КЛАССИФИКАЦИЯ ГРАФИЧЕСКИХ ОБЪЕКТОВ НА ИЕРОГЛИФИЧЕСКИХ ДОКУМЕНТАХ

К.Б. Стаценко,
аспирант Московского физико-технического института (Государственного Университета),
Д.Г. Дерягин,
руководитель группы обработки документов, ООО «Аби Продакшн» (Компания ABBYY).
А.В. Мякутин,
руководитель группы анализа документов, ООО «Аби Продакшн» (Компания ABBYY),
e-mail: Konstantin_S@abbyy.com.
Адрес: 129301, Россия, Москва, а/я 49, компания ABBYY.

Частью процесса распознавания текста является анализ структуры обрабатываемого документа. В ходе такого анализа возникает потребность в определении типов областей, найденных на изображении. В данной статье рассмотрен подход к решению подобной задачи на примере анализа иероглифических документов. В статье также приведены результаты экспериментальной апробации предложенного подхода.

Ключевые слова: классификация, анализ документа, распознавание текста.

Введение

Современное программное обеспечение для распознавания текста представляет собой сложный комплекс различных подсистем, отвечающих за те или иные его функции. Данная статья коснется одной из них, предназначенной для анализа документа. Она необходима для поиска и локализации графических объектов, таких как картинки, таблицы, рукописный и печатный текст, диаграммы и т.п. В процессе её работы выполняется две важные задачи: сегментация (разделение страницы на области, обладающие определенными свойствами) и классификация (определение типа каждой из областей). В ходе анализа документа эти задачи приходится выполнять неоднократно, начиная с поиска простейших объектов, таких как пунктуация, отдельные символы или разделители и заканчивая выделением таких высокоуровневых объектов, как бланк официального письма или текстовая колонка с врезками. В данной статье затрагиваются подходы к решению задачи классификации отдельных символов и их последовательностей на печатных иероглифических документах. Описываемые здесь классификаторы предназначены для работы с фрагментами текста, в то время как другие полезные объекты (картинки, разделители) должны быть обработаны отдельно, и будут рассматриваться как мусор.

Постановка задачи

Будем рассматривать черно-белую страницу с черным текстом на белом фоне (участки белого текста на черном фоне должны быть предварительно инвертированы). В процессе низкоуровневой сегментации подобного изображения на нем выделяются объекты, которые могут оказаться фрагментами текста, оформления (элементами картинок или диаграмм, черными или точечными разделителями и т.п.) или мусором (неоднородности бумаги, посторонние объекты за краями сканируемого листа, тени в области переплета и т.п.). Они, как правило, представляют собой связные черные области или группы таких областей, объединенных исходя из их размеров и геометрического положения. На *рис. 1* представлен пример типичного результата сегментации фрагмента страницы, прямоугольниками отмечены найденные объекты.

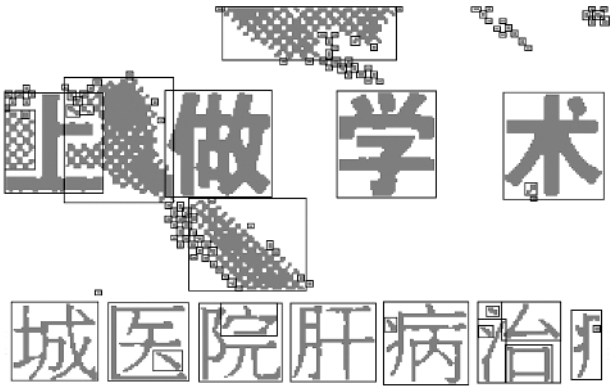

Рис. 1. Типичный результат сегментации фрагмента страницы. *Прямоугольниками обозначены найденные объекты.*

Каждый объект характеризуется *вектором свойств* $\boldsymbol{p}$ из шести элементов, соответствующих его ширине, высоте, количеству черных пикселей, количеству горизонтальных RLE штрихов, количеству вертикальных RLE штрихов и количеству белых дырок (RLE штрих представляет собой непрерывную последовательность черных пикселей и описывается двумя числами – начальной и конечной позициями). Свойства подсчитываются на этапе выделения и объединения черных областей, при этом в результирующем множестве могут одновременно присутствовать как составные, так и составляющие их объекты.

