

ЗАДАЧИ НЕСТАЦИОНАРНОЙ КЛАСТЕРИЗАЦИИ СОСТОЯНИЯ НЕФТЕХИМИЧЕСКОГО ОБОРУДОВАНИЯ

Гимаров В.А., Дли М.И., Битюцкий С.Я.

Рассмотрены постановка и возможные подходы к решению задач кластеризации состояния нефтехимического оборудования в условиях, когда с течением времени кластерная структура может претерпевать изменения из-за образования новых кластеров, слияния, расщепления или исчезновения существующих, а также из-за дрейфа их центров.

Введение. Достаточно часто в задачах кластеризации приходится учитывать не только текущие (статические) свойства классифицируемых (кластеризуемых) объектов, но и их временные, т.е. динамические изменения. Примером подобной задачи может служить следующая. Пусть производится наблюдение за состоянием группы однородных образцов нефтехимического оборудования, при этом состояние каждого j -го образца в произвольный момент времени t_k отображается вектором характеристик $x_j(t_k)$. В любой фиксированный момент времени в совокупности таких образцов $\{x_j(t)\}$ может быть выделено несколько групп (кластеров), внутри которых объекты обладают некоторыми общими свойствами (например, группа хорошо функционирующего оборудования, группа оборудования, имеющего какое-то определенное отклонение, группа оборудования, отличающегося плохим функционированием, и т.п.). Если количество кластеров, их размеры и центры с течением времени не изменяются, то имеем дело с так называемой статической кластерной структурой, выявлению которой (т.е. решению задачи кластеризации или распознаванию образов без обучения) посвящено много научных работ – см., например, [1-8]. Однако совсем иначе обстоит дело в ситуациях, когда с течением времени кластерная структура изменяется – т.е. какие-то объекты с течением времени начинают обладать некоторыми новыми свойствами, скажем, образуют группу объектов, функционирование которых находится на границе их катастрофического изменения, другие объекты подобные свойства теряют и т.п. В таком случае кластерная структура претерпевает временные изменения и становится динамической, при этом исследованию этой задачи кластеризации внимания до настоящего времени практически не уделялось (можно, пожалуй, отметить лишь статью [9]), хотя ее практическая важность очевидна – например, для целей прогнозирования развивающихся кризисных ситуаций в нефтехимических технических системах, прогноза отказа оборудования и т.д.

Настоящая статья посвящена формулировке задачи динамической кластеризации и возможным подходам к решению ее разновидностей.

Постановка задачи. Предположим, что производится мониторинг за совокупностью объектов (образцов оборудования), в каждый момент времени характеризующихся n -мерным вектором количественных признаков $x_j(t)$, где индекс j относится к номеру объекта. Измерение характеристик данных объектов осуществляется в дискретные, не обязательно равноотстоящие моменты t_k . Через некоторые (не обязательно равные) интервалы T проводится проверка стабильности структуры кластеров, и при необходимости, т.е. при наличии динамических изменений, – ее коррекция. С целью исключения влияния слишком старых наблюдений, мониторинг организуется по принципу временного окна, т.е.

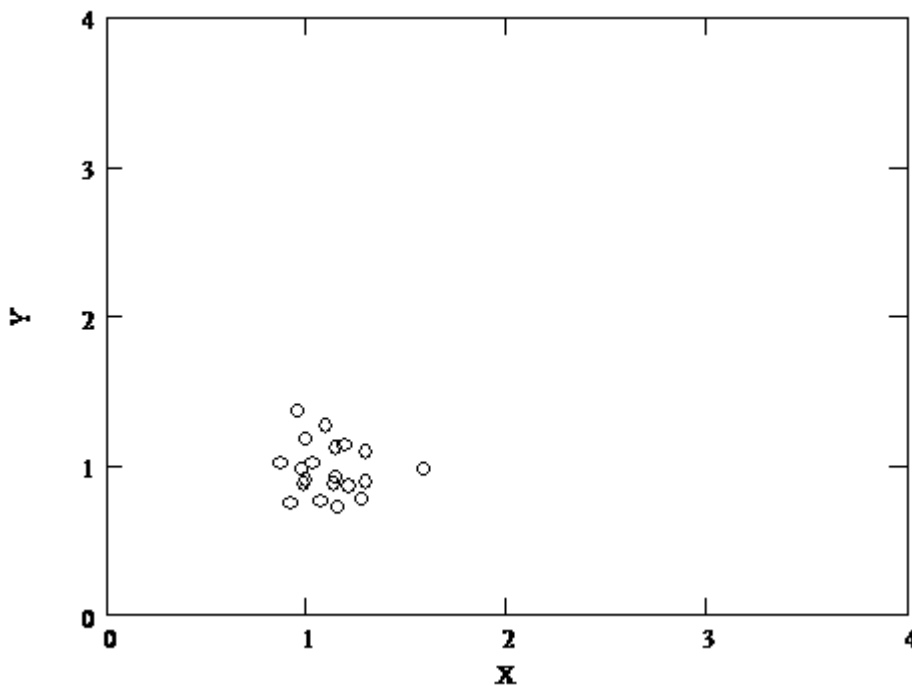
при анализе кластерной структуры учитываются только объекты, зафиксированные в последнем перед очередной проверкой интервале. Количество подобных объектов не предполагается фиксированным.

