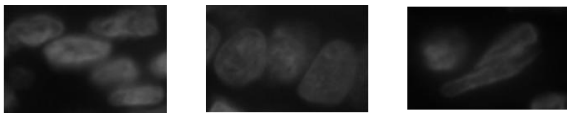
2.1. Алгоритм сегментации и методы анализа данных

В основе алгоритма сегментации лежит использование корреляционной зависимости между сигналами флуоресценции в R-, G-, B-каналах изображения. Представленный алгоритм реализуется поэтапно. На первом этапе происходит сегментация маски опухоли и оценка размеров ядер, на втором – сегментация ядер [24]. Сегментация ядер, в свою очередь, является многоэтапным процессом. В результате использования усовершенствованного алгоритма адаптивной сегментации полученное после бинаризации изображение содержит слившиеся объекты, которые разделяются водораздельным фильтром. Для устранения эффекта чрезмерной сегментации водораздельным методом, где полутоновое изображение линий перепадов и водоразделов предварительно сглаживается, применяется медианный фильтр.

В результате выполнения процедур сегментации изображения колоний раковых клеток и квантификации оценок статистических характеристик параметров ядер клеток получены N объектов-ядер $n_1, n_2, \ldots, n_N$, характеризующихся набором из K признаков (измеряемых оценок параметров объектов) $X_1, X_2, \ldots, X_K$.

Выходная характеристика, или зависимая переменная Y, представляет собой экспертную оценку состояния клеток. Например, для объекта n_i $y_i = 1$, если эксперт на основе визуальной оценки считает, что клетка раковая; $y_i = -1$ для нераковой клетки; $y = 0$ для объектов, подверженных значительным экспериментальным искажениям (рис. 1).

В ходе дальнейшего анализа требуется распределить объекты в группы раковых (больных) и нераковых (здоровых) по степени подобия, а также выявить факторы, обусловливающие различия статистических характеристик объектов, для решения задач определения фенотипа клеток, отделения шума от полезной информации, классификации стадий рака и т. д. В общем случае в ходе анализа решаются задачи снижения размерности признаков объектов, кластеризации, классификации и визуализации.



а) *б)* *в)*

Рис. 1. Примеры ядер: *а)* здоровые; *б)* больные; *в)* искаженные

В качестве характеристик объектов изображений выбраны десять наиболее значимых признаков [25]:

– координаты центра масс (*XM*, *YM*);
– угол между осью абсцисс и основной осью характеристического эллипса (Angle);
– эксцентриситет характеристического эллипса (Eccentricity);
– площадь исследуемого объекта совместно с площадью пустот, входящих в него (FilledArea);
– плотность как отношение площади объекта к его выпуклой площади (Solidity);
– средняя интенсивность в красном канале (MEAN_RED);
– средняя интенсивность в синем канале (MEAN_BLUE);
– среднеквадратичное отклонение в красном канале (STD_RED);
– медиана красного канала (MEDIAN_RED);
– среднее значение в зеленом канале (MEAN_GREEN).

Координаты центра масс и угол между осью абсцисс и основной осью характеристического эллипса не являются признаками объектов как таковыми, а описывают их местоположение и ориентацию на изображении. Однако данные параметры характеризуют свойство раковых клеток формировать колонии и различные пространственные ориентации относительно нераковых клеток. Использование координат центра масс в качестве признака позволяет усилить объединение в кластеры близкорасположенных объектов, а учет угла формирует кластеры одинаково ориентированных объектов.

Для оценки качества классификации и кластеризации использовалась ошибка неверно классифицированных объектов ERROR:

$$\text{ERROR} = 100\ \%*(\text{NUMBER\_1}+ \text{NUMBER\_2})/\text{TOTAL},$$

где NUMBER_1 – количество здоровых объектов, отнесенных к больным; NUMBER_2 – количество раковых объектов, отнесенных к здоровым; TOTAL – размер обучающей выборки.

Сравнение методов анализа проведено на экспериментальных изображениях.

На сегодняшний день не существует оптимального метода для решения задач классификации и кластеризации. В работе [19] проводится сравнение методов дискриминантного анализа, метода на основе опорных векторов и предлагается подход к усилению простых классификаторов путем их комбинирования в различные группы для исследования клеток линий HeLa и HT29. Наилучшие результаты классификации получены для метода опорных векторов (MOB). В [26] для классификации четырех линий раковых клеток Ovcar3, MiaPaCa2, MCF7 и MCF7-p53 используются комбинации искусственных нейронных сетей, метод *k*-ближайших соседей и метод главных компонент.

