УДК 004.9:519.23

В.В. Осипенко

Индуктивный алгоритм кластер-анализа в инструментарии системных информационно-аналитических исследований

Предложен оригинальный подход к решению задачи формирования экспертных групп в системных информационно-аналитических исследованиях, основанный на применении индуктивного алгоритма кластеризации. Метод можно применять во многих прикладных системно-аналитических исследованиях.

The unique approach to solving the problem of choosing experts groups in the system information-analytical researches is suggested. This approach is based on the use of an inductive modeling paradigm to solving the cluster analysis tasks. Such approach can also been used in many fields of applied researches pertaining to the problems of structuring, classification, clustering and modeling of complex systems.

Запропоновано оригінальний підхід до розв'язання задачі формування експертних груп у системних інформаційно-аналітичних дослідженнях, базований на застосуванні індуктивного алгоритму кластеризації. Метод можна застосовувати у багатьох прикладних системно-аналітичних дослідженнях.

Введение. С позиций методологии системного анализа [1], индуктивная технология системных информационно-аналитических исследований (ИТ СИАИ) [2] – это интеллектуальный инструментарий, нацеленный на решение некоторой проблемы, возникшей в сложной системе. Термин *проблема* не несет какого-то негативного оттенка. Проблемой в СИАИ может быть вполне позитивное задание разработки, например, международного меморандума о дружественных отношениях между двумя странами.

Для решения сложных проблем подобного класса, естественно, необходимы долгосрочные и высокобюджетные исследования, в которых должны быть задействованы специалисты нескольких, часто не смежных, направлений одновременно. Цель таких поисковых проектов предусматривает создание специфического документа, содержащего значимые выводы и оптимальные рекомендации по решению возникшей проблемы, качественные и количественные прогнозы принятых решений и много других необходимых для внедрения решений элементов.

Из сказанного следует, что системные информационно-аналитические исследования, – это существенный и неотъемлемый сегмент общей теории системного анализа для выполнения системно-аналитических проектов и конструирования результатов, нацеленных на поддержку принятия решений во многих прикладных областях.

С другой стороны, информационные (в том числе индуктивные) технологии системно-аналитических исследований сами по себе есть сложные системы с присущими им основными характеристиками, в частности свойствами эмерджентности, синергизма, иерархичности, открытости и другими. Действительно, одно из основных свойств ИТ СИАИ, как сложной системы, – ее открытость, что требует наличия внешней информационной среды как минимум в трех направлениях:

- постановке собственно проблемы и реципиента ее решения – наличия заказчика;
- получения внешней дополнительной информации целевого назначения;
- внешнего независимого оценивания результатов интеллектуальной деятельности аналитических групп в течение всего цикла исследований.

Третье из перечисленных направлений, очевидно, предусматривает участие экспертов высокого уровня – определенных субъектов из внешней среды по отношению к собственно системе, которая есть непосредственным исполнителем исследований.

Отсюда объективно возникает две важные и нетривиальные проблемы:

- отбор членов в экспертную комиссию высшего уровня (ЭКВУ), и для комплексных СИАИ они должны быть специалистами из разных предметных областей;

Ключевые слова: кластеризация, критерий, целевой признак, индуктивное моделирование.

• синтез формального прообраза (эталона) будущего результата [2] (учитывать следует выводы всех экспертов, приглашенных к конкурсным соревнованиям).

Известно, что выбор экспертов в проектах СИАИ – непростой и ответственный этап. Этим вопросам посвящены многочисленные труды ведущих ученых с мировым именем и практические рекомендации опытных специалистов. Сюда же относятся исследования такого направления, которое есть частью теории принятия решений под названием *коллективный выбор* [3, 4] и др. В [5] для определения компетентности экспертов предложен алгоритм, построенный на основе аксиомы несмещенности. В работе [6] для оценки компетентности экспертов и сопутствующих вопросов в области трансфера технологий предложен подход, основанный на методологии нечетких множеств. Распространение также получило применение метода парных сравнений, или метода анализа иерархий Т. Саати [7]. В приложении к упомянутой работе вполне уместно замечание о целесообразности и эффективности применения методов кластерного анализа для решения задач подобного направления [7, с. 272]. Очевидно, что перечень подходов к решению вопроса отбора экспертов в комиссии верхнего уровня можно продолжать.

