

УДК 004.932.2:004.93'1



С.И. Богучарский¹, А.Г. Каграманян², С.В. Машталир³

^{1,3} ХНУРЭ, г. Харьков, Украина, mashtalir_s@kture.kharkov.ua

² ХНУ им. В.Н. Каразина, г. Харьков, Украина

МОДИФИКАЦИЯ МЕТОДА J-СРЕДНИХ В ЗАДАЧАХ ФРАГМЕНТНОЙ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ

Рассмотрены методы пофрагментной кластеризации изображений. Предложена матричная модификация метода J-средних, в которой вариации центров получаемых кластеров позволяют получать лучшие по сравнению с традиционными подходами локальные экстремумы в смысле минимума суммы квадратов внутрикластерных расстояний или максимума межкластерных.

КЛАСТЕРИЗАЦИЯ, ИЗОБРАЖЕНИЕ, ФРАГМЕНТ, ЦЕНТРОИД

Введение

Задача кластеризации массивов многомерных наблюдений является одной из основных в рамках интеллектуального анализа данных, а для ее решения на сегодня разработано множество методов [1, 2] от строго математически формализованных до основанных на использовании тех или иных эвристик, привносящих в «искусственный» интеллект элементы «естественного». В общей «неформальной» постановке задача кластеризации может быть сформулирована следующим образом: пусть имеется неклассифицированная выборка, состоящая из N объектов, описываемых r -мерными векторами признаков $x(k) \in \mathbb{R}^r$, $k = 1, 2, \dots, N$, которые необходимо разбить на $1 < p < N$ непересекающихся классов Cl_l , $l = 1, 2, \dots, p$.

Одним из наиболее популярных подходов к решению задачи кластеризации является подход, основанный на разбиении пространства признаков и сводящийся к нахождению p центроидов $C(l)$, вокруг которых группируются классифицируемые наблюдения. Здесь наибольшее распространение получили методы, связанные с минимизацией суммы квадратов внутрикластерных расстояний между наблюдениями и центроидами. Среди подобных методов наиболее известны методы H -средних (H -means), K -средних (K -means), J -средних (J -means) [2–6] и их различные модификации. С формальной точки зрения данный подход может быть описан следующим образом: обозначим множество всех возможных разбиений массива данных

$X = \{x(1), x(2), \dots, x(N)\}$, $x(k) = (x_1(k), x_2(k), \dots, x_r(k))^T$ как P_p . Среди этого множества необходимо найти наилучшее разбиение $P_p = \{Cl_1, Cl_2, \dots, Cl_p\}$ с p непересекающимися кластерами $Cl_l \cap Cl_q = \emptyset$; $l \neq q$; $l, q \in \{1, 2, \dots, p\}$ с центроидами $C = \{C(1), C(2), \dots, C(p)\}$, обеспечивающее минимум целевой функции

$$\min_{P_p \in P_p} E = \min_{P_p \in P_p} E(x(k), C(l)) = \min \sum_{k=1}^N \sum_{l=1}^p \mu(x(k), C(l)) \|x(k) - C(l)\|^2, \quad (1)$$

где $\mu = \begin{cases} 1, & \text{если } x(k) \in Cl_l, \\ 0, & \text{в противном случае.} \end{cases}$

Решение данной задачи имеет простой вид

$$C(l) = \frac{1}{N_l} \sum_{x(k) \in Cl_l} x(k) = \frac{\sum_{k=1}^N \mu(x(k), C(l)) x(k)}{\sum_{k=1}^N \mu(x(k), C(l))}. \quad (2)$$

(здесь N_l – число наблюдений, отнесенных к кластеру Cl_l) и фактически сводится к нахождению центров тяжести соответствующих кластеров.

В общем случае это NP -полная переборная задача и ее реализация существенно усложняется при больших N и p . Поэтому эффективность того или иного конкретного алгоритма определяется тем, насколько удачно реализован этот перебор.

Одной из типичных задач, решаемых с помощью рассматриваемых методов, является сегментация изображений. При этом предполагается, что обрабатываемое изображение задано в виде $(M \times N)$ – числовой матрицы, которая должна быть разбита в процессе анализа на однородные в некотором смысле классы (кластеры, сегменты). Достаточно часто используется фрагментная сегментация изображений, когда исходная матрица разбивается на блоки размерности $(m \times n)$, $m \leq M, n \leq N$, каждый из которых далее преобразуется в $(mn \times 1)$ -векторы, общее количество которых составляет $\tilde{N} = MN(mn)^{-1}$. Именно этот набор векторов подвергается процедуре кластеризации. Ясно, что чем выше размерность обрабатываемых векторов, тем более трудоемкой является решаемая задача.

