Известия Гомельского государственного университета
имени Ф. Скорины, № (___), 2021

УДК 004.932.2

Методы кластеризации данных медицинских изображений в задачах компьютерной диагностики

Р.В. Козарь, А.А. Навроцкий, А. Б. Гуринович

В работе представлены результаты анализа существующих методов кластеризации данных медицинских изображений. Предложена модификация метода Виолы-Джонса с учетом кластеризации мягкого выхода.
Ключевые слова: медицинские изображения, метод Виолы-Джонса, кластеризация.

The clustering methods analysis results have presented for the medical imaging. A modification of the Viola-Jones method is proposed. It takes into account the clustering of «the soft output».

Введение. В настоящее время актуальны задачи принятия решений, зависящих от наличия на медицинских изображениях объекта, подлежащего классификации. Например, это — медицинские изображения, полученные с эндоскопической камеры. Способность распознавания считается основным свойством всех биологических существ. Компьютерные же системы этим свойством в полной мере не обладают. Основой алгоритма Виолы-Джонса для распознавания объектов является выделение локальных признаков каждого изображения и, далее, последующего обучения на них алгоритма. Для определения локальных признаков используются примитивы Хаара, которые, бесспорно, эффективны (см. [1, с. 145]).

В развитии метода были предложены примитивы с наклоном на 45 градусов и несимметричных конфигураций. Под данным признаком подразумевается трехмерный вектор следующего вида (1).

$$j = \{\text{маска}, \text{положение}, \text{размер}\} \quad (1)$$

Каждая такая маска характеризуется размером светлой и темной областей, определенными пропорциями и минимальным размером. Пример часто используемых масок из метода Виолы-Джонса изображен на рисунке 1.

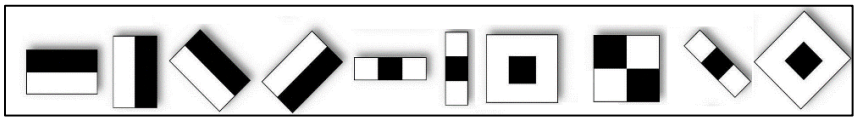


Рисунок 1 – Листинг набора масок для метода распознаваний Виолы-Джонса

Сохранение мелких деталей — исключительно важная составляющая при генерации изображений и в трехмерной компьютерной графике, и при обработке медицинских изображений (см. [2, с. 102]). Обработка медицинских изображений выполняется при помощи метода скользящего окна. Так как изначальный объект на изображении может иметь различный масштаб, следовательно, необходимо выполнить поиск объекта с признаками различного масштаба. Это позволит выполнять одинаковые и однотипные вычисления на разных областях исходного изображения, а также на разных масштабах их признаков. Скользящее окно U можно описать начальным размером Ux и Uy, вводится величина сдвига по осям координат. Первоначально в алгоритме минимальный размер скользящего окна был равен 24x24 пикселя. Однако использование базового алгоритма Виолы-Джонса для распознавания медицинских изображений имеет ряд недостатков:

– длительное время обучения при обработке медицинских изображений;

– большое количество близко расположенных друг к другу результатов, связанных с особенностями медицинских изображений (по причине применения различных масштабов и скользящего окна).

Распознавание медицинских изображений методом Виолы-Джонса.

Алгоритм Виолы-Джонса анализирует каждую область и находит однозначное решение о принадлежности искомого объекта к рассматриваемой области. Области изображения, которые прошли через весь каскад, классифицируются тогда, когда правильно классифицируются все прецеденты (см. [3, с. 21]). Отклик признака *fj(x)* вычисляется как разность интенсивностей пикселей как в светлой, так и в темной областях. Также важно отметить, что базовый алгоритм оперирует понятием слабого классификатора *hj(x)*, который вычисляется по формуле, представленной ниже:

$$h_j(x) = \begin{cases} 1, p_j, f_j(x) < p_j\theta_j \\ 0, \text{иначе} \end{cases} \quad (2)$$

В данном выражении параметром *Pj* является паритет, а параметром θj является граница. Данные параметры подпираются в процессе процедуры обучения.

Важно отметить, что окончательное решение принимается на основании значения сильного классификатора, значение которого вычисляется по следующей формуле:

$$H(x) = \begin{cases} 1, \sum_{t=1}^{T} a_t h_{j(t)}(x) \geq \frac{1}{2}\sum_{t=1}^{T} a_t \\ 0, \text{иначе} \end{cases} \quad (3)$$

Описанный выше алгоритм имеет существенный недостаток. Он заключается в том, результатом работы алгоритма имеется большое количество данных, обусловленное применением скользящего окна и спецификой самих медицинских изображений.

