УДК 004.048

КОНЦЕПТУАЛЬНЫЕ ОСНОВЫ И МЕТОДОЛОГИЯ СОЗДАНИЯ ИНДУКТИВНОЙ ТЕХНОЛОГИИ ОБЪЕКТИВНОЙ КЛАСТЕРИЗАЦИИ

С.А. Бабичев

Университет Яна Евангелиста Пуркинье в Усти на Лабе,
Чешская республика

sergii.babichev@ujep.cz

В статье представлены теоретические разработки по созданию методологии объективной кластеризации объектов сложной природы на основе методов индуктивного моделирования сложных систем. Разработана архитектура индуктивной технологии объективной кластеризации в виде подробной схемы пошаговой реализации процедуры индуктивного моделирования процесса кластеризации объектов сложной природы.
Ключевые слова. Индуктивное моделирование, объективная кластеризация, высокоразмерные данные.

У статті представлено теоретичні розробки по створенню методології об'єктивної кластеризації об'єктів складної природи на основі методів індуктивного моделювання складних систем. Розроблено архітектуру індуктивної технології об'єктивної кластеризації у вигляді детальної схеми покрокової реалізації процедури індуктивного моделювання процесу кластеризації об'єктів складної природи.
Ключові слова. Індуктивне моделювання, об'єктивна кластеризація, високорозмірні дані.

The paper presents the theoretical developments to create a methodology of objective clustering of complex nature objects based on the complex systems inductive modeling methods. The architecture of the objective clustering inductive modeling as a detailed scheme of step by step implementation of procedures of inductive modeling of the objects complex nature clustering is developed.
Keywords. Inductive modeling, objective clustering, high dimensional data.

Постановка проблемы. В настоящее время существует большое количество разнообразных алгоритмов кластеризации, каждый из которых имеет свои преимущества и недостатки и ориентирован на определенный тип данных. Традиционные алгоритмы кластеризации в случае обработки высокоразмерных данных сложной природы малоэффективны по причине высокой погрешности получения конечного результата. Под высокоразмерными будем понимать данные, размерность признакового пространства которых равна или больше количеству исследуемых объектов. Такими данными являются профили экспрессий генов нуклеотидов ДНК, энцефалограммы органов биологического организма, хроматограммы наркотических веществ и т.д. Особенностями исследуемых данных кроме

высокой размерности является уровень и специфичность шумовой компоненты, обусловленной биологическими процессами, протекающими в исследуемом объекте, и несовершенством системы их создания и формирования для последующей обработки. Одним из главных недостатков существующих алгоритмов кластеризации является их субъективизм, т.е. получение хороших результатов кластеризации объектов на одном множестве не гарантирует получение подобных результатов на другом аналогичном множестве. Повысить объективность кластеризации можно за счет разработки гибридных моделей на основе методов индуктивного моделирования сложных систем, являющимся логическим продолжением метода группового учета аргументов (МГУА) [1], реализация которого предполагает параллельную обработку данных на двух равномощных подмножествах, при этом окончательное решение по группировке объектов принимается на основании внешнего критерия баланса результатов кластеризации на двух подмножествах.

Анализ современных достижений и публикаций. Вопросы создания индуктивных методов объективной самоорганизации моделей сложных систем изложены в [2-4] и получили дальнейшее развитие в [5-8]. Авторами представлены исследования по реализации принципов индуктивного моделирования в области создания систем объективной самоорганизации объектов сложной природы на основе метода группового учета аргументов. В работах [9-12] представлены исследования по использованию методов индуктивного моделирования в индуктивных технологиях создания систем информационно-аналитических исследований в процессе анализа информации различной природы. Для оценки качества обработки информации двумя группами независимых экспертов автором вводится понятие критериев релевантности, корелевантности и баланса. Под релевантностью понимается уровень соответствия получаемых результатов целям поставленной задачи по мнению каждой группы в отдельности. Под корелевантностью понимается уровень взаимного соответствия результатов, полученных двумя независимыми группами экспертов. Критерий баланса объединяет обе группы критериев и позволяет выделить оптимальное решение с точки зрения как критерия релевантности, так и критерия корелевантности. Однако следует отметить, что исследования авторов ориентированы преимущественно на данные с невысокой размерностью признакового пространства, при этом индуктивным моделям на основе перебора кластеризаций с целью их самоорганизации с использованием внешних критериев баланса оценки качества кластеризации на равномощных подмножествах уделяется недостаточное внимание.

К нерешенным частям общей проблемы следует отнести отсутствие индуктивной технологии объективной кластеризации объектов сложной природы на основе принципов индуктивного моделирования сложных систем.

Целью статьи является разработка концептуальных основ и методологии создания индуктивной технологии объективной кластеризации высокоразмерных данных сложной природы.

