МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ В ПРИРОДНИЧИХ НАУКАХ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

А.А. ШУМЕЙКО, д.т.н., профессор
М.Ю. БРАТУТА, магистр
Днепровский государственный технический университет, г. Каменское

Использование метода главных компонент для кластеризации изображений лиц

В работе предложена кластеризация изображений лиц методом *k*-средних, использующим в качестве центра кластера главную компоненту элементов, входящих в кластер, а в качестве критерия близости, значения соответствующих коэффициентов. Рассмотрено приложение полученной кластеризации к решению задачи поиска лиц, наиболее похожих на тестовое изображение.

In the paper, clustering faces images by the *k*-means method is suggested. The center of the cluster using the main component of the Principal Component Analysis clusters elements. As a criterion of proximity, the values of the corresponding coefficients are used. The application of the resulting clusterization to the solution of the problem of finding the persons most similar to the test image is considered.

Введение

Круг задач, рассматриваемых в обработке изображений (Image Processing), достаточно большой, и среди них важную роль играют задачи, связанные с особенностями восприятия человеческих лиц [1—2]. Это нахождение на изображении человеческого лица, идентификация этого человека или анализ полученного изображения для оценки агрессивности данной личности, пола, возраста и других характеристик. Одной из задач является нахождение в имеющейся выборке человека похожего да данного. Что есть «похожесть» человеческих лиц. Это вопрос не только и не столько математический, сколько психологический, пример тому — тест Леопольда Сонди. Но, тем не менее, задача «похожести» всегда будет интересна человеку, вечная проблема — на кого похож ребенок. С другой стороны, известны многолетние наблюдения, которые говорят о том, что достаточно часто в счастливых семьях, супруги похожи друг на друга. Что первично — счастливы, потому, что похожи или похожи, потому что счастливы? Да и не важно. Важно то, что сама по себе задача похожести лиц имеет смысл и интересна.

Постановка задачи

Пусть дано множество изображений человеческих лиц $\Im = \{I_i\}_{i=1}^n$ и изображение I^0, для которого из $\Im$ нужно выбрать m наиболее похожих и, упорядочить их по рейтингу (степени похожести). Прежде всего, нужно заметить, что так как изображения представляют собой проекции трехмерного тела на плоскость в условиях разной освещенности, то первой задачей является выравнивание и масштабирование изображения (рис. 1), то есть использование аффинного преобразования, позволяющего получить изображения анфас в одном и том же масштабе. Сама по себе эта задача нетривиальная и не является предметом нашего исследования. Будем считать, что это уже проведено.

Для того, чтобы нивелировать искажения, внесенные освещением, будем использовать не полноцветную палитру, а люминисцентную составляющую [3].

$$Y = \frac{1}{256}\left(77\,\mathrm{Red} + 150\,\mathrm{Green} + 29\,\mathrm{Blue} + 128\right).$$

Вычисление степени похожести лиц

Самым простым подходом к определению рейтинга похожести лиц является использование метода наименьших квадратов

$$\left(I^0 - \sum_{i=1}^{n} \alpha_i I_i\right)^2 \xrightarrow[\alpha_i]{} \min.$$

Решение этой задачи сводится к нахождению корней системы линейных уравнений

$$\sum_{i=1}^{n} \alpha_i \left\langle I_i, I_j \right\rangle = \left\langle I^0, I_j \right\rangle (j = 1,2,...,n).$$

Большее значение α_i будет соответствовать степени «большей похожести», то есть более высокому рейтингу данного изображения.

В случае малого размера выборки этот подход дает приемлемые результаты, но увеличение выборки достаточно быстро приводит как к увеличению времени работы алгоритма, так и к неустойчивости решения.

Рис. 1. Выравненная и масштабированная база лиц

Кластеризация изображений

Одним из подходов решения этой проблемы является кластеризация исходного множества данных, то есть получение подвыборок, в которых собраны изображения людей, которые в данном кластере более похожи между собой, чем с любым другим из другого кластера [4].

Пусть $\Im = \{x_i\}_{i=1}^{n}$ множество объектов, представленных набором атрибутов $x_i = \{t_1^i, t_2^i, ..., t_m^i,\}$, где t_ν^i принимает значения из заданного множества T_ν^i. Задача кластеризации состоит в построении множества $C = \{c_\nu\}_{\nu=1}^{k}$ и отображения $F : \Im \to C$ заданного множества объектов на множество кластеров. Кластер содержит объекты из $\Im$ похожие (по заданному критерию) друг на друга

$$x_i \in c_\nu, x_j \in c_\nu \Rightarrow d(x_i, x_j) < \varepsilon,$$

где $d(\bullet,\circ)$ — мера близости между объектами (расстояние), а ε — максимальное значение порога, формирующего один кластер.

Задача кластеризация данных является достаточно популярной, существует множество разных методов и подходов, среди которых наиболее популярным является метод *k*-средних. В методах кластеризации краеугольной характеристикой является мера близости $d(\bullet,\circ)$, так для *k*-средних это евклидово расстояние. Приведем этот метод.

