Математическая модель статистического иерархического агломеративного метода кластеризации изображений

Башков Е.А., Вовк О.Л.

Кафедра ПМИ, ДонНТУ
vovk.olga@gmail.com

Abstract

Vovk O.L. The Mathematical Model of the Statistical Hierarchical Agglomerative Method of Images Clusterization. *This work is devoted to development of mathematical model of decision of images regions selection task by the statistical hierarchical agglomerative method of clusterization. The set theory is used for description of the offered mathematical model. Experimental estimation of clusterization quality of images by the offered method is given.*

Введение

Проблема выделения регионов изображений решается специалистами различных сфер: в медицинской диагностике, при анализе механических повреждений металлических поверхностей, при загрузке шихты при плавке, при контекстном поиске изображений. Постоянное увеличение количества областей, использующих архивы визуальной информации, ставит перед исследователями задачу разработки быстродействующих методов выделения регионов изображений. Одним из таких методов можно считать статистический иерархический агломеративный метод кластеризации изображений [1].

При кластеризации изображений исходными объектами являются пиксели, каждый из которых задан вектором цветовых составляющих. В ходе процедуры кластеризации происходит объединение пикселей в отдельные группы – кластеры (регионы), исходя из числовых значений цветовых характеристик.

Цель данной работы – построение математической модели статистического иерархического агломеративного метода выделения регионов изображений. В соответствии с поставленной целью можно выделить следующие основные задачи, решаемые в данной работе: обзор существующих методов выделения регионов изображений, разработка математической модели статистического иерархического агломеративного метода кластеризации изображений, адаптация рассматриваемого метода для построения системы контекстного поиска изображений с высокой энтропией.

1 Обзор основных статистических методов кластеризации

Согласно [2, 3], все статистические методы кластеризации принадлежат к одному из двух базовых видов: иерархическому или разделения (разбиения).

Иерархические методы представляют собой процедуры создания последовательности вложенных разбиений, исходя из данных матрицы близости [2]. Формально любой иерархический метод кластеризации состоит из следующих шагов [3]:

1. Расчет матрицы близости между всеми парами шаблонов. Изначально каждый объект – отдельный кластер.
2. Нахождение минимума в матрице близости и объединение кластеров с минимальным расстоянием. Обновление строк и столбцов матрицы, соответствующих объединенным кластерам.
3. Если все объекты принадлежат одному кластеру, то конец работы метода, иначе на шаг 2.

По способу формирования кластеров иерархические методы подразделяются на методы одиночной и полной связи [3]. При одиночной связи – в один кластер объединяются объекты с минимальным расстоянием, а при полной связи – в разные кластеры разносятся объекты с максимальным расстоянием.

Основная идея методов разбиения [3] – нахождение единственного разделения шаблонов по кластерам, вместо дендрограммы, полученной согласно иерархическим технологиям. Реализация методов разбиения предполагает выполнение следующих шагов [2, 3]:

1. Выбор начального распределения объектов по кластерам. Расчет “центров тяжести” полученных кластеров.
2. Перегруппировка объектов кластеров: отнесение каждого объекта к кластеру с минимальным расстоянием до центра.

Однако такая технология имеет следующие существенные ограничения [4]: при разных стартовых условиях такие методы выдают различные результаты, нет методики выбора количества кластеров. Предлагаемый в работе метод относится к иерархическому виду.

2 Математическая модель статистического иерархического агломеративного метода кластеризации для выделения регионов изображений

Рассматриваемый метод выделения регионов изображений использует битовую маску взаимосвязей и рангов цветовых составляющих центров кластеров [5].

Определим рассматриваемую маску в виде многомерного вектора для пространства цветов RGB, включающего в себя вектора рангов $\overline{S}_R, \overline{S}_G, \overline{S}_B$ и вектора взаимосвязей $\overline{S}_{GB}, \overline{S}_{RB}, \overline{S}_{RG}$:

$$\bar{S} = \{\bar{s}_R, \bar{s}_G, \bar{s}_B, \bar{s}_{GB}, \bar{s}_{RB}, \bar{s}_{RG}\}, \quad (1)$$

причем компоненты векторов могут принимать только два значения: 0 или 1.

