СИСТЕМНЫЙ АНАЛИЗ И УПРАВЛЕНИЕ

УДК 004.942

МЕТОДИКА ГРУППИРОВАНИЯ БАЗОВОЙ ИНФОРМАЦИИ ДЛЯ ИНФОРМАЦИОННЫХ ПРОЦЕССОВ СЛОЖНЫХ СИСТЕМ

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

В.И. Сумин[1], Т.Е. Смоленцева[2]
V.I.Sumin[1], T.E. Smolentceva[2]

[1] ФКОУ ВПО Воронежский институт ФСИН России, Россия, 394072, Воронеж , ул. Иркутская, 1-а
[2] ФГБОУ ВПО «Липецкий государственный технический университет», Россия, 398600, Липецк, ул. Московская, 30
[1] FSE VPO Voronezh Institute of the Federal penitentiary service of Russia, 1-A Irkutsk St., Voronezh, 394072, Russia
[2] FGBOU VPO "Lipetsk state technical University, 30 Moscow St., Lipetsk, 398600, Russia

e-mail: viktorsumin51@yandex.ru

Аннотация. 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.

Resume. 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.

Ключевые слова: итеративный метод кластерного анализа, базовая информация, центр тяжести кластеров.
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 наиболее схожие по своим характеристикам объекты.

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

Основная цель кластерного анализа – определение групп схожих объектов в выборке данных (кластеров). Сходство количественных данных оценивается на основе понятия метрики при определении точки координатного пространства на основе метрического расстояниями между ними. Причем размерность пространства определяется числом характеристик, которые описывают объект [5].

Групп схожих объектов в выборке данных использует следующие кластерные методы:
- иерархические алгомеративные и дивизимные методы;
- итеративные методы группировки;
- поиск модальных значений плотности;
- факторные методы;
- сгущений;

– использования теории графов.

Использование 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 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. Эти 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 этих методов позволяет ограничить число итераций при определении принадлежности рассматриваемых объектов к тому или другому классу. В этом случае определяется минимальная совокупность объектов, переходящих из класса в класс и тогда итерационный процесс прекращается и с определенной точностью рассматриваемые объекты объединены в кластеры.

Процесс группирования данных итеративных методов приведен на рис. 1:

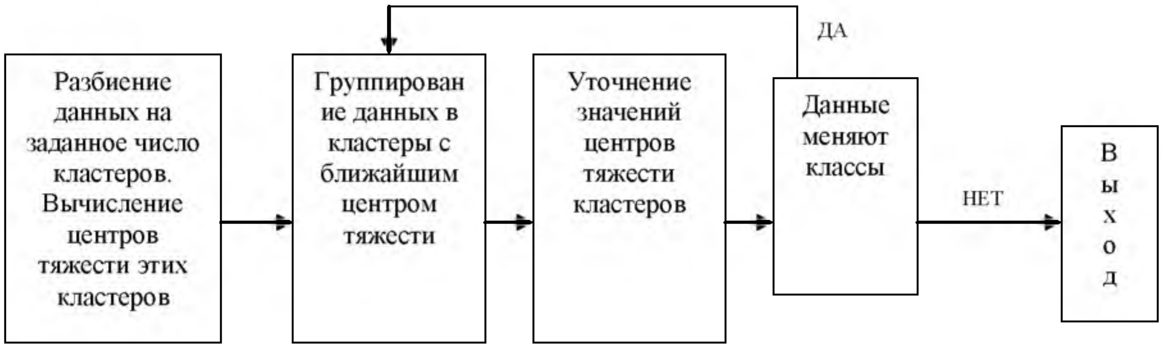

Рис. 1. Процесс группирования данных итеративными методами
Fig. 1. The process of grouping data by iterative methods

Рассмотрим применение итеративного метода кластерного анализа к группированию объектов базовой информации для информационных процессов сложных систем на основе значений 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-ой характеристики i-го объекта;

A_i - наименование i-го объекта;

B_j – наименование j-ой 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 тоже время уменьшается точность разбиения, определяемая лицом принимающих решения (ЛПР). Значение $K_э$ также определяется ЛПР в зависимости от требуемой точности получения классификации разбиения (меньше $K_э$ - грубее классификация) [3,7]. В том случае, 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 L количество классов уменьшается и наоборот.

