ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 004.032.26

НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ ПОТОКОВ ДАННЫХ С ПОМОЩЬЮ ЕМ-АЛГОРИТМА НА ОСНОВЕ САМООБУЧЕНИЯ ПО Т. КОХОНЕНУ

Е.В. БОДЯНСКИЙ, А.А. ДЕЙНЕКО, А.А. ЗАИКА, Я.В. КУЦЕНКО

В работе предложен мягкий вероятностный нечеткий алгоритм кластеризации многомерных данных, последовательно поступающих на обработку в режиме реального времени. Развиваемый подход предназначен для решения задач Dynamic Stream Mining в условиях перекрывающихся классов, по сравнению со своими прототипами значительно проще в вычислительной реализации, не использует никаких вероятностных предположений о природе обрабатываемых данных.

Ключевые слова: кластеризация, нечеткая логика, вычислительный интеллект, самообучение, нечеткая кластеризация, самоорганизующаяся карта Т. Кохонена.

ВВЕДЕНИЕ

Задача кластеризации больших массивов многомерных наблюдений (векторов-образов) часто встречается во многих реальных практических задачах, а для ее решения разработано множество алгоритмов [1–3], при этом в последние годы в рамках концепции Big Data особое внимание уделяется обработке информации, хранящейся либо в сверхбольших базах данных (VLDB), либо поступающей на обработку в on-line режиме в форме потока данных (data stream). Для решения этих задач с успехом может быть использован математический аппарат вычислительного интеллекта (computational intelligence) [4–7] и, прежде всего, искусственные нейронные сети и мягкие вычисления (soft computing), основанные на нечеткой логике. Понятно, что известные системы вычислительного интеллекта должны быть существенно модифицированы для обработки больших объемов информации, последовательно поступающей на обработку.

В наиболее общей постановке задачи кластеризации предполагается, что имеется массив (возможно растущий) из N многомерных наблюдений, описываемых n-мерными векторами признаков

$$x(k)=(x_1(k),x_2(k),...,x_n(k))^T\in R^n,$$
$$k=1,2,...,N,...,$$

который необходимо разбить на m кластеров, при этом это число может быть заранее не известно, т. е. $1<m<N$. Очевидно, что столь большое число известных методов решения задачи кластеризации связано с тем, что сегодня не существует универсального алгоритма пригодного для эффективного использования во всех возникающих ситуациях. Одна из таких возможных и достаточно сложных ситуаций связана с предположением, что каждый вектор наблюдений может одновременно относиться с различными уровнями возможностей, вероятностей или принадлежностей не к одному, а сразу к нескольким или ко всем формируемым кластерам. В этой ситуации на первый план выходят, так называемые, мягкие алгоритмы (soft algorithms) [8], среди которых наибольшее внимание привлечено к нечетким методам кластеризации [9–11] и вероятностным алгоритмам (probabilistic model-based algorithms) [8], среди которых для обработки больших массивов широкое распространение получил, так называемый, ЕМ-алгоритм (Expectation-Maximization algorithm) [12–17], в основе которого лежат сугубо вероятностные предпосылки. Каждый из отмеченных подходов имеет свои достоинства и недостатки, в связи с чем представляется целесообразным разработка численно простой мягкой процедуры кластеризации, объединяющей в себе достоинства вероятностного и фаззи-подходов и предназначенной для обработки потоков данных, поступающих в on-line режиме.

1. ЕМ-АЛГОРИТМ ВЕРОЯТНОСТНОЙ КЛАСТЕРИЗАЦИИ

Задача вероятностной кластеризации в общей постановке сводится к проблеме самообучения при неизвестном числе областей [18], при этом предполагается, что плотность распределения данных в каждом кластере подчиняется многомерному нормальному (гауссовскому) закону

$$p_j(x)=\frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det\Sigma_j}}\exp\left(-\frac{1}{2}(x-c_j)^T\Sigma_j^{-1}(x-c_j)\right),\quad j=1,2,...m, \tag{1}$$

а совместная плотность распределения всех данных описывается смесью

$$p(x)=\sum_{j=1}^{m}p_j p_j(x)=$$
$$=\sum_{j=1}^{m}\frac{p_j}{(2\pi)^{\frac{n}{2}}\sqrt{\det\Sigma_j}}\exp\left(-\frac{1}{2}(x-c_j)^T\Sigma_j^{=1}\left(x-c_j\right)\right),\quad j=1,2,...m, \tag{2}$$

где $c_j-(n\times1)$ – мерный вектор-центроид j-го кластера, $\sum_j-(n\times n)$ – корреляционная матрица j-го кластера такая, что

$$\Sigma_j = \frac{1}{N}\sum_{k=1}^{N}(x(k)-c_j)(x(k)-c_j)^T, \quad (3)$$

p_j – априорные вероятности (веса), подчиняющиеся естественному условию

$$\sum_{j=1}^{m} p_j = 1, \quad (4)$$

при этом предполагается, что c_j, Σ_j и $p_j \forall j = 1,2,...,m$ априори неизвестны и подлежат оцениванию в процессе кластеризации.