Следовательно, была поставлена цель: разработка эффективного метода объединения большого количества разобщённых данных на базе алгоритма Виолы-Джонса и анализа полученных результатов. Объединение будет осуществляться при помощи алгоритмов кластеризации.

Критерии оценки эффективности распознавания медицинских изображений.

Для сравнения традиционно используется статистические критерии, в частности, t-распределение (распределение Стьюдента) (см. [4, с. 375]). В исследовании используются известные критерии оценки, а также предложен новый критерий на базе алгоритмов эталонных данных (*Ground Truth*) и элементов ROC-анализа (*the Receiver Operator Characteristic*). Оба они представляют собой аппарат для анализа качества построенных моделей. Оба алгоритма активно используются для построения моделей в медицине и проведения клинических исследований.

Основа бинарной классификации в машинном обучении состоит в том, что все результаты эксперимента делятся на группы:

– истинно положительные примеры (*TP - True Positives*) – верно классифицированные примеры;

– истинно отрицательные примеры (*TN - True Negatives*) – верно классифицированные отрицательные примеры;

– положительные примеры, ошибочно принятые как отрицательные (*FN - False Negatives*) – ошибка I первого рода «ложный пропуск», вероятность этой ошибки - уровень значимости (пропуск классификатором присутствующего на изображении объекта);

– отрицательные примеры, классифицированные как положительные (*FP- False Positives*) – ошибка II второго рода «ложное срабатывание», вероятность которой - мощность критерия (определение классификатором объекта на изображении, которого в действительности там нет).

Важно понимать, что ошибка классификации происходит только тогда, когда классификатор относит входной объект к классу *Ci*, в то время, как для верного класса *Cj* выполняются следующие условия:

$$i \neq j \text{ и } C_i \neq C_j \tag{4}$$

К типовым оценкам эффективности распознавания медицинских изображений следует также отнести такие понятия, как:

– доля положительных примеров (*True Positives Rate*);

– доля ложноположительных примеров (*False Positives Rate*).

Выше описанные понятия рассчитываются по формулам, описанным ниже:

$$TPR = \frac{TP}{TP + FN} \times 100\% \tag{5}$$

$$FPR = \frac{FP}{TN + FP} \times 100\% \tag{6}$$

Дополнительный критерий оценки эффективности работы алгоритма распознавания медицинских изображений может быть сформулирован на основе идеи эталонных данных. За его основу взят алгоритм вычисления численного критерия оценки *er* как отношения непересекающихся частей к общей площади областей. Схема данного метода изображена на рисунке 2.

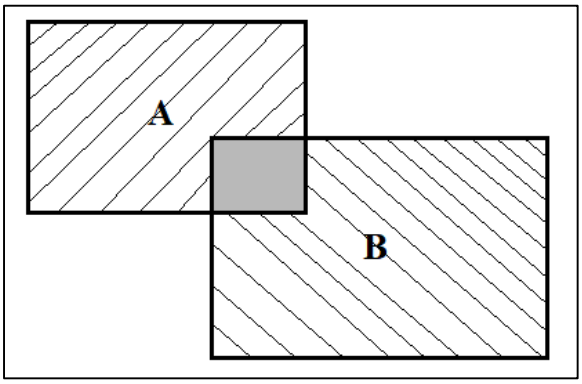


Рисунок 2 – Пример вычисления критерия на основании идеи эталонных данных

Формула для расчета данного критерия представлена ниже:

$$er(A, B) = \frac{|A \backslash B| + |B \backslash A|}{|A| + |B|} \tag{7}$$

Данный введенный критерий будет показывать ошибку несовпадения двух рассматриваемых областей, при этом он также учитывает взаимное положение объектов между собой. По сути, он представляет собой вероятность ошибки. Численное значение критерия ограничено интервалом [0, 1]. Когда значение достигается 1 можно сделать следующий вывод: объекты не пересекаются и получен результат FP. Когда достигается значение *er*, равное 0, тогда получаем полное совпадение с указанной областью. На основании данного вывода можно судить о результате TP, т.е. вероятность ошибки стремится к 0.

Модификация алгоритма распознавания медицинских изображений.