Требуется: на основании имеющейся информации провести анализ подобной (динамической) задачи кластеризации, выявить ее особенности, варианты и разработать соответствующие методы решения.

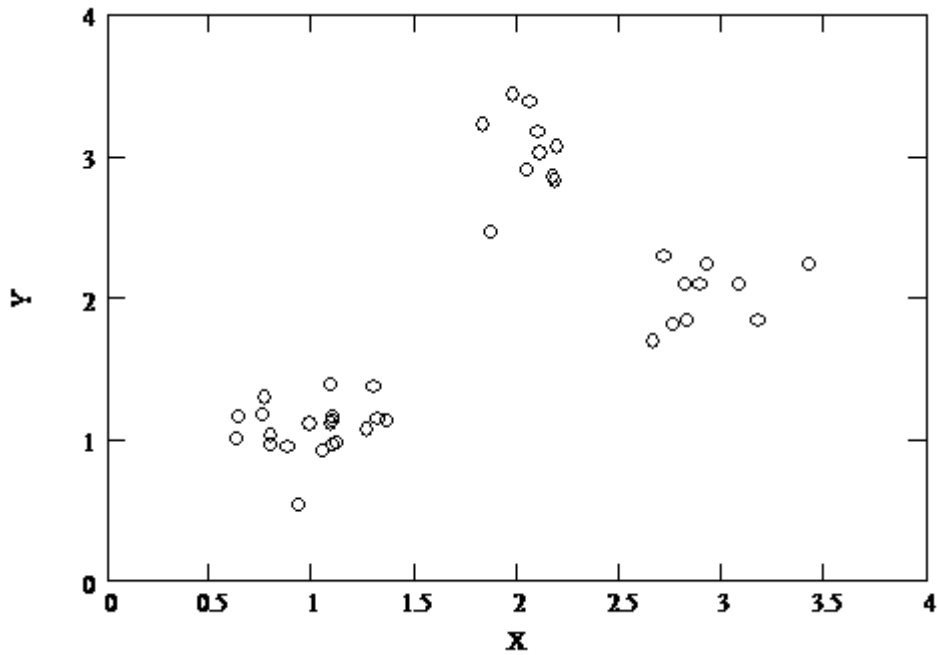
Анализ и варианты задачи динамической кластеризации. Ввиду учета (в последнем временном окне) новых объектов и "отбрасывании" старых, не попавших в данное окно, можно выделить, по крайней мере, следующие динамические (временные) изменения в структуре кластеров:

- 1) образование новых кластеров;
- 2) слияние кластеров;
- 3) расщепление или дробление кластеров;
- 4) исчезновение кластеров;
- 5) дрейф центров кластеров.