Условно обозначим вектора рангов $\bar{s}_R, \bar{s}_G, \bar{s}_B$ вектором $\bar{s}_\alpha$ ($\alpha = R,G,B$), который можно представить как:

$$\bar{s}_\alpha = (s_{\alpha 1}, s_{\alpha 2}, s_{\alpha 3}). \quad (2)$$

Компоненты вектора $\bar{s}_\alpha$ определяются по формулам:

$$\begin{aligned}
s_{\alpha 1} &= \begin{cases} 0, \alpha \in [x_l, GH - eps); \\ 1, \alpha \in [GH - eps, x_h]; \end{cases} \\
s_{a2} &= \begin{cases} 0, \alpha \in [x_l, GL - eps) \cup (GH + eps, x_h]; \\ 1, \alpha \in [GL - eps, GH + eps]; \end{cases} \\
s_{\alpha 3} &= \begin{cases} 0, \alpha \in (GL + eps, x_h]; \\ 1, \alpha \in [x_l, GL + eps]. \end{cases}
\end{aligned} \quad (3)$$

В формулах (3): [x_l, x_h] – пределы изменения числовых значений цветовых характеристик (для пространства цветов RGB: x_l=0, x_h=255), *GL*, *GH* – границы числовых значений цветовых характеристик для рассматриваемых рангов (автором работы предлагается выделять три ранга: низкий, средний и высокий и три, соответствующих перечисленным рангам интервалов: [x_l, *GL*], (*GL*, *GH*], (*GH*, x_h]), *eps* – параметр метода, введенный для возможности нескольких уровней для одного цветового компонента и нескольких взаимосвязей для одной пары компонентов.

Условно обозначим вектора взаимосвязей $\bar{s}_{GB}, \bar{s}_{RB}, \bar{s}_{RG}$ вектором $\bar{s}_{\alpha\beta}$ ($\alpha = R,G$ и $\beta = B,G$), который можно представить в виде:

$$\bar{s}_{\alpha\beta} = (s_{\alpha\beta 1}, s_{\alpha\beta 2}, s_{\alpha\beta 3}). \quad (4)$$

Компоненты вектора $\bar{s}_{\alpha\beta}$ определяются по формулам:

$$\begin{aligned}
s_{\alpha\beta 1} &= \begin{cases} 0, \alpha > \beta; \\ 1, \alpha \le \beta; \end{cases} \\
s_{a\beta 2} &= \begin{cases} 0, |\alpha - \beta| > eps; \\ 1, |\alpha - \beta| \le eps; \end{cases} \\
s_{\alpha\beta 3} &= \begin{cases} 0, \alpha < \beta; \\ 1, \alpha \ge \beta. \end{cases}
\end{aligned} \quad (5)$$

Описанная маска взаимосвязей и рангов цветовых компонентов центров кластеров предназначена для учета специфики характеристик

кластеризируемых объектов (особенностей анализируемого цветового пространства).

Перейдем к описанию статистического иерархического агломеративного метода кластеризации изображений.

На первом этапе метода (этапе полной связи) происходит уменьшение обрабатываемого числа кластеров путем распределения в отдельные кластеры пикселей с одинаковыми битовыми масками.

Предполагается, что изначально каждый пиксель изображения *t* представляет собой отдельный кластер (свойство агломеративности). Первоначально производится распределение точек изображения с одинаковыми битовыми масками в отдельные кластеры. Для этого каждому пикселю изображения ставится в соответствие битовая маска взаимосвязей и рангов, затем пиксели с различными масками разносятся в разные кластеры (соответственно, пиксели с одинаковыми масками объединяются в один кластер).