В оригинальном алгоритме решение о том, содержится ли основной объект в рассматриваемом скользящем окне, принимается однозначное решение. Либо объект присутствует, либо нет. Данное решение принимается на основании формулы (3). Модификация заключается в другом «нечетком» подходе, суть которого в том, что результат будет «нечетким» и на его основе можно будет принять другое решение (решение является неоднозначным).

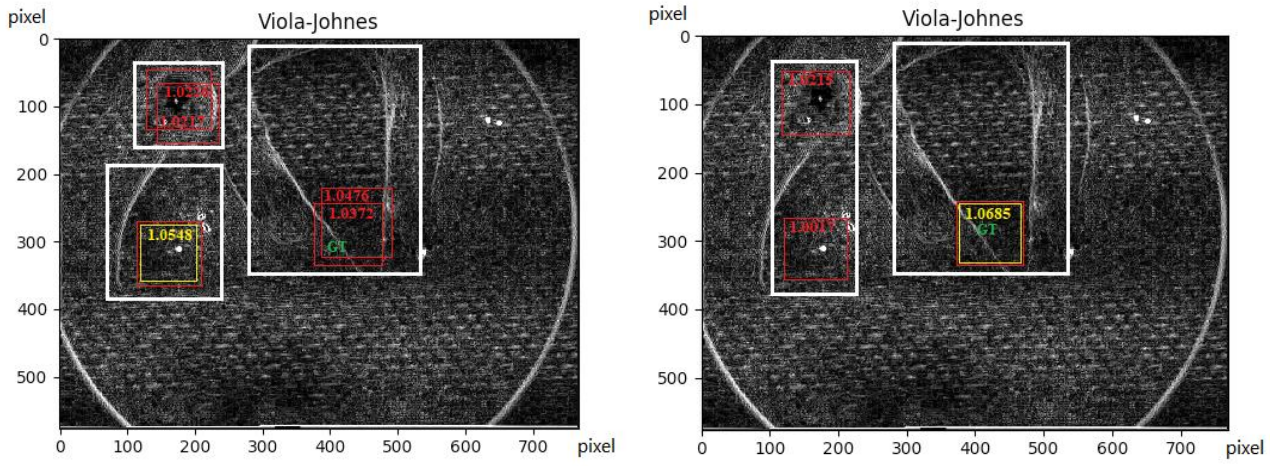
В проведенных опытах было обнаружено, что при обработке медицинских изображений в некоторых условиях алгоритмом Виолы-Джонса принималось решение о том, что искомого объекта нет на текущем скользящем окне, в то время как в реальности объект существовал. Основная идея заключается в «неточной» оценке нахождения объекта в рассматриваемой области. В оригинальном алгоритме для того, чтобы классификатор вынес решение о том, что в рассматриваемой области присутствует искомый шаблон, необходимо выполнение условия сильного классификатора. Данное утверждение описано формулой (3). Произведем модификацию данной формулы к следующему виду:

$$H(x) = \begin{Bmatrix} 1, \text{если } \dfrac{2 \times \sum_{t=1}^{T} a^{(t)} h_j{}^{(t)}}{\sum_{t=1}^{T} a^{(t)}} \geq 1 \\ 0, \text{иначе} \end{Bmatrix} \quad (8)$$

После этого необходимо избавиться от принятия решения по порогу. Для этого необходимо рассмотреть отношение в формуле (7) как отдельный числовой критерий, который назовем «мягким выходом». Описание мягкого выхода представлено формулой ниже:

$$Hs(x) = \frac{2 \times \sum_{t=1}^{T} a^{(t)} h_j{}^{(t)}}{\sum_{t=1}^{T} a^{(t)}} \quad (9)$$

Одним из возможных применений данного критерия является принятие однозначного решения о найденном объекте, исходя из значений *Hs(x)*. Этот критерий стал очень важным при обработке медицинских изображений, поскольку области определяются с гораздо меньшей вероятностью ошибки. На рисунке 3 показан пример работы оригинального алгоритма Виолы-Джонса с дополнительными значениями.



а) б)

Рисунок 3 – Результаты работы алгоритма распознавания Виолы-Джонса с критериями «мягкого выхода»: исследование №1 (а), исследование №2 (б)

Из рисунка б) можно видеть, что белым выделены области, найденные алгоритмом Виолы-Джонса. Белым отмечена необходимая истинная область. В каждой найденной области найдены красные подобласти. В каждой такой подобласти числом записано значение мягкого выхода. Также присутствует так называемая отметка *GT*, как область с наименьшей ошибкой *Ground Truth* по отношению к областям, выделенным белым цветом. Видно, что найденный алгоритмом объект в реальности имеет наибольшее значение мягкого выхода, равное 1.0685.