Изложение основного материала

Пусть исходная выборка данных представлена в виде матрицы $A=\{x_{ij}\}, i=1,\ldots,n;\ j=1,..m$, где n – количество строк или наблюдаемых объектов, m – количество признаков, характеризующих объект. Задача кластеризации сводится к разбиению множества объектов на непустые подмножества непересекающихся кластеров, при этом плоскость, разделяющая кластеры может принимать произвольную форму [5]:

$$K=\{K_s\}, s=1,\ldots,k;\ K_1\cup K_2\cup\ldots\cup K_k=A;$$
$$K_i\cap K_j=\varnothing, i\neq j;\ i,j=1,\ldots,k$$

где k – количество кластеров. Индуктивная модель объективной кластеризации предполагает последовательный перебор кластеризаций с целью выбора наилучшей [4,5]. Пусть W – множество всех допустимых кластеризаций на заданном множестве A. Наилучшей (оптимальной) по критерию качества $QC(K)$ является кластеризация, для которой:

$$K_{opt}=\arg\min_{K\subseteq W} QC(K) \text{ или } K_{opt}=\arg\max_{K\subseteq W} QC(K) \quad (1)$$

Кластеризация $K_{opt}\subseteq W$ является объективной, если по количеству кластеров, характеру распределения объектов в соответствующих кластерах и количеству несоответствий она меньше всего отличается от экспертной [5]. Технология создания индуктивной модели объективной кластеризации предполагает наличие следующих этапов:

1. определение функции аффинности исследуемых объектов, т.е. метрики, определяющей степень сходства объектов в m-мерном пространстве признаков;

2. разработка алгоритма разбиения исходного множества исследуемых объектов на два равномощных подмножества. Под равномощными в данном случае понимаются подмножества, содержащие одинаковые количества попарно близких объектов;

3. задание способа образования кластеров (сортировка, перегруппировка, объединение, разделение, и т.д.);

4. задание критерия оценки качества кластеризации QC, как меры сходства кластеров в различных кластеризациях;

5. организация движения к max, min или оптимальному значению критерия QC оценки качества кластеризации;

6. задание способа фиксации объективной кластеризации, соответствующей экстремальному или оптимальному значению критерия оценки качества кластеризации.

На рис. 1 изображен характер взаимодействия различных модулей индуктивной модели объективной кластеризации. Выбор функции аффинности в процессе оценки степени близости объектов или кластеров определяется характером признаков, характеризующих исследуемые объекты. В случае высокоразмерных данных сравниваются профили векторов признаков, при этом в качестве меры сходства могут использоваться различные меры оценки степени близости векторов в m-мерном пространстве признаков.

Рис. 1. Структурная схема взаимодействия модулей индуктивной модели объективной кластеризации

Способ формирования кластеров определяет алгоритм кластеризации, используемый в индуктивной модели для параллельной группировки объектов в двух равномощных подмножествах. Характер формирования двух равномощных подмножеств определяется выбранной мерой сходства объектов, которая в свою очередь, зависит от свойств признакового пространства исследуемых объектов. Для выбора объективной кластеризации необходимо на раннем этапе определить внешние и внутренние критерии, экстремальное или оптимальное значение которых в процессе перебора кластеризаций позволит зафиксировать объективную кластеризацию на исследуемой выборке данных.

Как известно [1-12], основу методологии индуктивного моделирования сложных систем составляют три фундаментальных принципа, позаимствованные с различных научных направлений и позволившие создать целостную, органично-взаимосвязанную теорию:

1. принцип эвристической самоорганизации, т.е. последовательного перебора различных моделей-претендентов с целью выбора наилучших моделей по априори определенным внешним критериям баланса;

2. принцип внешнего дополнения, основная идея которого состоит в необходимости для объективной верификации модели использования «свежей информации». При этом настройка модели и оценка качества её работы осуществляются на различных данных;

3. принцип неокончательности решений, идея которого заключается в генерации не одного, а определенного множества промежуточных результатов с последующим выбором из них наилучшего решения.

Реализация данных принципов в адаптированном варианте создает условия для создания методологии построения индуктивной модели объективной кластеризации сложных данных.

Принцип эвристической самоорганизации моделей. Индуктивная модель объективной кластеризации предполагает последовательный перебор кластеризаций на двух равномощных выборках, при этом на каждом шаге оценивается результат кластеризации посредствам расчета внешнего критерия баланса, определяющего разницу результатов кластеризаций объектов на двух подмножествах. Модель самоорганизуется таким образом, что, в зависимости от типа используемого алгоритма и мер сходства объектов и кластеров, лучшие кластеризации соответствуют экстремуму данного критерия или его оптимальному значению, соответствующему наиболее устойчивой с точки зрения данного критерия кластеризации. В процессе перебора кластеризаций возможна ситуация, когда значение внешнего критерия баланса имеет несколько локальных экстремумов, соответствующих различным кластеризациям объектов. Данное явление имеет место в случае иерархической кластеризации, когда в процессе группировки объектов или их последовательного разделения, кластеризации на двух подмножествах оказываются достаточно схожими, что приводит к возникновению локального минимума на данном уровне иерархии. В данном случае выбор оптимальной кластеризации определяется целями поставленной задачи, поскольку каждую кластеризацию, соответствующую экстремуму внешнего критерия баланса, можно считать объективной, а выбор определяется требуемым уровнем детализации разбиения или группировки объектов.

Исходя из вышесказанного можно сделать вывод, что индуктивная технология объективной кластеризации должна строиться в соответствии и по аналогии со схемами многорядных алгоритмов МГУА с использованием внешних критериев баланса для оценки степени объективности кластеризации в процессе работы модели.

