ОТБОР ЗНАЧИМЫХ ПРИЗНАКОВ МЕТОДАМИ ИЕРАРХИЧЕСКОГО КЛАСТЕРНОГО АНАЛИЗА

Лисица Е. В., Скакун В. В., Яцков Н. Н., Апанасович В.В.
Кафедра системного анализа и компьютерного моделирования, Белорусский государственный университет
Минск, Республика Беларусь
E-mail: ylisitsa@gmail.com

В работе рассматриваются методы иерархического кластерного анализа для выделения значимых признаков объектов, сегментированных на многоканальных люминесцентных изображениях. В ходе исследования были рассмотрены 42 характеристики объектов по форме и цвету. Оптимальная комбинация метода связывания и метрики были отобраны, используя кофенестический коэффициент.

Введение

Результатом сегментации изображения является его бинарная маска, которая содержит границы объектов, используя которую можно произвести их квантификацию, т. е. численно оценить признаки характеризующие объекты. В результате такого преобразования будут получены N объектов-ядер $n_1, n_2, \ldots, n_N$, характеризующиеся набором из K признаков (измеряемых оценок параметров объектов) $X_1, X_2, \ldots, X_K$. Одним из способ снижения количества используемых признаков явялется отбор наиболее значимых из них, например, методом иерархического кластерного анализа.

I. Материалы и методы

В данной работе рассматриваются микрочипы тканей опухолей молочной железы. Изображения представляют собой популяции клеток, окрашенные в зеленые, синие и красные цвета (трехканальные люминесцентные изображения в формате RGB). В цитоплазмах раковых клеток регистрируются процессы с участием белка цитокератина. Белок маркируется цианиновым красителем Cy3 (Cyanines) 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 использован краситель Cy5. Для маркировки ядер использован краситель 4,6-диамидино-2-фенилиндол дигидрохлорид (DAPI) и для него зарезервирован синий канал. Размер изображений 20482048 пикселей в каждом из трех каналов, размер пикселя 0.2 мкм/пиксель. Сегментация изображений проводилась улучшенным методом пороговой обработки с автоматической оценкой размеров сегментированных объектов[1]

Для описания формы сегментированных объектов выбраны опубликованные в литературе параметры [2, 3]: Площадь объекта (S);Периметр (Perimeter); Полная площадь (FilAR); Эквивалентный диаметр (*Deq*); Округлость (*Round*), ; Выпуклая площадь (*ConvexA*); Плотность (*Solid*); Длина (*Height*); Ширина (*Width*); Большая (*Major*) и малая (*Minor*) оси характеристического эллипса; Эллиптичность (*AR*); Округлость эллипса (*Circ*); Эксцентриситет характеристического эллипса (*Eccen*); Коэффициент заполнения (*Exten*); степень изогнутости границы объекта (*OUR*)

Для описания положения объектов в пространстве используются следующие параметры:Координаты центра масс *(X,Y)*; Координаты начала (верхний правый угол) характеристического прямоугольника *(BX,BY)*; Координаты граничной точки исследуемого объекта относительно правого верхнего угла изображения *(XSatrt, YSatrt)*; Угол между осью абсцисс и основной осью характеристического эллипса *(Angle)*.

Параметры интенсивности люминесценции в цветовых каналах. Для описания полученных распределений использовались стандартные параметры интенсивности флуоресценции: 1. Минимальное и максимальное значение интенсивностей в каждом цветовом канале (*MAX_R, MIN_R* – для красного канала R; *MAX_G, MIN_G* –для зелёного канала G; *MAX_B, MIN_B* – для синего канала B).

2. Среднее значение и среднеквадратичное отклонение интенсивности флуоресценции в цветовом канале (*MEAN_R, STD_R* – для красного канала R; *MEAN_G, STD_G* – для зелёного канала G; *MEAN_B, STD_B* – для синего канала B).

3. Медиана, 25, 75 % перцентили интенсивности флуоресценции (*MED_R, Q1_R, Q3_R* – для красного канала R; *MED_G, Q1_G, Q3_G* – для зелёного канала G; *MED_B, Q1_B, Q3_B* – для синего канала B).

Для сравнительного анализа разработанных методов ИАД произвольным образом выбраны девять экспериментальных изображений. Эталонная выборка сформирована экспертным путем. Для характеристики объектов ядер клеток выбрано 42 признака, показанных в таблице 1. На предварительной стадии анализа используется иерархический метод кластерного анализа для выделения основных групп признаков, характеризующих объекты.

Таблица 1 – Порядковые номера для обозначения признаков

Признак	N	Признак	N	Признак	N
S	1	X	2	Y	3
$Perimeter$	4	BX	5	BY	6
$Width$	7	$Height$	8	$Major$	9
$Minor$	10	$Angle$	11	$Circ$	12
AR	13	$Round$	14	$XSatrt$	15
$YSatrt$	16	$FilAR$	17	$ConvexA$	18
$Solid$	19	$Exten$	20	OUR	21
MAX_R	22	MIN_R	23	$MEAN\_R$	24
STD_R	25	MED_R	26	$Q1\_R$	27
$Q3\_R$	28	MAX_G	29	MIN_G	30
$MEAN\_G$	31	STD_G	32	MED_G	33
$Q1\_G$	34	$Q3\_G$	35	MAX_B	36
MIN_B	37	$MEAN\_B$	38	STD_B	39
MED_B	40	$Q1\_B$	41	$Q3\_B$	42

В методе иерархического кластерного анализа необходимо задать: меру сходства, способ кластеризации, число кластеров. Для сравнения двух объектов n_i и n_j используются расстояния: Евклидово (d_2); города (d_H); Минковского (d_{Mnkw}); косинусное (d_{cos}); корреляционное (d_{cor}). Для связывания объектов в кластеры существуют методы: ближнего соседа, дальнего соседа, средней связи, медианной связи. Кофенетический корреляционный коэффициент k служит для определения эффективности степени близости кластеров и методов связывания. Если коэффициент k близок к 1, то построение иерархического дерева считается успешным[4].

