Математические структуры и моделирование
2000. Вып. 5, с.133–139.

УДК 519.24

ДИНАМИЧЕСКИЙ ПОДХОД К ЗАДАЧЕ КЛАСТЕРИЗАЦИИ

Н.Ф. Жихалкина

In this article the dynamic approach to the cluster-analyze problems is suggested. Algorithm which based on the gravitational laws is used to look for clusters.

При решении задач классификации наблюдений рассматриваются взаимосвязи большого числа признаков. Увеличение размерности и, как следствие, потеря обозримости результатов приводят к задачам, требующим сведения множества характеристик к небольшому ряду обобщающих итогов. Объединение признаков и замена их одним, искусственно построенным на их основе, легло в основу такого направления математической статистики, как многомерный анализ. Наиболее значимыми его разделами являются кластерный анализ (классификация объектов) и факторный анализ (исследование связей).

В общем случае под кластерным анализом понимается методология проведения классификации неоднородных статистических совокупностей.

Пусть множество $I = \{I_1, \ldots, I_N\}$ обозначает N объектов, принадлежащих одному классу α, $C = \{C_1, \ldots, C_k\}$ – множество наблюдаемых показателей или характеристик, которыми обладает каждый элемент из I. Характеристики C_j, $j = 1, \ldots, k$, как правило, являются количественными и называются измерениями. Пусть r_{ij} – результат измерения j-й характеристики объекта I_i, тогда $\vec{r}_i = [r_{ij}]$ – вектор измерений для i-го объекта. Таким образом, для множества объектов I имеется множество векторов измерений $\mathcal{R} = \{\vec{r}_1, \ldots, \vec{r}_N\}$. Это множество может быть представлено как N точек в k-мерном евклидовом пространстве $\mathbb{R}^k$.

Задача кластерного анализа заключается в том, чтобы на основе данных, содержащихся в множестве $\mathcal{R}$, разбить множество объектов I на s, $s < N$ кластеров (подмножеств) $\alpha_1, \ldots, \alpha_s$, так, чтобы $\forall j \in [1, \ldots, N]$ $\exists!\ d \in [1, \ldots, s]$: $I_i \subset \alpha_d$ и чтобы объекты, принадлежащие одному кластеру, были «сходными» или «однородными», а разным кластерам – «разнородными» [2]. Далее необходимо количественно определить понятия «сходства» и «разнородности». Для

E-mail: zhihal@univer.omsk.su
Омский государственный университет

этой цели используется функция расстояния или метрика [2, 6]. Таким образом, задача кластеризации представляет собой проблему выделения однородных групп объектов. Существует несколько подходов к ее решению [6]:

1. Вероятностно-статистический подход – выделение групп, каждая из которых представляет реализацию некоторой случайной величины.

2. Вариационный подход – разделение совокупности по некоторому признаку на группы в соответствии с определенными интервалами.

3. Структурный подход (далее, согласно [6], под кластерным анализом будем понимать именно это направление) – выделение компактных групп объектов, удаленных друг от друга, а также поиск «естественного разбиения» совокупности на области скопления объектов (визуализация данных). Этот подход используется для двух видов исходных данных: матриц близости или расстояния между объектами и объектов, представляющих точки в многомерном пространстве.

Методы решения задач кластерного анализа [2, 3, 5, 6] подразделяются на три группы [6]:

1. Процедуры прямой классификации (эвристический подход) – выделение кластеров с заранее заданными свойствами (например, среднее межточечное расстояние внутри кластера меньше среднего расстояния от данных точек до остальных). Имеется набор определений кластера. Искомое разбиение зависит от выбора конкретного определения. В рамках этого направления развиваются иерархические процедуры, алгоритмы диагонализации, эталонные процедуры [6] и др.

2. Оптимизационное направление – решением задачи является разбиение, удовлетворяющее некоторому критерию оптимальности. Этот критерий представляет собой функционал, называемый целевой функцией. Примером такого функционала является внутригрупповая сумма квадратов отклонений [2].

3. Аппроксимационные методы – отношения, заданные в исходных данных, требуется наилучшим образом аппроксимировать отношением, отвечающим требованию о классификации.

Среди методов прямой классификации можно выделить подкласс алгоритмов, в которых вначале задается *число* кластеров, а затем происходит *разделение* объектов [6]. Тем самым решение задачи кластеризации разбивается на два этапа:

1) определение числа кластеров;

2) распределение объектов между кластерами.

В задачах классификации в зависимости от природы как рассматриваемых объектов, так и наблюдаемых показателей, *проблема определения числа кластеров* может представлять отдельный интерес.