Принцип конкуренции. Принцип внешнего дополнения в модели группового учета аргументов (МГУА) предполагает использование «свежей информации» для объективной верификации модели и выбора наилучшей

модели в процессе многорядной индуктивной процедуры синтеза оптимальной модели. В рамках индуктивной модели объективной кластеризации реализация данного принципа предполагает наличие двух равномощных подмножеств, содержащих одинаковое число попарно близких с точки зрения значений признаков объектов. В процессе работы алгоритма кластеризация осуществляется параллельно на двух подмножествах с последовательным сравнением результатов кластеризации по выбранным внешним критериям баланса. Идея алгоритма разделения исходного множества Ω исследуемых объектов на два равномощных подмножества Ω^A и Ω^B изложена в [4] и получила дальнейшее развитие в [5]. Реализация алгоритма предполагает наличие следующих этапов:

1. расчет $n\cdot(n-1)/2$ попарных расстояний между объектами в исходной выборке данных. Результатом данного шага является треугольная матрица расстояний. Если n – количество исследуемых объектов, то данная матрица имеет вид:

$$D=\begin{cases} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ & 0 & d_{23} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ & & & 0 & d_{n-1,n} \\ & & & & 0 \end{cases}; \qquad (2)$$

2. выделение пары объектов X_s, X_p, расстояние между которыми минимально:

$$d(X_s, X_p)=\min_{i,j} d(X_i, X_j);$$

3. распределение объекта X_s в подмножество Ω^A, а объекта X_p в подмножество Ω^B;

4. повторение шагов 2–3 для оставшихся объектов. Если количество объектов нечетное, последний объект распределяется в оба подмножества.

Реализация принципа конкуренции в рамках индуктивной модели объективной кластеризации предполагает наличие внешнего критерия баланса между результатами кластеризации на двух равномощных подмножествах, который может быть комплексным и учитывать как распределение кластеров в полученных кластеризациях, так и расположение объектов в соответствующих кластерах в различных кластеризациях.

Принцип неокончательности решений. Применительно к индуктивной модели объективной кластеризации реализация данного принципа предполагает фиксацию кластеризаций, соответствующих локальным минимумам или максимумам внешнего критерия баланса на различных уровнях иерархического

дерева. Каждый локальный экстремум соответствует объективной кластеризации при определенной степени детализации. Окончательный выбор, и как следствие, фиксация полученной кластеризации определяется целями поставленной задачи на данном этапе её решения.

Критерии релевантности в индуктивной технологии объективной кластеризации. В процессе реализации индуктивной технологии объективной кластеризации возникает необходимость оценки качества кластеризации на отдельных равномощных подмножествах данных, при этом отдельные оценки при использовании различных алгоритмов и различных оценочных функций для одних и тех же данных могут не совпадать друг с другом. Таким образом, возникает необходимость в оценке соответствия результатов моделирования целям поставленной задачи. Критерии оценки такого соответствия называются критериями релевантности (relevance – соответствие, адекватность). Данный термин широко используется в настоящее время в теории принятия решений, как уровень соответствия запрос – отклик. В [9,10] под релевантностью понимается степень соответствия результата информационно-аналитического исследования целям поставленной задачи. В индуктивной технологии объективной кластеризации под релевантностью будем понимать количественную меру адекватности группировки объектов на отдельных выборках данных. Иначе говоря, критерий релевантности представляет собой внутренний критерий оценки качества кластеризации данных. В реальных условиях в большинстве случаев количество кластеров неизвестно, поэтому в процессе работы алгоритма кластеризации выделяются наилучшие решения, соответствующие экстремумам используемых внутренних критериев. Очевидно, что хорошая кластеризация соответствует высокой разделяющей способности различных кластеров и высокой плотности сосредоточения объектов внутри кластеров. Поэтому внутренний критерий оценки качества кластеризации должен включать две составляющие: сумму квадратов отклонений положений объектов относительно соответствующих центроидов внутри кластеров (SSW) и сумму квадратов отклонений центроидов кластеров относительно общего центра масс между кластерами (SSB). Формулы расчета данных составляющих критерия релевантности можно записать следующим образом:

$$SSW = \sum_{j=1}^{K}\sum_{i=1}^{N_j}\left\|x_i^j - c_j\right\|^2 \tag{3}$$

$$SSB = \sum_{j=1}^{K} N_j \left\|c_j - \overline{C}\right\|^2 \tag{4}$$

где K – количество кластеров, N_j – количество объектов в кластере j, c_j – центроид кластера j: $c_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_i^j$, x_i^j – і-й объект в j-м кластере, $\overline{C}$ – общий

центроид исследуемых объектов, $\|\cdot\|$ – евклидова норма. Суммарное среднеквадратичное отклонение исследуемых объектов относительно общего центра масс SST можно определить по формуле:

$$SST = \sum_{i=1}^{N} \left\| x_i - \overline{C} \right\|^2 \tag{5}$$

где N – общее количество исследуемых объектов.

В [13-17] дано описание и проведен сравнительный анализ внутренних критериев оценки качества кластеризации при использовании различных комбинаций и разновидностей мер (3) и (4). В качестве основных критериев можно выделить следующие:

1. Calinski-Harabasz [17]:

$$CH = \frac{SSB_K \cdot (N-K)}{SSW_K \cdot (K-1)}. \tag{6}$$

N – количество объектов, K – количество кластеров.

2. Ball and Hall [18]:

$$BH = \frac{SSW_K}{K}. \tag{7}$$

3. Xu-index [19]:

$$Xu = D \cdot \log_2\left(SSW_K \Big/ \left(D \cdot N^2\right)\right) + \log_2 K. \tag{8}$$

Здесь D – размерность признакового пространства исследуемых объектов.