2.2. Кластеризация и классификация данных

Для *кластеризации* данных реализованы стандартные методы, такие как иерархический кластерный анализ, метод *k*-средних и сети Кохонена.

Для решения задачи кластерного анализа с помощью иерархических методов необходимо задать меру сходства, способ кластеризации и число кластеров. Для сравнения двух объектов n_i и n_j используются расстояния: евклидово (d_{euc}), города (d_{city}), Минковского (d_{Mink}), Махалонобби-

са (d_{mah}), косинусное (d_{cos}) и корреляционное (d_{cor}) [27, 28]. Методы иерархического кластерного анализа различаются по способу связывания объектов в кластеры. Наибольшее распространение получили следующие методы связывания: ближнего соседа, дальнего соседа, средней связи, медианной связи [29].

Для определения эффективности степени близости кластеров и методов связывания используется кофенетический корреляционный коэффициент κ [28]. Построение иерархического дерева считается успешным, если кофенетический корреляционный коэффициент близок к единице. Для наиболее успешной кластеризации строится дендрограмма иерархического дерева.

Суть метода *k*-средних заключается в следующем. Предположим, что заранее определено количество кластеров *k*, на которые необходимо разбить имеющиеся объекты. В качестве начальных центров кластеров можно выбрать любые *k* объектов. Для каждого следующего объекта рассчитываются расстояния до центров кластеров и данный объект относится к тому кластеру, расстояние до которого оказалось минимальным. После этого для кластера, в котором увеличилось количество наблюдений, рассчитывается новое положение центра кластера (как среднее по каждому признаку) по всем включенным в кластер объектам.

Нейронная сеть на основе самоорганизующихся карт состоит из компонентов, называемых узлами или нейронами. Изначально задается размерность карты, по ней некоторым образом строится первоначальный вариант карты. В процессе обучения векторы веса узлов приближаются к входным данным. Для каждого наблюдения выбирается наиболее похожий по вектору веса узел и значение его вектора веса приближается к наблюдению. Также к наблюдению приближаются векторы веса нескольких узлов, расположенных рядом. Таким образом, если в множестве входных данных два наблюдения были схожи, на карте им будут соответствовать близкие узлы. Кроме карт Кохонена существует также слой Кохонена, который представляет собой нейронную сеть с конкурирующей активационной функцией. В этом случае в качестве принадлежности к кластеру нейронная сеть выдает номер нейрона ближайшего к исследуемому объекту.

Для *классификации* данных реализованы наиболее эффективные алгоритмы, такие как дискриминантный анализ, МОВ и метод *k*-ближайших соседей (*k*ББ).

В дискриминантом анализе (ДА) рассматривается предположение о нормальном распределении признаков объектов. В линейном ДА (ЛДА) используется линейная комбинация признаков, позволяющая построить классифицирующее правило для объектов. Квадратичный ДА (КДА) работает аналогично с той лишь разницей, что предварительно производится преобразование пространства и в новом пространстве ищется линейная комбинация признаков. В исходном же пространстве эта комбинация будет нелинейной (квадратичной).

В основе работы метода *k*ББ лежит положение о том, что объект присваивается классу, который является самым распространенным среди его соседей. Для расчета расстояний между объектами можно использовать расстояния d_{euc}, d_{city}, d_{Mink}, d_{mah}, d_{cos}, d_{cor}.

2.3. Снижение размерности данных

Для визуализации данных и выделения значимой информации рассмотрен метод главных компонент [30]. В частности, для представления в наглядной форме измеренных характеристик и соответствующих им групп объектов (ядер и клеток) используются первые две компоненты, что представляет собой ортогональное проецирование многомерной системы на двухмерную плоскость.

2.4. Критерии оценки качества и результаты анализа

Основным критерием для оценки качества классификации и кластеризации выступала ошибка неверно классифицированных объектов. Для того чтобы исследовать, насколько устойчиво могут работать методы на независимых данных, использовался метод перекрестной проверки. За один этап перекрестной проверки происходит разделение имеющихся данных на обучающую часть и тестовый набор. Чтобы получить более точные результаты, разные циклы перекрестной проверки проводятся на разных разбиениях, затем результат усредняется по всем циклам [31].