Но в ИТ СИАИ на экспертов из ЭКВУ дополнительно возлагается параллельное решение еще нескольких задач, что предусмотрено, например, в методах «дельфийского типа» [8]:

• формирование требований к форме и содержанию (без семантического наполнения) будущего целевого результата – матрицы эталонного результата [2];

• независимого внешнего оценивания промежуточных и финальных результатов, наработанных независимыми аналитическими группами.

Цель данных исследований – разработка подхода к решению задачи выявления и формирования гомогенных относительно целевой проблемы СИАИ экспертных групп, основанного на индуктивном кластерном анализе с применением целевого признака, а также разработка критерия оценки качества кластеризации с регуляризирующим целевым элементом и описание многорядного индуктивного алгоритма кластер-анализа с выбором ансамбля информативных признаков.

Методы исследований

Описываемый далее подход, прежде всего, основывается на применении парадигмы индуктивного моделирования сложных систем к решению задач кластеризации [9, 10]. Алгоритмы, построенные по методологии [10], на практике достаточно эффективны, что показано на многочисленных примерах [11]. Но все-таки, индуктивные процедуры кластер-анализа относятся к задачам с многочисленными эвристиками, и этот факт причисляет такие алгоритмы к множеству некорректно поставленных с позиций строгого математического обоснования задач. Отсюда объективно возникает проблема регуляризации процедуры индуктивного кластер-анализа с целью получения устойчивых результатов. Впервые, в качестве определенного регуляризирующего элемента именно в индуктивном кластер-анализе с выбором информативных признаков, было предложено применение так называемого целевого (интегрального) признака [12], хотя это понятие было введено значительно раньше другими исследователями. Подход, предлагаемый в этой статье, также требует применения целевого признака как регуляризирующего элемента, но в несколько иной трактовке. В исследованиях использованы некоторые материалы аналитического обзора [13], методология МГУА [10, 11], а также ссылки на общеизвестные [14] и современные алгоритмы кластер-анализа [15].

Постановка задачи выбора групп экспертов в ИТ СИАИ

Согласно требованиям ИТ СИАИ, для участия в проекте системно-аналитического исследования, прежде всего, необходимо отобрать определенное количество (группу) экспертов $m^* < m$, m – стартовое количество участников экспертных соревнований в обязательной квалификационной сессии. Кроме того, еще до начала исследований, должно быть сформировано определенное информативное подмноже-

ство из $n^* < n$ признаков, которое будет, с одной стороны, характеризовать ЭКВУ как гомогенную экспертную группу, способную объективно оценивать достигнутые промежуточные результаты, а с другой – входить в информационный базис будущего финального результата исследования. Отметим, что такие факторы могут характеризовать лишь определенную, хотя и очень существенную, часть первичного информационного базиса [2].

Итак, с позиций общей постановки задачи кластерного анализа в широком смысле, т.е. выбора оптимальной кластеризации с одновременным выбором оптимального ансамбля информативных признаков в пределах заданной системы критериев [13], нашу задачу формально можно представить так.

Пусть задано множество $\{x_{0j}, j = 1,\ldots,m\}$ значений рейтингов (целевых признаков) для m экспертов, присвоенных им по результатам богатой предыдущей практики и опытом работы в близких по тематике к поставленной проблеме системных информационно-аналитических проектах.

Пусть в пространство признаков $x_{ij} \in X$ $(i = 1,\ldots,n, j = 1,\ldots,m)$ введены также такие, которые характеризуют начальный информационный базис [2] будущего исследования, заданный заказчиком поискового проекта.

Кроме упомянутых, во множество $\tilde{O}$ будут введены признаки тестовых заданий, по которым будущим экспертам необходимо будет выбрать ближайшие к значениям «ключей» варианты ответов на вопросы будущего исследования.