Из рисунка а) видно, что область с максимальным значением мягкого выхода 1.0548 наоборот имеет недостаточно положительное срабатывание и не может быть классифицирован как необходимый объект классификации.

Приведенные выше примеры наглядно демонстрируют то, что можно достоверно утверждать, что область с максимальным значением мягкого выхода является искомым объектом.

Далее будет показано, что выведенная формула мягкого выхода (8) может быть использована при кластеризации решений базового алгоритма Виолы-Джонса.

Кластеризация данных с учетом нового критерия «мягкого выхода».

Процесс кластеризации выполняет поиск структуры в коллекции немаркированных данных. Можно по-другому сказать, что это процесс организации объектов в группы, члены которой в некотором смысле похожи. Кластер – это множество похожих объектов, между тем, как непохожие принадлежат другим кластерам. К механизму кластеризации медицинских изображений предъявляются следующие требования:

– масштабируемость (однотипная работа на разных объемах данных);

– минимальные требования начальных знаний о природе объектов;

– способность отбрасывать шум и выбросы при распознавании.

Под указанные требования подходит алгоритм *DBSCAN*. Данный алгоритм прост в реализации и понимании. Этот алгоритм оперирует такими понятиями, как сосед, расстояние до соседей и количество соседей. Данные параметры на базе этих понятий могут быть использованы для повышения эффективности задачи распознавания медицинских изображений.

DBSCAN (Density-based spatial clustering of applications with noise) – это плотностной алгоритм пространственной кластеризации с присутствием шума. Также в его работе присутствует бустинг. Бустинг – комплекс методов, способствующих повышению точности аналитических моделей (см. [5, с. 775]). *DBSCAN* – это алгоритм кластеризации, основанной на плотности – если дан набор точек в некотором пространстве, алгоритм группирует вместе точки, которые тесно расположены (точки со многими близкими соседями), помечая как выбросы точки, которые находятся одиноко в областях с малой плотностью (ближайшие соседи которых лежат далеко). Также в своей работе он использует вейвлеты Хаара единичного масштаба и нулевого смещения (см. [6, с. 74]).

Из преимуществ данного алгоритма для обработки медицинских изображений можно выделить следующие:

– *DBSCAN* не требует спецификации числа кластеров в данных априори в отличие от метода *k*-средних;

– *DBSCAN* может найти кластеры произвольной формы;

– *DBSCAN* имеет понятие шума и устойчив к выбросам;

– *DBSCAN* требует лишь двух параметров и большей частью нечувствителен к порядку точек в базе данных.

В задаче кластеризации – задаче группировки множества объектов на подмножества [7] – данных также можно учитывать дополнительную информацию, которой является значение нового критерия мягкого выхода. Также, используя эту информацию, можно искусственно снизить порог распознавания и учитывать полученные распознанные области на этапе кластеризации данных. Очень важно понимать, что при снижении порога значительно увеличивается количество найденных областей.

На рисунке 4 отображено среднее количество найденных объектов в зависимости от значения критерия мягкого выхода. Идея предложенного метода заключается в проверке и подтверждении следующей гипотезы: при уменьшении порога срабатывания алгоритма распознавания и применяя данный метод кластеризации объектов становится возможным улучшить эффективность его распознавания.

Также при этом следует отметить, что параметры в данном алгоритме кластеризации подбираются эмпирически.

Также, проведя серию экспериментов был сформирован следующий график зависимости, по которому была точно определена оптимальная пара значений для алгоритма кластеризации при обработке данных медицинских изображений (а именно количество соседей, минимальное расстояние между соседями).