4. Krzanowski-Lai [20]:

$$KL = \frac{|diff_K|}{|diff_{K+1}|}, \tag{9}$$

$$diff_K = (K-1)^{\frac{2}{D}} SSW_{K-1} - K^{\frac{2}{D}} SSW_K.$$

5. Hartigan [21]:

$$H = \left(\frac{SSW_K}{SSW_{K+1}} - 1 \right)(N-K-1) \quad \text{или} \quad H = \log_2 \frac{SSB_K}{SSW_K} \tag{10}$$

6. Dunn’s index [14]:

$$Dunn = \frac{\min\limits_{i=1}^{K} \min\limits_{j=i+1}^{K} d(c_i, c_j)}{\max\limits_{k=1}^{K} diam(c_k)}, \tag{11}$$

$$d(c_i, c_j) = \min_{x \in c_i, x' \in c_j} \|x - x'\|^2,$$

$$diam(c_k) = \max_{x, x' \in c_k} \|x - x'\|^2.$$

7. Davies and Bouldin [15]:

$$DBI = \frac{1}{K} \sum_{i=1}^{K} R_i \ , \tag{12}$$

$$R_i = \max_{i, j=1,\ldots,k} R_{ij}, \ R_{ij} = \frac{S_i + S_j}{d_{ij}}, i \neq j,$$

$$S_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \|x_j - c_i\|^2, \ d_{ij} = \|c_i - c_j\|^2.$$

8. Отношение квадратов отклонений (R-squared) RS [22]:

$$RS_K = \frac{SSB_K}{SST_K} = \frac{\sum_{j=1}^{K} N_j \|c_j - \overline{C}\|^2}{\sum_{i=1}^{N} \|x_i - \overline{C}\|^2} \ . \tag{13}$$

9. Комплексный критерий на основе меры компактности объектов внутри кластеров и разделимости кластеров между собой (average scattering for clusters and total separation between clusters) SD [23]:

$$CD_K = w \cdot Scat(K) + Dis(K), \tag{14}$$

где $w-$ весовой коэффициент, равный $Dis(K_{\max})$, где $K_{\max}$ представляет максимальное количество получаемых кластеров. Термы в формуле (14) представляют собой среднее распределение объектов внутри кластеров $(Scat(K))$ и уровень суммарного различия кластеров между собой $(Dis(K))$:

$$Scat(K) = \frac{1}{K} \sum_{i=1}^{K} \|\sigma(c_i)\| / \|\sigma(C)\| \ , \tag{15}$$

$$Dis(K) = \frac{D_{\max}}{D_{\min}} \sum_{i=1}^{K-1} \left(\sum_{j=i+1}^{K} \|c_i - c_j\| \right)^{-1} , \tag{16}$$

где $D_{\max} = \max \|c_i - c_j\|$, $D_{\min} = \min \|c_i - c_j\|$, $\forall i, j = 1,\ldots,K$ представляют собой максимальное и минимальное расстояния между центрами кластеров соответственно. Дисперсия объектов в кластере рассчитывается следующим образом:

$$\sigma(c_i)=\frac{1}{n_i}\sum_{k=1}^{n_i}(x_i-c_i)^2. \tag{17}$$

Аналогичным способом рассчитывается дисперсия всех объектов выборки относительно их общего центра масс $\sigma(C)$.

10. Критерий на основе плотности распределения объектов и кластеров *S_Dbw* [23,24]:

$$S\_Dbw = Scat(K)+Dens\_bw(K), \tag{18}$$

где $Scat(K)$ определяет средний разброс объектов внутри кластеров (15), а второй терм в формуле (18) определяет меру различия кластеров:

$$Dens\_bw(K)=\frac{1}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{density(c_{ij})}{\max\left(density(c_i),density(c_j)\right)}, \tag{19}$$

где плотность распределения объектов по отношению к соответствующему центру определяется из условия:

$$density(c)=\sum_{i=1}^{N}f(x_i,c),\quad f(x_i,c)=\begin{cases}0, & if\ d(x_i,c)>stdev\\ 1, & otherwise\end{cases}, \tag{20}$$

$$stdev=\frac{1}{K}\left(\sum_{i=1}^{K}\left\|\sigma(c_i)\right\|\right)^{\frac{1}{2}}.$$

11. Коэффициент разбиения (partition coefficient) (PC) [25]:

$$PC=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}p(x_i)_j^2, \tag{21}$$

где $p(x_i)_j$ – степень принадлежности объекта x_i кластеру j.

12. Энтропийный коэффициент разбиения (Entropy partition coefficient) (EPC) [23]:

$$EPC=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}p(x_i)_j\log_2\left(p(x_i)_j\right). \tag{22}$$

Выбор того или иного коэффициента определяется используемым алгоритмом и характером исследуемых данных. Структурная блок-схема процесса определения количества кластеров на основе критериев релевантности представлена на рис. 2.

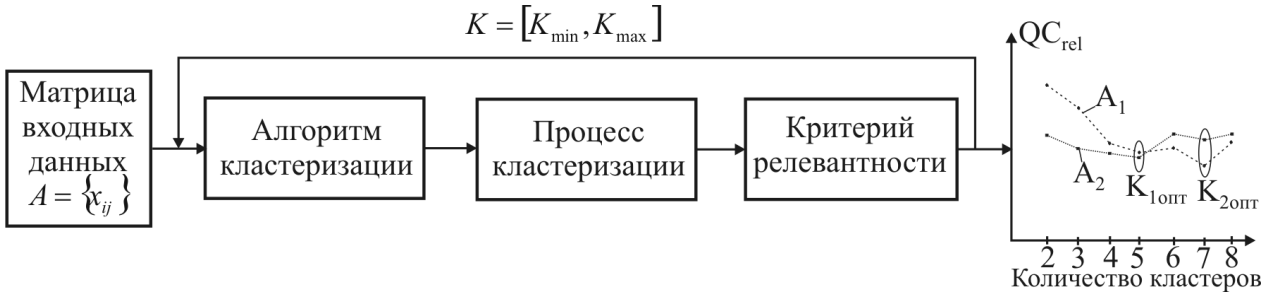

Рис. 2. Блок-схема процесса выбора оптимальных кластеризаций на основе критерия релевантности для данных A_1 и A_2

Реализация данного процесса предполагает наличие следующих шагов:

1. реализация выбранного алгоритма кластеризации для кластеризации K в пределах допустимого диапазона $K = [K_{\min}, K_{\max}]$;

2. фиксация полученной кластеризации, расчет центроидов выделенных кластеров;

3. вычисление критерия релевантности для полученной кластеризации;

4. повторение шагов 1-3 для получения требуемого количества кластеров в пределах заданного диапазона;

5. построение графиков зависимости критериев релевантности от количества полученных кластеров. Анализ полученных графиков, выделение оптимальных кластеризаций.