И наконец, пусть во множество $\tilde{O}$ введены признаки, отражающие личные предварительные выводы и умозаключения участвующих в конкурсе экспертов относительно предмета данного информационно-аналитического исследования.

Таким образом, общий массив входных данных в нашей задаче имеет следующий вид:

$$\tilde{X} = \left(x_{0j} : x_{ij} \in X\right),\ j = \overline{1,m}, i = \overline{1,n}. \qquad (1)$$

Необходимо:

• синтезировать подмножество $\left\{x_\eta^*\right\} = X^* \subset \subset X$, $\eta = 1,\ldots,n^*, n^* \le n$ упомянутых признаков, лучшее по заданному критерию оптимальности, которое позволило бы:

– классифицировать всех участников экспертных соревнований на $k < m$, $k = 1,\ldots,K$ однородных групп по результатам отборочной квалификационной сессии;

– отобрать единую группу (ЭКВУ) k^* из $m^* < m$ экспертов, соответствующих требованиям и проблематике данного информационно-аналитического проекта по заданным критериям.

Внесем несколько ремарок.

• В терминах индуктивного подхода к задачам кластер-анализа подмножество $\left\{x_\eta^*\right\} = X^*$ обычно трактуется как ансамбль информативных признаков [10].

• Упомянутые «ключи» как признаки в (1) обычно должны формироваться совместно представителем заказчика и модератором исследовательского проекта, а проблематика и методика их определения должны быть понятны всем предварительно приглашенным экспертам (этот вопрос нуждается в отдельном исследовании и выходит за рамки данной статьи).

• Следует отметить, что множество X должно быть достаточно представительным множеством признаков (квалификационных, тестовых, проблемных и др.), среди которых могут быть не только бесспорные требования к будущему результату, но и факторы, которые не есть императивными для запланированного исследования, но касательно могут иметь существенное отношение к его проблематике.

В большинстве современных информационно-аналитических технологий рейтинги, о которых идет речь, формируются в течение длительного периода, зачастую с применением довольно сложных процедур и присваиваются экспертам соответствующими уважаемыми сообществами профессионалов. Поэтому такие оценки достаточно устойчивы и общеприняты в определенных кругах экспертов, что добавляет

некоторую долю объективности такому понятию. Кроме таких рейтингов, в современной литературе часто встречаются такие термины как «самооценка», «коэффициент авторитета» и другие [4–6], которые также применимы при формировании понятия «целевого признака». Тем не менее, наверное следует согласиться с тем, что такие интегральные характеристики экспертов, которые должны выполнять роль внешнего независимого и объективного оценивания результатов будущих исследований, не всегда могут сами быть объективными по разного рода известным и неизвестным причинам. Однако в данной статье сознательно воспринимаются такие рейтинги как «лучшие» целевые признаки, хотя затронутая нетривиальная задача объективного синтеза целевой интегральной характеристики эксперта в данной проблематике остается открытой и весьма актуальной.

Критерии оптимальности в индуктивном кластер-анализе с целевым признаком

Прежде чем перейти к описанию алгоритма, сконструируем и дадим толкование критериям оптимальности решения сформулированной ранее в широком смысле задачи кластер-анализа, которые будут применяться в процедурах индуктивной кластеризации.

Известно, что среди основных характеристик k-го кластера (для удобства и без потери общности далее будем рассматривать евклидово пространство) есть его центр массы в пространстве признаков X:

$$\bar{m}_k(X)=\left\{\left(\frac{1}{r_k}\sum_{l=1}^{r_k}x_{l1}\right),\left(\frac{1}{r_k}\sum_{l=1}^{r_k}x_{l2}\right),\ldots,\left(\frac{1}{r_k}\sum_{l=1}^{r_k}x_{ln}\right)\right\}=\left\{\frac{1}{r_k}\sum_{l=1}^{r_k}x_{li},\ i=1,\ldots,n\right\},\ x_i\in X, \tag{2}$$

а среднее внутримножественное расстояние можно записать в виде

$$\overline{d_k^2(\omega_s^k,\omega_t^k)}=\frac{1}{r_k\left(r_k-1\right)}\sum_{s=1}^{r_k}\sum_{t=1}^{r_k}\sum_{i=1}^{n}\left(x_{is}-x_{it}\right)^2, \tag{3}$$

где r_k – количество исходных объектов ω^k в k-м кластере, n – начальное количество признаков пространства X.