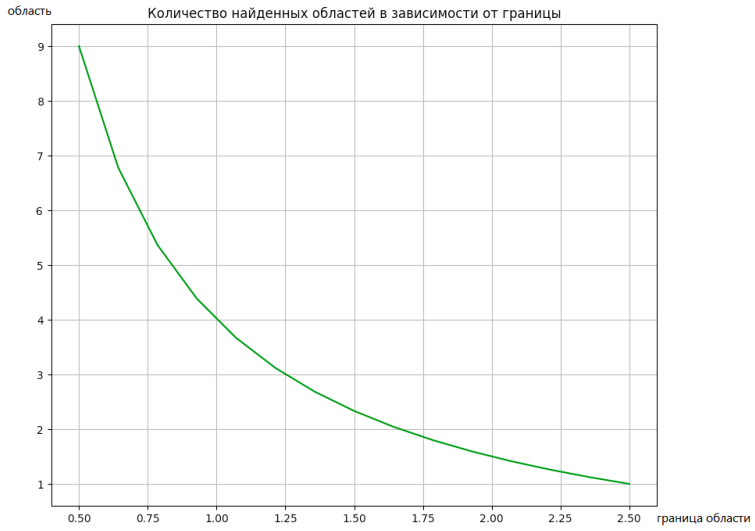


Рисунок 4 – Зависимость количества найденных объектов от порога срабатывания

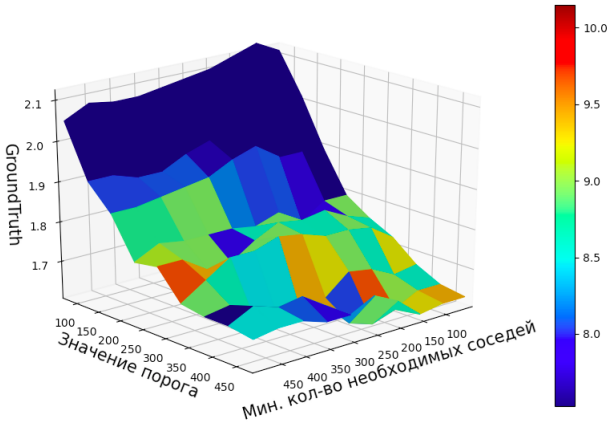


Рисунок 5 – График зависимости *GroundTruth* в зависимости от порога и параметра кластеризации

Также, учитывая вероятность ложного срабатывания, представленного в таблице 1, необходимо выбрать подходящие параметры.

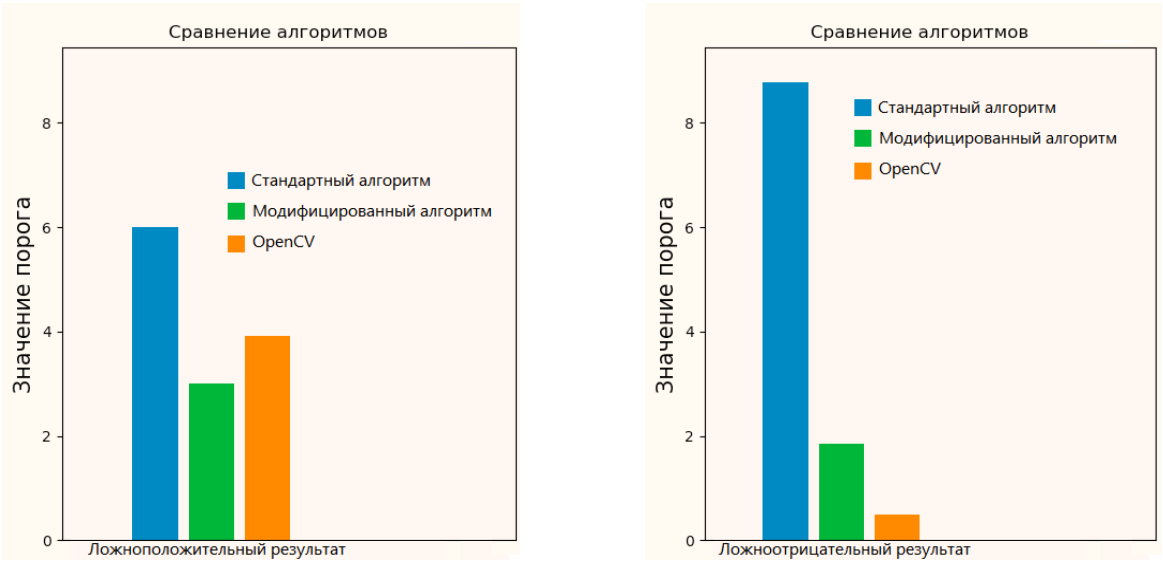
Таблица 1. Вероятность ложного срабатывания в зависимости от порога.