В результате классификации выделенных таким образом объектов, для каждого из них требуется определить, является ли он словом (цепочкой иероглифов), символом (отдельным иероглифом) или мусором, при этом отличать слово от символа не требуется (допускается одновременное отнесение объекта к обоим этим классам). Как правило, однозначно определить принадлежность объекта к одному из перечисленных классов на данном этапе не представляется возможным. Это связано со спецификой задачи, которая заключается в проведении предварительного, грубого и быстрого по времени анализа. Для описания принадлежности объектов к классам используется система качеств. Объект характеризуется двумя переменными, изменяющимися от нуля (худшее качество) до единицы (лучшее качество) которые соответствуют качествам символа и слова. Если обе они близки к нулю, то объект считается мусором.

Решение задачи

Чтобы вычислить искомые качества необходимо задать две функции вида $F(\boldsymbol{p})$: $\boldsymbol{X} \to [0, 1]$ на множестве $\boldsymbol{X}$ векторов свойств. Для этого предлагается несколько подходов, объединенных общей идеей. В каждом из них искомая функция $F(\boldsymbol{p})$ представляется с помощью набора простых функций $f_i(\boldsymbol{p})$, зависящих от некоторых параметров. Эти параметры предлагается определять при помощи различных алгоритмов неконтролируемой кластеризации.

Метод, основанный на алгоритме ISODATA

Описание метода.

В первом из предлагаемых методов каждая функция $F(\boldsymbol{p})$ представлялась с помощью набора простых функций $f_i(\boldsymbol{p})$, которые в свою очередь представлялись в виде произведения пороговой функции $t(p, a, b, c, d)$ следующим образом:

$$F(\boldsymbol{p}) = \max_{i \in [1, N_c]} \left(f_i(\boldsymbol{p}) \right);$$
$$f_i(\boldsymbol{p}) = \prod_j^K t\left(p_j, a_{ij}, b_{ij}, c_{ij}, d_{ij} \right); \quad (1)$$

где: p_j – $j^{\text{ая}}$ компонента вектора $\boldsymbol{p}$;
a_{ij}, b_{ij}, c_{ij} и d_{ij} – параметры пороговой функции;
K – количество компонент $\boldsymbol{p}$.

Другими словами, $F(\boldsymbol{p})$ помещает объект с вектором свойств $\boldsymbol{p}$ в один или несколько кластеров N_c, описываемых при помощи функций $f_i(\boldsymbol{p})$. Пороговая функция задавалась следующим образом:

$$t\underset{a\le b\le c\le d}{(p,a,b,c,d)} = \begin{cases} 0 : p \in (-\infty; a] \cup [d; +\infty) \\ \dfrac{p-a}{b-a} : p \in (a; b) \\ 1 : p \in [b; c] \\ \dfrac{d-p}{d-c} : p \in (c; d) \end{cases} \quad (2)$$

Отсюда видно, что каждый кластер представлял собой *K*-мерный параллелепипед с нечеткой границей, размеры и ширина границы которого задавались при помощи параметров a_{ij}, b_{ij}, c_{ij} и d_{ij}. Область, занимаемая такими кластерами в пространстве свойств, может иметь достаточно сложный вид, особенно, при большем их количестве. Это позволяет достаточно точно описывать подмножества объектов рассматриваемых типов (символы и слова), однако чрезмерное увеличение числа кластеров может привести к переобученности (overfitting) и замедляет работу классификатора.

