УДК 004.9
С. И. Богучарский
С. В. Машталир, к. т. н.

МОДИФИЦИРОВАННЫЙ МЕТОД КЛАСТЕРИЗАЦИИ X -СРЕДНИХ В ЗАДАЧАХ СЕГМЕНТАЦИЯ ИЗОБРАЖЕНИЙ

Аннотация. *Предложен модифицированный матричный метод кластеризации X -средних для решения задачи сегментации изображений в очень больших базах данных. Особенностью предложенного подхода является возможность обработки матричных сигналов в отсутствии информации о статистических характеристиках самих сигналов, а также числа кластеров, которое автоматически определяется в процессе анализа исходного массива данных.*

Ключевые слова: *Сегментация зображений, методы кластеризации, очень большие базы данных, матричные модификации методов, статистические характеристики*

С. І. Богучарський
С. В. Машталір, к. т. н.

МОДИФІКОВАНИЙ МЕТОД КЛАСТЕРИЗАЦІЇ X -СЕРЕДНІХ В ЗАДАЧАХ СЕГМЕНТАЦІЇ ЗОБРАЖЕНЬ

Анотація. *Запропоновано модифікований матричний метод кластеризації X -середніх для вирішення задачі сегментації зображень в дуже великих базах даних. Особливістю запропонованого підходу є можливість обробки матричних сигналів у відсутності інформації про статистичні характеристики самих сигналів, а також числа кластерів, яке автоматично визначається в процесі аналізу вихідного масиву даних.*

Ключові слова: *Сегментація збережений, методи кластеризації, дуже великі бази даних, матричні модифікації методів, статистичні характеристики*

S. I. Bogucharskiy
S. V. Mashtalir, PhD

MODIFIED X -MEANS CLUSTERING METHODS IN IMAGE SEGMENTATION PROBLEMS

Abstract. *The modified matrix X -means clustering method to solve the image segmentation problem in very large databases (VLDB) are proposed. A special feature of this approach is the ability to process matrix signals in the absence of information on the signals statistical characteristics, as well as the number of clusters is automatically determined during the original data set analysis.*

Keywords: *The image segmentation, clustering methods, VLDB, the matrix modification techniques, statistical characteristics*

Введение. Задача кластеризации больших массивов данных занимает важное место в общей проблеме интеллектуального анализа данных, а для ее решения к настоящему времени разработано множество разнообразных алгоритмов [1]. Особое место среди этого множества занимают методы, ориентированные на обработку информации, содержащиеся в очень больших базах данных, к которым предъявляются повышенные требования по быстродействию, простоте численной и программной реализации, требуемому объему памяти. Именно такие методы чаще всего используются при обработке изображений различной природы и, прежде всего, их сегментации. Здесь наиболее популярным является метод K-средних [2], благодаря, прежде всего, своей простоте, наглядности и интерпретируемости результатов. Метод K-средних относится к алгоритмам кластеризации, основанным на разбиении, которые разделяют массив, содержащий N объектов, описываемых n-мерными векторами признаков $x(k) \in R^n$, $k = 1, 2, \ldots N$ на p кластеров, где p задается априори. Именно от этого параметра зачастую зависит качество получаемого решения, при этом

субъективизм, привносимый пользователем, зачастую снижает его уровень. Конечно, при использовании иерархических алгоритмов кластеризации эта проблема решается автоматически, однако этот подход плохо приспособлен для обработки изображений, особенно, если речь идет о потоке видео. «Ближайшим родственником» метода K-средних, позволяющим автоматически определять количество кластеров p, содержащихся в обрабатываемом массиве $\mathrm{X} = \{x(1), x(2), ..., x(\mathrm{N})\} \subset R^n$, явление X-средних [5], однако его применение в задачах сегментации изображений ограничивается рядом обстоятельств. Во-первых, обработку изображений удобнее производить, представляя исходную информацию не в форме векторов, а в виде последовательности «окон», содержащих несколько расположенных рядом пикселей. В этом случае при использовании K-средних и X-средних эти «окна» предварительно необходимо векторизовать, а полученное решение – девекторизовать. Во-вторых, стандартная процедура X-средних, основанная на байессовском подходе, весьма громоздка с вычислительной точки зрения. И наконец, в третьих, априорное предположение о гауссовском распределении данных в кластерах и их сферичности естественно ограничивают применимость этого подхода в задачах сегментации изображений.

