УДК 004.942

МЕТОДИКА ГРУППИРОВАНИЯ БАЗОВОЙ ИНФОРМАЦИИ ДЛЯ ИНФОРМАЦИОННЫХ ПРОЦЕССОВ СЛОЖНЫХ СИСТЕМ

[1]Сумин В.И., [1]Дыбова М.А., [2]Смоленцева Т.Е.

[1]ФКОУ ВПО «Воронежский институт ФСИН России», Воронеж, e-mail: viktorsumin51@yandex.ru;

[2]ФГБОУ ВПО «Липецкий государственный технический университет», Липецк

В данной статье рассматривается методика группировки объектов для информационных процессов сложных систем на классы, к которым относятся наиболее схожие по своим характеристикам объекты. Рассмотрен процесс группирования данных итеративными методами, а также применение итеративного метода кластерного анализа к объединению объектов базовой информации для информационных процессов сложных систем на основе значений анализируемых характеристик, включающий следующие этапы: формирование исходного разбиения на нужное число классов, проверка принадлежности объекта к классу, вычисление порогового значения, по которому определяется принадлежность объекта классу. В работе выявлены исходные данные разбиения объектов на классы.

Ключевые слова: итеративный метод кластерного анализа, базовая информация, центр тяжести кластеров

A METHOD OF GROUPING BASIC INFORMATION FOR INFORMATION PROCESSES OF COMPLEX SYSTEMS

[1]Sumin V.I., [1]Dybova M.A., [2]Smolentseva T.E.

[1]FSE VPO «Voronezh Institute of the Federal penitentiary service of Russia», Voronezh, e-mail: viktorsumin51@yandex.ru;

[2]FGBOU VPO «Lipetsk state technical University», Lipetsk

This article discusses the technique of grouping objects for information processes of complex systems into classes, which are most similar in their characteristics to the objects. The process of grouping data iterative methods, and the use of iterative cluster analysis method for grouping objects of basic information for information processes of complex systems based on the values of the analyzed characteristics, comprising the following steps: forming the source partition to the desired number of classes to check if the object class, the calculation of the threshold value, which is determined by the identity of the object class. Determined the source data partitioning objects into classes.

Keywords: iterative method of cluster analysis, basic information, the center of gravity of the clusters

Рассмотрим методику группировки объектов для информационных процессов сложных систем на классы, к которым относятся объекты, наиболее близкие по своим характеристикам.

Эффективным механизмом объединения в группы объектов разнообразного функционирования по общим характеристикам является кластерный анализ [1, 2]. С использованием компьютерной техники кластерный анализ является одним направлений статистической науки.

Основной задачей кластерного анализа является определение групп схожих объектов в выборке данных *(кластеров)*. Сходство количественных данных оценивается на основе понятия метрики при определении точки пространства на основе метрического расстояния между ними. Причем размерность пространства определяется числом характеристик, которые описывают объект [5].

Группа схожих объектов при выборке данных использует следующие кластерные методы:

– иерархические алгомеративные и дивизимные методы;
– итеративные методы группировки;
– факторные методы;
– поиск модальных значений плотности;
– сгущений;
– использования теории графов.

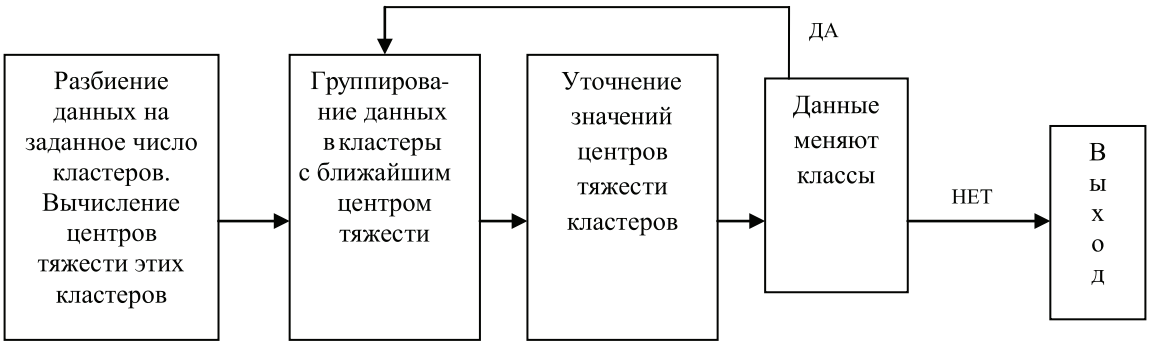
Применение различных методов к одним и тем же объектам может привести к сильно различающимся результатам.