Чтобы ограничить количество кластеров, необходимых для определения искомых функций и, в то же время, не утратить возможность описывать геометрически сложные области на множестве $\boldsymbol{X}$ векторов свойств $\boldsymbol{pX}$, пороговая функция t вычислялась не от компонент $\boldsymbol{p}$, а от значений признаков q_j. Каждый признак задавался при помощи подбираемой экспериментально функции $q_j(\boldsymbol{p})$. В итоге выражение (1) приобрело следующий вид:

$$F(\boldsymbol{p}) = \max_{i\in[1,N_c]}\left(f_i(\boldsymbol{p})\right);$$
$$f_i(\boldsymbol{p}) = \prod_{j}^{K'} t\left(q_j(\boldsymbol{p}), a_{ij}, b_{ij}, c_{ij}, d_{ij}\right); \quad (3)$$

где K'– количество признаков, вообще говоря, не совпадающее с размерностью $\boldsymbol{p}$. Другими словами, функция F задавалась не на множестве свойств $\boldsymbol{X}$, а на некотором множестве $\boldsymbol{Y}$. Введение признаков, имеющих четкий физический смысл упрощало настройку и отладку классификатора.

Подбор признаков

Поскольку число исходных свойств было изначально фиксированным и небольшим, количество признаков также предполагалось ограничить пятью – семью. В этом случае, в отличие от [5], создание автомата для подбора признаков было признано нецелесообразным.

На первом этапе подбора был сформирован ориентировочный набор признаков, которые казались наиболее полезными. В него вошли:

1. Отношение количества черного к полусумме количеств вертикальных и горизонтальных RLE штрихов (средняя толщина черты иероглифа).
2. Отношение ширины к высоте.
3. Количество белых связных областей (дублировал соответствующее свойство).
4. Отношение количества черного к площади объекта.
5. Полусумма количеств вертикальных и горизонтальных RLE штрихов (длинна границы черное/белое).
6. Отношение количества горизонтальных RLE штрихов к площади объекта (плотность горизонтальных RLE штрихов).
7. Отношение количества вертикальных RLE штрихов к площади объекта (плотность вертикальных RLE штрихов).

Все признаки нормировались так, чтобы все они были безразмерными величинами. Для выяснения значимости признаков проводилось сравнение качества классификации при использовании всех семи признаков с качеством классификации при использовании только шести из них. *На рис. 2* приведены результаты описанного сравнения, количество ошибок при классификации по семи признакам принято за 100%, номер столбика гистограммы соответствует номеру исключаемого признака в приведенном выше перечислении. По результатам сравнения худший признак (в данном случае №3) был удален. В результате повторного выполнения описанной процедуры был удален признак №7. В третий раз малоинформативных признаков процедура не выявила.

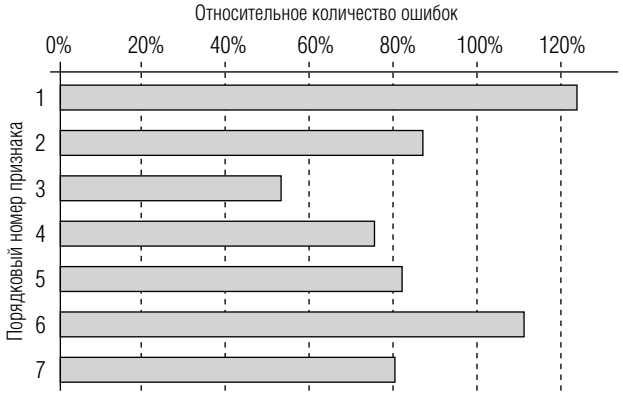

Рис. 2. Зависимость качества классификации от исключаемого признака.

(Количество ошибок при классификации по семи признакам принято за 100%, номер столбика гистограммы соответствует номеру исключаемого признака в перечислении, приведенном на странице 59.)

На втором этапе был рассмотрен набор новых признаков – кандидатов на включение, в него вошли:

1. Отношение количеств вертикальных и горизонтальных RLE штрихов.
2. Ширина объекта.
3. Высота объекта.
4. Отношение количества белых дырок к количеству черного (плотность дырок).
5. Отношение количества дырок к полусумме количеств вертикальных и горизонтальных RLE штрихов.
6. Отношение количества черного к произведению высоты объекта на количество вертикальных RLE штрихов (относительная средняя длинна вертикального RLE штриха).
7. Отношение количества черного к произведению ширины объекта на количество горизонтальных RLE штрихов (относительная средняя длинна горизонтального RLE штриха).

