Букачев Д.С. Об одном методе многофакторной кластеризации объектов // Международный журнал информационных технологий и энергоэффективности. – 2016. – Т.1 №1 с. 31-33

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.021

ОБ ОДНОМ МЕТОДЕ МНОГОФАКТОРНОЙ КЛАСТЕРИЗАЦИИ ОБЪЕКТОВ

Букачев Д.С.
ФГБОУ ВПО Смоленский государственный университет, Смоленск, Россия (21400, г. Смоленск, ул. Пржевальского, 4), e-mail: dsbuka@yandex.ru

Настоящая статья посвящена построению конструктивного алгоритма многофакторного разбиения множества объектов на группы в соответствии со схемой иерархической классификации. В результате функционирования алгоритма определяется количество кластеров и групповая принадлежность каждого элемента.

Ключевые слова: кластеризация, метрика, диаметр кластера.

ABOUT ONE MULTI-FACTOR CLUSTERING OBJECTS METHOD

Bukachev D.S.
Federal State Educational Institution of Higher Professional Education Smolensk State University, Smolensk, Russia (21400, Smolensk, street Przewalski, 4), e-mail: dsbuka@yandex.ru

This article is dedicated to building a constructive algorithm for multivariate partitioning a set of objects into groups in accordance with the hierarchical classification scheme. As a result of the operation of the algorithm determines the number of clusters and group affiliation of each element.

Keywords: clustering, metric, diameter of the cluster.

Спектр применений кластерного анализа весьма широк [1-3]: его используют в археологии, медицине, психологии, химии, биологии, государственном управлении, филологии, антропологии, маркетинге, социологии и других дисциплинах. В данная статья предлагается конструктивный алгоритм многофакторного разбиения множества объектов на группы в соответствии со схемой иерархической классификации.

При реализации алгоритма используются следующие входные данные:

- количество объектов n, их характеристики $X_i \in \Omega_i$ $(i = \overline{1, m})$, упорядоченные по убыванию значимости;
- метрики ρ_i, определенные на множествах Ω_i;
- допустимые метрические диаметры $\rho_{i\max}$ кластера по каждой из метрик ρ_i.

В результате функционирования алгоритма по входным данным определяется набор выходных значений следующих параметров:

- N_g – количество кластеров;

- g_k – номер кластера объекта с номером k, $k = \overline{1, n}$.

Описание алгоритма

Шаг 1. Будем считать, что изначально объекты принадлежат одной группе. Устанавливаем начальные значения выходных параметров: $N_g = 1,\ g_k = 1$ ($k = \overline{1, n}$).

Шаг 2. Последовательное разбиение объектов по каждой из характеристик. Введем в рассмотрение массив меток *L* длиной *n*.

Для всех $n_\rho = \overline{1, m}$:

Для всех $n_g = \overline{1, N_g}$:

Шаг 2.1. Первичная кластеризация группы n_g (проверка парной совместимости).

1. Формируем массив *M* номеров объектов, принадлежащих группе с номером n_g.

2. Если количество *K* элементов массива *M* больше 1, переходим к пункту 2.1.3, в противном случае рассматриваем следующую группу.

3. Присваиваем L_j ($j = \overline{1, n}$) начальное значение 0. Переменной r (количество образованных подкластеров в процессе разбиения) также присваиваем начальное значение 0.

4. Для всех $i_1,\ i_2 = \overline{1, K},\ i_2 > i_1$:

   Если $\rho_{n_\rho}(M_{i1}, M_{i2}) \le \rho_{n_\rho \max}$:

   1. $L_{\min} = Min(L_{M_{i_1}}, L_{M_{i_2}});\ L_{\max} = Max(L_{M_{i_1}}, L_{M_{i_2}})$;
   2. если $L_{\max} = 0$, то $r = r + 1;\ l = r$;
   3. если $((L_{\min} = 0) \wedge (L_{\max} > 0)) \vee ((L_{\min} > 0) \wedge (L_{\min} = L_{\max}))$, то $l = L_{\max}$;
   4. если $(L_{\min} > 0) \wedge (L_{\min} \ne L_{\max})$, то
      a. для всех $j = \overline{1, K}$:
         если $L_{M_j} \ge L_{\max}$, то
            если $L_{M_j} = L_{\max}$, то $L_{M_j} = L_{\min}$,
            иначе $L_{M_j} = L_{M_j} - 1$;
      b. $r = r - 1;\ l = L_{\min}$;
   5. $L_{M_{i_1}} = l;\ L_{M_{i_2}} = l$.