Итеративные методы используют первичные данные, т.е. вычисления и хранения матрицы сходств между объектами, которые не требуется хранить. Следовательно, итеративные методы группировки позволяют обрабатывать большое множество при этом, осуществляют несколько просмотров данных и поэтому могут компенсировать последствия плохого исходного разбиения данных, что позволяет исключить главный недостаток иерархических алгомеративных методов. Данные методы формируют кластеры одного ранга, которые не могут быть вложенными, а следовательно, не могут быть частью иерархии и к тому же они не допускают перекрытия этих кластеров [2, 4].

Использование вышеописанных методов позволяет ограничить число итераций

при определении принадлежности рассматриваемых объектов к тому или иному классу. В данном случае определяется минимальная совокупность объектов, переходящих из класса в класс, и тогда итерационный процесс прекращается и с определенной точностью рассматриваемые объекты объединены в кластеры.

На рисунке представлен процесс группирования данных итеративными методами.



Процесс группирования данных итеративными методами

Рассмотрим применение итеративного метода кластерного анализа к процессу группирования объектов базовой информации для информационных процессов сложных систем на основе значений параметров анализируемых характеристик [6].

Представим базовую информацию, циркулирующую в системе, в виде множеств

$$\{P_{i,j}, A_i, B_j\},$$

где $i = \overline{1,I}$ – индекс объектов, носителя первичной информации; $j = \overline{1,J}$ – индекс выбранных или всех характеристик объектов; $P_{i,j}$ – количественное значение j-й характеристики i-го объекта; A_i – наименование i-го объекта; B_j – наименование j-й характеристики.

Все элементы $P_{i,j}$ необходимо с точностью T_I разбить на $K_э$ классов.

Значение T_I определяет количество итераций. При увеличении количество итераций уменьшается количество шагов, но в то же время уменьшается точность разбиения, определяемая ЛПР. Значение $K_э$ также определяется ЛПР в зависимости от требуемой точности получения классификации разбиения (меньше $K_э$ – грубее классификация) [3, 7].

Чтобы в процессе исходного разбиения получить требуемое количество классов не меньше величины $K_э$ *необходимо* ввести масштабный коэффициент L. Масштабному коэффициенту L вначале присваивается значение равное 1.0. В том случае, если увеличить количество классов, L уменьшается и наоборот.

Алгоритм разбиения на классы представлен ниже:

0. Для формирования исходного разбиения на нужное число классов необходимо:

Сформировать множество $\{r_i, d_i, a_i, k_i\}$ размерностью $I = 1,I$:

r_i – смешанный момент корреляции Карла Пирсона или угловая мера

$$r_i = \sum_{j=1}^{J}\left(P_{i,j} - \overline{P}_{i,j}\right) \Big/ \sqrt{\sum_{j=1}^{J}\left(P_{i,j} - \overline{P_{i,j}}\right)^2},$$

где $\overline{P}_{i,j} = \frac{1}{J}\sum_{j=1}^{J} P_{i,j}$;

d_i – евклидово расстояние от начала координат до $P_{i,j}$

$$d_i = \sqrt{\sum_{j=1}^{J}\left(0 - P_{i,j}\right)^2};$$

a_i – индекс объекта в соответствии с $P_{i,j}$; k_i – номер класса, к которому будет принадлежать i-й объект, первоначально все $k_i = 0$.

1. Первоначальное разбиение на классы.

1.1. Для начала итеративного процесса:

– первоначально $C_k = 0, k_1 = 1, i = 1, C_i = 1$;

– вычисляется среднее расстояние s между всеми элементами d_i:

$$s = \sum_{i=1}^{I-1}\left(d_i - d_{i+1}\right) \Big/ I -.$$

1.2. Вычисляется пороговое значение α, по которому определяется принадлежность i-го объекта к C_k классу:

– $\alpha = s \cdot L$;

– $i = C_i$.

1.3. Вычисляется расстояние между очередным элементом и следующим:

$$\Delta d = d_i - d_{(i+1)}.$$

1.4. Проверяется принадлежность $i + 1$ – го объекта к классу C_k.