Данные обстоятельства делают целесообразной разработку эффективной модификации метода X-средних, ориентированного на решение задачи сегментации изображений в условиях, когда исходная информация задана в виде массива «окон»-матриц, а количество возможных классов-сегментов априори неизвестно.

Матричная модификация метода X-средних. Исходной информацией для решения задачи сегментации-кластеризации является массив матричных наблюдений $\mathrm{X} = \{x(1), x(2), ... x(\mathrm{N})\}$, $x(k) = \{x_{i_1 i_2}(k)\} \in R^{m \times n}$; $i_1 = 1, 2, ..., m$; $i_2 = 1, 2, ..., n$; $k = 1, 2, ..., \mathrm{N}$, который должен быть разделен в процессе самообучения на p сегментов-кластеров, описываемых центроидами $C = \{C(1), C(2), ..., C(p)\}$. При этом число кластеров p априорно не задается, а должно быть определено непосредственно в процессе обработки информации в диапазоне $2 \le p_{\min} \le p \le p_{\max} \le N - 1$.

Основная идея метода X-средних состоит в многократном, применении к исходному массиву данных X алгоритма K-средних с разными значениями p и оценивании получаемых результатов с помощью того или иного критерия, основанного на байесовском оценивании и, прежде всего, оценках Акаике, Касса-Вассермана, Шварца [6-10].

В общем случае процедура K-средних состоит из последовательности двух чередующихся этапов. При этом на первом этапе для случайно заданного набора исходных центроидов $C(1), ..., C(\mathrm{l}), ... C(p)$ имеющееся множество всех наблюдений $x(k)$ разбивается на p групп так, что каждое наблюдение $x(k)$ приписывается к ближайшему в смысле расстояние $D(x(k), \mathrm{C(l)}) = (Sp(x(k) - C(l))(x(k) - C(l))^T)^{1/2}$ центроиду, а на втором – происходит перерасчет всех центроидов, при этом в качестве новых оценок принимаются средние арифметические координат всех наблюдений, приписанных к конкретному центроиду. Эта процедура продолжается до тех пор, пока не стабилизируются координаты всех центроидов.

Можно показать, что описанная процедура минимизирует целевую функцию

$$E(x(k), C(l)) = \sum_{k=1}^{N} \sum_{l=1}^{p} \mu(x(k), C(l)) D^2(x(k), C(l)) =$$

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

(здесь $\mu(x(k), C(l)) = \begin{cases} 1, \text{ если } x(k) \in Cl_l, \\ 0, \text{ в противном случае,} \end{cases}$

Cl_l – обозначение l-го кластера), при этом координаты результирующих центроидов определяются соотношением

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

где N_l – число точек, отнесенных к l-му кластеру.

Процедура X-средних также состоит из последовательности двух этапов, один из которых называется «Улучшение параметров», а второй – «Улучшение структуры» [6-10]. При этом под улучшением параметров понимается нахождение координат центроидов с помощью стандартных K-средних при фиксированном числе кластеров p, а под улучшением структуры – наращивание числа кластеров до достижения требуемого качества сегментации.

Процесс обработки начинается с задания $p = p_{\min}$ нахождения $p_{\min}$ центроидов с помощью стандартного алгоритма K-средних. После окончания того этапа производится оценка полученного результата и реализуется второй этап, состоящий в изменении числа кластеров путем расщеплением уже сформированных $p_{\min}$ сегментов.

На этом этапе используется два возможных варианта: расщепление одного случайно выбранного кластера на два с начальными центроидами, совпадающими с двумя произвольно выбранными точками из исходного кластера, и расщепление половины из первоначального сформированных сегментов. Таким образом, в результате первого варианта образуется $p_{\min} + 1$ кластеров, а в результате второго $1{,}5 p_{\min}$ сегментов. К вновь сформированным классам применяется та же процедура K-средних и производится оценка качества полученного результата. Процесс расщепления-кластеризации продолжается до достижения $p = p_{\max}$. Далее из множества полученных результатов выбирается наилучший с точки зрения принятого критерия.

Каждый из полученных результатов с позиции байесовского оценивания трактуется как альтернативная кластерная модель M_l, $l = 1, 2, \ldots$, $p_{\max}$, среди набора которых должна быть выбрана наилучшая, для чего предлагается использовать критерий Шварца [3, 6] в форме

$$BIC(M_l) = L(l) - \frac{h(l)}{2} \log N,$$

где $L(l)$ – логарифмическая функция правдоподобия для l-той модели, $h(l)$ – число параметров модели кластера M_l.