Пусть $C = \{c_i\}_{i=1}^{k}$ множество кластеров с центрами

$$\mu_i = \frac{\sum\{x_j \mid x_j \in c_i\}}{\sum\{1 \mid x_j \in c_i\}} = \frac{\sum\limits_{j=1}^{n} u_j^i x_j}{\sum\limits_{j=1}^{n} u_j^i},$$

где u_j^i индикаторная функция, то есть

$$u_j^i = \begin{cases} 1, & \text{если } x_j \in c_i, \\ 0, & \text{в противном случае.} \end{cases}$$

Целевая функция

$$S(C, \Im) = \sum_{i=1}^{k}\sum_{j=1}^{n} u_j^i d(x_j, \mu_i),$$

и условия

$$\sum_{i=1}^{n} u_j^i = 1, 0 < \sum_{j=1}^{k} u_j^i \le n,$$

то есть, каждый элемент может быть только в одном кластере, и, кластер не может быть пустым или содержать элементов больше, чем их исходное количество.

Условие остановки выполнения алгоритма после ν-го шага будет иметь вид

$$\left| S^\nu(C, \Im) - S^{\nu-1}(C, \Im) \right| < \varepsilon,$$

где ε выбранный порог.

Непосредственное использование метода *k*-средних для кластеризации изображений лиц весьма проблематично, во-первых, что есть центром тяжести данного кластера, а во-вторых, что есть расстояние между двумя изображениями. Решению этой задачи посвящена данная работа.

Для решения этой задачи будем использовать метод главных компонент, то есть нахождение минимума $\varepsilon\left(e_1,...,e_k,\alpha_{1,1},...,\alpha_{n,k}\right)$

$$\underbrace{\varepsilon\left(e_1,...,e_k,\alpha_{1,1},...,\alpha_{n,k}\right)}_{unknowns}=\sum_{j=1}^{n}\varepsilon_j^2=\sum_{j=1}^{n}\left\|x_j-\sum_{i=1}^{k}\alpha_{j,i}e_i\right\|_2^2. \quad (1)$$

по всем $e_1,...,e_k$ и $\alpha_{1,1},...,\alpha_{n,k}$.

Точное решение этой задачи сводится к нахождению собственных чисел и векторов ковариационной матрицы, размер которой совпадает с размером выборки. Ясно, что решение этой задачи неустойчиво и для больших выборок весьма проблематично. Поэтому будем использовать итерационный подход, позволяющий, хотя и не точно, но зато гарантировано получить решение этой задачи [4].

Для случая i =1 задача (1) сводится к определению одной компоненты e_1, которая наилучшим образом восстанавливает все исходные данные $\{x_1,...,x_n\}$

$$\varepsilon\left(e_1,\alpha_{1,1},...,\alpha_{n,1}\right)=\sum_{j=1}^{n}\left\|x_j-\alpha_{j,1}e_1\right\|_2^2\to\min \quad (2)$$

по всем e_1 и $\{\alpha_{i,1}\}_{i=1}^{n}$ при условии $\sum_{i=1}^{n}\alpha_{i,1}^2=1.$

Если $\{\widetilde{\alpha}_{i,1}\}_{i=1}^{n}$ и $\widetilde{e}_1$ есть решение этой задачи и $\Delta x_j=x_j-\widetilde{\alpha}_{j,1}\widetilde{e}_1$ — ошибка восстановления данных одной первой главной компонентой, то решая задачу

$$\sum_{j=1}^{n}\left\|\Delta x_j-\alpha_{j,2}e_2\right\|_2^2\to\min$$

по всем e_2 и $\{\alpha_{i,2}\}_{i=1}^{n}$ при условии $\sum_{i=1}^{n}\alpha_{i,1}^2=1$, получаем вторую главную компоненту $\widetilde{e}_2$ и соответствующий вектор $\{\widetilde{\alpha}_{i,2}\}_{i=1}^{n}$ и т.д.

При фиксированных $\{\alpha_{i,1}\}_{i=1}^{n}$ задача (2) решается методом наименьших квадратов. В силу того, что функция цели представляет собой квадратичный функционал, необходимое и достаточное условия экстремума совпадают. Таким образом, решение задачи сводится к поиску решения уравнения

$$\frac{\partial}{\partial e_1}\varepsilon\left(e_1,\alpha_{1,1},...,\alpha_{n,1}\right)=-2\sum_{j=1}^{n}\left(x_j-\alpha_{j,1}e_1\right)\alpha_{j,1}=$$

$$=-2\left(\sum_{j=1}^{n}x_j\alpha_{j,1}-\sum_{j=1}^{n}\alpha_{j,1}^2e_1\right).$$

