УДК 512.8
DOI: 10.20310/1810-0198-2016-21-2-459-464

КЛАСТЕРИЗАЦИЯ ОКРЕСТНОСТНОЙ СТРУКТУРЫ

© **А. М. Шмырин, Н. М. Мишачев, А. С. Косарева**

В статье определяется метрика на множестве узлов окрестностной структуры и рассматривается задача кластеризации окрестностной структуры по этой метрике или, что то же самое, задача построения *окрестностной фактор-структуры*. Определяемая метрика учитывает как связи между узлами структуры, так и имеющиеся экспериментальные данные – кортежи состояний и управлений.
Ключевые слова: окрестностная структура; метрика; кластеризация; фактор-структура.

Окрестностная структура (см. [1]) является удобным средством формализации связей между элементами моделируемой системы. Эту структуру можно рассматривать как «дискретный каркас» на основе которого можно строить математические модели следующего уровня, например линейные или билинейные. В книге [1] математические модели, надстроенные над окрестностной структурой, называются *окрестностными системами*. При моделировании систем с большим количеством элементов возникает необходимость построения фактор-структуры, содержащей меньшее количество укрупненных элементов (кластеров). Для этого нужно определить метрику на множестве элементов системы и правило преобразования связей между ними в связи между кластерами.

1. Окрестностные структуры.

Для того, чтобы корректно определить кластеризацию окрестностной структуры, т. е. переход к фактор-структуре, нам пришлось, по сравнению с [1], модифицировать основное определение и сопутствующие ему обозначения.

Окрестностной структурой мы называем взвешенный ориентированный граф с n вершинами и ребрами двух сортов, назовем их s-ребрами и u-ребрами. Эти ребра символизируют два типа связей между вершинами: связи «по состоянию» и связи «по управлению». Две вершины a_i и a_j могут быть соединены не более чем двумя s-ребрами (от a_i к a_j и от a_j к a_i) и, аналогично, не более чем двумя u-ребрами. Можно считать, что два таких s-ребра (или u-ребра) – это одно ребро с двумя ориентациями: от a_i к a_j и от a_j к a_i. Вершины графа называются *узлами* окрестностной структуры. Вес каждого ребра – это пара чисел $(\widetilde{c},\overline{c})$ из интервала $[0,1]$, таких что $\widetilde{c}+\overline{c}>0$. Мы считаем, что ребро с весом $(\widetilde{c},\overline{c})$,ведущее из узла a_i в узел a_j, описывает две связи a_i с a_j: *мягкую* связь с весом $\widetilde{c}$ и *жесткую* связь с весом $\overline{c}$. Если $\overline{c}=0$, то имеется только мягкая связь узлов, если $\widetilde{c}=0$, то имеется только жесткая связь узлов. Мягким связям будут соответствовать *вычисляемые* коэффициенты математической модели, жестким – *заданные* коэффициенты.

Окрестностью узла a_i по состоянию (управлению) называется множество всех концевых узлов всех s-ребер (u-ребер) исходящих из узла a_i. Мы будем описывать окрестностную структуру *матрицей смежностей* $[S,\widetilde{S},U,\overline{U}]$, составленной из четырех матриц порядка $n\times n$. Строка $\widetilde{S}_i=[\widetilde{s}_{i1},\ldots,\widetilde{s}_{in}]$ матрицы $\widetilde{S}$ задает веса мягких (вычисляемых) связей узла a_i по состояниям: $\widetilde{s}_{ij}\in(0,1]$ (вес связи), если узел a_j входит в окрестность узла a_i по состоянию и $\widetilde{s}_{ij}=0$ если узел a_j не входит в эту окрестность. Аналогично определяются строки

остальных матриц $\overline{S}, \widetilde{U}, \overline{U}$. Обычно $\widetilde{s} + \overline{s}_{ij} \neq 0$ и $\widetilde{u} + \overline{u}_{ij} \neq 0$, т. е. каждый узел входит в свои окрестности по состояниям и по управлениям (этому соответствует петли графа), но, вообще говоря, мы не исключаем случай, когда $\widetilde{s} + \overline{s}_{ij} \neq 0$ или $\widetilde{u} + \overline{u}_{ij} \neq 0$.