Целью работы является построение метода фрагментной сегментации изображений на основе перехода алгоритма J -средних к матричной форме.

1. Матричные модификации методов K -средних и H -средних

С содержательной точки зрения более естественным представляется вместо $(mn \times 1)$ -векторов описывать исходную информацию, подлежащую кластеризации, в форме $(m \times n)$ -матриц «окон» $x(k) = \{x_{i_1 i_2}(k)\} \in \mathbb{R}^{(m \times n)}$; $i_1 = 1, 2, \dots, m$; $i_2 = 1, 2, \dots, n$;

$k = 1, 2, \dots, \tilde{N} = MN(mn)^{-1}$; при этом вместо целевой функции (1) естественно ввести ее матричную модификацию, основанную на сферической норме

$$E = E(x(k), C(l)) =$$

$$= \sum_{k=1}^{\tilde{N}} \sum_{l=1}^p \mu(x(k), C(l)) Sp(x(k) - C(l))(x(k) - C(l))^T, \quad (3)$$

при этом получаемое решение имеет форму (2) с той разницей, что центроид $C(l)$ описывается $(m \times n)$ -матрицей.

В этой ситуации метод K -средних может быть описан следующим образом. Пусть задано некоторое начальное в общем случае произвольное разбиение $P_p = \{Cl_1, Cl_1, \dots, Cl_p\}$. Пусть после расчета соответствующих центроидов $C = \{C(1), C(2), \dots, C(p)\}$ возникает ситуация, при которой некоторое наблюдение $x(k)$ $k = 1, 2, \dots, \tilde{N}$, которое попало в Cl_l , расположено ближе к $C(q)$ чем к $C(l)$, т.е.

$$Sp(x(k) - C(q))(x(k) - C(q))^T < \\ < Sp(x(k) - C(l))(x(k) - C(l))^T.$$

Это означает, что образ $x(k)$ должен быть перенесен из Cl_l в Cl_q . В этом случае центроиды кластеров должны быть откорректированы в соответствии с выражениями

$$C(l) = (N_l - 1)^{-1} (N_l C(l) - x(k)), \\ C(q) = (N_q + 1)^{-1} (N_q C(q) + x(k)).$$

Такая коррекция приводит к изменению целевой функции

$$\Delta E = \Delta E(x(k), C(l), C(q)) = \\ = \frac{N_q}{N_q + 1} Sp(AA^T) - \frac{N_l}{N_l - 1} Sp(BB^T),$$

где $Sp(\circ)$ обозначает след матрицы, $A = x(k) - C(q)$, $B = x(k) - C(l)$, $x(k) \in Cl_q$.

Такие приращения рассчитываются на каждой итерации алгоритма для всех имеющихся наблюдений всех пересчитываемых центроидов. Если на какой-то итерации выяснится, что все приращения неотрицательны, полагается, что достигнуто оптимальное решение и процесс останавливается. В противном случае процесс коррекции центроидов продолжается.

Весьма близок к данной процедуре алгоритм H -средних, работу которого удобно записать в виде последовательности шагов.

Шаг 1. Задание начального достаточно произвольного разбиения в форме набора центроидов $C(1), C(2), \dots, C(p)$.

Шаг 2. Отнесение каждого из имеющихся наблюдений $x(k)$ к кластеру Cl_q с «ближайшим» центроидом $C(q)$, т.е. $x(k) \in Cl_q$, если $\forall l = 1, 2, \dots, p$

$$Sp(x(k) - C(q))(x(k) - C(q))^T < Sp(x(k) - C(l))(x(k) - C(l))^T.$$

Шаг 3. Пересчет центроидов согласно выражению

$$C(q) = \frac{1}{N_q} \sum_{x(k) \in Cl_q} x(k).$$

Шаг 4. Возврат к шагу 2, если координаты центроидов изменяются. В противном случае — окончание алгоритма.

Несложно заметить, что на каждой итерации необходимо проводить $\tilde{N}p$ сравнений, что при больших объемах данных значительном числе кластеров требует существенных вычислительных затрат.