Используем целевой признак как регуляризирующий элемент и вычислим центр k-го кластера только по значениям целевого признака вошедших в него объектов ω^k. Это можно трактовать как проекцию центра k-го кластера с n-мерного евклидова пространства X в одномерное евклидово пространство $\Re^1$, т.е. на ось действительных чисел x_0. Выражения (2) и (3) при этом преобразуются в более простой вид:

$$\bar{m}_k(x_0)=\hat{m}_k=\frac{1}{r_k}\sum_{l=1}^{r_k}x_{0l}, \tag{4}$$

$$\overline{d_k^2(\omega_s^k,\omega_t^k)_{x^0}}=\hat{d}_k^2=$$

$$=\frac{1}{r_k\left(r_k-1\right)}\sum_{s=1}^{r_k}\sum_{t=1}^{r_k}\left(x_{0s}-x_{0t}\right)^2,\ s\neq t. \tag{5}$$

Известно, что применение методологии индуктивного моделирования сложных систем [10] для получения оптимальной кластеризации требует разделения входного множества объектов $\omega_k\in\Omega$, подлежащих кластеризации, не менее чем на два непересекающиеся подмножества Ω^A и Ω^B, при этом: $\Omega^A\cup\Omega^B=\Omega$, $\Omega^A\cap\Omega^B=\varnothing$.

Пусть на подмножествах Ω^A и Ω^B по одной из выбранных процедур кластер-анализа получены кластеризации $s_t^A\in S^A$ и $s_t^B\in S^B$ с одинаковым количеством кластеров $k_t^A=k_t^B=K_t$ (t – номер кластеризации, соответствующий некоторому подпространству признаков $X_t\subset X$, $k_t^{(\cdot)}$ – количество кластеров в t-й кластеризации) в евклидовом подпространстве признаков $X_t\subset X$, и пусть для всех K_t кластеров из s_t^A и s_t^B вычислены их центры $\hat{m}_k^A$ и $\hat{m}_k^B$, $k=1,\ldots,K_t$, по оси целевого признака x_0.

Тогда критерий оптимальности регуляризованной кластеризации можно записать в простейшем и более общем виде как

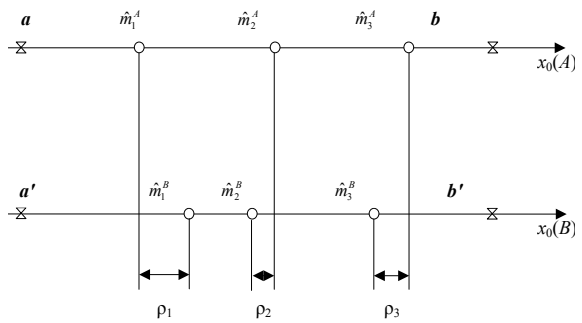
$$\rho^2(\hat{m})=\sum_{k=1}^{K}(\hat{m}_k^A-\hat{m}_k^B)^2\to\min. \tag{6}$$

Критерий (6) требует, чтобы сумма квадратов отклонений между центрами кластеров на оси целевого признака, установленных на подмножествах Ω^A и Ω^B, была минимальной. По-

этому такой критерий в данном случае можно еще назвать *критерием наименьших межцентровых отклонений* (НМО), который, очевидно, есть представителем класса *внешних критериев непротиворечивости* (НП) в МГУА, достаточно широко применяемых в процедурах индуктивного моделирования. Для того, чтобы значения критерия (6) изменялись в пределах некоторого заданного интервала, например [0, 1], его можно записать в виде

$$\rho^2(\hat{m}) = \sum_{k=1}^{K}(\hat{m}_k^A - \hat{m}_k^B)^2 \Big/ \sum_{k=1}^{K}(\hat{m}_k^A + \hat{m}_k^B)^2 \to \min. \quad (7)$$

Это не принципиально, но такая процедура часто применяется в программных реализациях алгоритмов. На рисунке показан принцип работы критерия $\rho^2(\hat{m})$ при $n_k = 3$.