Здесь можно отметить, что показатель степени экспоненты в (1), (2) есть расстояние Махаланобиса между c_j и наблюдением $x(k)$, т.е.

$$d_M^2(c_j, x(k)) = (x(k)-c_j)^T \Sigma_j^{-1}(x(k)-c_j), \quad (5)$$

а условие (4) совпадает с условием, накладываемым на сумму принадлежностей в популярном алгоритме нечетких С-средних (FCM) Дж. Бездека [9]

$$\sum_{j=1}^{m} u_j(k) = 1, \quad (6)$$

(здесь $u_j(k)$ – уровень нечеткой принадлежности наблюдения $x(k)$ j-му кластеру), в связи с чем алгоритмы кластеризации, связанные с ограничением (6), называются вероятностными алгоритмами нечеткой кластеризации.

Работа EM-алгоритма состоит из повторяющейся последовательности двух шагов, при этом на шаге E (expectation step) производится оценивание параметров совместного распределения (2), а на шаге M (maximization step) максимизируется критерий самообучения в виде логарифмической функции правдоподобия

$$E(c_j, \Sigma_j, p_j, x(k)) = \sum_{k=1}^{N} \log\left(\sum_{j=1}^{m} p_j p_j(x(k-1))\right),$$

для чего могут быть использованы как традиционные градиентные, так и квазиньютоновские процедуры оптимизации [1].

И, наконец, для оценки вероятности принадлежности наблюдения j-му кластеру используется выражение

$$p_j(x(k)) = \frac{p_j \exp\left(-\frac{1}{2}\left(x(k)-c_j\right)^T \Sigma_j^{-1}\left(x(k)-c_j\right)\right)}{\sum_{l=1}^{m} p_l \exp\left(-\frac{1}{2}\left(x(k)-c_j\right)^T \Sigma_j^{-1}\left(x(k)-c_l\right)\right)} \quad (7)$$

удовлетворяющее условию (4).

Частным случаем EM-алгоритма является популярный метод кластеризации K-средних и совпадающий с ним при $p_j = m^{-1}$ и единичных корреляционных матрицах Σ_j. При этом метод K-средних является четкой процедурой, что означает, что каждое наблюдение может быть отнесено к единственному кластеру. При этом метод K-средних существенно проще с вычислительной точки зрения, чем EM-алгоритм и связан с минимизацией критерия самообучения, основанного на евклидовых расстояниях

$$E(c_j, x(k)) = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j(k)\left\|x(k)-c_j\right\|^2, \quad (8)$$

где

$$u_j(k) = \begin{cases} 1, \text{если } x(k) \text{ принадлежит} \\ \quad j\text{-му кластеру}, \\ 0 \text{ в противном случае}. \end{cases} \quad (9)$$

EM-алгоритм также относится к процедурам, основанным на расстояниях, и в этом смысле близок к, так называемому, методу K-средних Махаланобиса, являющемуся четкой процедурой, минимизирующей целевую функцию

$$E(c_j, \Sigma_j, x(k)) = \\ = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j(k)(x(k)-c_j)^T \Sigma_j^{-1}\left(x(k)-c_j\right), \quad (10)$$

где Σ_j и $u_j(k)$ определяется выражениями (3), (9).

Принципиальная разница между EM-алгоритмом и методом K-средних Махаланобиса состоит в том, что последний дает однозначное решение о принадлежности каждого наблюдения единственному кластеру, в то время как вероятностная процедура учитывает возможность перекрытия классов с расчетом вероятностей согласно выражению (7).

Наряду с несомненными достоинствами EM-алгоритм обладает и рядом существенных ограничений. Во-первых, исходные данные должны иметь случайную природу и подчиняться нормальному закону распределения, во-вторых, на M-этапе возможно «застревание» процесса оптимизации в локальных экстремумах (эта проблема может быть преодолена с помощью использования процедур многоэкстремальной оптимизации), в-третьих, с вычислительной точки зрения это все-таки довольно громоздкая процедура [8] и, наконец, подразумевается, что весь массив данных, подлежащих кластеризации, задан заранее и не изменяется в процессе обработки.