Кроме того, необходимо отметить и такой существенный недостаток обоих методов, как возникновение в процессе расчетов вырожденных решений, когда к некоторому из случайно выбранных центроидов не будет отнесено ни одного наблюдения. Для устранения этого недостатка был введен метод H -means+ [7], суть которого состоит в следующем. Пусть на некоторой итерации алгоритма сформировано $p - t$ «непустых» кластеров, где $t > 0$ — количество кластеров, не содержащих ни одного наблюдения $x(k)$. Далее находится t образцов из непустых кластеров, наиболее удаленных от своих центроидов, т.е. вносящих наибольший «вклад» в приращение целевой функции. Далее эти образцы-наблюдения формируют центроиды новых «непустых» кластеров, к которым приписываются наблюдения из соседних кластеров, близко расположенные к этим новым сформированным центроидам. Далее все расчеты производятся аналогично обычным H -средним.

Метод достаточно прост в численной реализации однако, подобно K -средним и обычным H -средним обладает существенным недостатком: возможностью остановки процесса поиска в локальных экстремумах. Конечно, можно было бы организовать реализацию поиска в случае его «застывания», однако это может существенно увеличить время обработки информации, особенно, если речь идет о больших базах данных. Выход из сложившейся ситуации может быть найден путем использования эвристик, используемых в задачах многоэкстремальной оптимизации.

2. Матричная модификация J -средних

Такой процедурой кластеризации, обладающей глобальными свойствами, является метод J -means введенный [4]. Суть метода состоит в том, что при «застывании» K - или H -means в локальном экстремуме, в его окрестности совершаются «скачки» выводящие процедуру из этой окрестности в области притяжения более «глубоких» экстремумов. Эти скачки по сути есть те случайные блуждания, которые производятся в алгоритмах многоэкстремального случайного поиска [8]. Для описания эвристики, связанной с J -means, вводится два основных понятия: «занятые точки» — наблюдения, совпадающие с центроидами или расположенные в их малых окрестностях, и «окрестность скачка» — область, в которой совершаются случайные движения, выводящие алгоритм кластеризации из локальных экстремумов.

Суть J -means состоит в том, что при нахождении некоторого центроида $C(l)$, в котором процедура кластеризации останавливается, организуется процесс перемещения этого центроида в незанятые точки, находящиеся в его окрестности. Такие скачки «центроид — незанятая точка» производятся до тех пор, пока не будет найден более глубокий экстремум, соответствующий меньшему значению целевой функции E .

С формальной точки зрения суть J -means состоит в следующем: пусть получено некоторое текущее разбиение $P_p = \{Cl_1, Cl_2, \dots, Cl_p\}$ с соответствующим набором центроидов $C = \{C(1), C(2), \dots, C(p)\}$. В рассмотрение вводится набор областей соседства $H_l(P_p)$, где $l \in \{1, 2, \dots, p\}$, а $H_l(P_p)$ определяет все возможные перемещения центроида $C(l)$ в точки $x(s)$, принадлежащие его окрестности, где $s \in \{1, 2, \dots, \tilde{N}\}$. Понятно, что при таких перемещениях постоянно корректируется и набор центроидов C , приобретая вид

$$C_s = \{C(1), \dots, C(l-1), x(s), C(l+1), \dots, C(p)\}.$$

Все соседства от $l=1$ до $l=p$ образуют множество соседств текущего разбиения P_p . При этом поиск не останавливается, пока не будет достигнут более глубокий экстремум в окрестности любого из текущих центроидов. При этом осуществляется минимизация не критерия (3), а целевой функции вида

$$E(x(k), C(q), x(s)) = \sum_{k=1}^{\tilde{N}} \left(\sum_{q=1}^p \mu(x(k), C(q)) S p(AA^T) - \mu(x(k), x(s)) S p(BB^T) \right), \quad (4)$$

где $A = x(k) - C(q)$, $B = x(k) - x(s)$,

$$\mu(x(k), C(q)) = \begin{cases} 1, & \text{если } x(k) \in Cl_q, \\ 0, & \text{в противном случае,} \end{cases}$$

$$\mu(x(k), x(s)) = \begin{cases} 1, & \text{если } x(k) \in H_l, \\ 0, & \text{в противном случае.} \end{cases}$$

Работу метода J -средних удобно записать в виде последовательности шагов.