Отсюда получаем

$$e_1=\frac{\sum_{j=1}^{n}x_j\alpha_{j,1}}{\sum_{j=1}^{n}\alpha_{j,1}^2},$$

учитывая условие нормирования единицей, то есть $\sum_{i=1}^{n}\alpha_{i,1}^2=1$, имеем

$$e_1=\sum_{j=1}^{n}x_j\alpha_{j,1}.$$

Следующий шаг будем делать исходя из предположения, что в задаче (2) нам известна компонента e_1 и требуется найти экстремум по $\{\alpha_{i,1}\}_{i=1}^{n}$

$$\frac{\partial}{\partial\alpha_{\nu,1}}\varepsilon\left(e_1,\alpha_{1,1},...,\alpha_{n,1}\right)=-2\left(x_\nu-\alpha_{\nu,1}e_1\right)e_1=$$

$$=-2\left(\left\langle x_\nu,e_1\right\rangle-\alpha_{\nu,1}\left\langle e_1,e_1\right\rangle\right)=0,$$

то есть

$$\alpha_{\nu,1}=\frac{\left\langle x_\nu,e_1\right\rangle}{\left\langle e_1,e_1\right\rangle},$$

где, как обычно, $\langle x,y\rangle$ — скалярное произведение векторов x и y.

Далее, считая найденные $\{\alpha_{i,1}\}_{i=1}^{n}$ известными, повторяем весь процесс, пока не произойдет стабилизация ошибки. Полученные e_1 будем считать первой главной компонентой $\widetilde{e}_1$. Тогда $\Delta x_j=x_j-\widetilde{\alpha}_{j,1}\widetilde{e}_1$ — ошибка восстановления данных одной первой главной компонентой.

Применяя этот алгоритм к ошибке восстановления Δx_j, находим вторую главную компоненту e_2 вместе с коэффициентами $\alpha_{j,2}$, и т.д.

Приведем алгоритмизацию этого алгоритма.

Вначале центрируем данные, вычитая из исходных данных среднее значение и в дальнейшем считаем, что данные в среднем равны нулю.

1.Положим номер итерации $\nu=1$.

2.Выбираем стартовые значения $\{\alpha_{i,1}^{\nu}\}_{i=1}^{n}$, например, пусть все они между собой равны, то есть $\alpha_{i,1}^{\nu}=\frac{1}{\sqrt{n}},i=1,2,...,n.$

3.Вычисляем $e_1^{\nu}=\sum_{j=1}^{n}x_j\alpha_{j,1}^{\nu}$.

4.Далее находим $\beta_i=\frac{\left\langle x_i,e_1^{\nu}\right\rangle}{\left\langle e_1^{\nu},e_1^{\nu}\right\rangle}$, и, нормируя единицей, получаем

$$\alpha_{i,1}^{\nu+1}=\frac{\beta_i}{\sqrt{\sum_{j=1}^{n}\beta_j^2}}.$$

5.Полагаем $\nu=\nu+1$.

6.Проводим проверку критерия остановки, в качестве этого может быть либо стабилизация коэффициентов $\{\alpha_{i,1}^{\nu}\}_{i=1}^{n}$, либо стабилизация главной компоненты e_1^{ν}, либо проверка на заранее заданное фиксированное число итераций. Если условие окончания итерационного процесса не выполнено, то переходим к пункту 3.

Применяя итерационный метод нахождения главных компонент в модифицированном методе k-средних, и используя в качестве центра кластера главную компоненту, получаем (рис. 2)

Рис. 2. Центры полученных кластеров

Заметим, что в качестве критерия близости, используется значение соответствующего масштабирующего коэффициента.

Выводы

Предложенный метод кластеризации изображений лиц людей показал достаточно высокую эффективность для решения задачи похожести лиц.

ЛИТЕРАТУРА

1. Щеголева Н.Л. Применение алгоритмов двумерного анализа главных компонент для задач распознавания изображений лиц / Н.Л.Щеголева, Г.А.Кухарев // Бизнес-информатика .– 2011 .– №№4(18) .– С.31–38 .– Режим доступу: https://cyberleninka.ru/article/n/primenenie-algoritmov-dvumernogo-analiza-glavnyh-komponent-dlya-zadach-raspoznavaniya-izobrazheniy-lits
2. Прокошев В.Г. Проблема автоматического распознавания лиц с одним эталонным изображением / В.Г.Прокошев, М.М.Рожков, П.Шамин // Научно-технические ведомости СПбГПУ .– 2010 .– №5 .– С. 13–18 .– Режим доступу: https://cyberleninka.ru/article/n/problema-avtomaticheskogo-raspoznavaniya-lits-s-odnim-etalonnym-izobrazheniem
3. Лигун А.О. Комп’ютерна графіка (Обробка та стиск зображень): навч.посіб./ А.О.Лигун, О.О.Шумейко.– Біла К.О., 2010.– 114 с .– Режим доступу: http://pzs.dstu.dp.ua/Data/U_CompGraph.pdf
4. Шумейко А.А. Интеллектуальный анализ данных (Введение в Data Mining) / А.А.Шумейко, С.Л.Сотник .– Днепропетровск: Белая Е.А., 2012 .– 212 с .– Режим доступу: http://pzs.dstu.dp.ua/DataMining/bibl/DataMining.pdf

пост. 15.11.2017