Для ее решения предлагается «гравитационный» метод, основанный на динамике системы точек в k-мерном евклидовом пространстве, координаты которых определяются векторами измерений $\vec{r}_i$, $i = 1, \ldots, N$. При этом в качестве меры близости или однородности объектов рассматривается евклидова метрика $d(\vec{r}_1, \vec{r}_2) = [\sum_{j=1}^{k}(r_{1j} - r_{2j})^2]^{1/2}$.

В целом, стандартные методы кластеризации характеризуются тем, что

множество данных, полученное в процессе наблюдений, статично. Существуют алгоритмы, в которых динамически меняется положение кластера (алгоритм эталонного типа «Форэль-1», эвристические приемы перемещения объектов в исходном пространстве [6] и др.). Оказывается, что привнесение динамики в систему исходных данных также может быть полезным как при определении числа кластеров, так и при последующей группировке объектов.

Предлагаемый метод кластеризации основан на аналогиях с задачей N тел [7], известной из механики, что послужило причиной выбора названия - «гравитационный». В методе также происходит преобразование координат k-мерного вектора, названное «отражением», которое представляет собой аналог оператора рекомбинации в эволюционных алгоритмах оптимизации [9].

Шаг 0: Будем рассматривать множество векторов измерений $\vec{r}_i$, $i = 1, \ldots, N$ как систему материальных точек с равными массами в пространстве $\mathbb{R}^k$, распределенных внутри некоторого k-мерного гиперкуба. Пусть n – номер итерации, $n := 0$.

Шаг 1: Происходит взаимодействие материальных точек под действием гравитационных сил на основе закона тяготения (задача N тел). Строится модель взаимодействия частиц без столкновений. Используя разностную схему «с перешагиванием», осуществляется переход от дифференциальных уравнений к конечно-разностной аппроксимации производных [7].

Алгоритм основан лишь на аналогиях с реальными процессами. При этом обязательным требованием остается прямая зависимость силы взаимодействия между частицами от их массы. Обратная зависимость этой силы от степени расстояния может варьироваться, что приводит к различным реализациям предлагаемого метода (в классической механике сила взаимного притяжения обратно пропорциональна расстоянию между частицами, возведенному в степень $(k-1)$, где $k > 2$ – размерность пространства). Вводится ряд ограничений. Во-первых, при вычислении координат берется среднее значение скорости за последние два шага по времени, во-вторых, начальные скорости частиц равны нулю, и это условие сохраняется на каждом шаге, тем самым скорости не накапливаются [4]. Окончательно, опираясь на законы движения Ньютона, формулы расчета координат взаимодействующих частиц при работе алгоритма принимают вид [4, 7]:

$$\begin{cases} \vec{a}_i{}^{n+1} = \sum_{j=1, j\neq i}^{N} \frac{m_j(\vec{r}_j{}^n - \vec{r}_i{}^n)}{|\vec{r}_j{}^n - \vec{r}_i{}^n|^{k_1}} \\ \vec{r}_i{}^{n+1} = \vec{r}_i{}^n + \frac{\lambda_i^n}{2}(\vec{a}_i{}^n + 2\vec{a}_i{}^{n+1}) \\ \vec{r}_i{}^0 = \vec{r}_{i_0}. \end{cases}$$

Настраиваемыми параметрами алгоритма являются:

1) «шаг по времени» λ_i^n;

2) радиус взаимодействия R_1, который либо задается на основе некоторых эвристических соображений, либо зависит от размерности пространства и числа взаимодействующих материальных точек [4]. Каждая точка $\vec{r}_i{}^n$ взаимодействует лишь с теми точками $\vec{r}_j{}^n$, которые попали в сферу с центром в $\vec{r}_i{}^n$ радиуса R_1;

3) $k_1 \in [1, \ldots, k]$.

Шаг 2: ($k \geq 2$) «Отражение» в координатной плоскости Oxy: x-координата точки вычисляется через y-координаты «соседей». Ось Ox выбирается случайным образом из k координатных осей используемой системы координат, ось Oy задается аналогично. Таким образом, выбор плоскости «отражения» происходит по схеме Бернулли с вероятностью успеха $p_{cross} = \frac{1}{k(k-1)}$ [8].

$$\begin{cases} x_i^{n+1} = x_i^n + \mu_i^n \sum_{j=1, j\neq i}^{N} \frac{m_j (y_j^n - y_i^n)}{||\vec{r}_j^n - \vec{r}_i^n||^{k_2}} \\ y_i^{n+1} = y_i^n - \mu_i^n \sum_{j=1, j\neq i}^{N} \frac{m_j (x_j^n - x_i^n)}{||\vec{r}_j^n - \vec{r}_i^n||^{k_2}}. \end{cases}$$

Как и на шаге 1, настраиваемыми параметрами являются μ_i^n, радиус взаимодействия R_2 и $k_2 \in [1, 2]$.

