КЛАСТЕРИЗАЦИЯ ИЗОБРАЖЕНИЙ МЕТОДОМ «K-СРЕДНИХ»

С.А. Леонтьева, А.Ю.Демин
Руководитель: к.т.н. А.Ю. Демин
Томский политехнический университет

Введение

Стремительное развитие вычислительных мощностей и постоянное снижение их стоимости сделали возможным хранение больших объемов оцифрованных изображений. В настоящее время электронные коллекции изображений используются все чаще и чаще. Наиболее перспективные области применения баз данных изображений следующие [1]:

- криминалистика (хранение архивов визуальных доказательств);
- медицинская диагностика (определение наиболее точного диагноза путем сравнения с уже установленными диагнозами);
- журналистика и реклама (иллюстрация статей и рекламных проектов);
- интеллектуальная собственность (примером может служить процесс регистрации новых торговых марок предприятий).

В связи с многообразием отраслей использования коллекций оцифрованных изображений очевидна необходимость разработки эффективного механизма поиска информации в таких базах данных. На сегодняшний день наиболее распространенным является поиск изображений по текстовым описаниям [1]. Существенным недостатком этого метода является неоднозначность соответствия между визуальным содержанием и текстовым описанием изображения, которое является субъективным. Поэтому возникает проблема организации средств поиска изображений по визуальному содержимому.

Поиск изображений по визуальному содержимому [2] – набор технологий для извлечения из базы данных изображений, наиболее подобных заданному изображению-образцу по некоторому набору числовых значений характеристик сравниваемых изображений. Одной из самых простых и понятных характеристик является гистограмма изображения. Она характеризует, сколько пикселей каждого цвета встретилось в изображении (цветность изображения). По данной характеристике можно судить на сколько близки друг к другу изображения. Для расчета расстояния между объектами часто берут Евклидово расстояние: $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$,
где n – количество пикселей, x и y – изображения между которыми находят расстояние.

По близости между изображениями можно разбить их на группы (кластеризовать изображения). Одним из методов кластеризации является метод «k-средних».

Метод кластеризации «k-средних»

Метод «k-средних» кластерного анализа ставит себе целью разделить все имеющиеся наблюдения на k кластеров, при этом каждое наблюдение относится к тому кластеру, к центру которого оно ближе всего.

В качестве меры близости используется Евклидово расстояние.

Итак, рассмотрим ряд наблюдений
(x $(x^{(1)}, x^{(2)}, ..., x^{(m)}), x^{(j)} \in R^m$.

Метод k-средних разделяет m наблюдений на k групп (или кластеров) (k ≤ m)

$S = \{S_1, S_2, ..., S_k\}$, чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центроидов этих кластеров:

$$\min\left[\sum_{i=1}^{k}\sum_{x}(j) \in S_i \parallel x^{(j)} - \mu_i \parallel^2\right],$$

где

$x^{(j)} \in R^n, \mu_i \in R^n$

μ_i – центроид для кластера S_i.

Алгоритм метода сводится к следующему: если мера близости до центроида определена, то разбиение объектов на кластеры сводится к определению центроидов этих кластеров. Число кластеров *k* задается исследователем заранее.

Рассмотрим первоначальный набор k средних (центроидов) $\mu_1, ..., \mu_k$ в кластерах $S_1, S_2, ..., S_k$. На первом этапе центроиды кластеров выбираются случайно или по определенному правилу (например, выбрать центроиды, максимизирующие начальные расстояния между кластерами).

Относим наблюдения к тем кластерам, чье среднее (центроид) к ним ближе всего. Каждое наблюдение принадлежит только к одному кластеру, даже если его можно отнести к двум и более кластерам.

Затем центроид каждого *i*-го кластера перевычисляется по следующему правилу:

$$\mu_i = \frac{1}{S_i}\sum_{x^{(j)} \in S_i} x^{(j)}.$$

Таким образом, алгоритм k-средних заключается в перевычислении на каждом шаге центроида для каждого кластера, полученного на предыдущем шаге.

Алгоритм останавливается, когда значения μ_i не меняются: $\mu_i^{шаг(T)} = \mu_i^{шаг(T+1)}$.

Важно отметить что, неправильный выбор первоначального числа кластеров k может привести к некорректным результатам. Именно поэтому при использовании метода «k-средних» важно сначала провести проверку подходящего числа кластеров для данного набора данных.

В ходе исследования разработана программа. Входными данными такой программы являются изображения, которые нужно разбить на кластеры и количество кластеров. Так же для корректной работы программы нужно указать количество загружаемых изображений.

Алгоритм полученной программы включает в себя следующие этапы.

Первым этапом в текстовые поля заносятся данные о количестве изображений и кластеров. Далее для каждого изображения строится гистограмма, на основе которых находится Евклидово расстояние между изображением и предполагаемым центром кластера. Далее выбирается, к какому кластеру ближе изображение, и пересчитывается центр кластера путем усреднения всех точек входящих в кластер. Если центры кластера изменились, то снова рассчитываем расстояние и формируем новые кластеры. Это продолжается до тех пор, пока центры кластеров не перестанут изменяться.

Результаты и обсуждения

Программа тестировалась на изображениях лиц людей. На рис. 1 представлены образцы этих изображений.

Рис. 1. Образцы изображений лиц.

Каждое лицо представлено двадцатью изображениями, полученными под разными ракурсами, с различной освещенностью и с различной мимикой лица.

На вход программе поступали 120 изображений и количество нужных кластеров равное 6. В результате программа безошибочно распределила изображения на кластеры.

Кроме того, программа была протестирована на изображениях предметов. На рис. 2 представлены образцы этих изображений.

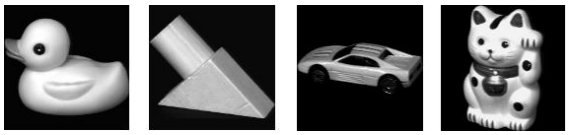

Рис. 2. Образы изображений предметов.

В данном случае только часть изображений была отнесена к правильным кластерам.

Заключение

В ходе проведения исследования были подтверждено:

- качество кластеризации методом «k-средних» зависит от первоначального разбиения;
- в качестве метрики можно использовать Евклидово расстояние;
- число кластеров заранее не известно и выбирается исследователем заранее.

Следует отметить, что часто для качественной кластеризации изображений недостаточно только цветности изображения, нужно вводить дополнительные параметры.

Список использованных источников

1. Eakins J. P., Graham M. E. «A report to the JISC Technology Applications Programme», Institute for Image Data Research, University of Nothumbria at Newcastle, Jan. 1999. 54 p.
2. Wang J. Z., Li J., Wiederhold G. «SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries», IEEE Transactions on Pattern Analysis and Machine Intelligence. – Sept. 2001. – V. 23, № 9. – P. 947-963.
3. Башков Е.А., Вовк О.Л. Кластеризация изображении методом дейтограмм [Электронный ресурс] // Электронный архив Донецкого национального технического университета г. Донецк. 2003. С. 1-10. URL:http://ea.donntu.org:8080/bitstream/123456789/5571/1/13.pdf.