Методы классификации и кластеризации исследованы на восьми экспериментальных изображениях. В качестве эталонных масок изображений приняты контуры ядер, выделенные

экспертным путем. Рассмотрены следующие размеры обучающей выборки: 15, 30, 50, 70, 100, 150 и 200 объектов. Результаты сравнительного анализа наиболее эффективных алгоритмов кластерного анализа представлены в таблице.

Ошибка кластеризации ERROR для иерархического кластерного анализа и метода *k*-средних

Метод	Изображение								
	1	2	3	4	5	6	7	8	Среднее значение
Иерархический (d_{cos})	1,4	4,3	28,5	1,7	5,6	3,5	2,8	1,7	6,2
Иерархический (d_{cor})	3	4,3	39,6	4	3,7	4,4	4,8	2,3	8,3
k-средних (d_{euc})	1,9	3,2	30,7	2,2	9	6,7	48,4	1,1	12,9
k-средних (d_{city})	2,8	3,1	42,8	2,5	8,6	8,5	38,5	1,2	13,5
k-средних (d_{cos})	6,9	7,6	47,3	2,2	11,9	22,7	40,5	24,5	20,5
k-средних (d_{cor})	7,1	6,7	49,8	2,7	9,8	22	36,2	23,9	19,8

Для иерархического кластерного анализа две метрики и центроидный метод связывания показали наилучшие результаты с наименьшей ошибкой кластеризации. Наилучшие результаты получены для центроидного метода связывания. Средние значения ошибок кластеризации, вычисленные по восьми изображениям, составляют 6,2 и 8,3 % для косинусного и корреляционного расстояний соответственно.

Наименьшая ошибка для исследуемых изображений – 1,4 %, а максимальная – 39,6 %. Более подробно результаты анализа другими методами вычисления расстояния и связывания показаны в приложении А (http://www.sstcenter.com/dsa/Staff/Lisitsa/index.html).

Наименьшая ошибка кластеризации для метода *k*-средних в два раза больше, чем ошибка для иерархического кластерного анализа, и составляет ERROR = 12,9 %. Две метрики показали сопоставимые результаты кластеризации: евклидово расстояние и метрика города (таблица).

Наихудшие результаты получены для нейронных сетей. Ошибка кластеризации для слоя Кохонена составляет 20,2 %, а для карты Кохонена наименьшая ошибка кластеризации – 20,3 %, что в два раза больше, чем ошибка кластеризации для метода *k*-средних.

Результаты сравнительного анализа алгоритмов классификации в зависимости от объема обучающей выборки представлены на рис. 2. Для получения достоверных результатов использовался метод перекрестной проверки, из имеющихся данных семь раз независимо выбирались обучающие выборки. Наилучшие результаты получены для линейного дискриминантного классификатора.

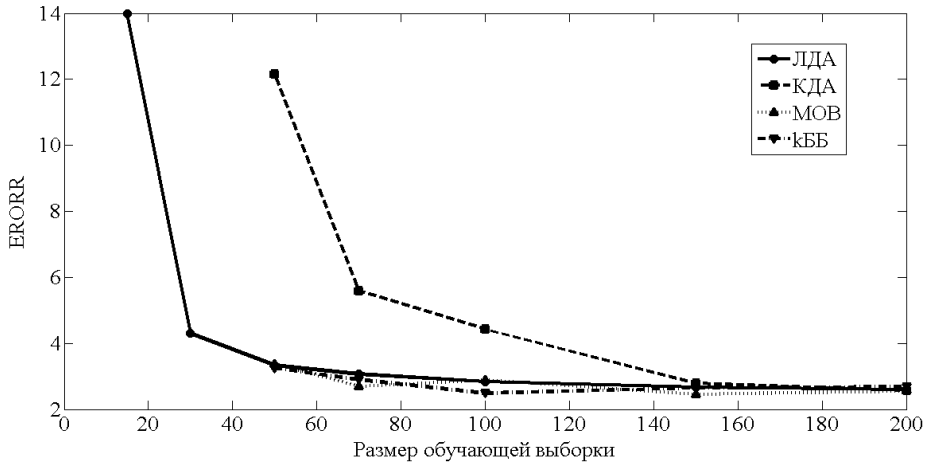


Рис. 2. Зависимость ошибки классификации от размера обучающей выборки для методов линейного и квадратичного дискриминантных анализов, опорных ветров и *k*-ближайших соседей

Ошибка классификации уменьшается по мере увеличения размера обучающей выборки, при размере выборки от 50 объектов и более она не превосходит 4 %.