Изображение *t* размером [*w*x*h*] пикселов определяем в виде, аналогичном виду (6):

$$t = \{\ p_{jk}=\{r_{jk},\ g_{jk},\ b_{jk}\} \mid j \in [1,w],\ k \in [1,h]\ \ \},\ \ j \in N, k \in N. \qquad (6)$$

Тогда формально объединение пикселов с одинаковыми битовыми масками в кластеры можно записать в следующем виде:

$$\exists m_1, m_2, ..., m_q :\ \overline{S}_{j_l k_l} = \overline{S}_{j_e k_e}$$
$$\forall l, e \le m_v, l \ne e, v \in [1,q], m_v \in [1, w \cdot h], m_v \in N, v \in N. \qquad (7)$$

В формуле (7): m_v – количество элементов группы (кластера) с индексом *v*; q – количество кластеров; $\overline{S}_{jk}$ – битовая маска пикселя p_{jk} с координатами (*j, k)* изображения *t*; *l, e* – индексы пикселов внутри кластера.

Результатом объединения пикселов с одинаковыми битовыми масками будет набор кластеров:

$$A = \{a_1, a_2, ..., a_q\}:\ a_v = \cup p_{jk} :\ \overline{S}_{j_l k_l} = \overline{S}_{j_e k_e}$$
$$\forall l, e \le m_v, l \ne e, v \in [1,q], m_v \in [1, w \cdot h], m_v \in N, v \in N. \qquad (8)$$

На втором этапе метода (этапе одиночной связи) создаем новые кластеры путем объединения кластеров с минимальным расстоянием. Этап повторяется до тех пор, пока выполняется условие “сравнимости” кластеров, основанное на битовой маске взаимосвязей и рангов и описанное ниже.

Рассмотрим более подробно этап одиночной связи предлагаемого метода.

Для полученных кластеров на первом этапе кластеров строится матрица расстояний (близости) между центрами кластеров. В качестве расстояния между кластерами используется среднее Евклидово расстояние

между всеми парами точек, входящих в кластеры. В построенной матрице происходит поиск наиболее “близких” кластеров (т.е. минимумов матрицы близости). Если расстояния между несколькими парами кластеров являются одинаковыми и минимальными, то, в первую очередь, в качестве “близких” кластеров, выбираются кластеры с минимальной площадью. Найденные “близкие” кластеры объединяются, образуя новые кластеры, для которых производится перерасчет центров. Из матрицы расстояний удаляются строки и столбцы, соответствующие объединенным кластерам, и добавляется строка и столбец, соответствующие полученному кластеру. Объединение кластеров с минимальным расстоянием производится до тех пор, пока удовлетворяется условие “сравнимости” претендующих на объединение кластеров.

В основе условия “сравнимости” кластеров – подсчет числа эквивалентных ненулевых бит масок кластеров, претендующих на объединение.

Для формального описания метода каждому элементу a_v множества A ставится в соответствие многомерный вектор вида:

$$\overline{a}_v = \{\{p_{vjk}\}, \overline{S}_v\}, \quad v \in [1, q], v \in N. \tag{9}$$

где $\{p_{vjk}\}$ – набор пикселов v-го кластера, $\overline{S}_v$ – маска v-го кластера.

Для последующего иерархического объединения кластеров строится симметричная относительно главной диагонали матрица расстояний Ψ размером [qxq]:

$$\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} & \dots & \psi_{1q} \\ \psi_{21} & \dots & & \\ \dots & & & \\ \psi_{q1} & \psi_{q2} & \dots & \psi_{qq} \end{pmatrix}, \tag{10}$$

элементы которой рассчитываются как:

$$\psi_{\gamma\lambda} = \frac{\sum_{y}^{K_\gamma}\sum_{z}^{K_\lambda}\sigma_{yz}}{K_\gamma \cdot K_\lambda}, \gamma \in [1, q], \lambda \in [1, q], \gamma \in N, \lambda \in N, \tag{11}$$

$$\sigma_{yz} = \sqrt{(r_y - r_z)^2 + (g_y - g_z)^2 + (b_y - b_z)^2}.$$

В формулах (11): K_γ – количество пикселов кластера с индексом γ, K_λ – количество пикселов кластера с индексом λ, σ_{yz} – Евклидово расстояние между цветовыми компонентами пикселя с индексом y кластера с индексом γ и пикселя с индексом z кластера с индексом λ.