Как видно из рис. 2, оптимальная кластеризация соответствует локальному минимуму (или максимуму) значения критерия релевантности, при этом в пределах заданного диапазона кластеризаций возможно несколько экстремумов. Каждый из локальных минимумов или максимумов соответствует адекватной группировке объектов при различной степени детализации процесса. Однако следует отметить, что на основании внутреннего критерия релевантности невозможно оценить объективность кластеризации, поскольку данная оценка возможна при наличии «свежей» информации на основании внешнего критерия оценки различия соответствующих кластеризаций на двух равномощных подмножествах.

Критерии корелевантности в индуктивной технологии объективной кластеризации. Как говорилось выше, одним из существенных недостатков существующих методов и моделей кластеризации является ошибка воспроизводимости, т.е. высокая точность работы соответствующего алгоритма на одной выборке не гарантирует подобных результатов на другой подобной выборке данных. В предложенной индуктивной модели данная проблема решается посредством использования двух равномощных подвыборок данных, при этом кластеризация производится параллельно на двух подвыборках с одновременным сопоставлением промежуточных результатов. Таким образом, реализуется принцип конкуренции, один из трех фундаментальных принципов

индуктивной модели объективной кластеризации. Применение данного принципа вызывает необходимость создания еще одного критерия – аналога критерия непротиворечивости в теории индуктивного моделирования сложных систем. Технология использования данного критерия в индуктивном моделировании сложных систем предполагает, чтобы модели одинаковой структуры на двух частях исследуемой выборки данных давали максимально близкие выходы. В [9,10] автором для оценки данного свойства при разработке индуктивной технологии информационно-аналитических исследований вводится понятие корелевантности, как мера различия результатов исследований двух параллельных однотипных групп при выполнении аналогичного задания. Критерием корелевантности является количественная мера данного различия.

В индуктивной модели объективной кластеризации результатом моделирования на отдельных равномощных подмножествах является матрица промежуточных результатов, имеющая следующий вид:

$$W\left(R\left(K\right)\right)=\left\|QC_{ij}^{rel}\right\|=\begin{pmatrix} QC_{11}^{rel} & \dots & QC_{1k}^{rel} \\ \dots & \dots & \dots \\ QC_{n1}^{rel} & \dots & QC_{nk}^{rel} \end{pmatrix}, \tag{23}$$

где $k=K_{\min},...,K_{\max}$ – количество получаемых кластеризаций в процессе работы модели, n – количество используемых критериев релевантности. При использовании одного критерия релевантности матрица (23) вырождается в вектор-строку:

$$W\left(R\left(K\right)\right)=\left(QC_{1}^{rel},...,QC_{k}^{rel}\right). \tag{24}$$

Введем понятие матрицы различий критериев релевантности, элементы которой являются критериями корелевантности на данном этапе процесса кластеризации:

$$\Delta_{corel}^{2}=\left\|\delta_{ij}^{2}\right\|=\begin{pmatrix} \delta_{11}^{2} & \dots & \delta_{1k}^{2} \\ \dots & \dots & \dots \\ \delta_{n1}^{2} & \dots & \delta_{nk}^{2} \end{pmatrix}. \tag{25}$$

Элементы матрицы (25) представляют разницу квадратов соответствующих критериев релевантности, полученных по результатам кластеризации на подмножествах A и B соответственно:

$$\delta_{ij}^{2}=\left(\left(QC_{ij}^{rel}\right)_{A}-\left(QC_{ij}^{rel}\right)_{B}\right)^{2}. \tag{26}$$

Корелевантностью промежуточных результатов в индуктивной процедуре объективной кластеризации называется квадрат разницы критериев релевантности на различных этапах кластеризации объектов на двух равномощных подмножествах:

$$QC_{corel} = \left(W\left(R(K)\right)_A - W\left(R(K)\right)_B\right)^2. \quad (27)$$

Объективная кластеризация соответствует минимуму критерия корелевантности, при этом следует отметить, что в процессе работы модели возможно получение нескольких локальных близких экстремумов, т.е. объективных кластеризаций может быть несколько. В данном случае оптимальная кластеризация выбирается на основании комплексного анализа критериев релевантности и корелевантности, иначе говоря, находится баланс критериев релевантности и корелевантности.

Критерий баланса в индуктивной технологии объективной кластеризации. Необходимость третьего критерия в индуктивной модели объективной кластеризации вызвана возможными несовпадениями результатов кластеризации по критериям релевантности и корелевантности, поскольку объективная кластеризация по результатам критерия корелевантности не всегда является оптимальной по результатам критерия релевантности и наоборот. Поэтому в индуктивной модели объективной кластеризации кроме критериев селекции кластеризаций на двух равномощных подмножествах возникает необходимость в создании регулятора, целью которого является нахождение баланса между критериями релевантности и корелевантности для выбора наилучшей с точки зрения поставленных целей кластеризации.