Для КДА наименьший размер обучающей выборки должен составлять как минимум 50 объектов. При размере обучающей выборки 150 объектов и более качество классификации методом КДА не отличается от результатов для линейного классификатора.

Использование метода *k*-ближайших соседей является эффективным при размере обучающей выборки от 50 объектов и более и метрики города. В этом случае происходит успешное обучение классификатора, аналогичные результаты получены для линейного метода опорных векторов. Ошибка классификации методами МОВ и *k*ББ варьируется около 3 % (рис. 2).

2.5. Интерфейс CellDataMiner

Главное окно программного пакета CellDataMiner показано на рис. 3. На первом этапе пользователю необходимо загрузить выбранное изображение кнопкой Load Image. Сегментация объектов на изображении начнется при нажатии кнопки Process Image. После выделения объектов на изображении их границы будут показаны в окне программы SegementationAxes, при этом ядра, относящиеся к раковым клеткам, будут подсвечены бирюзовым цветом, а ядра здоровых клеток – фиолетовым.

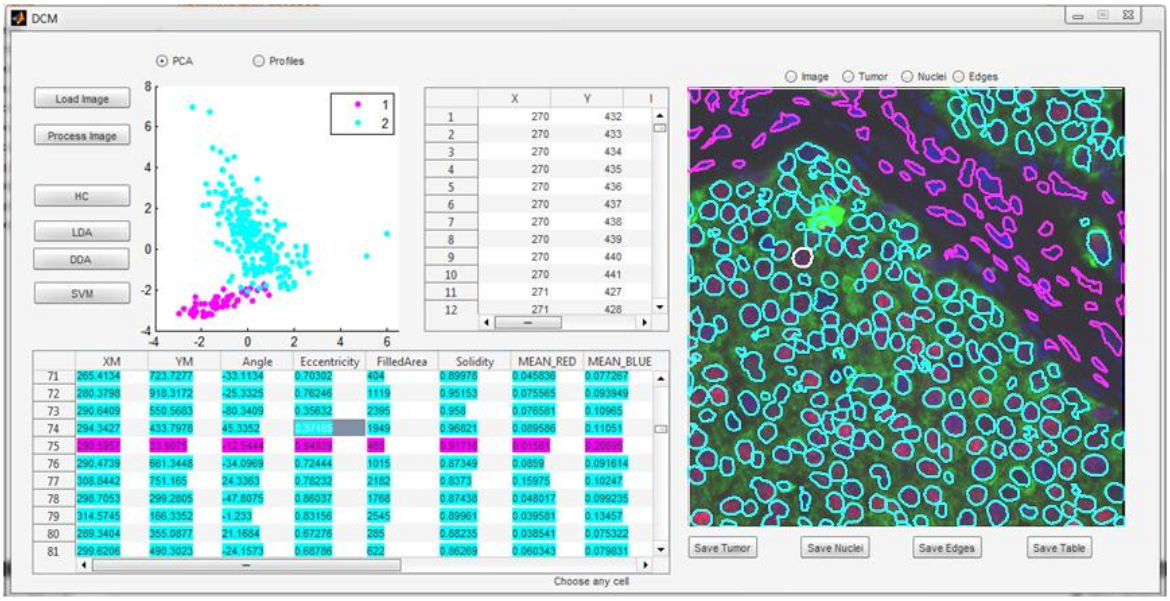
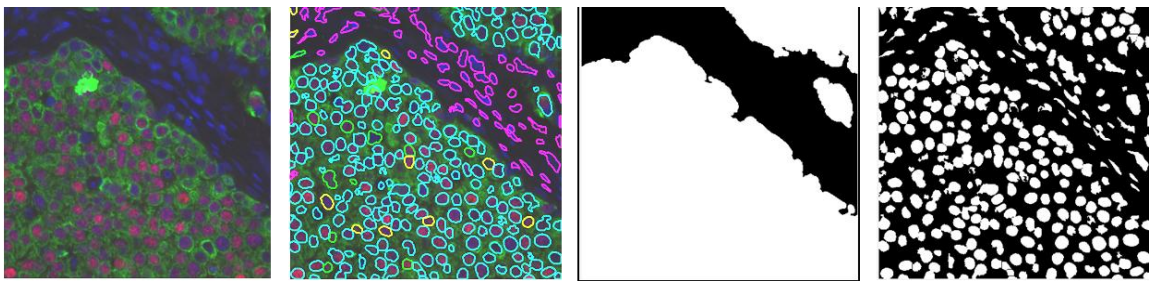