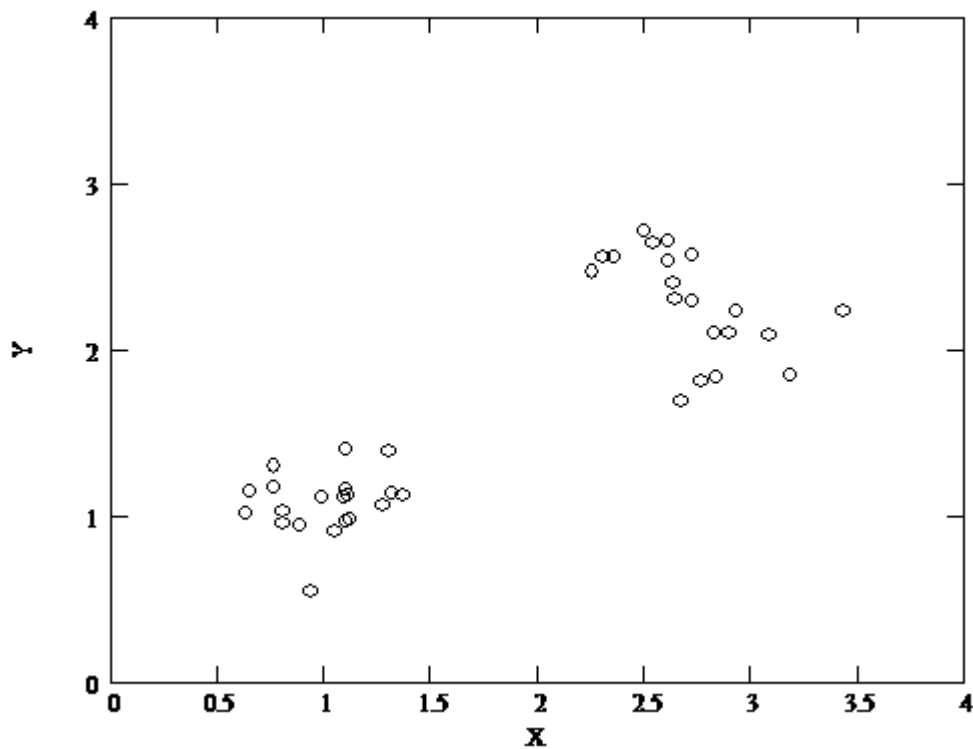
Некоторые иллюстрации к данным изменениям для случая объектов, характеризующихся двумя признаками – X и Y – приведены на рис. 1.



а) в момент $t = 10$ имеется только один кластер;



б) к моменту $t = 50$ сформировались два новых кластера



в) слияние двух новых кластеров в один к моменту $t = 100$;

Рис. 1. Динамические изменения в структуре кластеров

Первые четыре типа структурных изменений представляют собой резкие (скачкообразные) изменения кластерной структуры. Во многих приложениях

такие резкие изменения бывают связаны с недостатками поведения наблюдаемой системы, причем как можно более раннее выявление подобных изменений помогает избежать различных нежелательных последствий (например, фатальных повреждений оборудования в нефтехимическом производстве). Пятый тип структурных изменений носит непрерывный (и обычно незначительный) характер, который не всегда легко обнаружить, но который также имеет большое значение на практике.

В связи с изложенным, задачу исследования можно конкретизировать теперь таким образом:

1) требуется разработать методы обнаружения временного изменения структуры кластеров;

2) требуется разработать методы коррекции этой структуры на основе выявленных изменений.

Методы решения задач динамической кластеризации. За основу решения задач динамической кластеризации выберем один из алгоритмов статической кластеризации, а именно, алгоритм нечеткой самоорганизации C-means [10], в соответствии с которым центры кластеров c_i (при их заданном числе – K и объеме выборке N) находятся по выражениям:

$$c_i = \frac{\sum_{j=1}^N u_{ij}^p x_j}{\sum_{j=1}^N u_{ij}^p} \quad (1)$$

и

$$u_{ij} = \frac{1}{\sum_{k=1}^K \left(\frac{d_{ij}^2}{d_{kj}^2} \right)^{\frac{1}{p-1}}}, \quad (2)$$

где p - это весовой коэффициент, который принимает значения из интервала $[1, \infty)$, d_{ij} - евклидово расстояние между центром c_i и вектором x_j , т.е. $d_{ij} = \|c_i - x_j\|$, u_{ij} - степень принадлежности j -го объекта (вектора x_j) к i -му кластеру (с центром c_i), при этом справедливо

$$0 \leq u_{ij} \leq 1, \quad \sum_{i=1}^K u_{ij} = 1 \quad (3)$$

для $j = 1, 2, \dots, N$.