Принцип действия критерия баланса в индуктивной модели объективной кластеризации следующий:

– после проведенных кластеризаций на подмножествах А и В рассчитывается критерий корелевантности в соответствии с формулой (27);

– рассчитывается среднее значение полученной матрицы по столбцам (полученным кластеризациям) в пределах от K_{min} до K_{max}.

$$QC*_{corel} = \frac{1}{n}\sum_{i=1}^{n}\left(QC_{corel}\right)_{ik},\ k = K_{\min},\ldots,K_{\max}. \quad (28)$$

При использовании одного критерия релевантности данный шаг пропускается;

– определяется критерий баланса, как минимум комплексного критерия корелевантности:

$$QC_{bal} = \arg\min_{K_{\min}\le K\le K_{\max}} QC*_{corel}\left(K\right). \quad (29)$$

Следует отметить, что критерий баланса, рассчитываемый по формуле (29) является вспомогательным критерием. Его применение актуально при использовании нескольких критериев релевантности и наличии нескольких локальных минимумов комплексного критерия корелевантности. При использовании одного критерия релевантности и наличии одного локального минимума принятие решения по выбору наилучшей кластеризации осуществляется на основании данного критерия корелевантности, т.е. в данном

случае он является и критерием баланса. Кроме того, для получения адекватной информации на выходе критерии корелевантности должны быть пронормированы соответствующим образом для приведения их к одинаковому диапазону.

Комплексная критериальная оценка качества группировки объектов в индуктивной технологии объективной кластеризации. Как говорилось выше, одним из основных факторов, способствующих высокой эффективности работы индуктивной модели объективной кластеризации, является адекватный выбор критериев оценки качества кластеризации на различных этапах работы модели. Данные критерии должны учитывать как расположение объектов в соответствующих кластерах относительно соответствующего центра масс, так и положение центроидов соответствующих кластеров по отношению друг к другу в различных кластеризациях. Пример возможного расположения объектов в трехкластерной индуктивной модели объективной кластеризации представлен на рис. 3. Положение центроида k-го кластера определяется как среднее значение признаков объектов, входящих в данный кластер:

$$c_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ij}\,, \tag{30}$$

где n_k – число объектов, входящих в k-й кластер, $j = 1,\ldots,m$ – количество признаков, характеризующих данный объект.

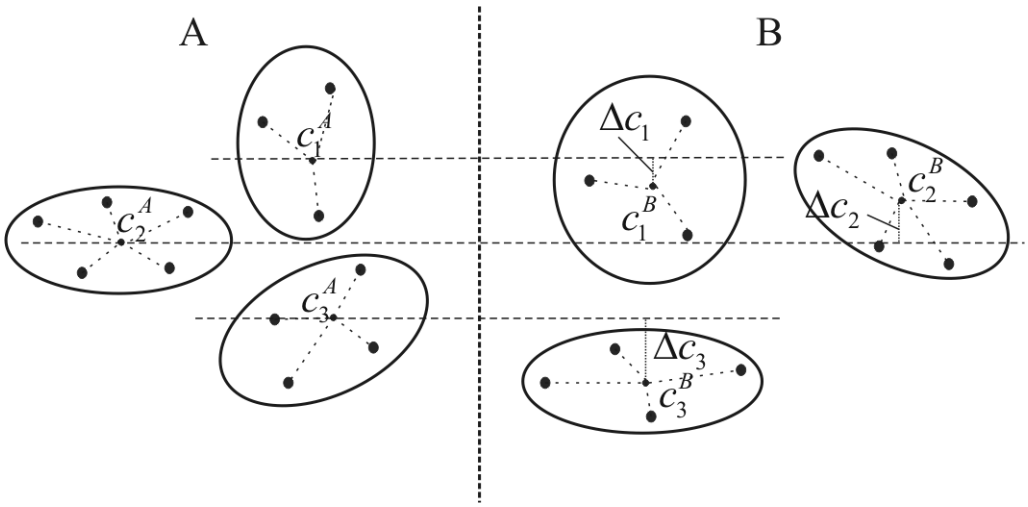

Рис. 3. Пример расположения объектов и кластеров
в индуктивной технологии объективной кластеризации

Первая составляющая комплексного критерия основана на предположении, что в случае объективной кластеризации среднее значение суммарного смещения центров масс соответствующих кластеров в различных кластеризациях должно быть минимальным:

$$QC_1(A,B)=\sqrt{\sum_{j=1}^{m}\left(\frac{1}{k}\sum_{i=1}^{k}\left(c_i(A)-c_i(B)\right)^2\right)^2}\rightarrow\min, \tag{31}$$

где k – количество кластеров в кластеризациях. В случае нормализации значения критерия формула (30) принимает вид:

$$QC_1(A,B)=\sqrt{\sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{k}\left(c_i(A)-c_i(B)\right)^2}{\sum_{i=1}^{k}\left(c_i(A)+c_i(B)\right)^2}\right)^2}\rightarrow\min. \tag{32}$$

Вторая составляющая комплексного критерия учитывает разницу в характере распределения кластеров и объектов в соответствующих кластерах в различных кластеризациях. Расстояние между объектом и центроидом кластера, в котором находится объект, определим по формуле Евклида:

$$d(X_a,c_k)=\sqrt{\sum_{j=1}^{m}\left(x_{aj}-c_k\right)^2}. \tag{33}$$

Тогда среднее расстояние от объектов до центроидов соответствующих кластеров может быть рассчитано следующим образом:

$$D_W=\frac{1}{k}\sum_{s=1}^{k}\left(\frac{1}{n_s}\sum_{i=1}^{n_s}d(X_i,c_s)\right), \tag{34}$$

где $s=1,\ldots,k$ – количество кластеров, n_s – количество объектов в кластере s, c_s – центроид кластера s.