Шаг 1. Задание начального, достаточно произвольного, разбиения $P_p = \{Cl_1, Cl_2, \dots, Cl_p\}$, определение его центров $C(1), C(2), \dots, C(p)$ и расчет целевой функции $E(x(k), C(l))$, соответствующей этому разбиению.

Шаг 2. Определение всех незанятых точек, т.е. наблюдений, не совпадающих с центроидами или не находящихся в их непосредственной близости.

Шаг 3. Скачки в окрестностях центроидов, т.е. их замена незанятыми точками $x(s)$ с расчетом целевых функций $E(x(k), C(q), x(s))$ вида (4).

Шаг 4. Принятие решений об остановке или дальнейшем поиске. Если

$$E(x(k), C(q), x(s)) > E(x(k), C(q)),$$

то полагается, что оптимальное решение было найдено на предыдущих итерациях и процесс кластеризации останавливается. В противном случае вы-

бирается решение, соответствующее наименьшему значению целевой функции $E(x(k), C(q), x(s))$ для всех реализованных $x(s)$. Полученное разбиение полагается начальным и производится возврат к шагу 2.

Главная особенность J -means, отличающая его от всех методов кластеризации с непользованием центроидов, — это наличие шага 3 (скачки в окрестностях), удачная реализация которого определяет эффективность полученного решения в целом. Качество кластеризации может быть улучшено, если шаг 3 реализовать в форме последовательности подшагов.

Шаг 3а. Формирование нового центроида $C(p+1)$ в незанятой точке $x(s)$ и нахождению «наилучшего» центроида $C(l)$, дающего наибольшее значение целевой функции $E(x(k), C(q), x(s))$.

Шаг 3б. Удаление центроида $C(l)$ и замена его на $x(s)$.

Шаг 3в. Формирование нового разбиения P_p^{new} , в котором $C^{new}(l) \equiv x(s)$.

Можно отметить [4], то в ряде случаев качество получаемого решения может быть улучшено, если после шага 3 включить стандартную процедуру K или H -means. Авторы статьи [4] назвали метод J -means+.

Наибольшие проблемы, связанные с практическим использованием метода J -means, возникают при задании размеров (диаметра) окрестности скачка вокруг каждого центроида. Слишком малая окрестность может не обнаружить незанятые точки с меньшими значениями целевой функции, а слишком большие окрестности резко увеличивают вычислительную сложность алгоритма, особенно в случаях когда эти окрестности перекрываются.

В связи с этим целесообразно ввести в рассмотрение окрестности переменных размеров, при этом процесс кластеризации начинается в малых окрестностях и если в них улучшение не происходит, их размер начинает увеличиваться, захватывая новые незанятые точки. В этом случае свободным параметром алгоритма является S_{max} — число, которое определяет сколько раз окрестность может увеличиваться, «захватывая» больше незанятых точек. Работу метода J -средних с расширяющимися окрестностями можно записать в виде последовательности шагов.

Шаг 1. Задание начального разбиения $P_p = \{Cl_1, Cl_2, \dots, Cl_p\}$, определение его центроидов $C = \{C(1), C(2), \dots, C(p)\}$, расчет целевой функции $E(x(k), C(l))$, соответствующий этому разбиению, выбор условий останова и значения параметра S_{max} .

Шаг 2. Проверка условий останова. Если условия останова выполняются (в смысле значений целевой функции $E(x(k), C(l))$) — процесс кластеризации заканчивается.

Шаг 3. Расчеты в окрестности минимального диаметра. Положить $S = 1 \leq S_{max}$.

Шаг 4. Если $S > S_{\max}$, вернуться к шагу 2.

Шаг 5. Выбрать случайным образом незанятую точку $x(r)$ в окрестности заданного диаметра и провести разбиение, используя это наблюдение вместо центроида $C(l)$.

Шаг 6. Использовать J -means метод, используя в качестве начального разбиения набор центроидов $C_s = \{C(1), \dots, C(l-1), x(s), C(l+1), \dots, C(p)\}$.

Шаг 7. Если обеспечено меньшее значение целевой функции – закончить работу алгоритма. Иначе положить $S = S + 1$ и идти к шагу 4.

При работе данного алгоритма дополнительно в качестве условий останова могут быть использованы также время работы процедуры, максимальное число итераций алгоритма между улучшениями целевой функции, ограничения на число расширений окрестностей соседства.