Выявление динамических изменений и методы соответствующих коррекций кластерной структуры рассмотрим отдельно для каждого из перечисленных типов этих изменений.

1. Выявление новых кластеров. После проведения классификации новых объектов (при известной по предыдущему временному окну кластерной структуре) в момент t необходимо определить, не привело ли появление новых объектов к появлению и новых кластеров. Представляется, что для объявления возникновения новых кластеров необходимо выполнение следующих условий:

1) наличие новых объектов с малыми степенями принадлежности ко всем существующим кластерам ("свободных" объектов);

2) достаточно большое количество N_{cb} таких объектов, существенно превышающее число существующих кластеров;

3) компактность этих объектов: они должны образовывать компактную группу.

Первое условие представляется достаточно очевидным. Проверку второго условия можно свести к проверке выполнения неравенства

$$N_{cb} \geq d_1 N_{min}, \quad (4)$$

где $d_1 \in [0,1]$ – заданная пороговая величина, $N_{min} = \min(N_1^*, N_2^*, \dots, N_K^*)$ – минимальный размер кластера, при вычислении которого учитываются только объекты с достаточно большими значениями степеней принадлежности, т.е.

$$N_i^* = |\{ \mathbf{x}_j / u_{ij} \geq d_2 \}| \quad (5)$$

(иначе говоря, N_i^* – число "хороших" объектов, надежно относимых к i -му кластеру – при превышении степеней их принадлежности к этому кластеру заданного порога $d_2 \in [0,1]$).

Соответственно, возможное количество новых кластеров K_t^H в момент t определяется соотношением

$$K_t^H = \text{int} \left\{ \frac{N_{cb}}{d_1 N_{min}} \right\}, \quad (6)$$

где $\{\cdot\}$ означает операцию определения целой части.

Проверка третьего условия может быть выполнена, исходя предложенной в статье [11] меры компактности кластеров:

$$Mk(i) = \frac{u_i^*}{V_i}, \quad i = 1, 2, \dots, K, \quad (7)$$

где u_i^* – сумма степеней принадлежности "хороших" объектов, отнесенных к i -му кластеру, т.е.

$$u_i^* = \sum_{j=1}^N u_{ij}, \quad \forall \mathbf{x}_j \in \{ \mathbf{x}_j / u_{ij} \geq d_1 \}, \quad (8)$$

а V_i – гиперобъем i -го кластера, определяемый выражениями

$$V_i = \sqrt{\det(\mathbf{F}_i)}, \quad (9)$$

$$\mathbf{F}_i = \frac{\sum_{j=1}^N u_{ij} (\mathbf{x}_j - \mathbf{c}_i)(\mathbf{x}_j - \mathbf{c}_i)^T}{\sum_{j=1}^N u_{ij}}. \quad (10)$$

Третье условие можно считать выполненным для k -го потенциально нового кластера, если справедливо

$$Mk^H(k) \geq d_3 Mk_{cp}, \quad k = 1, 2, \dots, K_t^H, \quad (11)$$

где

$$Mk_{cp} = \frac{1}{K} \sum_{i=1}^K Mk(i) \quad (12)$$

– средняя мера компактности уже существующих кластеров, $d_3 \in [0,1]$ – заданный порог.

Таким образом, необходимо вводить новые кластеры, если $K_t^H > 0$ и для некоторых k ($k = 1, 2, \dots, K_t^H$) выполняется неравенство (11).

2. Выявление сливающихся кластеров. По аналогии с предыдущим, можно сформулировать три условия, выполнение которых позволяет объявить два кластера сливающимися:

- 1) наличие объектов, имеющих высокие степени принадлежности одновременно для двух кластеров;
- 2) достаточно большое число таких объектов;
- 3) близость центров двух соответствующих кластеров.

Данные условия должны проверяться в процессе мониторинга над исследуемыми объектами, при этом задачу выявления сливающихся кластеров можно упростить, если в первую очередь рассматривать только кластеры, размеры которых в текущем временном окне увеличились по сравнению с их размерами в предыдущем таком окне, т.е. если для какого-то i -го кластера справедливо

$$N_i^*(t) > N_i^*(t-1), \quad (13)$$

то он рассматривается как претендент на сливающийся кластер.