Порог алгоритма распознавания	0,82	0,84	0,86	0,88	0,9	0,92	0,94	0,96	0,98	1
Вероятность ложного срабатывания	98%	78%	18%	15%	15%	14%	8%	2%	1%	0%

Анализ полученных результатов.

Проводимые эксперименты анализируют медицинские изображения, полученные путем фотографирования медицинским эндоскопом. Сравнение производилось с использованием 10 снимков. При сравнении использовались следующие критерии оценки: ложный пропуск (*FN*) и ложное обнаружение (*FP*). Оценка эффективности распознавания медицинских изображений в данной статье велась с двумя альтернативными реализациями алгоритмов Виолы-Джонса: типовой алгоритм из *OpenCV* и модифицированный алгоритм с дополнительной кластеризацией. На рисунке 6 показан результат по обоим критериям сравнения. Видно, что представленная модификация алгоритма Виолы-Джонса гораздо эффективнее обрабатывает ошибки 1 рода, чем стандартный алгоритм и его реализация в *OpenCV*. В то же время ошибки 2 рода обрабатываются гораздо хуже, чем в *OpenCV*, но заметно лучше, чем в оригинальном алгоритме. Это позволяет сделать вывод, что предложенная модификация метода распознавания медицинских изображений работает гораздо эффективнее оригинального метода.

При различных значениях параметров алгоритма кластеризации изменяется процентное соотношение ошибок 1 и 2 рода. Также необходимо уточнить, что подобранные параметры алгоритма кластеризации и пороговое значение критерия мягкого выхода показывают одинаковые результаты на различных тестовых выборках. Данное утверждение позволяет сделать следующий вывод: предложенный метод инвариантен к рассматриваемым исходным данным.



а) б)

Рисунок 6 – Сравнение результатов работы алгоритмов распознавания медицинских изображений: ложноположительный результат (а), ложноотрицательный результат (б)

Выводы.

В статье исследована эффективность использования алгоритма Виолы-Джонса для распознавания медицинских изображений. Определено, что при использовании этого метода при распознавании и обработке медицинских изображений необходимо большое количество выходных данных алгоритма распознавания и длительное время работы, требуемое для обучения. Установлено то, что использование нового критерия оценки, который был назван «мягким выходом», дает однозначный ответ о принадлежности рассматриваемой области к искомому объекту.

Предложен метод распознавания объектов, который использует измененное значение «мягкого выхода» и дополнительную постобработку в виде кластеризации полученных областей. Метод высоко результативен при распознавании медицинских изображений, что подтверждено приведенными тестами.

Полученные результаты убедительно показали, что предложенная модификация алгоритма на 58% эффективнее базового алгоритма Виолы-Джонса справляется с ошибками 1 рода и на 79% эффективнее справляется с ошибками 2 рода. По сравнению с известной реализацией алгоритма в библиотеке *OpenCV*, предложенный алгоритм на 32% эффективнее справляется с ошибками 1 рода и на 65% хуже с ошибками 2 рода.

В дальнейшем, планируется улучшить эффективность работы предложенного алгоритма для уменьшения значения вероятности ошибки 2 рода.

Литература

1. Viola P., Jones M.J. Robust real time face detection // International Journal of Computer Vision. – 2004. – V. 57. – № 2. – P. 137–154.

2. Гонсалес Р., Вудс Р. Цифровая обработка изображений. – М.: Техносфера, 2005. – 1072 с.

3. Местецкий Л.М. Математические методы распознавания образов. – М.: МГУ, ВМиК, 2002–2004. – С. 20–24.

4. Шапиро Л. Компьютерное зрение / Л. Шапиро, Д. Стокман. – М.: Изд. Дом «Лаборатория знаний», 2015. – 763 с.

5. Freund Y., Schapire R.E. A Short Introduction to Boosting // Journal of Japanese Society for Artifical Intelligence – September 1999. – V. 14. – № 5. – P. 771–780.

6. Буй Тхи Тху Чанг, Спицын В.Г. Разложение цифровых изображений с помощью двумерного дискретного вейвлет-преобразования и быстрого преобразования // Известия Томского политехнического университета. – 2011. – Т. 318. – № 5. – С. 73–76.

7. Кластеризация. Университет ИТМО. [Электронный ресурс]. – Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=Кластеризация. – Дата доступа: 20.05.2021.

Medical Image Data Clustering Techniques in problems of computer diagnostics

R.V. KOZAR, A.A. NAUROTSKY, A. B. GOURINOVITCH