З а м е ч а н и е. В книге [1], если использовать нашу терминологию, все ребра графа являются одинарными, т. е. описывают либо мягкую либо жесткую связь. В наших обозначениях это соответствует тому, что одно из двух чисел $\widetilde{c}$ или $\overline{c}$ является нулем. Наличие весовых множителей для связей i-го узла с остальными узлами в книге [1] интерпретируется как *нечеткость* соответствующих окрестностей.

2. Кортежи данных.

Статистику наблюдений реальной системы, описываемой окрестностной структурой, запишем в виде $K \times 2n$-матрицы данных $[X_1, \ldots, X_n, V_1, \ldots, V_n]$, где K – количество наблюдений системы. Строки $[X^k, V^k]$ матрицы данных являются результатами k-го наблюдения (эксперимента): $X^k = [x_1^k, \ldots, x_n^k]$ и $V^k = [v_1^k, \ldots, v_n^k]$ – это состояния и управления в узлах структуры, наблюдавшиеся в k-м эксперименте. Мы будем называть эти строки *кортежами данных*. Столбцы матрицы данных X_i или V_i – это векторы данных всех наблюдений соответствующего узла. В общем случае состояние или управление в узлах структуры могут быть многопараметрическими и потому элементы матрицы данных могут быть векторами (вообще говоря, разных размерностей). В простейшем случае, который мы далее и будем рассматривать, эти элементы являются *числами*. Итак, мы будем считать состояния и управления в узлах структуры *скалярами*. Это ограничение не является принципиальным и все дальнейшее нетрудно переписать для случая векторных состояний и управлений.

З а м е ч а н и е. Мы будем считать, что единицы измерения состояний и управлений в каждом узле выбраны таким образом, что выборки $\{x_i^k\}, k = 1, \ldots, K$ и $\{v_i^k\}, k = 1, \ldots, K$ (здесь i – номер узла) центрированы и нормированы. Переход к таким единицам измерения всегда возможен, но, как правило, требует пересчета заданных («жестких») коэффициентов математической модели.

3. Окрестностные системы.

Простейшей математической моделью, ассоциированной с окрестностной структурой (и описывающей работу соответствующей реальной системы), является *линейная симметричная окрестностная система (модель)*. В рамках нашего подхода к определению метрики и дальнейшей кластеризации окрестностной структуры можно было бы рассматривать и более сложные модели, например билинейные (см [1]), но мы для простоты ограничимся здесь линейным случаем.

Мы будем использовать далее обозначение $A \circ B$ для поэлементного произведения (произведения Адамара) матриц A и B. Линейная симметричная окрестностная система имеет вид

$$(\widetilde{\Omega} \circ \widetilde{S})X + (\widetilde{T} \circ \widetilde{U})V = (\overline{\Omega} \circ \overline{S})X + (\overline{T} \circ \overline{U})V, \tag{1}$$

где $[S, \widetilde{S}, U, \overline{U}]$ – определенная выше $n \times 4n$ матрица смежности окрестностной структуры, $\widetilde{\Omega} = \{\widetilde{\omega}_{iq}\}$ и $\widetilde{T} = \{\widetilde{\tau}_{iq}\}$ – $n \times n$-матрицы вычисляемых коэффициентов системы, $\overline{\Omega} = \{\overline{\omega}_{iq}\}$ и $\overline{T} = \{\overline{\tau}_{iq}\}$ – $n \times n$-матрицы заданных коэффициентов системы, X и V – n-кортежи состояний и управлений, являющиеся неизвестными окрестностной системы. Вычисляемые коэффициенты системы соответствуют мягким связям, заданные коэффициенты соответствуют жестким связям. Напомним, что мы считаем состояния и управления в узлах структуры скалярами, и потому элементы матриц $\widetilde{\Omega}, \widetilde{T}, \overline{\Omega}, \overline{T}$ являются скалярами. В общем случае, когда состояния и управления являются векторами, элементы матриц $\widetilde{\Omega}, \widetilde{T}, \overline{\Omega}, \overline{T}$ будут матрицами.

З а м е ч а н и я.