Рис. 3. Общий вид интерфейса программного пакета CellDataMiner

Одновременно в таблице NucleiTable, в столбце PRE-TYPE устанавливаются значения 1 для ядер здоровых клеток и 2 для ядер раковых клеток. В окне отображения границ объектов можно также посмотреть исходное изображение до выделения границ, границы объектов, бинарные изображения сегментации, маски опухоли и ядер (рис. 4).



а) *б)* *в)* *г)*

Рис. 4. Примеры изображений в окне SegementationAxes: *а)* исходное изображение; *б)* границы объектов; *в)* маска опухоли; *г)* маска ядер

В отображении результатов сегментации реализована интерактивная работа пользователя с объектами. Выбрав определенный объект на изображении и щелкнув по нему клавишей мыши, можно посмотреть его параметры в поле Status, а сам объект будет выделен белым цветом. В окне DMAxes можно отобразить сами объекты в пространстве главных компонент, выбрав режим PCA, и посмотреть профили объектов, выбрав режим Profiles.

Интерактивное взаимодействие с пользователем также осуществляется при работе с таблицей NucleiTable. В таблице отображены признаки сегментированных объектов, при этом цветом обозначается принадлежность объектов к классам. При нажатии на объект в таблице он будет подсвечен белым цветом в окне изображения, а значения интенсивностей его пикселов показаны в таблице PixelLevelTable (рис. 3).

Отдельным блоком в ПО вынесены методы анализа данных. С целью разбиения данных на группы реализован метод иерархического кластерного анализа. Для запуска метода необходимо нажать на кнопку HC, после прогона алгоритма индексы классов для объектов можно посмотреть в таблице NucleiTable, объекты классов будут отображены в пространстве главных компонент, расположение объектов классов отобразится на изображении SegemntationAxes. Объекты первого класса отобразятся бирюзовым цветом, объекты второго класса – фиолетовым.

В ПО реализованы два метода кластеризации данных – ЛДА и КДА. Методы требуют ввода обучающей выборки для настройки параметров классификатора. Ввод обучающей выборки осуществляется в таблице NucleiTable в столбце USER-TYPE. При этом для обозначения объектов обучающей выборки первого класса используется цифра 3, сами объекты подсвечиваются на изображении PixelLevelTable желтым цветом, а для обозначения объектов обучающей выборки второго класса используется цифра 4, объекты выделяются зеленым цветом. После того как объекты установлены, запуск анализа осуществляется нажатием кнопки LDA для линейного классификатора и DDA для квадратичного (рис. 3).

Сохранить полученные результаты можно при помощи блока кнопок: SaveTumor – сохранение результатов сегментации опухоли на изображении, SaveNuclei – сохранение результатов сегментации ядер, SaveEdges – сохранение цветного изображения с границами объектов на нем, SaveTable – сохранение результатов анализа данных. Пример исследования изображения показан на рис. 3.

3. Анализ экспериментальных изображений с использованием пакета CellDataMiner

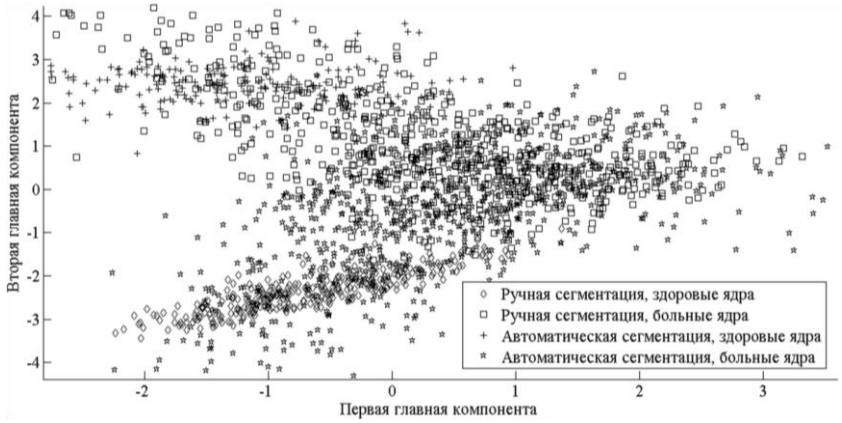
Работоспособность разработанного ПО проверена на примере обработки больших объемов данных, полученных в результате эксперимента по исследованию опухоли молочной железы. Для анализа отобраны 137 изображений. Для каждого экспериментального изображения определено количество нераковых и раковых клеток, а также их профили в пространстве характерных признаков. Общее количество клеток варьируется в пределах 1000–3000, количество раковых клеток – в пределах 100–1700.