С учетом того, что внутрикластерное распределение данных полагается подчиненным нормальному закону и при этом весь массив данных «разбит» на p кластеров, можно записать выражение для внутрикластерной дисперсии

$$\sigma_l^2 = \frac{1}{N - p} \sum_{x(k) \in Cl_l} Sp(x(k) - C(l))(x(k) - C(l))^{\mathrm{T}},$$

условной вероятности

$$P(x(k) \mid x(k) \in Cl_l) = \frac{N_l}{N} \frac{1}{\sqrt{2\pi}\sigma_l^{mn}} \cdot \cdot \exp\left(-\frac{1}{2\sigma_l^2} Sp(x(k) - C(l))(x(k)C(l))\right)^{\mathrm{T}}),$$

логарифмической функции правдоподобия

$$L(l) = \log \prod_{k=1}^{N} P(x(k)) = \sum_{k=1}^{N} (\log \frac{1}{\sqrt{2\pi}\sigma_l^{mn}} \cdot \cdot \exp(-\frac{1}{2\sigma_l^2} Sp(x(k) - C(l))(x(k)C(l))^{\mathrm{T}}) + \log \frac{N_l}{N}),$$

при этом $h(l) = \mathrm{p}$.

Кластерная модель с максимальным значением критерия Шварца полагается наилучшей, а число соответствующих ей кластеров – оптимальным значением количества сегментов в обрабатываемом изображении.

Вместе с тем, как уже отмечалось выше, существенные недостатки этого подхода, связанные прежде всего, с достаточно жесткими вероятностными предположениями, ограничивают его использование в задачах обработки изображений.

Матричная модификация X-средних для обработки изображений. Среди критериев качества кластеризации, не опирающихся на статистические предложения и пригодных для нахождения количества кластеров в массиве данных, в качестве весьма эффективного показал себя критерий Цалиньского-Харабаша [2], который может быть адаптирован на случай матричных объектов.

Итак, пусть задан массив данных $\mathrm{X}=\{x(1),\dots x(N)\}$, $x(k)=\{x_{i_1 i_2}(k)\}\in R^{m\times n}$, $k=1,2,\dots,N$, который некоторым образом разбит на p кластеров с набором центроидов $C=\{C(1),C(2),\dots C(p)\}\subset R^{m\times n}$. Для оценки качества такого разбиения критерий Цалиньского-Харабаша может быть записан в виде

$$CH(p)=\frac{\frac{1}{p-1}SpS_B(p)}{\frac{1}{N-p}SpS_w(p)},$$

где $S_B(p)=\sum_{l=1}^{p}N_l(C(l)-\overline{C})(C(l)-\overline{C})^{\mathrm{T}}$ – матрица междукластерного рассеивания,

$$S_w(p)=\sum_{l=1}^{p}\sum_{k=1}^{N}\mu(x(k),\mathrm{C}(l))(x(k)-C(l))\cdot$$
$$\cdot(x(k)-C(l))^{\mathrm{T}}$$

– матрица внутрикластерного рассеивания,

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

– центроид (центр тяжести) кластера Cl_l,

$\overline{C}=\frac{1}{N_l}\sum_{l=1}^{p}N_lC(l)$ – матричный центр тяжести массива X.

Таким образом, предлагаемая модификация метода X-средних для обработки изображений на основе K-средних и критерия Цалиньского-Харабаша может быть реализована в виде последовательности этапов:

– задание $p=p_{\min}$, решение задачи с помощью K-средних и расчет $CH(p_{\min})$;

– расщепление любого из кластеров, решение задачи с помощью K-средних и расчет $CH(p_{\min}+1)$;

– при $CH(p+1)\le CH(p)$ полагается, что p есть наилучшая оценка количества кластеров в массиве X.

Заключение. В статье предложен модифицированный метод кластеризации X-средних для решения задачи сегментации изображений. Особенностью предложенной модификации является возможность обработки матричных сигналов в отсутствии информации о статистических характеристиках этих сигналов и числа кластеров, которое автоматически определяется в процессе анализа исходного массива данных. Алгоритмическая реализация метода характеризуется вычислительной простотой и высоким быстродействием.