Количественным критерием для объявления двух кластеров сливающимися может служить мера их сходства [12]:

$$I_{ij} = \frac{\sum_{k=1}^N \min(u_{ik}, u_{jk})}{\sum_{k=1}^N u_{ik}}, \quad (14)$$

где u_{ik} и u_{jk} – степени принадлежности k -го объекта к кластерам i и j соответственно. Данная мера, как можно показать, однако, не является симметричной, поэтому для выявления указанного сходства предпочтительней использовать меру

$$M_{c_{ij}} = \max(I_{ij}, I_{ji}), \quad (15)$$

в соответствии с которой кластеры объявляют сливающимися, если значение $M_{c_{ij}}$ превышает некоторый порог h (при $h = 0$ все кластеры будут считаться слившимися, при $h = 1$ слившихся кластеров никогда не будет).

Использование выражений (13)-(15) как количественных показателей сливания кластеров имеет, конечно, определенные недостатки (не учитывается форма кластеров и близость их центров), но, тем не менее, как показывают результаты моделирования, в большинстве случаев они дают правильный результат.

3. Выявление расщепляющихся кластеров. По аналогии со слиянием кластеров, причиной их расщепления является (в последнем временном окне) отнесение к какому-либо из уже существующих кластеров большого числа новых объектов, которые могут привести к неоднородности его внутренней структуры. Соответственно, в данном случае справедливо условие (13) выявления подобных (потенциально расщепляемых) кластеров. Более тонким критерием для такого расщепления является многоэкстремальность гистограмм признаков, построенных для объектов, относящихся к подобным кластерам.

Пример неоднородности, отражающий наличие двух максимумов гистограммы для признака X , приведен на рис. 2.

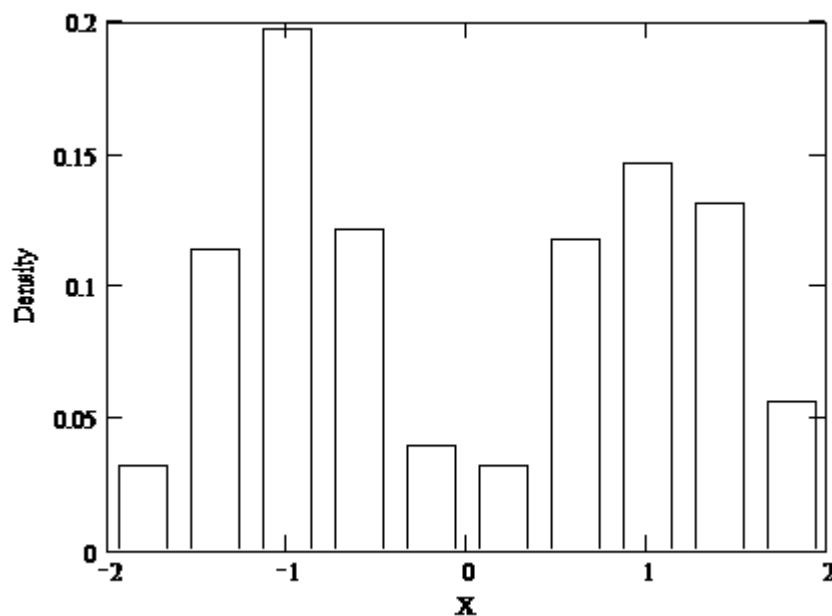


Рис. 2. Пример многоэкстремальной гистограммы для признака X

Будем полагать, что кластер может рассматриваться как расщепляемый, если отношение

$$R = \frac{f_{\max} - f_{\min}}{f_{\max}}, \quad (16)$$

где f_{\min} , f_{\max} – значения глобальных минимума и максимума гистограммы хотя бы для одного из признаков (компонентов вектора \mathbf{x}) превышают некоторое заданное пороговое значение $g \in [0, 1]$.

Заметим, что данный критерий, вообще говоря, не является наилучшим, и можно предложить другие подходы к нахождению расщепляемых кластеров, например, с использованием метода главных компонент [8] и т.д.

4. Выявление исчезающих кластеров является достаточно простой операцией: кластер объявляется исчезнувшим, если к нему не был отнесен ни один объект из последнего временного окна.