Полезность признаков нового набора оценивалась так же, как и предыдущего, с той лишь разницей, что признаки не исключались, а добавлялись. Исходным вариантом служил набор из пяти признаков, полученный в ходе предыдущего этапа настройки. На *рис. 3* приведены результаты оценки новых признаков, в соответствии с которыми набор пополнился признаком №7.

Таким образом, был получен окончательный перечень используемых признаков. Аналогичная процедура выполнялась и для классификатора на базе алгоритма EM, который будет описан ниже.

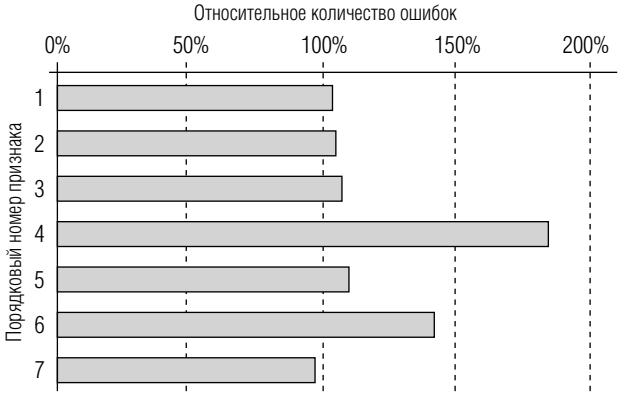

Рис. 3. Зависимость качества классификации от добавляемого признака из набора, приведенного на странице 5.
(Количество ошибок при классификации по семи исходным признакам принято за 100%.)

Обучение.

Определение количества кластеров и их параметров осуществлялось автоматически в процессе обучения классификатора. Для этого использовался алгоритм неконтролируемой кластеризации ISODATA [1]. Его основная идея заключается в минимизации суммы квадратов расстояний до центров кластеров. При этом количество кластеров может меняться в ходе итеративного процесса вследствие разделения и слияния. Разделение кластера выполняется в том случае, если его дисперсия вдоль одной из координат превышает порог разделения. Слияние кластеров осуществляется, когда расстояние между их центрами оказывается меньше порога слияния.

После того как при помощи описанного алгоритма определялись координаты центров кластеров в пространстве признаков и наборы объектов из обучающей выборки, вошедших в каждый кластер, подбирались значения параметров *aij, bij, cij* и *dij*. Это делалось таким образом, чтобы $\forall j: j \in [1; K]$ в пограничные интервалы $(a_{ij}; b_{ij})$ и $(c_{ij}; d_{ij})$ попал заданный небольшой процент точек *i*-го кластера, в то время как в интервал $(a_{ij}; d_{ij})$ должно было попасть большинство точек этого кластера. На *рис. 4* показан пример результата обучения предложенного метода в случае двумерного пространства признаков. Обучающая выборка разбита на три группы, вокруг каждой из которых установлены границы. Пунктирными прямоугольниками обозначены внешние границы кластеров, за которыми $f_i(\boldsymbol{p}) = 0$, сплошными прямоугольниками отмечены внутренние границы, в пределах которых $f_i(\boldsymbol{p}) = 1$.

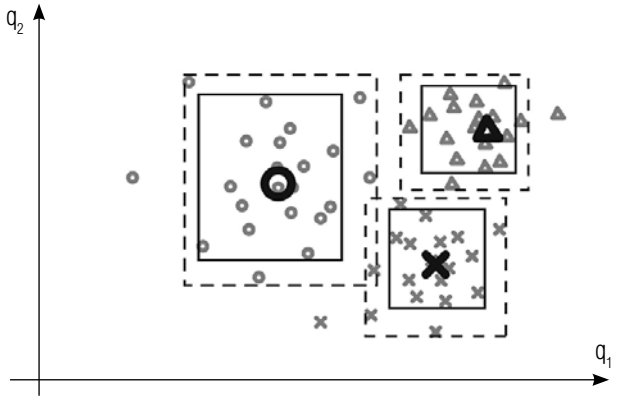

Рис. 4. Кластера, выделенные при помощи алгоритма ISODATA.
Черными крупными значками отмечены центры кластеров. Пунктирными линиями показаны внешние границы, а сплошными внутренние границы кластеров.