Основной эвристикой описанных выше процедур является перемещение центроидов в незанятые точки и нахождение на этой основе разбиений с лучшими значениями целевой функции. Основной же проблемой является необходимость многократного полного перебора всех сочетаний наблюдений и рассчитанных центроидов. Понятно, что при больших $\tilde{N}$ и p подход, основанный на J -means оказывается неэффективным.

В связи с этим целесообразно ввести в рассмотрение матричную модификацию метода известного как «Быстрые J -средние» (Fast J -means) [6]. Основная цель этой модификации – уменьшение вычислительной сложности J -means алгоритма. При этом поиск производится в окрестностях не всех центроидов $C(l)$, $l = 1, 2, \dots, p$, а только в одном единственном кластере, для выбора которого может быть использовано либо минимальное внутрикластерное расстояние

$$\overline{D}^2 = \frac{1}{N_l} \sum_{x(k) \in Cl_l} Sp(x(k) - C(l))(x(k) - C(l))^T,$$

либо максимальное межкластерное расстояние

$$\overline{\overline{D}}^2 = \frac{1}{p} \sum_{x(k) \in Cl_l} Sp(C(q) - C(l))(C(q) - C(l))^T.$$

Понятно, что наиболее удаленный от других кластер максимального диаметра имеет больше других шансы содержать более глубокий экстремум. Понятно также, что Fast J -means с точки зрения вычислительной реализации в p раз проще стандартного J -means.

Выводы

Предложена матричная модификация метода J -means предназначенная для решения задач сегментации цифровых изображений. Особенностью

введенной модификации является возможность обработки фрагментов изображений без их предварительной векторизации, введение расширяющихся окрестностей поиска, а также организация этого поиска в единственной окрестности, где с наибольшей вероятностью может содержаться искомое решение. По сравнению со стандартным методом кластеризации J -means предполагаемая модификация проще с вычислительной точки зрения и требует существенно меньшие вычислительные затраты.

Список литературы: 1. Gan G., Ma C., Wu J. Data Clustering: Theory, Algorithms and Applications. – Philadelphia, Pennsylvania: SIAM, 2007. – 455 p. 2. Xu R., Wunsch D.C. Clustering. – Hoboken, New Jersey: John Wiley&Sons, Inc., 2009. – 341 p. 3. MacQueen J.B. Some methods for classification and analysis of multivariate observations // Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability. 1967. – v. 2. – P. 281-297. 4. Hansen P., Mladenovic N. J-Means: a new local search heuristic for minimum sum of squares clustering // Pattern Recognition. – 2001. – 34. – P. 405-413. 5. Rodriguez-Sotelo J.L., Cuesta-Frau D., Castellanos-Dominguez G. An improved method for unsupervised analysis of ECG beats based on WT features and J-means clustering // Computers in Cardiology. – 2007. – 34. – P. 581-584. 6. Xinye Zh., Jubai A. Fast J-means algorithm: a constrained local search heuristic // Int. J. of Advancements in Computing Technology. – 2012. – 4. – P. 97-105. 7. Cooper L. Location-allocation problems // Operation Research. – 1963. – 11. – P. 331-343. 8. Расстригин Л.А. Статистические методы поиска. – М.: Наука, 1968. – 376 с.

Поступила в редколлегию 6.02.2015

УДК 004.932.2:004.93'1

Модифікація методу j-середніх в задачах сегментації зображень / С.І. Богучарський, А.Г. Каграманян, С.В. Машталір // Біоніка інтелекту: наук.-техн. журнал. – 2015. – № 1 (84). – С. 19–22.

Розглянуто методи пофрагментної кластеризації. Запропоновано матричну модифікацію методу j-середніх, в якій варіації центрів кластерів, що отримуються, дозволяють отримувати кращі в порівнянні з традиційними підходами локальні екстремуми.

Бібліогр.: 8 найм.

UDC 004.932.2:004.93'1

An induced agreement of relations in problems of information granulation / S.I. Bogucharskiy, A.G. Kagramanyan, S.V. Mashtalir // Bionics of Intelligence: Sci. Mag. – 2015. – № 1 (84). – P. 19–22.

Multi-frame clusering methods are considered. A matrix modification of J-mean in which clusters centers variations allows get better local extrema in compare to traditional approaches is offered.

Ref.: 8 items.