5. Выявление дрейфа центров кластеров. С течением времени новые регистрируемые объекты могут вызывать медленные изменения положений центров кластеров (дрейф центров кластеров). Такие изменения могут быть достаточно незначительными (носить непрерывный характер), но их целесообразно выявлять и отслеживать, поскольку они, в конце концов, могут быть причиной скачкообразных изменений кластерной структуры. Основываясь на результатах работы [1], выявление указанного дрейфа будем проводить, используя понятие компактности P_i нечеткого кластера i по отношению к отнесенным к нему объектам:

$$P_i = \frac{\sum_{j=1}^{N_i^*} u_{ij}^2 \|\mathbf{x}_j - \mathbf{c}_i\|^2}{u_i^*} \quad (17)$$

(значения N_i^* и u_i^* рассчитываются, соответственно, по формулам (5) и (8)).

Будем полагать, что дрейф имеет место, если компактность кластера, рассчитанная для двух последних временных окон, уменьшается, т.е. если справедливо неравенство

$$\Delta P_i(t) = P_i(t) - P_i(t-1) < q < 0, \quad (18)$$

где q – некоторое заданное пороговое значение.

При выполнении (18) принимается гипотеза о наличии дрейфа, и координаты центра кластера необходимо пересчитывать.

С учетом изложенного, общий алгоритм динамической кластеризации может быть описан следующим образом:

1) организуется мониторинг за совокупностью исследуемых объектов путем фиксации признаков этих объектов во временных окнах; объекты, зафиксированные в первом временном окне кластеризуются с помощью алгоритма нечеткой самоорганизации C-means, при этом начальное число кластеров подбирается итеративно или задается из каких-либо априорных соображений;

2) во втором и последующих временных окнах осуществляется проверка динамических изменений кластерной структуры (выявление новых кластеров, сливающихся, расщепляющихся, исчезающих кластеров, а также дрейфа центров существующих кластеров) – в соответствии с предложенными процедурами;

3) при необходимости проводится коррекция структуры кластеров – в соответствии с выявленными изменениями;

4) после выявления изменений и соответствующей коррекции – использование найденной кластерной структуры по прикладному назначению (распознаванию объектов, прогнозированию поведения системы, образованной совокупностью наблюдаемых объектов и т.п.).

Результаты вычислительных экспериментов по изложенному алгоритму (моделирование проводилось в среде Mathcad) подтвердили работоспособность изложенного подхода и его эффективность, особенно, если все кластеры имели форму гиперсфер.

Список литературы

1. Ту Дж., Гонсалес Р. Принципы распознавания образов. М.: Мир, 1978.
2. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. М.: Наука, 1974.
3. Фукунага К. Введение в статистическую теорию распознавания образов. М.: Наука, 1979.
4. Кузин Л.Т. Основы кибернетики: В 2-х т.т. Т.2. Основы кибернетических моделей. М.: Энергия, 1979.
5. Патрик Э. Основы теории распознавания образов. М.: Сов. радио, 1980.
6. Распознавание образов: состояние и перспективы / К.Верхаген, Р.Дёйн, Ф.Грун и др. М.: Радио и связь, 1985.
7. Васильев В.И. Распознающие системы. Справочник. Киев: Наукова думка, 1969.
8. Гайдышев И. Анализ и обработка данных: специальный справочник. СПб.: Питер, 2001.
9. Гимаров В.А., Дли М.И., Круглов В.В. Временная изменчивость образов // Вестник МЭИ. 2003. № 2. С. 91-95.

10. Осовский С. Нейронные сети для обработки информации. М.: финансы и статистика, 2002.
11. Gath I., Geva A.B. Unsupervised optimal Fuzzy Clustering // IEEE Trans. On Pattern Analysis and Machine Intelligence. 1989. № 11. P. 773-781.
12. Setnes M., Kaymak U. Extended Fuzzy c-Means with Volume Prototypes and Cluster Merging / Proceedings of the 6th European Conference on Intelligent Techniques and Soft Computing (EUFIT'98). Aachen, Germany, 1998. P. 1360-1364.
13. Bensaid A.M. et al. Validity-Guided (Re)Clustering with Applications to Image segmentation // IEEE Trans. On Fuzzy Systems. 1996. Vol. 4 (2). P. 112-123.