Тестирование.

Результаты работы построенного таким образом классификатора проверялись на тестовой выборке, состоящей из заранее разбитых по классам экземпляров слов, букв и мусора. В выборку входили тексты, напечатанные различными шрифтами на японском, китайском и корейском языках. Слово считалось классифицированным «уверенно» если соответствующая оценка классификатора превышала 0.5, если оценка лежала в промежутке [0.5; 0.1] объект считался классифицированным «неуверенно», иначе «потерянным». Этот же принцип действовал и для символов. Мусор считался «отвергнутым» если наибольшая из двух оценок не превышает 0.1, если же она оказывается в промежутке [0.5; 0.1] то объект считался классифицированным «неуверенно», иначе «пропущенным». В *табл. 1* приведены результаты проверки.

Таблица 1.

Результаты работы классификаторов

Примененный классификатор	Слов потеряно	Слов неуверенно	Символов потеряно	Символов неуверенно	Мусора пропущено	Мусора неуверенно	Количество кластеров	Независимых переменных
На основе алгоритма ISODATA	4.8%	5.1%	0.35%	1.8%	1.1%	1.8%	17	408
На основе алгоритма EM	0.12%	1.0%	0.25%	1.1%	1.0%	1.1%	25	1075
На основе алгоритма EM с упрощенными признаками	0.15%	0.9%	0.31%	1.2%	1.2%	0.9%	30	2190
Упрощенный, на основе алгоритма EM	0.18%	1.4%	0.47%	1.0%	1.8%	1.4%	50	600

Классификатор показал неплохие результаты для различения символов и мусора, однако потерял много слов что, скорее всего, связано с особенностью распределения этих объектов в пространстве признаков. Слова, в отличие от символов, не концентрируются вокруг определенных значений, а имеют достаточно большой и неоднородный разброс по некоторым признакам. В этом случае ситуацию можно скорректировать использованием разных наборов признаков для классификации символов и слов. Также оказалось, что качество классификации сильно зависит от соотношения дисперсий признаков и становится наилучшим в случае одинаковых дисперсий. Всё вышеперечисленное накладывает излишние требования на используемые признаки, чем затрудняет подбор их оптимального состава.

Метод, основанный на алгоритме EM

Описание метода.

Для снижения требований к признакам и повышения надежности классификации был предложен второй подход. Он основан на аппроксимации искомых функций при помощи кластеров, которые описываются не как многомерные параллелепипеды, а при помощи многомерного нормального распределения:

$$\varphi_i(\boldsymbol{q}) = \frac{1}{(2\pi)^{K'/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{q}-\mu_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{q}-\mu_i)} \quad (4)$$

Где μ_i – радиус-вектор центра, а $\boldsymbol{\Sigma}_i$ – матрица ковариации i-ого кластера. В этом случае искомые функции $F(\boldsymbol{p})$ записываются как:

$$F_{w,c}(\boldsymbol{p}) = \frac{\phi_{w,c}(\boldsymbol{q}(\boldsymbol{p}))}{\phi_w(\boldsymbol{q}(\boldsymbol{p})) + \phi_c(\boldsymbol{q}(\boldsymbol{p})) + \phi_t(\boldsymbol{q}(\boldsymbol{p}))} \quad (5)$$

$$\phi_{w,c,t}(\boldsymbol{q}) = \sum_i^{N_{w,c,t}} \alpha_i \varphi_{w,c,ti}(\boldsymbol{q}) \quad (6)$$

Где α_i – вес i-ого кластера, индексы w, c и t соответствуют функциям для слов, символов и мусора, а $\boldsymbol{q}(\boldsymbol{p})$ – векторная функция с координатами $q_j(\boldsymbol{p})$. Как видно из (5), в отличие от (3), в расчете $F(\boldsymbol{p})$ участвуют не только положительные, но и отрицательные примеры (в виде кластеров мусора $\phi_t(\boldsymbol{q})$).