Расстояние между центроидами кластеров определим как среднее расстояние от центроидов до центра масс объектов исследуемого множества данных:

$$D_B=\frac{1}{k}\sum_{s=1}^{k}d(c_s,\overline{C}) \tag{35}$$

Очевидно, что кластеризация будет более качественной при более плотном расположении объектов внутри кластеров и большем расстоянии между центроидами кластеров относительно общего центра масс:

$$D_W\rightarrow\min,\ D_B\rightarrow\max. \tag{36}$$

Тогда отношение

$$D=D_W/D_B \tag{37}$$

может быть показателем качества группировки объектов в соответствующей кластеризации. Вторую составляющую комплексного критерия

корелевантности можно представить как модуль разности показателя (37) для различных кластеризаций. В нормализованном варианте данная формула имеет вид:

$$QC_2(A,B)=\frac{|D(A)-D(B)|}{D(A)+D(B)} \tag{38}$$

Объективная кластеризация выбирается на основании анализа локальных минимумов критериев (32) и (38) в процессе переборе всех доступных кластеризаций.

Архитектура и концептуальное описание пошаговой процедуры индуктивной технологии объективной кластеризации. На рис. 4 представлена общая архитектура реализации индуктивной технологии объективной кластеризации. На вход системы подается матрица данных, строки которой являются исследуемыми объектами, а столбцы представляют атрибуты или признаки, определяющие свойства данных объектов. Выходом являются совокупность кластеров, каждый из которых включает в себя группу объектов, признаки которых обладают высокой аффинностью для данных объектов. Реализация данной технологии предполагает наличие следующих этапов.

Этап I.

1. Постановка проблемы. Формирование целей кластеризации.
2. Анализ исследуемых данных, определение характера признаков исследуемых объектов, приведение данных к виду матрицы $A=\{x_{ij}\}, i=1,\ldots,n; j=1,..m$, где n – количество наблюдаемых объектов, m – количество признаков, характеризующих соответствующий объект.
3. Предобработка данных, включающая в себя восстановление пропущенных значений (при необходимости), фильтрацию с целью удаления неинформативных столбцов, фильтрацию с целью удаления «белого шума», нормализацию столбцов для приведения их к единому распределению и диапазону.
4. Редуцирование размерности признакового пространства исследуемых объектов (при необходимости).
5. Определение функции аффинности для дальнейшей оценки степени близости объектов.

Рис. 4. Архитектура индуктивной технологии объективной кластеризации признакового пространства исследуемых объектов (при необходимости).

6. Формирование двух равномощных подмножеств A и B в соответствии с ранее представленным алгоритмом.

7. Выбор и настройка алгоритма кластеризации. Инициализация исходных параметров алгоритма.

Этап II.

1. Кластеризация данных на подмножествах A и B. Формирование кластеров в пределах выбранного диапазона $K_{\min} \le K \le K_{\max}$. Если количество кластеров в различных кластеризациях различается, процесс прекращается по причине неудачно выбранного алгоритма или неправильной начальной инициализации параметров алгоритма. В этом случае необходимо применить другой алгоритм из множества допустимых или изменить начальные параметры текущего алгоритма.

2. Формирование оценок частной кластеризации, расчет центроидов $C(K)^A$, $C(K)^B$ и критериев релевантности $QC_{rel}(K)^A$, $QC_{rel}(K)^B$ для текущей кластеризации на подмножествах A и B.

Этап III.

Расчет критериев корелевантности в соответствии с формулами (32), (38) для данной кластеризации.

Этап IV.

1. Построение графиков зависимости рассчитанных критериев корелевантности от количества полученных кластеров в пределах заданного диапазона $K_{\min} \le K \le K_{\max}$.

2. Расчет критерия баланса по формуле (29).

Этап V.

1. Анализ полученных результатов. В случае отсутствия локальных минимумов критериев корелевантности или превышения значений данных критериев допустимых норм (на рис 4 знак «–») выбор другого алгоритма кластеризации или повторная начальная инициализация текущего алгоритма. Повторение этапов II-V данной процедуры.

2. При наличии локальных минимумов при условии перебора всех кластеризаций в заданном диапазоне фиксация объективной кластеризации, соответствующей минимуму критерия баланса.

Выводы

В данной статье получила дальнейшее развитие методология создания индуктивных технологий на основе индуктивных методов моделирования сложных систем. Технология объективной кластеризации представлена как определенный аналог стратегии интеллектуального проекта исследования сложных систем, что позволило сформулировать данное исследование как задачу объективной кластеризации высокоразмерных объектов сложной природы в условиях неполноты информации.

Рассмотрена возможность применения основных принципов методологии индуктивного моделирования сложных систем: принципа самоорганизации, принципа внешнего дополнения и принципа свободы выбора в индуктивной технологии объективной кластеризации объектов сложной природы, что позволило сформулировать основные принципы данной технологии: принцип эвристической самоорганизации модели, принцип конкуренции и принцип неокончательности решений. Данные принципы легли в основу индуктивной технологии объективной кластеризации объектов.