В связи с этим представляется целесообразной разработка числено простого алгоритма кластеризации, основанного на метрике Махаланобиса, учитывающего возможность взаимного перекрытия формируемых кластеров и позволяющего анализировать поток данных, последовательно поступающих на обработку в on-line режиме.

2. НЕЧЕТКАЯ ВЕРОЯТНОСТНАЯ КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ WTA-ПРАВИЛА САМООБУЧЕНИЯ

Результат минимизации критерия самообучения (8) метода K-средних имеет вид среднего арифметического данных каждого кластера

$$c_j = \frac{\sum_{k=1}^{N} u_j(k)x(k)}{\sum_{k=1}^{N} u_j(k)} = \frac{1}{N_j}\sum_{x(k)\in Cl_j} x(k), \quad (11)$$

где N_j – количество наблюдений $x(k)$, отнесенных к кластеру Cl_j.

Если же данные поступают на обработку последовательно в on-line режиме, решение (11) может быть получено с помощью кластеризирующей нейронной сети Т. Кохонена [19], настройка синаптических весов которой, являющихся по сути компонентами векторов-центроидов, производится с помощью тех или иных алгоритмов конкурентного самообучения, наиболее популярным из которых является WTA-правило (“Winner Takes All”). При этом сама процедура настройки подобно EM-алгоритму состоит из последовательности двух этапов: конкуренции (соответствует E-шагу) и синаптической адаптации [20] (соответствует M-шагу).

Суть конкурентного обучения по Кохонену состоит в том, что если к моменту поступления наблюдения $x(k)$ рассчитаны координаты центроидов $c_1(k-1),...,c_j(k-1),...,c_m(k-1)$, среди них выбирается «победитель», ближайший в смысле евклидова расстояния

$$d_E^2\left(c_j(k-1),x(k)\right)=\left\|x(k)-c_j(k-1)\right\|^2 \quad (12)$$

к $x(k)$ (E-шаг), который и уточняется на M-шаге с помощью рекуррентной процедуры

$$c_j(k)=\begin{cases}c_j(k-1)+\eta(k)\left(x(k)-c_j(k-1)\right),\\ \quad \text{если } c_j(k-1)-\text{"победитель"},\\ c_j(k-1) \text{ в противном случае},\end{cases} \quad (13)$$

где $\eta(k)$ параметр шага обучения, выбираемый обычно из эмпирических соображений, хотя несложно показать, что результат (11) может быть получен при $\eta(k)=k^{-1}$.

Несложно заметить, что первое соотношение (13) есть не что иное, как градиентная минимизация (8), т.е. может быть переписано в виде

$$c_j(k)=\begin{cases}c_j(k-1)-\eta(k)\nabla_{c_j}d_E^2\left(c_j(k-1),x(k)\right),\\ \quad \text{если } c_j(k-1)-\text{"победитель"},\\ c_j(k-1) \text{ в противном случае}.\end{cases} \quad (14)$$

Аналогично (14) может быть введена градиентная процедура минимизации критерия (10) на основе метрики Махаланобиса (5) в виде

$$c_j(k)=\begin{cases}c_j(k-1)-\eta(k)\nabla_{c_j}d_M^2\left(c_j(k-1),x(k)\right),\\ \quad \text{если } c_j(k-1)-\text{"победитель"},\\ c_j(k-1) \text{ в противном случае},\end{cases}$$

или

$$c_j(k)=\begin{cases}c_j(k-1)+\eta(k)\times\\ \times\Sigma_j^{-1}(k-1)\left(x(k)-c_j(k-1)\right),\\ \quad \text{если } c_j(k-1)-\text{"победитель"},\\ \Sigma_j(k-1)=\dfrac{1}{k-1}\times\\ \times\displaystyle\sum_{l=1}^{k-1}\left(x(l)-c_j(k-1)\right)\left(x(l)-c_j(k-1)\right)^T,\\ c_j(k-1) \text{ в противном случае}.\end{cases} \quad (15)$$

Несложно заметить, что алгоритм самообучения (15) является по сути последовательной реализацией метода K-средних Махаланобиса, т.е. позволяет получить только четкое решение.