Шаг 3: $n := n + 1$. Переход на шаг 1. В общем случае в качестве критерия остановки задается число итераций.

Согласно классификации методов кластерного анализа, приведенной в [6], основными отличительными чертами предложенного метода являются:

– по характеру отношения, которое отыскивается как результат кластеризации: метод строит разбиение с пересекающимися кластерами;

– по степени участия человека в процедуре выделения кластеров: человек участвует в процессе построения разбиения;

– по объему необходимых априорных сведений для работы алгоритма (задание параметров): число кластеров неизвестно, но задано пороговое значение величины близости (радиус взаимодействия).

В программной реализации гравитационного метода каждая точка наделяется массой, равной единице ($m_i \equiv 1$). Важным параметром является радиус взаимодействия. Естественно предположить, что его величина должна быть мала в сравнении с линейными размерами рассматриваемой области, λ_i^n и μ_i^n не зависят ни от номера итерации, ни от вектора состояния и равны константам. Степени расстояния k_1, k_2 выбираются равными двум независимо от размерности задачи.

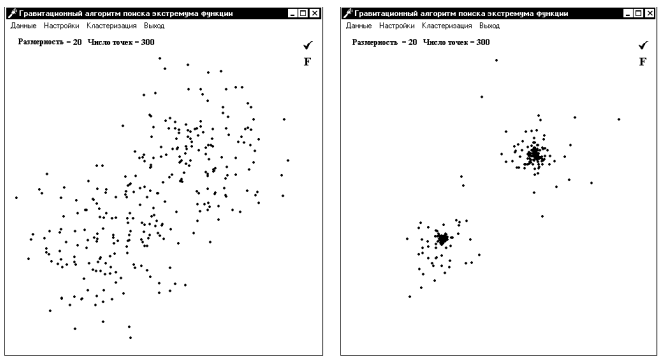

Рис. 1. Размерность пространства = 20, число точек = 300

Первый этап тестирования алгоритма проходил на выборках с покоординатным нормальным законом распределения [8]. На рисунках 1, 2, 3 представлены начальные (слева) и конечные (справа) конфигурации. В силу того,

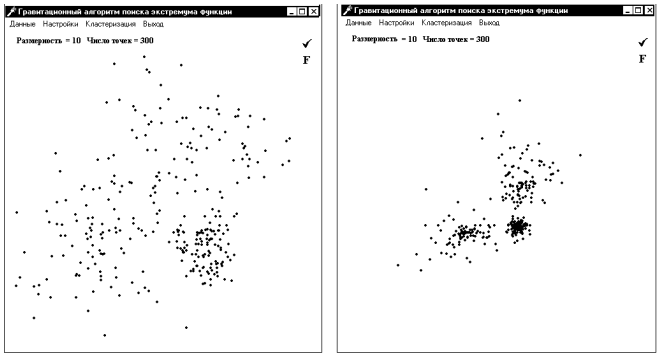

Рис. 2. Размерность пространства = 10, число точек = 300

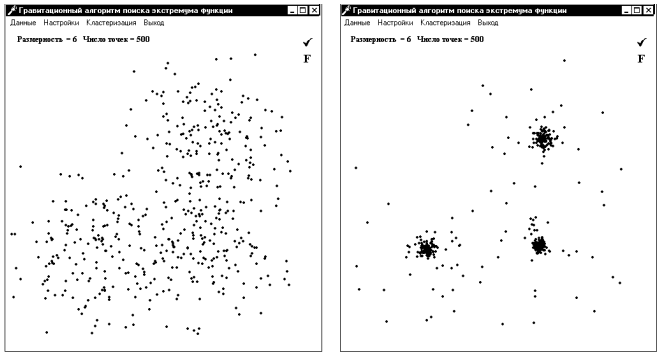

Рис. 3. Размерность пространства = 6, число точек = 500

что координаты точек имеют независимое распределение, результаты расчетов приведены для одной координатной плоскости. Использовалось несколько выборок (рис. 1 – две, рис. 2, 3 – три), отличающихся друг от друга математическим ожиданием и дисперсией. Число полученных кластеров совпадает с исходными данными, а центры областей скопления точек согласуются с математическими ожиданиями заданных распределений.