Если $\Delta d \le \alpha$, то $k_{(i+1)} = C_k$, $i = i + 1$ и:

– если $i \le I$,то переход к пункту 1.3;

– если $i > I$, то переход к пункту 1.5.

Если $\Delta d > \alpha$, то $C_k = C_k + 1$, $i = i + 1$, $Ci = i$ и переход к пункту 1.2.

1.5. Объединяются с использованием смешанного момента корреляции Карла Пирсона r_i.

1.6. В начале:

– элементы $\{r_i, k_i\}$ переупорядочиваются по возрастанию элементов k_i и r_i соответственно;

– первоначально определяются $C_k = 1$, $C_{kt} = 1$, $k_1 = 1$, $i = 1$, $C_i = 1$.

1.7. Вычисляется пороговое значение α, по которому определяется принадлежность i + объекта C_i классу:

$$\alpha = (r_i - r_{(i+1)}) \cdot L.$$

Если $\alpha = 0$, то $i = i + 1$ и а вычисляется заново.

Если $|\alpha| > 0$, то $\alpha = |\alpha|$ и $i = C_i$.

1.8. Проверяется, закончились ли объекты C_{kt} класса.

Если $C_{kt} = k_i$, то переход к пункту 1.9.

Если$C_{kt} \ne k_i$, то $C_{kt} = C_{kt} + 1$, $C_k = C_k + 1$, $i = i + 1$, $C_i = I$ и:

– если $i > I$, то переход к пункту 1.11;

– если $i \le I$, то переход к пункту 1.7.

1.9. Вычисляется расстояние между очередным и следующим элементами

$$\Delta r = r_i - r_{(i+1)}.$$

1.10. Проверяется принадлежность i + объекта к C_k классу:

Если $\Delta r \le \alpha$, то $k_{(i+1)} = C_k$, $i = i + 1$, и:

– если $i \le I$, то переход к пункту 1.9;

– если $i > I$, то переход к пункту 1.11.

Если $\Delta d > \alpha$, то $C_k = C_k + 1$, $i = i + 1$, $C_i = i$ и переход к пункту 1.7.

1.11. Результаты $P_{i,j}$ разбиты на «К» классов.

Если $K = K_э$, то требуемое разбиение получено и переход к пункту 2 [1].

2. Если $K > K_э$, то увеличивается параметр L и переход к пункту 1.1. Если $K < K_э$, то уменьшается параметр L и переход к пункту 1.1.

3. Вычисляются $\overline{P_{k,j}}$ – центры тяжести полученных классов:

$$\overline{P}_{k,j} = \sum_{(1,k_{k_i})}^{I} P_{a_i j} \Bigg/ \sum_{(1,k_{k_i})}^{I} 1, \quad k = \overline{1, K_э}$$ – индекс полученных классов.

4. Проверяется, находится ли каждый объект в ближайшем классе.

4.1. Первоначально $i = 1$, $n = 0$.

4.2. Вычисляется квадрат отклонения объекта a_i от центра тяжести всех классов:

$$F_{ka_i} = \sum_{j=1}^{J} \left(P_{a_i,j} - \overline{P_{k,j}} \right)^2,$$

где $k = \overline{1, K_э}$ – индекс полученных классов; $j = \overline{1,J}$ – индекс характеристики, участвовавшей в формировании результата $P_{i,j}a_i$ объекта.

4.3. Если $\min(F_{kai})$ достигается при $k = k_i$, то объект a_i находится в ближайшем классе, изменения его класса не происходит.

Если $\min(F_{kai})$ достигается при $k \ne k_i$, то объект a_i не находится в ближайшем классе, поэтому $k_i = k$ (класс объекта заменился на ближайший) и $n = n + 1$ (объект a_i перешел в другой класс).

4.4. Увеличивается $i = i + 1$ и проверяется:

– если $i > I$, то закончился просмотр всех объектов и переход к пункту 5;

– если $i \le I$, то переход к пункту 4.2.

5. Если $\frac{n}{I} \cdot 100 > T_I$, то требуемая точность итеративного процесса не достигнута и переход к пункту 2.

Если $\frac{n}{I} \cdot 100 \le T_I$, то требуемая точность итеративного процесса достигнута. Получено окончательное разбиение $P_{i,j}$ по классам.