Затем в построенной матрице Ψ производится поиск минимального элемента:

$$\psi_{\min} = \psi_{v1v2} : \psi_{v1v2} \le \psi_{\gamma\lambda} \quad \forall \quad \gamma, \lambda \ne v1, v2. \tag{12}$$

Далее для кластеров с минимальным расстоянием ($\overline{a}_{v1}$, $\overline{a}_{v2}$) проверяется удовлетворение условия “сравнимости”, в основе которого подсчет числа эквивалентных ненулевых бит масок *Kb*(*v1*,*v2*):

$$Kb(v1, v2) = \overline{S}_{v1} \cdot \overline{S}_{v2} - \hat{\overline{S}}_{v1} \cdot \tilde{\overline{S}}_{v2}, \tag{13}$$

где “·” – знак скалярного произведения векторов, вектора $\hat{\overline{S}}_{v1}$, $\tilde{\overline{S}}_{v2}$ вычисляются по формулам:

$$\hat{\overline{S}}_{v1} = \sum_{\chi=1}^{17} \{S_{v1\chi} \cdot S_{v1\chi+1}\}, \quad \tilde{\overline{S}}_{v2} = \sum_{\chi=2}^{18} \{S_{v2\chi-1} \cdot S_{v2\chi}\}. \tag{14}$$

Для удовлетворения условия “сравнимости” необходимо выполнение условия:

$$Kb(v1,v2) \ge K_{opt}, \tag{15}$$

где K_{opt} является вторым параметром предлагаемого метода (первым обозначен параметр *eps*) и подбирается экспериментально в зависимости от анализируемого набора изображений. Методика подбора K_{opt} и *eps* основана на введенном в [6] условии, что энтропия по площади выделенных кластеров возрастает с увеличением числа кластеров.

Теперь переходим к анализу полученного для рассматриваемой пары кластеров числового значения параметра $K_{opt.}$ Если для кластеров с минимальным расстоянием ($\overline{a}_{v1}$, $\overline{a}_{v2}$) удовлетворяется условие (15), то происходит объединение данных кластеров и пересчет маски для нового кластера $\overline{S}_{v1v2}$ согласно формул (1-5)

$$\overline{a}_{v1v2} = \{\{p_{v1jk} \cup p_{v2jk}\}, \overline{S}_{v1v2}\}, \tag{16}$$

и уменьшение общего числа кластеров:

$$q = q - 1. \tag{17}$$

Если для кластеров с минимальным расстоянием ($\overline{a}_{v1}$, $\overline{a}_{v2}$) не удовлетворяется условие (15), то производится поиск другого минимума:

$$\psi_{\min} = \psi_{v3v4} : \psi_{v1v2} \le \psi_{v3v4} \le \psi_{\gamma\lambda} \quad \forall \quad \gamma, \lambda \ne v1, v2, \; v3 \ne v1, v4 \ne v2. \tag{18}$$

Затем производится переназначение переменных:

$$v1=v3,\ v2=v4. \tag{19}$$

И повторяются действия, описанные формулами (13)-(17).

Вычисления по методу прекращаются, когда выполняется условие:

$$\exists q : \; Kb(v1, v2) < K_{opt} \quad \forall \quad v1, v2 \le q. \tag{20}$$

3 Оценка качества кластеризации изображений

Предлагается экспериментальная оценка качества кластеризации изображений по критерию оптимальности разбиения изображений по цветовому подобию пикселей внутри полученных кластеров [7], которая базируется на минимизации следующего функционала:

$$Z(I) = \sum_{i=1}^{R} \left(\alpha \frac{1}{A_i} + \beta D_i + \gamma \sum_{j=1, j \neq i}^{R} \frac{1}{D_{i-j}} \right), \qquad (21)$$

где I – анализируемое изображение, D_i – среднее расстояние между цветовыми характеристиками i-ого региона, $D_{i\text{-}j}$ – среднее расстояние между цветовыми компонентами i-ого и j-ого регионов, α, β, γ – контрольные параметры, соответствующие уровню приоритета того или иного компонента критерия (21).