Алгоритм успешно сегментировал 120 слайдов. Результат неверной сегментации объектов остальных 17 слайдов вызван недооценкой размеров ядер вследствие наличия малого количества объектов на них при низкой средней интенсивности люминесценции изображения.

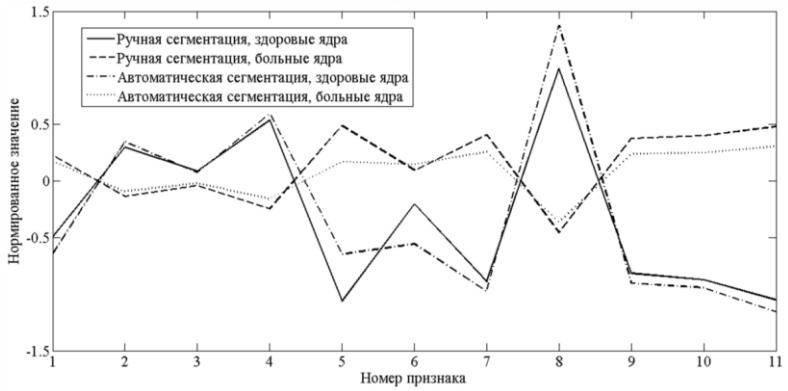
В результате анализа определено, что все виды изображений можно условно разделить на три класса:

1. Изображения, где количество раковых клеток значительно меньше количества здоровых клеток. Количество раковых клеток не превосходит 100.
2. Изображения, где количества двух классов клеток сопоставимы. Количество раковых и здоровых клеток составляет около 500.
3. Изображения, где количество раковых клеток превышает количество здоровых клеток. Количество здоровых клеток не превосходит 100.

На рис. 5 показано отображение найденных объектов в пространстве первых двух главных компонент и средние профили признаков объектов для одного из рассматриваемых изображений, определенные автоматическим методом на основе результатов сегментации маски опухоли и экспертным путем.



а)



б)

Рис. 5. Результаты анализа экспериментальных изображений с использованием пакета CellDataMiner: *а)* объекты, найденные экспертным путем и автоматическим методом в пространстве первых двух главных компонент; *б)* средние профили признаков объектов, найденных экспертным путем и автоматическим методом

Отображения объектов в пространстве первых двух главных компонент, найденные автоматическим методом и экспертным путем, имеют различия (рис. 5, *а*). Это можно объяснить тем, что у автоматического метода существуют ошибки локализации и сегментации, которые вносят искажения. Однако средние профили признаков объектов достаточно близки, за исключением нескольких признаков, для которых характерно значительное количество значений-выбросов.

Результаты анализа демонстрируют, что отобранные признаки позволяют выделять кластеры объектов на изображениях. При этом необходимо отметить, что для классификации объектов достаточным является использование параметров формы. Ядра раковых клеток в большинстве своем обладают бо́льшими размерами, чем ядра здоровых клеток. Для неракковых объектов характерна бо́льшая экспрессия в синем канале, чем для раковых клеток. Необходимо отметить, что зеленый краситель используется только для цитоплазмы, тем не менее его накопление в ядрах раковых клеток происходит интенсивнее, чем в ядрах здоровых клеток.

Заключение

В работе представлено программное средство для анализа люминесцентных изображений раковых клеток. Разработанный программный пакет CellDataMiner интегрирует объектно-ориентированную методологию анализа биообъектов люминесцентных изображений на пиксельном уровне с последующим использованием методов анализа данных для выделения групп раковых клеток или исследования стадий развития онкологического заболевания.