К исходным данными разбиения объектов на классы относятся: $i = \overline{1,I}$ – индекс объекта; $j = \overline{1,J}$ – индекс характеристики объекта; $P_{i,j}$ – количественное значение j-й характеристики i-го; A_i – наименование i-го объекта; B_j– наименование j-й характеристики; T_I – точность разбиения в процентах; $K_э$ – количество классов разбиения.

Результатом разбиения объектов на классы являются: K – количество полученных классов; $\overline{P_{k,j}}$ – центры тяжести полученных классов; k_i – номер класса, к которому принадлежит i-й объект; a_i – индекс объекта в соответствии с $P_{i,j}$.

Список литературы

1. Вагин В.Н. Головина Е.Ю. Достоверный и правдоподобный вывод в интеллектуальных системах. – М. :Физматлит, 2004.

2. Жиляков Е.Г., Ломазов В.А., Ломазова В.И. Компьютерная кластеризация совокупности аддитивных математических моделей взаимосвязанных процессов // Вопросы радиоэлектроники. Сер. ЭВТ. – 2011. – Вып. 1. – С. 115–119.

3. Журавлев Ю.И., Рязанов В.В., Сенько О.В. «Распознавание». Математические методы. Программная система. Практические применения. – М.: Фазис, 2006.

4. Жиляков Е.Г., Скубилин В.В. О некоторых моделях краткосрочного прогнозирования // Научные ведомости Белгород. гос. ун-та. Сер. История Политология Экономика Информатика. – 2013. – № 22 (165). – Вып. 28/1.

5. Мендель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с

6. Сумин В.И., Смоленцева Т.Е. Моделирование обучения с использованием временных рядов наблюдений: монография. – М.: Издательско-полиграфический центр «Научная книга», 2014. – 104 с.

7. Сумин В. И., Цветков В.В. Об алгоритмах и моделях, данных в решениях задач принятия решения // Научные ведомости Белгород. гос. ун-та. Сер. История Политология Экономика Информатика. – 2010. – № 13 (84). – Вып. 15/1. – С. 120–128.

References

1. Vagin V.N. Golovina E.Ju. Dostovernyj i pravdopodobnyj vyvod v intellektual'nyh sistemah. M. :Fizmatlit, 2004.

2. Zhiljakov E.G., Lomazov V.A., Lomazova V.I. Komp'juternaja klasterizacija sovokupnosti additivnyh matematicheskih modelej vzaimosvjazannyh processov // Voprosy radiojelektroniki. Ser. JeVT. 2011. Vyp. 1. pp. 115–119.

3. Zhuravlev Ju.I., Rjazanov V.V., Sen'ko O.V. «Raspoznavanie». Matematicheskie metody. Programmnaja sistema. Prakticheskie primenenija. M.: Fazis, 2006.

4. Zhiljakov E.G., Skubilin V.V. O nekotoryh modeljah kratkosrochnogo prognozirovanija // Nauchnye vedomosti Belgorod. gos. un-ta. Ser. Istorija Politologija Jekonomika Informatika. 2013. no. 22 (165). Vyp. 28/1.

5. Mendel', I.D. Klasternyj analiz. M.: Finansy i statistika, 1988. 176 p.

6. Sumin V.I., Smolenceva T.E. Modelirovanie obuchenija s ispol'zovaniem vremennyh rjadov nabljudenij: monografija. M.: Izdatel'sko-poligraficheskij centr «Nauchnaja kniga», 2014. 104 p.

7. Sumin V. I., Cvetkov V.V. Ob algoritmah i modeljah, dannyh v reshenijah zadach prinjatija reshenija // Nauchnye vedomosti Belgorod. gos. un-ta. Ser. Istorija Politologija Jekonomika Informatika. 2010. no. 13 (84). Vyp. 15/1. pp. 120–128.

Рецензенты:

Филатов Г.Ф., д.ф.-м.н., профессор кафедры математики Военного учебного научного центра Военно-воздушных сил, «Военно-воздушная академия имени профессора Н.Е. Жуковского и Ю.А. Гагарина», г. Воронеж;

Чопоров О.Н., д.т.н., профессор, проректор по научной работе Воронежского института высоких технологий, г. Воронеж.

Работа поступила в редакцию 02.03.2015.