Принцип работы критерия наименьших межцентровых отклонений (непротиворечивости кластеризации: $\rho_\Sigma = |\rho_1| + |\rho_2| + |\rho_3|$)

Вторым критерием качества кластеризации целесообразно назначить классический *критерий минимума средних внутримножественных расстояний*. В общем случае вычисление такого критерия – достаточно трудоемкая процедура. Но в предлагаемом подходе эти расстояния однозначно идентифицируются по оси x_0, т.е. в одномерном пространстве $\Re^1$, что существенно упрощает вычисление такого критерия, а именно:

$$\delta^2(\hat{d}^2) = \sum_{k=1}^{K}\hat{d}_k^2 \to \min. \quad (8)$$

Критерий (8) не обладает свойством внешнего дополнения, но характеризует качество кластеризаций, полученных независимо на подмножествах Ω^A и Ω^B, и может применяться системно для более четкого отбора решений, в которых получены очень близкие значения критерия наименьших межцентровых отклонений – $\rho^2(\hat{m})$.

Алгоритм индуктивного кластер-анализа с целевым признаком в ИТ СИАИ

Ш а г 1. Выбор или предварительный синтез целевого признака x_0 для всех $\omega_k \in \Omega$.

Ш а г 2. Разделение (1) на две части A и B (Ω^A и Ω^B) согласно требований МГУА [9, 10]. Подготовленная общая матрица данных $\tilde{X}$ будет иметь такой условный вид:

$$\tilde{X} = \left[\left(x_{0j} \vdots X\right)^A \cdot\cdot \left(x_{0j} \vdots X\right)^B\right], \quad j = 1,...,m^A = m^B, \ m^A + m^B = m. \quad (9)$$

Ш а г 3. Настройка одной из процедур кластеризации (например, классического иерархического агломеративного алгоритма Ланса-Уильямса [14] или одного из современных алгоритмов типа [15] и др.). На этом этапе роль природы $\tilde{X}$ существенна.

Ш а г 4. Кластеризация объектов $\omega_k \in \Omega$ с помощью выбранного и настроенного алгоритма независимо на подмножествах Ω^A и Ω^B в пространстве $\tilde{X}$ по одной из классических схем алгоритмов МГУА с индуктивным наращиванием количества признаков в их ансамблях. Многорядная индуктивная процедура кластеризации, например, могла бы быть такой.

Первый ряд селекции.

1. Кластеризация объектов на подмножествах Ω^A и Ω^B по ансамблям $\{x_i\}, i = 1,...,n$.

2. Проецирование центров полученных кластеров на ось x_0.

3. Для кластеризаций, в которых выполняется условие $k_t^A = k_t^B = K_t$ (t – текущий номер кластеризации, $k_t^{(\cdot)}$ – количество кластеров в t-й кластеризации), вычисляются значения критерия оптимальности $\rho^2(\hat{m})$.

4. Для таких же вариантов решений вычисляются значения критерия $\hat{\delta}(\hat{d}^2)$.

Второй ряд селекции.

1. Кластеризация объектов на подмножествах Ω^A и Ω^B по ансамблям $\{x_i, x_j\}$, $i, j = 1, \ldots, n$, $i \neq j$.

2. Выполняются пп. (2 – 4) первого ряда селекции и по системе критериев (7), (8) отбираются F ($F \le n$) лучших кластеризаций S_f, $f = 1,\ldots,F$ и соответствующих ансамблей признаков $X_f, f = 1,\ldots,F$.

Третий и последующие ряды селекции:

1. Кластеризация объектов на подмножествах Ω^A и Ω^B по ансамблям $\{X_f, x_l\}$, $f = 1,\ldots,F$, $l = 1,\ldots,n$ при условии, что признак с индексом l не присутствует в уже созданных ансамблях X_f.

2. Выполняется п. 2 второго ряда селекции.

