УДК 004.93'1; 004.932

АНСАМБЛЕВЫЙ АЛГОРИТМ КЛАСТЕРИЗАЦИИ БОЛЬШИХ МАССИВОВ ДАННЫХ*

И. А. Пестунов[1], В. Б. Бериков[2], Е. А. Куликова[1], С. А. Рылов[1]

[1] *Учреждение Российской академии наук*
Институт вычислительных технологий Сибирского отделения РАН,
630090, г. Новосибирск, просп. Академика Лаврентьева, 6
E-mail: pestunov@ict.nsc.ru
[2] *Учреждение Российской академии наук*
Институт математики им. С. Л. Соболева Сибирского отделения РАН,
630090, г. Новосибирск, просп. Академика Коптюга, 4

Предложен и теоретически обоснован ансамблевый алгоритм кластеризации ECCA (Ensemble of Combined Clustering Algorithms) для обработки больших массивов данных. Представлены результаты экспериментального исследования алгоритма на модельных и реальных данных, подтверждающие его эффективность.

Ключевые слова: ансамблевый алгоритм кластеризации, сеточный подход, большие массивы данных.

Введение. В последние годы усилия многих исследователей направлены на создание эффективных алгоритмов кластеризации для анализа больших массивов данных (генетических данных, многоспектральных изображений, интернет-данных и т. п.) [1, 2]. Потребность в таких алгоритмах непрерывно возрастает в связи со стремительным прогрессом в области создания средств и технологий автоматизированного получения и хранения данных, а также бурным развитием интернет-технологий.

Одним из наиболее эффективных подходов к кластеризации больших массивов данных является так называемый сеточный (grid-based) подход [3], при котором происходит переход от кластеризации отдельных объектов к кластеризации элементов сеточной структуры (ячеек или клеток), формируемой в пространстве признаков. В этом подходе предполагается, что все объекты, попавшие в клетку, принадлежат к одному кластеру. Поэтому процесс формирования сеточной структуры является важным этапом работы алгоритма.

По способам построения сеточной структуры алгоритмы кластеризации условно можно разделить на две группы [4]: с адаптивной и фиксированной сеткой.

Алгоритмы с адаптивной сеткой анализируют распределение данных, чтобы как можно точнее описать границы кластеров, образованных исходными объектами [5]. Адаптивная сетка позволяет уменьшить сеточный (граничный) эффект, но её построение сопряжено, как правило, со значительными вычислительными затратами.

Алгоритмы с фиксированной сеткой отличаются высокой вычислительной эффективностью, однако из-за сеточного эффекта качество кластеризации в большинстве случаев оказывается низким, а получаемые результаты — неустойчивыми, поскольку зависят от масштаба сетки. На практике эта неустойчивость существенно затрудняет настройку параметров алгоритма.

Для решения данной проблемы в последние годы активно развиваются сеточные методы, основанные на использовании не одной, а нескольких сеток с фиксированным шагом

*Работа выполнена при поддержке Российского фонда фундаментальных исследований (гранты № 11-07-00346-а, № 11-07-00202-а).

[6–8]. Основные трудности такого подхода заключаются в разработке метода комбинирования результатов, полученных на разных сетках, поскольку сформированные кластеры не всегда однозначно сопоставимы друг с другом. В [6] представлен алгоритм, который выполняет кластеризацию на последовательности сеток до тех пор, пока не будет получен повторяющийся (устойчивый) результат. В алгоритмах [7, 8] выполняются две кластеризации на сетках разного размера. Конечный результат формируется путём объединения пересекающихся кластеров, построенных по каждой из этих сеток.

В данной работе для повышения качества и устойчивости решений предлагается алгоритм кластеризации ECCA (Ensemble of Combined Clustering Algorithms), использующий ансамбль алгоритмов с фиксированными равномерными сетками, в котором итоговое коллективное решение строится на основе попарной классификации элементов сеточной структуры.

1. Постановка задачи. Пусть множество классифицируемых объектов X состоит из векторов, лежащих в пространстве признаков R^d: $X = \{x_i = (x_i^1, \ldots, x_i^d) \in R^d,\ i = \overline{1, N}\}$. Векторы x_i лежат в прямоугольном гиперпараллелепипеде $\Omega = [l^1, r^1] \times \ldots \times [l^d, r^d]$, где $l^j = \min\limits_{x_i \in X} x_i^j$, $r^j = \max\limits_{x_i \in X} x_i^j$. Под сеточной структурой будем понимать разбиение пространства признаков гиперплоскостями $x^j = (r^j - l^j)i/m + l^j$, $i = 0, \ldots, m$ (m — число разграниченных участков по каждой размерности). Минимальным элементом этой структуры является клетка (замкнутый прямоугольный гиперпараллелепипед, ограниченный гиперплоскостями). Введём общую нумерацию клеток (последовательно от одного слоя клеток к другому).