В ходе экспериментов проводится сравнение разработанного статистического иерархического агломеративного метода с методом k-средних в интерпретации Ванга для выделения регионов изображений на основании цветовых характеристик [6]. Для метода k-средних в качестве стадии предварительной сегментации используется разбиение на блоки размером 4x4 пикселя [6]. Были выбраны следующие значения параметров рассматриваемого критерия $\alpha=\beta=\gamma=1$ (все компоненты критерия равнозначны).

На рис.1 приведены результаты экспериментального сравнения характеристик качества кластеризации изображений с помощью статистического иерархического агломеративного (СИА) метода и метода k-средних. По оси абсцисс – энтропия как мера цветовой сложности изображений [8].

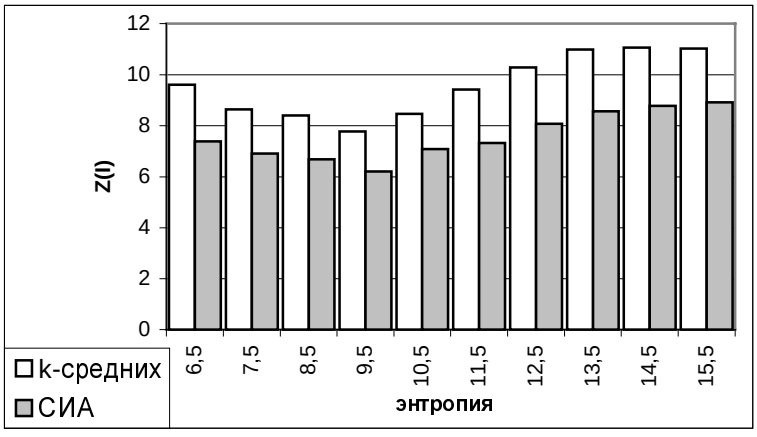

Рисунок 1 – Результаты оценки качества кластеризации согласно критерию оптимальности разбиения по цветовому подобию

Выводы

В работе решена актуальная прикладная задача – разработана математическая модель статистического иерархического агломеративного

метода кластеризации изображений. Предлагаемый метод позволяет выделять однородные по цветовому подобию области изображений и может применяться в системах удаленного наблюдения, при анализе механических повреждений металлических поверхностей, в медицинской диагностике при анализе снимков ультразвуковых исследований.

Проведенные эксперименты по оценке качества выделения кластеров изображений показали превосходство разработанного метода в сравнении с методом k-средних.

В перспективе планируется внедрение разработанного метода при разработке систем слежения.

Литература

1. Башков Е.А., Вовк О.Л. Оценка эффективности нового статистического иерархического агломеративного алгоритма кластеризации для распознавания регионов изображений // Системні дослідження та інформаційні технології. – Інститут прикладного системного аналізу НАН України, Київ. – 2005. – №2. – С. 117-130.

2. Chen C.H., Pau L.F., Wang P.S.P. The Handbook of Pattern Recognition and Computer Vision (2nd Edition). – World Scientific Publishing Co., 1998. – 1004 p.

3. Jain A.K., Murty M.N., Flynn P.J. Data Clustering: A Review // ACM Computing Surveys. – 1999. – vol. 31, №3. – P. 264-323.

4. Ким Д. О., Мьюллер Ч. У., Клекка У. Р. Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – 215 с.

5. Вовк О.Л. Новый подход к выделению визуально подобных цветов изображений // Проблемы управления и информатики. – 2006. – №6. – С. 100-105.

6. Wang J. Z., Li J. Wiederhold G. SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2001. – vol. 23, №9. – P. 947-963.

7. Del Bimbo A. Visual Information Retrieval. – San Francisco: Morgan Kaufmann Publishers, 1999. – 270 p.

8. Прэтт У. Цифровая обработка изображений. – М.: Мир, 1982. – Кн. 1 – 312 с.