Таким образом, если функции $\phi_{w,c,t}(\boldsymbol{q}(\boldsymbol{p}))$ будут описывать плотности вероятности $\boldsymbol{p}_{w,c,t}$ в пространстве X с точностью до ненулевого множителя, зависящего только от $\boldsymbol{p}$, то искомые оценки $F_w(\boldsymbol{p})$ и $F_c(\boldsymbol{p})$ будут иметь простой физический смысл. Они окажутся равными вероятностям того, что объект, имеющий свойства $\boldsymbol{p}$, принадлежит к клас-

сам слов и символов соответственно. Если функция $\boldsymbol{q}^{-1}: \boldsymbol{Y} \to \boldsymbol{X}$ инъективная, непрерывно дифференцируемая и якобиан $\boldsymbol{J}(\boldsymbol{q}^{-1}) \neq 0$, то для выполнения вышеуказанного условия достаточно, чтобы функция (6) описывала плотность вероятности случайной величины $\boldsymbol{q}$ в пространстве $\boldsymbol{Y}$.

Обучение.

Пусть иероглифы делятся на группы в зависимости от особенностей их начертания (например, от количества штрихов) с вероятностями α_i. При условии, что иероглиф принадлежит одной из таких групп, будем рассматривать описывающую его случайную величину $\boldsymbol{q}$ как линейную комбинацию N независимых, нормально распределенных случайных величин (скрытых параметров) g_i. Другими словами, будем описывать плотность иероглифов i-ой группы в пространстве признаков как (4), а относительную частоту встречаемости этой группы как α_i. В этом случае (6) будет соответствовать плотности вероятности $\boldsymbol{q}$.

Используя метод наибольшего правдоподобия для нахождения неизвестных α_i, μ_i и $\boldsymbol{\Sigma}_i$, входящих в выражение (6), необходимо максимизировать функцию правдоподобия

$$f(\boldsymbol{\Theta}) = P(\boldsymbol{\Theta}) \prod_{j}^{M} \phi(\boldsymbol{q}_j) = P(\boldsymbol{U}, \boldsymbol{\Theta}) = \sum_{\boldsymbol{J} \in \boldsymbol{G}} P(\boldsymbol{U}, \boldsymbol{\Theta}, \boldsymbol{J}) \quad (7)$$

Здесь $\boldsymbol{\Theta}$ – совокупность искомых переменных α_i, μ_i и $\boldsymbol{\Sigma}_i$, $i \in [1;N]$, $P(\boldsymbol{U}, \boldsymbol{\Theta})$ – плотность вероятности совместного распределения U и $\boldsymbol{\Theta}$, U – обучающая выборка, состоящая из M величин $\boldsymbol{q}_j$, $\boldsymbol{J}$ – скрытый параметр, описывающий распределение точек $\boldsymbol{q}_j$ по классам. Для этого применялся алгоритм EM (Expectation Maximization) [2,3], суть которого заключается в итеративной оптимизации значений искомых переменных в присутствии скрытых параметров, от которых необходимо избавиться интегрированием. Каждая итерация алгоритма состоит из двух шагов. На первом шаге (expectation) производится оценка распределения скрытых параметров, исходя из оценки значений искомых переменных, полученной в предыдущей итерации. На втором шаге (maximization) выполняется поиск новой оценки для искомых переменных, доставляющей максимум функции правдоподобия с учетом нового распределения скрытых параметров. Можно доказать [4], что данный алгоритм сойдется к локальному максимуму функции $P(\boldsymbol{U}, \boldsymbol{\Theta})$.

Тестирование.

Предложенный классификатор проверялся при помощи того же метода и на тех же данных что и предыдущий. Во второй строке *табл. 1* приведены результаты проверки. Видно существенное увеличение качества классификации слов, а также улучшения по остальным параметрам в сравнении с результатами предыдущего классификатора. Для достижения этого результата использовалось шесть безразмерных признаков, подобранных при помощи описанного выше алгоритма. В третьей строке упомянутой таблицы приведены результаты классификации, при условии использования восьми признаков. Шесть из них были равны значениям свойств, а оставшиеся два: площади объекта и произведению количества горизонтальных RLE штрихов на длину. Поскольку классификатор способен обучатся произвольной линейной комбинации признаков, результаты лишь незначительно ухудшились. Два дополнительных признака представляли собой нелинейные компоненты наиболее значимых признаков, использовавшихся в предыдущем тестировании (относительного количества черного и относительной длины горизонтального RLE штриха).