Клетки B_i и B_j $(i \neq j)$ являются смежными, если их пересечение не пусто. Множество смежных с B клеток обозначим через A_B. Плотностью D_B клетки B назовём отношение $D_B = N_B/V_B$, где N_B — количество элементов множества X, попавших в клетку B; V_B — объём клетки B. Клетку B будем считать непустой, если $D_B \geq \tau$, где τ — величина заданного порога. Все точки множества X, попавшие в клетки с плотностью меньше τ, будем относить к «шуму». Обозначим множество всех непустых клеток через $\aleph$. Непустая клетка B_i непосредственно связна с непустой клеткой B_j $(B_i \to B_j)$, если B_j — максимальная по номеру клетка, удовлетворяющая условиям $B_j = \arg\max\limits_{B_k \in A_{B_i}} D_{B_k}$ и $D_{B_j} \geq D_{B_i}$. Непустые клетки B_i и B_j непосредственно связны $(B_i \rightleftarrows B_j)$, если $B_i \to B_j$ или $B_j \to B_i$. Непустые клетки B_i и B_j связны $(B_i \sim B_j)$, если существуют $k_1, \ldots, k_l$ такие, что $k_1 = i$, $k_l = j$ и для всех $p = 1, \ldots, l-1$ выполнено $B_{k_p} \rightleftarrows B_{k_{p+1}}$.

Введение отношения связности порождает естественное разбиение множества непустых клеток на компоненты связности $\{G_i,\ i = 1, \ldots, S\}$. Под компонентой связности будем понимать максимальное множество попарно связных клеток. Представителем компоненты связности G назовём клетку $Y(G)$, удовлетворяющую условию $Y(G) = \arg\max\limits_{B \in G} D_B$ (если несколько клеток удовлетворяют данному условию, то $Y(G)$ выбирается из них случайным образом). Компоненты связности G' и G'' смежные, если существуют смежные клетки B' и B'' такие, что $B' \in G'$ и $B'' \in G''$. Смежные компоненты связности G_i и G_j связны $(G_i \sim G_j)$, если существует набор клеток (путь) $P_{ij} = \{Y_i = B_{k_1}, \ldots, B_{k_t}, \ldots, B_{k_l} = Y_j\}$ такой, что:

1) для всех $t = 1, \ldots, l-1$ клетка $B_{k_t} \in G_i \cup G_j$ и B_{k_t}, $B_{k_{t+1}}$ — смежные клетки;

2) $\min\limits_{B_{k_t} \in P_{ij}} D_{B_{k_t}} / \min(D_{B_{Y_i}}, D_{B_{Y_j}}) > T$, $T > 0$, — порог объединения.

Определение. Кластером C назовём максимальное множество попарно связных компонент связности: 1) для любых компонент связности G_i, $G_j \in C$ выполнено $G_i \sim G_j$, 2) для любых $G_i \in C$, $G_j \notin C$ верно $G_i \not\sim G_j$.

С учётом изложенного задача кластеризации заключается в разбиении множества $\aleph$ на совокупность кластеров $\{C_i,\ i = 1, \ldots, M\}$ таких, что $\aleph = \bigcup\limits_{i=1}^{M} C_i$ и $C_i \cap C_j = \oslash$ при $i \neq j$; число кластеров M заранее неизвестно.

Далее опишем эффективный метод решения этой задачи, основанный на ансамблевом подходе.

2. Описание метода. Предлагаемый метод опирается на сеточный алгоритм $\mathrm{CCA}(m, T, \tau)$ [9], где m — число разбиений, T — порог объединения компонент связности, τ — порог шума. В работе этого алгоритма можно выделить три основных этапа.

1. Формирование клеточной структуры. На этом этапе для каждой точки $x_i \in X$ определяется содержащая её клетка, вычисляются плотности D_B всех клеток и выявляются непустые клетки.

2. Выделение компонент связности $G_1, \ldots, G_S$ и поиск их представителей $Y(G_1), \ldots, Y(G_S)$.

3. Формирование кластеров $C_1, \ldots, C_M$ в соответствии с указанным выше определением на основе выделенных компонент связности.

Алгоритм $\mathrm{CCA}(m, T, \tau)$ является вычислительно эффективным в пространстве признаков небольшой размерности ($\leqslant 6$) [9], его сложность составляет $O(dN + dm^d)$, где N — число классифицируемых объектов, d — размерность пространства признаков.