Правило останова: индуктивная процедура прерывается при условии

$$\rho^2(\hat{m})_s \le \rho^2(\hat{m})_{s+1}, \qquad (10)$$

где s – ряд селекции в терминах МГУА. При этом фиксируется значение $k^{*(A)} = k^{*(B)} = K^*$, $K^* \le m/2$ и подпространство информативных признаков $\{x_l^*\} = X^*$, $l = 1,...,n^*$, $n^* \le n$.

Таким образом, результат работы описанной индуктивной процедуры кластеризации таков:

- оптимальная кластеризация с гомогенными экспертными группами, одна из которых имеет достаточно прав стать в дальнейшем единственной ЭКВУ;
- подпространство информативных признаков, куда по определению должны входить как важнейшие формализованные атрибуты будущего результата СИАИ, так и параметры исходного информационного базиса.

Заключение. Ценность предложенного подхода к решению задачи обнаружения и формирования однородных экспертных групп дополняется возможностью одновременного выбора важных информационных компонент будущего результата уже на этапах проектирования СИАИ. Таким образом, решена также задача кластер-анализа в широком смысле применительно к ИТ СИАИ. Описанный вычислительный алгоритм нашел применение в практических задачах.

Описанный метод, как вычислительная процедура, носит универсальный характер и, несомненно, может претендовать на применение во многих сферах прикладных системно-аналитических исследований, соприкасающихся с проблемами структуризации, классификации, кластеризации и моделирования в сложных системах, а также как независимый и самостоятельный инструмент обработки статистических данных.

1. *Згуровський М.З.*, *Панкратова Н.Д.* Основи системного аналізу. – К.: Вид. гр. BHV, 2007. – 544 с.
2. *Осипенко В.В.* Оценивание релевантности результатов в индуктивных процедурах системно-аналитических исследований. – УСиМ. – 2012. – № 1. – С. 29–37.
3. *Ларичев О.И.* Теория и методы принятия решений. – М.: Логос, 2002. – 392 с.
4. *Эрроу К.Дж.* Коллективный выбор и индивидуальные ценности. – М.: Изд. дом ГУ ВШЭ, 2004. – 204 с.
5. *Снитюк В.Е.*, *Рифат М.А.* Модели и методы определения компетентности экспертов на базе аксиомы несмещенности // Вісн. ЧІТІ, 2000. – № 4. – С. 121–126.
6. *Герасимов Б.М.*, *Евтухова Т.И.* Информационно-аналитическое обеспечение трансфера технологий // Автоматизація виробничих процесів. – 2004. – № 2 (19). – С. 118–124.
7. *Саати Т.* Принятие решений. Метод анализа иерархий. – М.: Радио и связь, 1993. – 279 с.
8. *Delphi* Method: Techniques and Applications // Harold A. Linstone, Murray Turoff (Eds.). – Addison-Wesley Educ. Publ. Inc., 1975. – 621 p.
9. *Ивахненко А.Г.* Индуктивный метод самоорганизации моделей сложных систем. – Киев: Наук. думка, 1982. – 296 с.
10. *Ивахненко А.Г.* Объективная кластеризация на основе теории самоорганизации моделей // Автоматика. – 1987. – № 5. – С. 6–15.
11. *Применение* алгоритмов многоальтернативного распознавания образов и метода группового учета аргументов для обработки экспертных оценок проектов глобальной инвестиции капитала / А.Г. Ивахненко, Е.А. Савченко, Г.А. Ивахненко и др. // Кибернетика и выч. техника. – 2001. – **133**. – С. 3–7.
12. *Осипенко В.В.* Решение задачи двойной кластеризации на основе самоорганизации // Автоматика. – 1988. – № 5.– С. 74–79.
13. *Сарычева Л.В.* Объективный кластерный анализ данных на основе МГУА // Проблемы управления и информатики. – 2008. – № 2. – С. 86 –104.
14. *Мандель И.Д.* Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.
15. *Литвиненко В.И.* Кластерный анализ данных на основе модифицированной иммунной сети // УСиМ. – 2009. – № 1. – С. 54– 61, 85.

Тел. для справок: +38 044 530-5949, +38 050 411-7277 (Киев)
E-mail: *vvo7@ukr.net*