Алгоритм разбиения на классы представлен ниже:

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

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

r_i -смешанный момент корреляции Карла Пирсона или угловая мера

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

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 $P_{i,j}$;

ki - номер класса, к которому будет принадлежать i-ый объект, первоначально все $к_i = 0$.

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

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

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

- α = s × 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$\Delta d \leq \alpha$, то$k_{(i+1)} = C_k$, $i = i + 1$и:

- если $i \leq I$,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:

α = $(r_i - r_{(i+1)}) \times L$.

Если α = 0, то $i=i+1$ и a 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$C_{kt} \neq k_i$, 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 $i \leq I$, то переход к пункту 1.7.

1.9. Вычисляется расстояние между очередным элементом и следующим $\Delta r = r_i - r_{(i+1)}$/

1.10. Проверяется принадлежность i+ объекта к C_k классу:

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

- если $i \leq I$, 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 $K=Kэ$, 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 $K<Kэ$, то уменьшается параметр L и переход к пункту 1.1.

2. Вычисляются $\overline{P_{i,j}}$ - центры тяжести полученных классов:

$$\overline{P}_{k,j} = \sum_{(1,k_{k_i})}^{I} P_{a,j} / \sum_{(1,k_{k_i})}^{I} 1 \quad k = \overline{1, K_Э}$$ - индекс полученных классов.

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

3.1. Первоначально i = 1, n=0.

3.2. Вычисляется квадрат отклонения объекта a_iот центра тяжести всех классов:

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

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.

3.3. Если min(F_{kai}) достигается при k = k_i, то объект a_iнаходится в ближайшем классе, изменения его класса не происходит.

Если min(F_{kai}) достигается при $k \neq k_i$, 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).

3.4. Увеличивается i = i +1 и проверяется:

- если i>I, то закончился просмотр всех объектов и переход к пункту 4;
- если $i \leq I$, то переход к пункту 3.2.

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

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

Исходными данными разбиения объектов на классы являются:

$i = \overline{1, I}$ - индекс объекта;

$j = \overline{1, J}$ - индекс характеристики объекта;

$P_{i,j}$ - количественное значение j-ой характеристики i-го;

A_i- наименование i-го объекта;

B_j– наименование j-ой характеристики;

T_I - требуемая точность разбиения в процентах;

$K_Э$ - требуемое 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-ый объект;

a_i - индекс объекта в соответствии с $P_{i,j}$.

Список литературы
References

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

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

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

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

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

4. Жиляков, Е.Г. О некоторых моделях краткосрочного прогнозирования / Е.Г. Жиляков, В.В. Скубилин // Научные ведомости БелГУ. Сер. История. Политология. Экономика. Информатика. – 2013. – № 22 (165). – Вып. 28/1. – С. 144–147.

Zhilyakov, E.G. O nekotoryh modeljah kratkosrochnogo prognozirovanija / E.G. Zhilyakov, V.V. Skubilin // Nauchnye vedomosti BelGU. Ser. Istorija. Politologija. Jekonomika. Informatika. – 2013. – № 22 (165). – Vyp. 28/1. – S. 144–147.

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

Mendel', I.D. Klasternyj analiz / I.D. Mendel'. – M. : Finansy i statistika, 1988. – 176 s.

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

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

7. Цветков, В.В. Об алгоритмах и моделях, данных в решениях задач принятия решения / В.В. Цветков, В.И. Сумин // Научные ведомости БелГУ. Сер. История. Политология. Экономика. Информатика. – 2010. – № 13 (84). – Вып. 15/1. – С. 138–142.

Cvetkov, V.V. Ob algoritmah i modeljah, dannyh v reshenijah zadach prinjatija reshenija / V.V. Cvetkov, V.I. Sumin // Nauchnye vedomosti BelGU. Ser. Istorija. Politologija. Jekonomika. Informatika. – 2010. – № 13 (84). – Vyp. 15/1. – S. 138–142.