К недостаткам описанного алгоритма можно отнести замедление процесса классификации приблизительно в 3 раза вследствие усложнения формул для расчета $F(\boldsymbol{p})$ и опасность возникновения переобученности из-за большого количества независимых переменных, параметризующих решающее правило.

Метод, основанный на алгоритме EM (упрощенный)

О*писание метода.*

Для того чтобы устранить недостатки предыдущего метода, а именно, увеличить скорость классификации и уменьшить риск переобучения, но в то же время сохранить такие его достоинства как ясный физический смысл результата классификации и отсутствие необходимости особым образом нормировать признаки, был предложен третий метод. Так же как и предыдущий, он основан на аппроксимации искомых функций при помощи кластеров, которые описываются многомерными нормальными распределениями. Однако теперь, случайные величины, описываемые этими рас-

пределениями, считаются независимыми. В этом случае *формула (4)* будет выглядеть так:

$$\varphi_i(\boldsymbol{q}) = \frac{1}{(2\pi)^{K'/2}\prod_j^{K'}\sigma_{ji}} e^{-\frac{1}{2}\sum_j^{K'}\frac{(\boldsymbol{q}_j-\mu_{ji})}{\sigma_{ji}^2}} \quad (8)$$

Где μ_{ji} – j-ая компонента радиус-вектора центра i-ого кластера, σ_{ji}^2 – дисперсия j-ой случайной величины, при условии, что она принадлежит i-ому кластеру, $\boldsymbol{q}_j$ – j-ая компонента вектора $\boldsymbol{q}$. Обучение данного классификатора основано на алгоритме EM и полностью аналогично описанному выше. В четвертой стоке *табл. 1* приведены результаты тестирования этого классификатора. Результаты говорят о том, что упрощенный EM в отличие от EM, не способен обучаться произвольным линейным комбинациям признаков, соответствующий их подбор вновь становится актуальным. Тем не менее, в отличие от классификатора на базе алгоритма ISODATA, отсутствует необходимость в специальном нормировании признаков. Альтернативой тщательному их подбору является увеличение количества кластеров.

Заключение

В данной статье были описаны три подхода к классификации графических объектов на странице. Метод на основе алгоритма ISODATA обладает большей скоростью, показывает неплохие результаты при различении отдельных иероглифов и мусора, но хуже справляется с классификацией слов и требует более тщательной подборки признаков. Метод на основе алгоритма EM более универсален, неприхотлив и точен, способен обучатся линейным комбинациям признаков, хотя требует большего времени для своей работы и может приводить к переобучению. Упрощенный метод на основе алгоритма EM сочетает в себе ряд достоинств и недостатков двух предыдущих, занимая промежуточную позицию. Он не способен обучатся линейным комбинациям признаков, однако не требует их нормирования, скорость его работы сравнима с первым классификатором, а количество обучаемых переменных существенно меньше чем у второго. ■

Литература

1. A.K. Jain, M.N. Murty, P.J. Flynn. Data Clustering: A Review: ACM Computing Surveys. 1999. Т. 31, № 3.
2. С. Рассел, П. Норвиг. Искусственный интеллект. Современный подход: М. Вильямс, 2007 г.
3. F. Dellaert. The Expectation Maximization Algorithm: Georgia Institute of Technology. Technical Report number GIT-GVU-02-20. 2002.
4. G. McLachlan, T. Krishnan. The EM algorithm and extensions: Wiley series in probability and statistics. John Wiley & Sons. 1997.
5. Y. Zheng, H. Li, D. Doermann. Machine Printed Text and Handwriting Identification in Noisy Document Images: Ieee transactions on pattern analysis and machine intelligence. 2004, Т. 26, № 3, С. 337-353.