Разработана технология выбора объективной кластеризаций на основе критериев релевантности, корелевантности и баланса, разработан комплексный критерий корелевантности, учитывающий в качестве составляющих как положение центроидов соответствующих кластеров в различных кластеризациях, так и характер распределения объектов в кластерах относительно соответствующих центроидов и центров кластеров относительно общего центра масс в различных кластеризациях.

Разработана архитектура индуктивной технологии объективной кластеризации, представленная в виде упрощенной подробной схемы реализации процедуры индуктивного моделирования объективной кластеризации, представлено концептуальное описание пошаговой процедуры реализации индуктивной технологии объективной кластеризации, основанной на существующих методах и алгоритмах кластеризации и позволяющей повысить объективность группировки объектов за счет комплексного использования внутренних и внешних критериев оценки качества группировки исследуемых объектов.

Перспективами дальнейших исследований авторов является практическая реализация представленных разработок для кластеризации высокоразмерных данных сложной биологической природы на основе различных алгоритмов кластеризации объектов.

Список литературы.

1. Івахненко О.Г. Метод групового урахування аргументів – конкурент методу стохастичної апроксимації // Автоматика, 1968. – №3. – С. 58-72.

2. Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем.− Киев: Наукова думка,1982.− 296 с

3. Ивахненко А.Г. Объективная самоорганизация на основе теории самоорганизации моделей // Автоматика, 1987. – №5. – С. 6-15.

4. Madala H.R., Ivakhnenko A.G. Inductive Learning Algorithms for Complex Systems Modeling.− CRC Press, 1994. – 365 p.

5. Сарычева Л.В. Объективный кластерный анализ данных на основе метода группового учета аргументов // Проблемы управления и автоматики, 2008. – №2. –С. 86-104.

6. Степашко В.С. Елементи теорії індуктивного моделювання / Стан та перспективи розвитку інформатики в Україні: монографія / Колектив авторів. — Київ: Наукова думка, 2010. – 1008 с. / – С. 471-486.

7. Степашко В.С. Самоорганизация прогнозирующих моделей сложных процессов и систем. – ХУ Всероссийская научно-техническая конференция “Нейроинформатика-2013”: Лекции по нейроинформатике / Ю.В.Тюменцев — отв. ред. – М.: НИЯУ МИФИ, 2013. – 320 с. / – С. 150-170.

8. Степашко В.С. Теоретические аспекты МГУА как метода индуктивного моделирования / В.С. Степашко // УСиМ. – 2003. – №2. – 31-38.

9. Осипенко В.В. Система критеріїв в індуктивних процедурах системних інформаційно-аналітичних досліджень / В.В. Осипенко // Системні технології. Міжвузівський збірник наукових праць. Випуск 6(71). – Дніпропетровськ. – 2011. – С. 106-113.

10. Осипенко В.В. Оценка релевантности результатов в индуктивных процедурах системно-аналитических исследований / В.В. Осипенко // Управляющие системы и машины, № 1, 2012. – С. 26-32.

11. Осипенко В.В. Розроблення конкурентних маркетингових стратегій за індуктивною технологією системних інформаційно-аналітичних досліджень / В.В. Осипенко // Вісник Нац. університету «Львівська політехніка». Сер. Комп’ютерні науки та інформаційні технології. Вип. № 732, — 2012. — С. 351-358.

12. Osypenko V.V. Inductive technologies of system-information-analytical research as an intelligent tools of the effective management in retail business / V.V. Osypenko // В кн.: Індуктивне моделювання складних систем. — К.: МННЦІТіС НАН України, 2012. — С. 11-20.

13. Milligan G, Cooper M. An examination of procedures for determining the number of clusters in a data set // Psychometrika. – 1985. – №50. – P. 159–179.

14. Dimitriadou E, Dolnicar S, Weingassel A. An examination of indexes for determining the number of clusters in binary data sets // Psychometrika. – 2002. – №67(1). – P. 137-160.

15. Dunn J. Well separated clusters and optimal fuzzy partitions // Journal of Cybernetica. – 1974. – №4. – P. 95-104.

16. Davies D., Bouldin D. Cluster separation measure // IEEE Trans. on Pattern Analysis and Machine Intelligence. – 1979. – №1. – P. 95-104.

17. Xie X., Beni G. A validity measure for fuzzy clustering // IEEE Trans. on Pattern Analysis and Machine Intelligence. – 1991. – №13. – P. 841-847.

18. Halkidi M., Vazirgiannis M. Clustering validity assessment: Finding the optimal partitioning of a data set // Proc. of the 2001 IEEE Int. Conf. on Data Mining (ICDM'01). – 2001. – P. 187-194.

19. Hartigan J. Clustering algorithms // New York, NY: Wiley. – 1975. – 369p.

20. Still S., Bialek W. How many clusters? An information theoretic perspective // Neural Computation. – 2004. – №16. – P. 2483-2506.

21. Ball G. Hubert L. ISODATA, A novel method of data analysis and pattern classification // Menlo Park, CA: Standford Research Institute. –1965. – 365 p.

22. Krzanowski W., Lai Y. A criterion for determining the number of groups in a data set using sum-of-squares clustering // Biometrics. – 1985. – №44. – P. 23-34.

23. Batistakis M. H. Y., Vazirgiannis M. Clustering validity checking methods: Part I // ACM SIGMOD Record. – 2002. – №31. – P. 40-45.

24. Calinski T., Harabasz J. A dendrite method for cluster analysis // Communication in statistics. – 1974. – №3. – P. 1–27.

25. Halkidi M., Batistakis Y., Vazirgiannis M. Clustering validity checking methods: Part II // ACM SIGMOD Record. – 2002. – №31. – P. 19-27.