Дальнейшая апробация алгоритма проводилась на базе социологических данных обобщенной активности избирателей по различным параметрам для городов Надым и Омск [1]. Каждому избирательному участку отвечала материальная точка в пространстве, размерность которого совпадала с количеством оценочных параметров. Векторы наблюдений принадлежали отрезку $[0, 1]$, иными словами, область распределения точек представляла собой единичный гиперкуб. Отметим, что поиск кластеров производился как для полного объема данных, так и для различных проекций. В ходе численных экспериментов получены следующие результаты: для обоих городов (Омск – 447 избирательных участков, 15 параметров активности; Надым – 38 избирательных участков, 22 параметра активности) при достаточно малом шаге по времени (0.005) и зависимости радиуса взаимодействия от размерности пространства разделения на кластеры не наблюдается (рис. 4, 5), конечная конфигурация представляет собой один кластер со сгущением плотности в центре. Наличие одного кластера является подтверждающим фактором одномодальности рас-

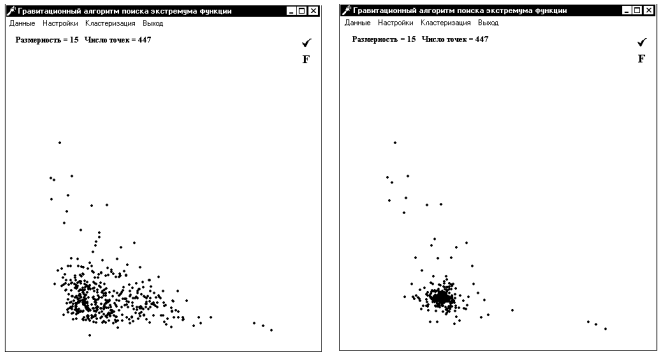

Рис. 4. Данные по Омску

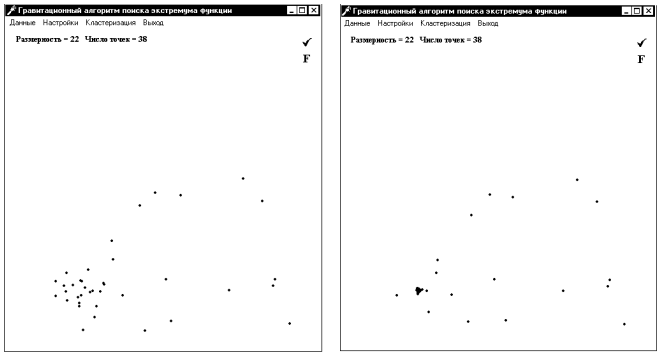

Рис. 5. Данные по Надыму

пределения исходной выборки [1].

Необходимо отметить, что на данном этапе разработки метода обработка полученных результатов (число кластеров, их взаимное расположение и т.д.) проводилась визуально. Подобная практика применяется на начальном шаге решения задач прямой классификации. В ряде случаев визуализация данных позволяет выявить некоторые регулярные структуры и получить априорную информацию об имеющихся кластерах. Кроме того, динамический подход к задаче кластеризации дает возможность проследить формирование новых кластеров в процессе эволюции системы, что может оказаться полезным при решении задач прогнозирования.

Литература

1. Гуц А.К., Файзуллин Р.Т. *Математическое исследование методики организации одного социального мониторинга* // Наст. сборник.

2. Дюран Б., Оделл П. *Кластерный анализ.* М.: Статистика, 1977. 128 с.

3. Жамбю. М. *Иерархический кластер – анализ и соответствия.* М.: Финансы и статистика, 1988. 342 с.

4. Жихалкина Н. Ф., Файзуллин Р. Т. *Гравитационные аналогии в задаче оптимизации* // Математические структуры и моделирование. 1998. Омск: ОмГУ. Вып. 2. С.60–76.

5. *Классификация и кластер* / Ред. Дж. Вэн Ройзин. М.: Мир, 1980.

6. Мандель И.Д. *Кластерный анализ.* М.: Финансы и статистика, 1988. 176 с.

7. Поттер П. *Вычислительные методы в физике.* М.: Мир, 1975. С.162–193.

8. Чистяков В. П. *Курс теории вероятностей.* М.: Наука, 1982. 156 с.

9. Reeves C. R. *Genetic Algorithms for the Operations Reseacher* // INFORMS Journal on Computing. V.9. N 3. 1997. P.231–250.