1. Отличие системы (1) от аналогичной системы в книге [1] состоит в том, что узел a_j, участвующий в записи i-го уравнения системы, может порождать слагаемое и в левой и в правой частях уравнения. Слагаемое в левой части содержит синтезируемый (вычисляемый) коэффициент и соответствует мягкой связи узла a_i с узлом a_j. Слагаемое в правой части содержит заданный коэффициент и соответствует жесткой связи узла a_i с узлом a_j. В книге [1] узел a_j (в уравнении для узла a_i) мог порождать слагаемое только в одной из частей.

2. Система (1) записана в виде, удобном для дальнейшей идентификации. После идентификации синтезированные и заданные коэффициенты при одинаковых неизвестных x_i (или u_i) можно объединить.

3. Не следует путать *коэффициенты* $\widetilde{\Omega}, \widetilde{T}, \overline{\Omega}, \overline{T}$ и *веса* $[S, \widetilde{S}, U, \overline{U}]$. В окрестностной системе (1) *заранее заданы* матрицы $\overline{\Omega}, \overline{T}$ (заданные коэффициенты) и $[S, \widetilde{S}, U, \overline{U}]$ (веса). Матрицы $\widetilde{\Omega}, \widetilde{T}$ – синтезируемая часть системы.

4. Вообще говоря, левая часть системы может еще содержать неизвестный вектор $\widetilde{C}$, а правая – заданный вектор $\overline{C}$. Но такую систему всегда можно привести к виду (1). Для этого нужно формально ввести в систему два дополнительных узла, которые входят в окрестности всех узлов по состояниям, а их собственные окрестности – пустые. Все остальные узлы нужно считать мягко связанными с первым формальным узлом и жестко – со вторым. При этом соответствующий столбец матрицы $\widetilde{\Omega}$ (столбец вычисляемых коэффициентов) равен $\widetilde{C}$ и соответствующий столбец матрицы $\overline{\Omega}$ (столбец заданных коэффициентов) равен $\overline{C}$. Все данные по состояниям формальных узлов – единицы ($x^k_{n+1} = x^k_{n+2} = 1$) и все данные по управлениям – нули ($v^k_{n+1} = v^k_{n+2} = 0$). Такие формальные узлы не порождают дополнительные уравнения системы.

В координатной записи линейная симметричная окрестностная система имеет вид

$$\sum_{q=1}^{n}[(\widetilde{\omega}_{iq}\widetilde{s}_{iq})\ x_j + (\widetilde{\tau}_{iq}\widetilde{u}_{iq})\ v_q] = \sum_{q=1}^{n}[(\overline{\omega}_{iq}\overline{s}_{iq})\ x_q + (\overline{\tau}_{iq}\overline{u}_{iq})\ v_q], \tag{2}$$

где $\widetilde{\omega}_{iq} = 0$ если $\widetilde{s}_{iq} = 0$, $\widetilde{\tau}_{iq} = 0$ если $\widetilde{u}_{iq} = 0$, $\overline{\omega}_{iq} = 0$ если $\overline{s}_{iq} = 0$ и $\overline{\tau}_{iq} = 0$ если $\overline{u}_{iq} = 0$.

4. Синтез (идентификация) линейной окрестностной системы.

Рассмотрим задачу нахождения коэффициентов $\widetilde{\omega}_{iq}$ и $\widetilde{\tau}_{iq}$ окрестностной системы (2) по экспериментальным данным – кортежам $[X^k, V^k], k = 1, \ldots, K$. Подстановка кортежей данных в уравнении окрестностной модели (2) приводит к системе nK линейных уравнений для неизвестных коэффициентов $\widetilde{\Omega}$ и $\widetilde{T}$ окрестностной модели. Вообще говоря, в окрестностной модели могут быть дополнительные условия, связывающие искомые коэффициенты разных уравнений модели, например условие симметричности или антисимметричности матриц $\widetilde{\Omega}$ и $\widetilde{T}$. Если таких условий нет, то система уравнений для нахождения коэффициентов окрестностной модели распадается на n систем, по одной системе для каждого узла модели:

$$\sum_{q=1}^{n}[\widetilde{\omega}_{iq}(\widetilde{s}_{iq}x^k_j) + \widetilde{\tau}_{iq}(\widetilde{u}_{iq}v^k_q)] = b^k_i, \tag{3}$$