5. Для всех $j = \overline{1, K}$:

если $L_{M_j} = 0$: $r = r + 1;\ L_{M_j} = r$.

Шаг 2.2. Вторичная кластеризация (обеспечение групповой совместимости).

Для всех $i = \overline{1, r}$:

1. Формируем массив *M* номеров объектов группы с номером n_g, для которых соответствующее значение метки *L* равно *i* (то есть группируем элементы, относящиеся к подклассу с номером *i*, образованному при первичной кластеризации).
2. Если количество *K* элементов массива *M* больше 1, переходим к пункту 2.2.3, в противном случае рассматриваем следующий подкласс.

3. Логической переменной *TrueCluster* присваиваем значение (*K*<3); переменным *nL* и *nR* присваиваем 1 и 2 соответственно.
4. Пока не *TrueCluster*:
   1. $R_{\max} = 0$;
   2. Для всех $i_1,\ i_2 = \overline{1, K},\ i_2 > i_1$:

      Если $L_{M_{i1}} = L_{M_{i2}}$:

      $\rho = \rho_{n_\rho}(M_{i1}, M_{i2})$;

      если $R_{\max} < \rho$, то $R_{\max} = \rho,\quad nL = i_1,\quad nR = i_2$;
   3. Если $R_{\max} = 0$, то

      $TrueCluster = "истина"$;

      иначе

      $TrueCluster = \left(\rho_{n_\rho}(M_{nL}, M_{nR}) \le \rho_{n_\rho \max}\right)$;
   4. Если не *TrueCluster*:
      a) $r = r + 1;\quad L_{M_{nR}} = r$;

      b) для всех $j = \overline{1, K},\ j \ne nL,\ j \ne nR,\ L_{M_j} = i$:

      $\rho_L = \rho_{n_\rho}(M_j, M_{nL})$;

      $\rho_R = \rho_{n_\rho}(M_j, M_{nR})$;

      если $\rho_R \le \rho_{\max}$:

      если $\rho_L \le \rho_{\max}$:

      если $\rho_R < \rho_L$, то $L_{M_j} = r$;

      иначе $L_{M_j} = r$;

Шаг 2.3. Если количество подклассов r, полученных при разбиении группы с номером n_g, больше 1, переопределяем характеристики N_g и g_k:

1. $N_g = N_g + r - 1$
2. Для всех $k = \overline{1, n}$:

   если $g_k > n_g$, то $g_k = g_k + r - 1$;

   если $g_k = n_g$, то $g_k = g_k + L_k - 1$.

Таким образом, получили значения заявленных выходных параметров: N_g (количество кластеров) и g_k (индекс принадлежности объекта), $1 \le k \le n$.

Предложенный конструктивный метод обеспечивает многофакторное разбиение множества объектов на группы в соответствии со схемой иерархической классификации.

Список литературы

1. Воронцов К.В. Алгоритмы кластеризации и многомерного шкалирования. Курс лекций. МГУ, 2007.
2. Котов А., Красильников Н. Кластеризация данных, 2006.

3. Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. 1999. Vol. 31, no. 3.

References

1. Voroncov K.V. Algoritmy klasterizacii i mnogomernogo shkalirovanija. Kurs lekcij. MSU, 2007.
2. Kotov A., Krasil'nikov N. Klasterizacija dannyh, 2006.
3. Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. 1999. Vol. 31, no. 3.