Однако CCA относится к классу алгоритмов с фиксированной сеткой, поэтому результаты его работы существенно зависят от параметра m, который определяет масштаб элементов сеточной структуры. На практике эта неустойчивость результатов значительно затрудняет настройку параметров алгоритма.

Известно [10–12], что устойчивость решений в задачах кластеризации может быть повышена благодаря формированию ансамбля алгоритмов и построению на его основе коллективного решения. При этом используются результаты, полученные разными алгоритмами либо одним алгоритмом с различными значениями параметров. Кроме того, для формирования ансамбля могут быть применены разные подсистемы переменных. Ансамблевый подход является одним из наиболее перспективных направлений в кластерном анализе [1].

В данной работе для формирования ансамбля предлагается использовать результаты выполнения алгоритма $\mathrm{CCA}(m, T, \tau)$ с различными значениями параметра m, а для формирования итогового коллективного решения — применить метод, основанный на нахождении согласованной матрицы подобия (или различия) объектов [13]. Этот метод может быть описан следующим образом.

Пусть с помощью некоторого алгоритма кластеризации ($\mu = \mu(\Theta)$), зависящего от случайного вектора параметров $\Theta \in \mathbf{\Theta}$ (здесь $\mathbf{\Theta}$ — некоторое допустимое множество параметров), получен набор частных решений $\mathbb{Q} = \{Q^{(1)}, \ldots, Q^{(l)}, \ldots, Q^{(L)}\}$, где $Q^{(l)}$ — l-й вариант кластеризации, содержащий $M^{(l)}$ кластеров.

Обозначим через $H(\Theta_l)$ бинарную матрицу $H(\Theta_l) = \{H_{i,\,j}(\Theta_l)\}$ размера $N \times N$, которая для l-й группировки вводится как

$$H_{i,\,j}(\Theta_l) = \begin{cases} 0, & \text{если объекты отнесены в один кластер,} \\ 1 & \text{иначе.} \end{cases}$$

После построения L частных решений можно сформировать согласованную матрицу различий

$$\mathbf{H} = \{\mathbf{H}_{i,\,j}\}, \quad \mathbf{H}_{i,\,j} = \frac{1}{L}\sum_{l=1}^{L} H_{i,\,j}(\Theta_l),$$

где $i,\ j = 1, \ldots, N$. Величина $\mathbf{H}_{i,\,j}$ равна частоте классификации x_i и x_j в разные группы в наборе группировок $\mathbb{Q}$. Близкое к нулю значение величины означает, что данные объекты имеют большой шанс попадания в одну и ту же группу. Близкое к единице значение этой величины свидетельствует о том, что шанс оказаться в одной группе у объектов незначителен.

В нашем случае $\mu = \mathrm{CCA}(m, T, \tau)$, где число разбиений $m \in \{m_{\min},\, m_{\min} + 1, \ldots$ $\ldots,\, m_{\min} + L\}$, а объектами классификации будут представители компонент связности $Y(G_1), \ldots, Y(G_S)$.

После вычисления согласованной матрицы различий для нахождения коллективного решения применим стандартный агломеративный метод построения дендрограммы, который в качестве входной информации использует попарные расстояния между объектами [14]. При этом расстояния между группами будем определять по принципу «средней связи», т. е. как среднее арифметическое попарных расстояний между объектами, входящими в группы. Процесс объединения продолжается до тех пор, пока расстояние между ближайшими группами не превысит заданное пороговое значение T_d, принадлежащее отрезку [0, 1]. Этот метод позволяет выделять иерархическую структуру кластеров, которая упрощает процесс интерпретации результатов.

3. Теоретическое обоснование метода. Для исследования свойств предложенного метода формирования коллективного решения рассмотрим его вероятностную модель.

Предположим, что имеется некоторая скрытая (непосредственно ненаблюдаемая) переменная U, которая задаёт принадлежность каждого объекта к некоторому из M классов (кластеров). Рассмотрим следующую вероятностную модель генерации данных. Пусть каждый класс характеризуется определённым законом условного распределения $p(x \,|\, U = i) = p_i(x)$, где $x \in R^d$, $i = 1, \ldots, M$. Для каждого объекта определяется класс, к которому он будет относиться, в соответствии с априорными вероятностями $P_i = P(U = i)$, $i = 1, \ldots, M$, где $\sum\limits_{i=1}^{M} P_i = 1$. Затем в соответствии с распределением $p_i(x)$ вычисляется наблюдаемое значение x. Указанная процедура проводится независимо для каждого объекта; в результате получаем случайную выборку объектов.