В качестве личного вклада авторов следует отметить:

– разработку объектно-ориентированной методологии анализа биообъектов на пиксельном уровне изображения;

– разработку и программную реализацию автоматического алгоритма сегментации с учетом статистической взаимосвязи между каналами изображения, позволяющего повысить точность сегментации;

– полную автоматизацию программных средств анализа больших объемов данных;

– реализацию средств интерактивного анализа сегментированных объектов изображений;

– проведение сравнительного анализа наиболее эффективных алгоритмов классификации и кластерного анализа сегментированных объектов биоизображений с целью интеграции наилучших алгоритмов в программный пакет.

Реализованные в ПО методы исследованы на экспериментальных данных. Наилучшие результаты кластеризации получены для иерархического кластерного анализа, наихудшие результаты получены для слоя Кохонена. Для классификации данных наилучшим методом является ЛДА; при больших размерах обучающей выборки КДА, *k*ББ и МОВ показывают такие же результаты, как и ЛДА. Их ошибка классификации не превосходит 5 %.

С использованием разработанного программного средства выполнен анализ 137 изображений раковых клеток, полученных в ходе экспериментов по исследованию микрочипов тканей опухолей молочной железы. Успешная сегментация была выполнена для 120 слайдов. Установлено точное количество раковых и нераковых клеток на каждом из исследуемых изображений. Общее количество клеток варьируется в пределах 1000–3000. Количество раковых клеток изменяется в пределах 100–1700. Рассчитаны основные характеристики раковых клеток: интенсивность биомаркера рака, параметры формы и структуры биообъектов и т. д. Выбранные характеристики позволяют выделить два класса объектов. Для разделения объектов анализируемых изображений на классы достаточно использовать признаки формы. Результаты анализа экспериментальных изображений позволяют сделать следующий вывод: автоматически сегментированные контуры объектов хорошо согласуются с контурами, выделенными вручную, а подобранные параметры объектов могут использоваться для классификации ядер клеток.

Список литературы

1. Spatial quantitative analysis of fluorescently labeled nuclear structures: problems, methods, pitfalls / O. Ronneberger [et al.] // Chromosome Res. – 2008. – Vol. 16(3). – P. 523–562.

2. Molecular Biology of the Cell / B. Alberts [et al.]. – N. Y. : Garland Science, 2012.

3. Rueden, C. VisBio: a computational tool for visualization of multidimensional biological image data / C. Rueden, K. Eliceiri, J. White // Traffic. – 2004. – Vol. 5. – P. 411–417.

4. BioImageXD: an open, general-purpose and high-throughput image-processing platform / P. Kankaanpaa [et al.] // Nat Methods. – 2012. – Vol. 9(7). – P. 683–689.

5. Computer control of microscopes using μManager / A. Edelstein [et al.] // Current Protocols in Molecular Biology. – 2010. – Vol. 14(20). – P. 1–17.

6. The FARSIGHT Project: Associative 4D/5D Image Analysis Methods for Quantifying Complex and Dynamic Biological Microenvironments / B. Roysam [et al.] // Microscopy and Microanalysis. – 2008. – Vol. 14 (Supplement S2). – P. 60–61.

7. Schneider, C.A. NIH Image to ImageJ: 25 years of image analysis / C.A. Schneider, W.S. Rasband, K.W. Eliceiri // Nat Methods. – 2012. – Vol. 9(7). – P. 671–675.

8. Out, W.A. A new method for morphometric analysis of opal phytoliths from plants / W.A. Out, J.F. Pertusa Grau, M. Madella // Microsc Microanal. – 2014. – Vol. 20(6). – P. 1876–1887.

9. Fiji: an open-source platform for biological-image analysis / J. Schindelin [et al.] // Nat Methods. – 2012. – Vol. 9(7). – P. 676–682.

10. Optimized digital counting colonies of clonogenic assays using ImageJ software and customized macros: comparison with manual counting / Z. Cai [et al.] // Int. J. Radiat. Biol. – 2011. – Vol. 87(11). – P. 1135–1146.

11. Designing a wearable navigation system for image-guided cancer resection surgery / P. Shao [et al.] // Ann Biomed Eng. – 2014. – Vol. 42(11). – P. 2228–2237.

12. Computer-aided Image Processing of Angiogenic Histological / M. Sprindzuk [et al.] // J. Clin. Med. Res. – 2009. – Vol. 1(5). – P. 249–261.

13. Bray, M.A. Using CellProfiler for Automatic Identification and Measurement of Biological Objects in Images / M.A. Bray, M.S. Vokes, A.E. Carpenter // Current Protocols in Molecular Biology. – 2015. – Vol. 109. – P. 14 17 1–14 17 13.