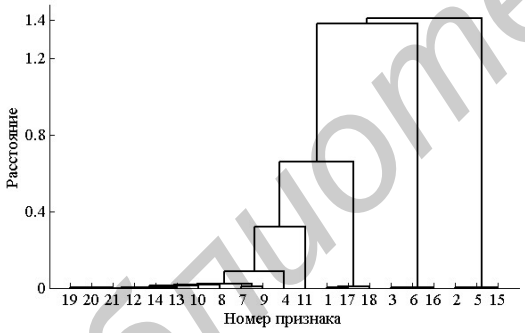

Рис. 1 – Дендрограмма признаков формы

II. Результаты

Для определения наиболее оптимальной комбинации метода связывания и метрики для различных способов кластеризации данных рассчитаны кофенестические коэффициенты,наилучше построение дендрограммы для признаков объектов изображения получено для расстояния Минковского и метода ближнего соседа $k = 0,99913$. Так как часть групп содержит только признаки цвета, а другая часть групп содержит только признаки формы, то для дальнейшего анализа целесообразно исследовать отдельно признаки формы и признаки цвета. Наилучшее значение кофенестического коэффициента для параметров формы были получены при использовании d_{cos} и метода средней связи. На рис. 1 показана дендрограмма признаков формы объектов построенная по всем изображениям.

Анализ дендрограммы признаков формы позволяет выделить 8 групп параметров:1 - $Width, Major$; 2 - $Height, Minor$; 3 - $Round, Circ, AR, OUR, Solid, Exten$; 4 - $Perimeter$; 5 - $S, FilAR, ConvexA$; 6 - $Angle$; 7- $Y, BY, YStart$; 8 - $X, BX, XStart$.

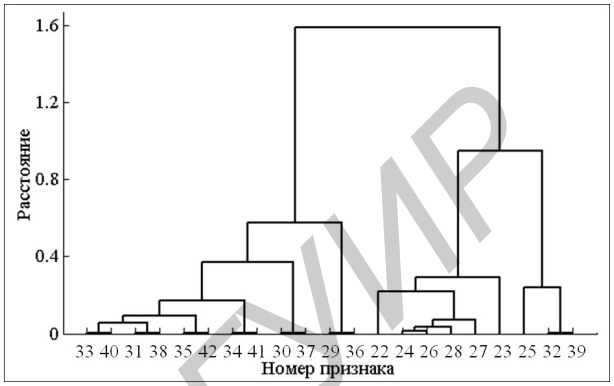

Рис. 2 – Дендрограмма признаков цвета

Для построения дендрограммы признаков цвета необходимо использовать расстояние dcor и метод средней связи, их кофенестически коэффициент слставляет $k = 0,99693$. Анализ дендрограммы признаков цвета позволяет выделить 4 группы параметров: 1 - MAX_G, MAX_B; 2 - STD_R, STD_G, STD_B; 3 - $MAX\_R, MIN\_R, MEAN\_R, MEDIAN\_R, Q3\_R, Q1\_R$; 4 - $MED\_G, MED\_B, MEAN\_G, MEAN\_B, Q3\_G, Q3\_B, Q1\_G, Q1\_B, MIN\_G, MIN\_B$. Параметры с равномерным законом распределения $(Y, BY, YSatrt, X, BX, XSatrt, Angle)$ не являются информативными поэтому их можно исключить из дальнейшего анализа. Из каждой группы признаков для дальнейшего описания объектов достаточно отобрать по одному признаку, таким образом, для описания объектов можно использовать только 9 признаков из имеющихся 42: $S, Perimeter, Width, Height, Circ, MAX\_G, STD\_B, MED\_R, MED\_G$.

III. Литература

1. Алгоритм автоматической сегментации границ ядер раковых клеток на трех-канальных люминесцентных изображениях / Лисица Е. В., Яцков Н. Н., Апанасович В. В., Апанасович Т. В. // :Журнал прикладной спектроскопии. – 2015. – Номер. 82(4). – Стр. 598-607.
2. Schindelin J. R. C. T., Hiner M.C., Eliceiri K.W. The ImageJ ecosystem: An open platform for biomedical image analysis // Mol Reprod Dev. 2015 T. 82. C. 518-529.
3. Kamentsky L. J. T. R., Fraser A., Bray M.A., Logan D.J., Madden K.L., Ljosa V., Rueden C., Eliceiri K.W., Carpenter A.E. Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software // Bioinformatics. 2011. T. 27, № 8. C. 1179-80.
4. Интеллектуальный анализ данных/ Н. Н. Яцков – Минск: БГУ, 2014. – 151 с.