Список использованной литературы

1. Gan G., Ma Ch,. Wu J. (2007), Data clustering. Theory, algorithms, and applications. *Philadelphia, Pennsilvania: SIAM*. 455 p. DOI:http://dx.doi.org/10.1137/1.9780898718348.fm

2. Xu R., Wunsch D.C. (2008), Clustering. *Hoboken.* John Wiley & Sons. 358 p.

3. Pelleg D., Moore A. (2000), X-means: Extending K-means with efficient estimation of the number of clusters. *Proc. of the 17th int. conf. on Machine Learning.* San Francisco: Morgan Kaufmann. pp. 727-730. url:https://www.cs.cmu.edu/~dpelleg/download/xmeans.pdf

4. Ishioka T. (2000), Extended K-means with an efficient estimation of the number of clusters. *Proc. of second int. conf. Intelligent Data Engineering and Automated Learning IDEAL. 2000.* Hong Kong, China. pp.17-22.

5. Ishioka T. (2005), An expansion of X-means for automatically determining the optimal number of clusters. *Proc. of the 4th IASTED Int. conf. Computational intelligence.* Calgary, Alberta, Canada. pp. 91-96.
url: http://www.rd.dnc.ac.jp/~tunenori/doc/487-053.pdf

6. Schwarz G. (1978), Estimation the dimension of a model. *The Annals of Statistics*, Vol. 6(2). pp. 461-464.
url:http://www.andrew.cmu.edu/user/kk3n/simplicity/schwarzbic.pdf

7. Bozdogan H. (1987), Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychametrica.* Vol. 52. pp. 345-370.
url:http://www.ics.uci.edu/~staceyah/201/Bozdogan_1987-AIC.pdf

8. Jolion J.M., Meer P., Bataouche S.(1991) Robust clustering with applications in computer vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* Vol 13. pp. 291-802.
url:http://soe.rutgers.edu/~meer/OLD/robcluster.pdf

9. Krishnapuram R., Freg C.P. (1992), Fitting and unknown numbers of lines and planes to image to image data through compatible cluster merging. *Pattern Recognition*. Vol. 25. pp. 385-400.
url:http://ac.els-cdn.com/003132039290087Y/1-s2.0-003132039290087Y-main.pdf?_tid=7d5975fc-7e10-11e5-ac47-00000aab0f27&acdnat=1446104679_6557f44949554bc4308305b67036dfd8

10. Kass B., Wasserman L. (1995), A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. of the Amer. Statistical Association.*Vol.90 pp.773-795

References

1. Gan G., Ma Ch,. Wu J. (2007), Data clustering. Theory, algorithms, and applications. *Philadelphia, Pennsilvania: SIAM*. 455 p.

2. Xu R., Wunsch D.C. (2008), Clustering. *Hoboken.* John Wiley & Sons. 358 p.

3. Pelleg D., Moore A. (2000), X-means: Extending K-means with efficient estimation of the number of clusters. *Proc. of the 17th int. conf. on Machine Learning.* San Francisco: Morgan Kaufmann. pp. 727-730.

4. Ishioka T. (2000), Extended K-means with an efficient estimation of the number of clusters. *Proc. of second int. conf. Intelligent Data Engineering and Automated Learning IDEAL. 2000.* Hong Kong, China. pp.17-22.

5. Ishioka T. (2005), An expansion of X-means for automatically determining the optimal number of clusters. *Proc. of the 4th IASTED Int. conf. Computational intelligence.* Calgary, Alberta, Canada. pp. 91-96.

6. Schwarz G. (1978), Estimation the dimension of a model. *The Annals of Statistics*, Vol. 6(2). pp. 461-464.

7. Bozdogan H. (1987), Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychametrica.* Vol. 52. pp. 345-370.

8. Jolion J.M., Meer P., Bataouche S.(1991) Robust clustering with applications in computer vision. *IEEE Trans.on Pattern Analysis and Machine Intelligence.* Vol 13 pp. 291 - 802.

9. Krishnapuram R., Freg C.P. (1992), Fitting and unknown numbers of lines and planes to image to image data through compatible cluster merging. *Pattern Recognition*. Vol. 25. pp. 385-400.

10. Kass B., Wasserman L. (1995), A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. of the Amer. Statistical Association.* Vol. 90. pp.773-795.

Богучарский
Сергей Игоревич,
аспирант каф. информатики Харьковского нац. ун-та радиоэлектроники, Украина, Харьков, пр. Ленина, 14, ауд. 288
sbogucharskiy@rambler.ru

Машталир
Сергей Владимирович,
к.т.н., доцент каф. информатики Харьковского нац. ун-та радиоэлектроники Украина, Харьков, пр. Ленина, 14, ауд. 288
sergii.mashtalir@nure.ua