14. CellProfiler Analyst: data exploration and analysis software for complex image-based screens / T.R. Jones [et al.] // BMC Bioinformatics. – 2008. – Vol. 9. – 482 p.

15. Gonzalez, W. Eddins, Digital Image Processing Using MATLAB / W. Gonzalez. – 2nd edition. – Gatesmark Publishing, 2009.

16. Novoselova, N. Supervised Clustering of Genes for Multi-Class Phenotype Classification / N. Novoselova, I. Tom // Modeling and Simulation (MS'2012). – Minsk : BSU, 2012. – P. 32–36.

17. Alilou, M. Segmentation of cell nuclei in heterogeneous microscopy images: a reshapable templates approach / M. Alilou, V. Kovalev, V. Taimouri // Comput Med Imaging Graph. – 2013. – Vol. 37(7–8). – P. 488–499.

18. Segmentation of microscope cell images via adaptive eigenfilters / S. Kumar [et al.] // Image Proc. ICIP'04. Intern. Conf. – Singapore, 2004. – Vol. 1. – P. 135–138.

19. Abbas, S.S. A comparative study of cell classifiers for image-based high-throughput screening / S.S. Abbas, T.M. Dijkstra, T. Heskes // BMC Bioinformatics. – 2014. – Vol. 15. – 342 p.

20. Quantitative analysis of estrogen receptor heterogeneity in breast cancer / G.G. Chung [et al.] // Lab. Invest. – 2007. – Vol. 87(7). – P. 662–669.

21. Camp, R.L. Automated subcellular localization and quantification of protein expression in tissue microarrays / R.L. Camp, G.G. Chung, D.L. Rimm // Nat. Med. – 2002. – Vol. 8(11). – P. 1323–1327.

22. Simulation Model for Three-Channel Luminescent Images of Cancer Cell Populations / E.V. Lisitsa [et al.] // Journal of Applied Spectroscopy. – 2015. – Vol. 81(6). – P. 996–1003.

23. Review of free software tools for image analysis of fluorescence cell micrographs / V. Wiesmann [et al.] // Journal of Microscopy. – 2015. – Vol. 257, iss. 1. – P. 39–53.

24. Алгоритм автоматической сегментации границ ядер раковых клеток на трехканальных люминесцентных изображениях / Е.В. Лисица [и др.] // Журнал прикладной спектроскопии. – 2015. – № 82(4). – С. 598–607.

25. Разработка методов цифровой обработки люминесцентных изображений биологических объектов / В.В. Апанасович [и др.]. – Минск : Белорусский фонд фундаментальных исследований, 2013.

26. High-content phenotypic profiling of drug response signatures across distinct cancer cells / P.D. Caie [et al.] // Mol Cancer Ther. – 2010. – Vol. 9(6). – P. 1913–1926.

27. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян [и др.]. – СПб. : БХВ-Петербург, 2004. – 336 с.

28. Uragun, B. The discrimination of interaural level difference sensitivity functions: development of a taxonomic data template for modelling / B. Uragun, R. Rajan. – Clayton : Monash University, 2013. – P. 1471–2202.

29. Мандель, И.Д. Кластерный анализ / И.Д. Мандель. – М. : Финансы и статистика, 1988. – 176 с.

30. Прикладная статистика. Классификация и снижение размерности / С.А. Айвазян [и др.]. – М. : Финансы и статистика, 1989. – 607 с.

31. Воронцов, К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов / К.В. Воронцов // Математические вопросы кибернетики. – М. : Физматлит, 2004. – Vol. 13. – С. 5–36.

Поступила 21.10.2015

[1]*Белорусский государственный университет,*
Минск, пр. Независимости, 4
e-mail: Lisitsa@bsu.by

[2]*Университет Джорджа Вашингтона,*
1922 F str NW, Old Main
e-mail: apanasovich@gwu.edu

Y.U. Lisitsa, M.M. Yatskou, V.V. Apanasovich, T.V. Apanasovich

THE SOFTWARE PACKAGE CellDataMiner FOR DATA ANALYSIS OF FLUORESCENT IMAGES OF CANCER CELLS

The paper presents the software package CellDataMiner for data analysis of lumencent images of cancer cells. The comparative analysis of classification and clustering methods is carried out. The most sufficient of them are implemented in the software. The software package is tested on the dataset of the experimental images of breast cancer.