где $b^k_i = \sum_{q=1}^{n}[\overline{\omega}_{iq}(\overline{s}_{iq}x^k_q) + \overline{\tau}_{iq}(\overline{u}_{iq}v^k_q)]$. Подчеркнем, что неизвестными в системе (3) являются искомые коэффициенты $\widetilde{\omega}_{iq}$ и $\widetilde{\tau}_{iq}$ i-го уравнения окрестностной модели, а коэффициентами – числа $\widetilde{s}_{iq}x^k_j$ и $\widetilde{u}_{iq}v^k_q$ Обозначим через r_i количество (ненулевых) неизвестных $\widetilde{\omega}_{iq}$ и $\widetilde{\tau}_{iq}$ в системе (4). В невырожденном случае, т. е. когда ранг матрицы системы максимален, коэффициенты $\widetilde{\omega}_{iq}$ и $\widetilde{\tau}_{iq}$ i можно найти как:

а) нормальное решение системы (решение с минимальной нормой), если $K < r_i$;

b) решение определенной системы при $K = r_i$;
c) псевдорешение переопределенной системы (вектор неизвестных, минимизирующий норму вектора невязки), если $K > r_i$.
Если ранг матрицы системы не максимален, то искомые коэффициенты можно найти как псевдорешение системы: минимальный по норме вектор, минимизирующий невязку.

5. Метрика на множестве узлов окрестностной структуры.

Для каждого узла a_i окрестностной структуры определим *взвешенную матрицу данных*

$$D_i = [\widetilde{s}_{i1}X_1, \ldots, \widetilde{s}_{in}X_n, \overline{s}_{i1}X_1, \ldots, \overline{s}_{in}X_n, \widetilde{u}_{i1}V_1, \ldots, \widetilde{u}_{in}V_n, \overline{u}_{i1}V_1, \ldots, \overline{u}_{in}V_n,] \tag{4}$$

этого узла. Ненулевыми столбцами матрицы D_i являются взвешенные столбцы матрицы данных $[X_1, \ldots, X_n, V_1, \ldots, V_n]$, соответствующие всем узлам из окрестностей узла a_i. Зададим метрику на множестве узлов окрестностной структуры формулой

$$r_{ij} = \sqrt{||D_i - D_j||^2 + ||X_i - X_j||^2 + ||V_i - V_j||^2}. \tag{5}$$

Первое слагаемое под корнем оценивает сходство окрестностей узлов *с учетом экспериментальных данных и взвешенных связей*. Узел, входящий в окрестности по состояниям (управлениям) узлов a_i и a_j, или не входящий ни в одну из этих окрестностей, даст нулевой вклад в $||D_i - D_j||$. Узел a_q, входящий только в одну из этих окрестностей, даст вклад, пропорциональный взвешенным данным наблюдений в a_q. Сумма двух следующих слагаемых под корнем – это квадрат расстояния между векторами данных узлов a_i и a_j. Таким образом, определенная метрика вычисляет близость узлов по данным наблюдений и по их взвешенным связям (окрестностям) в окрестностной системе.

З а м е ч а н и я.

1. Вместо взвешенной матрицы данных D_i в формуле (5) можно использовать нормированные строки весов $\widehat{D}_i = C_i/||C_i||$, где

$$C_i = [\widetilde{s}_{i1}, \ldots, \widetilde{s}_{in}, \overline{s}_{i1}, \ldots, \overline{s}_{in}, \widetilde{u}_{i1}, \ldots, \widetilde{u}_{in}, \overline{u}_{i1}, \ldots, \overline{u}_{in},]. \tag{6}$$

Определенная таким образом метрика $\widehat{r}_{ij}$ в некоторых случаях может оказаться удобнее чем r_{ij}. Например, метрика $\widehat{r}_{ij}$ оценивает расстояния между узлами окрестностями структуры и в том случае, когда данных нет.

2. Другой подход к определению меры близости узлов окрестностной структуры по связям и данным обсуждался в [2]. Заметим, что определенная в [2] мера d_{ij}, вообще говоря, не является метрикой (не выполнено неравенство треугольника).

6. Окрестностные фактор-структуры.