Для оценки уровня принадлежности отдельных наблюдений в случае перекрывающихся классов вместо громоздкого выражения (7) целесообразно воспользоваться оценкой, принятой в стандартном FCM-алгоритме Дж. Бездека, используя вместо расстояния $d_E^2\left(c_j(k),x(k)\right)$ метрику Махаланобиса $d_M^2\left(c_j(k),x(k)\right)$ в виде

$$u_j(k)=\frac{d_M^{-2}\left(c_j(k),x(k)\right)}{\displaystyle\sum_{l=1}^{m}\left(d_M^{-2}\left(c_l(k),x(k)\right)\right)}=$$
$$\frac{\left(\left(x(k)-c_j(k)\right)\Sigma_j^{-1}\left(x(k)-c_j(k)\right)\right)^{-1}}{\displaystyle\sum_{l=1}^{m}\left(\left(x(k)-c_l(k)\right)^T\Sigma_l^{-1}\left(x(k)-c_l(k)\right)\right)^{-1}}. \quad (16)$$

Таким образом, процедура нечеткой вероятностной кластеризации (15), (16), являясь своеобразным гибридом EM-алгоритма, метода K-средних Махаланобиса, алгоритмов нечеткой кластеризации Бездека [9] и Гата-Гевы [10] и нечеткой кластеризующей нейронной сети Кохонена в ее адаптивном варианте [21, 22], характеризуется вычислительной простотой и позволяет анализировать данные, последовательно поступающие на обработку в on-line режиме.

ВЫВОДЫ

Предложен мягкий вероятностный нечеткий алгоритм кластеризации многомерных данных, последовательно поступающих на обработку в режиме реального времени. Развиваемый подход предназначен для решения задач Dynamic Stream Mining в условиях перекрывающихся классов, по сравнению со своими прототипами значительно проще в вычислительной реализации, не использует никаких вероятностных предположений о природе обрабатываемых данных.

Литература

[1] *Gan, G., Ma, Ch. and Wu, J.*, Data Clustering: Theory, Algorithms and Applications, Philadelphia: SIAM, 2007. — 466 p.

[2] *Xu, R. and Wunsch, D.C.*, Clustering, IEEE Press Series on Computational Intelligence. Hoboken, NJ: John Wiley & Sons, Inc., 2009. — 370 p.

[3] *Aggarwal, C.C. and Reddy, C.K.*, Data Clustering. Algorithms and Application, Boca Raton: CRC Press, 2014. — 648 p.

[4] *Rutkowski, L.*, Computational Intelligence. Methods and Techniques, Berlin-Heidelberg: Springer-Verlag, 2008. — 514 p.

[5] *Mumford, C. and Jain, L.*, Computational Intelligence. Collaboration, Fuzzy and Emergence, Berlin: Springer-Vergal, 2009. — 726 p.

[6] *Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M. and Held, P.*, Computational Intelligence. A Methodological Introduction, Berlin: Springer, 2013. — 488 p.

[7] *Du, K.-L. and Swamy, M.N.S.*, Neural Networks and Statistical Learning, London: Springer-Verlag, 2014. — 824 p.

[8] *Aggarwal, C.C.*, Data Mining, Cham: Springer, Int. Publ., Switzerland, 2015. — 734 p.

[9] *Bezdek, J.-C.* Pattern Recognition with Fuzzy Objective Function Algorithms, N.Y.: Plenum Press, 1981, 272 p.

[10] *Gath, I. and Geva, A.B.*, Unsupervised optimal fuzzy clustering, Pattern Analysis and Machine Intelligence, 1989. vol. 2, no.7. — P. 773–787.

[11] *Bezdek, J.C. Keller, J., Krishnapuram, R. and Pal, N.*, Fuzzy Models and Algorithms for Pattern Recognition andImage Processing. The Handbooks of Fuzzy Sets, Kluwer, Dordrecht, Netherlands: Springer, 1999. vol. 4. — 776 p.

[12] *Dempster, A.P., Laird, N.M. and Rubin, D.B.*, Maximum likelihood from incomplete data via the EM algorithm, J. of the Royal Statistical Society, 1977, Ser.B, vol. 39, no.1. — P. 1–38.

[13] *Hathaway, R.*, Another interpretation of the EM algorithm for mixture distributions. J. of Statistics & Probability Letters, 1986, vol. 4. — P. 53–56.

[14] *Meng, X.L. and Rubin, D.B.*, Maximum likelihood estimation via the ECM algorithm:a general framework, Biometrica, 1993, vol. 80. — P. 267–278.

[15] *Liu, C. and Rubin, D.B.*, The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence, Biometrica, 1994, vol. 81. — P. 633–648.

[16] *Fessler, J.A. and Hero, A.O.*, Space – alternating generalized EM algorithm, *IEEE Trans. on Signal Processing,* 1994. vol .42. — P. 2664–2677.

[17] *Friedman, J., Hastie, T. and Tibshirani, R.*, The Elements of Statistical Learning. Data Mining, Inference and Prediction, Berlin: Springer, 2003. – 552 p.

[18] *Tsypkin, Ya.Z.*, *Foundations of learning systems theory*, M.: Nauka, 1970. (in Russion)

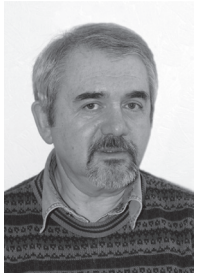
[19] *Kohonen, T.* Self-Organizing Maps, Berlin: Springer-Verlag, 1995. – 362 p.

[20] *Haykin, S.*, Neural Networks. A Comprehensive Foundation, Upper Saddle River, N.J.: Prentice Hall, Inc., 1999. – 842 p.

[21] *Gorshkov, Ye., Kolodyaznhiy, V. and Bodyanskiy, Ye.*, New recursive learning algorithms for fuzzy Kohonen clustering network, Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronic Systems, Rapperswil, Switzerland, 2009. — P. 58–61.

[22] *Bodyanskiy, Ye., Kolchygin, B.V. and Pliss, I.P.*, Adaptive neuro-fuzzy Kohonen network with variable fuzzifier, International Journal «Information Theories and Applications», Sofia: ITHEA, 2011. vol.18, no3. — P. 215–223.

Поступила в редколлегию 11.03.2016



Бодянский Евгений Владимирович, руководитель ПНИЛ АСУ, д.т.н., проф., профессор кафедры искусственного интеллекта Харьковского национального университета радиоэлектроники. Научные интересы: гибридные системы вычислительного интеллекта.

Дейнеко Анастасия Александровна, канд. техн. наук, научный сотрудник ПНИЛ АСУ Харьковского национального университета радиоэлектроники. Научные интересы: гибридные системы вычислительного интеллекта.

Куценко Яна Владимировна, аспирантка кафедры искусственного интеллекта Харьковского национального университета радиоэлектроники. Научные интересы: нейронные сети, мягкие вычисления, вычислительный интеллект, нейро-фаззи системы.

Заика Александр Анатольевич, студент кафедры искусственного интеллекта Харьковского национального университета радиоэлектроники. Научные интересы: нейронные сети, мягкие вычисления, вычислительный интеллект.

УДК 004.032.26

Нечітке кластерування потоків даних за допомогою ЕМ-алгоритму на основі самонавчання за Т. Когоненом / Є. В. Бодянський, А. О. Дейнеко, А. О. Заїка, Я. В. Куценко // Прикладна радіоелектроніка: наук.-техн. журнал. – 2016. – Том 15. – № 1. – С. 80–83.

У статті запропоновано м'який ймовірнісний нечіткий алгоритм кластерування багатовимірних даних, які послідовно надходять на опрацювання в режимі реального часу. Розглянутий підхід призначений для вирішення завдань Dynamic Stream Mining за умов перетинання класів, порівняно зі своїми прототипами значно простіше в обчислювальній реалізації, не використовує ніяких ймовірнісних припущень про природу оброблюваних даних.

Ключові слова: кластерування, нечітка логіка, обчислювальний інтелект, самонавчання, нечітке кластерування, самоорганізовна мапа Т. Когонена.

Бібліогр.: 22 найм.

UDC 004.032.26

Fuzzy clustering of data streams by EM-algorithm based on T. Kohonen's self-learning / Ye.V. Bodyanskiy, A.O. Deineko, O.O. Zayika, Ya.V. Kutsenko // Applied Radio Electronics: Sci. Journ. — 2016. — Vol. 15. — № 1. — P. 80–83.

In the paper a soft probabilistic fuzzy clustering algorithm of multidimensional data sets that are sequentially fed to processing in on-line mode is proposed. The approach under investigation is designed for solving Dynamic Stream Mining problems in conditions of overlapped classes and is simpler in computation realization in comparison with its prototypes, it does not use any probabilistic assumptions about the nature of data processed.

Keywords: clustering, fuzzy logic, computational intelligence, self-leaning, fuzzy clustering, Kohonen self-organizing map.

Ref.: 22 items.