Используя определенную выше метрику r_{ij} (или $\widehat{r}_{ij}$), можно разбить узлы окрестностной структуры на кластеры с помощью какого-либо алгоритма кластеризации. В любом случае, далее нужно пересчитать матрицу смежности окрестностной структуры и матрицу данных. В качестве вектора данных (по управлению или состоянию) кластера можно взять среднее арифметическое векторов данных всех вошедших в кластер узлов. Далее, окрестность кластера I по состояниям (управлениям) будем считать состоящей из всех кластеров J, таких что хотя бы один из узлов кластера J входит в окрестность хотя бы одного из узлов кластера I. Вес соответствующей мягкой связи кластера I с кластером J положим равным среднему арифметическому весов всех мягких связей узлов из кластера I с узлами из кластера J.

Аналогично, вес соответствующей жесткой связи кластера I с кластером J положим равным среднему арифметическому весов всех жестких связей узлов из кластера I с узлами из кластера J. Заданные коэффициенты жестких связей кластеров можно определить как средние арифметические соответствующих заданных коэффициентов жестких связей узлов из I с узлами из J.

СПИСОК ЛИТЕРАТУРЫ

1. *Блюмин С.Л., Шмырин А.М.* Окрестностные системы. Липецк: ЛЭГИ, 2005.

2. *Shmyrin, A.M., Kosareva, A.S.* The measure of similarity in solving the problem of clustering neighborhood structures // Modern informatization problems in the technological and telecommunication systems analysis and synthesis: Proceedings of the XXI-th International Open Science Conference (Yelm, WA, USA, Januar 2016), 2016. P. 341–346.

БЛАГОДАРНОСТИ: Работа поддержана грантом РФФИ (код проекта 16-07-00854 а).

Поступила в редакцию 21 марта 2016 г.

Шмырин Анатолий Михайлович, Липецкий государственный технический университет, г. Липецк, Российская Федерация, доктор технических наук, профессор, заведующий кафедрой высшей математики, e-mail: amsh@lipetsk.ru

Мишачёв Николай Михайлович, Липецкий государственный технический университет, г. Липецк, Российская Федерация, кандидат физико-математических наук, доцент кафедры высшей математики, e-mail: nmish@lipetsk.ru

Косарева Анастасия Сергеевна, Липецкий государственный технический университет, г. Липецк, Российская Федерация, аспирант, кафедра высшей математики, e-mail: kosarewanastya@yandex.ru

UDC 512.8
DOI: 10.20310/1810-0198-2016-21-2-459-464

CLUSTERING OF NEIGHBORHOOD STRUCTURE

© **A. M. Shmyrin, N. M. Mishachev, A. C. Kosareva**

We define a metric on the set of nodes of neighborhood structure and consider the clustering problem for the structure with respect to the metric, or, what is the same, the problem of constructing the factor-structure. Determined metric takes into account both the connections between the nodes of the structure and the experimental data at the nodes.
Key words: neighborhood structures; metrics; clustering; factor-structure.

ACKNOWLEDGEMENTS: The work is supported by the Russian Fund for Basic Research (project № 16-07-00854 a)

REFERENCES

1. *Blyumin S.L., Shmyrin A.M.* Neighborhood system. Lipetsk: LEGI, 2005.

2. *Shmyrin, A.M., Kosareva, A.S.* The measure of similarity in solving the problem of clustering neighborhood structures // Modern informatization problems in the technological and telecommunication systems analysis and synthesis: Proceedings of the XXI-th International Open Science Conference (Yelm, WA, USA, Januar 2016), 2016. P. 341–346.

Received 21 March 2016.

Shmyrin Anatoliy Mikhailovich, Lipetsk State Technical University, Lipetsk, the Russian Federation, Doctor of Techniques, Professor, the Head of the Higher Mathematics Department, e-mail: amsh@lipetsk.ru

Mishachev Nikolay Mikhailovich, Lipetsk State Technical University, Lipetsk, the Russian Federation, Candidate of Physics and Mathematics, Associate Professor of the Higher Mathematics Department, e-mail: nmish@lipetsk.ru

Kosareva Anastasia Sergeevna, Lipetsk State Technical University, Lipetsk, the Russian Federation, graduate student of the Higher Mathematics Department, e-mail: kosarewanastya@yandex.ru