Пусть с помощью некоторого алгоритма кластерного анализа μ строится разбиение множества объектов на M подмножеств. Поскольку нумерация кластеров не играет роли, удобнее рассматривать отношение эквивалентности, т. е. указывать, относит ли алгоритм μ каждую пару объектов в один и тот же класс либо в разные классы. Рассмотрим произвольную пару a, b различных объектов и определим для неё величину

$$h_{\mu,\,a,\,b} = \begin{cases} 0, & \text{если объекты отнесены в один класс,} \\ 1 & \text{иначе.} \end{cases}$$

Пусть $P_U = P(U(a) \neq U(b))$ — вероятность принадлежности объектов к различным классам. Обозначим вероятность ошибки, которую может совершить алгоритм μ при классификации a и b, через $P_{\mathrm{err},\,\mu}$, где

$$P_{\mathrm{err},\,\mu} = \begin{cases} P_U, & \text{если } h_{\mu,\,a,\,b} = 0, \\ 1 - P_U, & \text{если } h_{\mu,\,a,\,b} = 1. \end{cases}$$

Легко заметить, что

$$P_{\mathrm{err},\,\mu} = P_U + (1 - 2P_U) h_{\mu,\,a,\,b}. \tag{1}$$

Предположим, алгоритм μ зависит от случайного вектора параметров $\Theta \in \mathbf{\Theta}$: $\mu = \mu(\Theta)$. Чтобы подчеркнуть зависимость результатов работы от параметра Θ, в дальнейшем будем обозначать $h_{\mu(\Theta),\,a,\,b} = h(\Theta)$, $P_{\mathrm{err},\,\mu(\Theta)} = P_{\mathrm{err}}(\Theta)$.

Пусть в результате L-кратного применения алгоритма μ со случайно и независимо отобранными параметрами $\theta_1, \ldots, \theta_L$ получен набор решений $h(\theta_1), \ldots, h(\theta_L)$. Для определённости будем считать, что L нечётно. Коллективным (ансамблевым) решением назовём функцию

$$H(h(\theta_1), \ldots, h(\theta_L)) = \begin{cases} 0, & \text{если } \dfrac{1}{L}\displaystyle\sum_{l=1}^{L} h(\theta_l) < \dfrac{1}{2}, \\ 1 & \text{иначе.} \end{cases}$$

Необходимо исследовать поведение коллективного решения в зависимости от размера ансамбля L. Заметим, что каждый отдельный алгоритм также можно рассматривать как вырожденный случай ансамбля с $L = 1$.

Утверждение 1. Начальный момент k-го порядка для величины вероятности ошибки алгоритма $\mu(\Theta)$ равен

$$\nu_k = (1 - P_h)P_U^k + P_h(1 - P_U)^k,$$

где $P_h = P(h(\Theta) = 1)$.

Доказательство. Справедливость выражения следует из того, что

$$\nu_k = \mathbf{E}_\Theta P_{\mathrm{err}}^k(\Theta) = \mathbf{E}_\Theta(P_U + (1 - 2P_U)h(\Theta))^k = \mathbf{E}_\Theta \sum_{m=0}^{k} C_k^m P_U^m (1 - 2P_U)^{k-m} h^{k-m}(\Theta) =$$

$$= \sum_{m=0}^{k} C_k^m P_U^m (1 - 2P_U)^{k-m} \mathbf{E}_\Theta h^{k-m}(\Theta).$$

Так как $\mathbf{E}_\Theta h^q(\Theta) = \mathbf{E}_\Theta h(\Theta) = P_h$ при $q > 0$, получим

$$\nu_k = P_U^k + \sum_{m=1}^{k} C_k^m P_U^m (1 - 2P_U)^{k-m} P_h = P_U^k - P_h P_U^k + P_h \sum_{m=0}^{k} C_k^m P_U^m (1 - 2P_U)^{k-m} =$$

$$= P_U^k - P_h P_U^k + P_h(P_U + 1 - 2P_U)^k = P_U^k - P_h P_U^k + P_h(1 - P_U)^k = (1 - P_h)P_U^k + P_h(1 - P_U)^k,$$

что и требовалось доказать.

Следствие 1. Математическое ожидание и дисперсия величины вероятности ошибки для алгоритма $\mu(\Theta)$ равны соответственно

$$\mathbf{E}_\Theta P_{\mathrm{err}}(\Theta) = P_U + (1 - 2P_U)P_h, \quad \mathbf{Var}_\Theta P_{\mathrm{err}}(\Theta) = (1 - 2P_U)^2 P_h(1 - P_h).$$

Доказательство. Справедливость выражения для математического ожидания следует из доказанного утверждения для момента ν_1, а также непосредственно из (1). Рассмотрим выражение для дисперсии. По определению

$$\mathbf{Var}_\Theta P_{\mathrm{err}}(\Theta) = \nu_2 - \nu_1^2.$$

Отсюда

$$\mathbf{Var}_{\Theta} P_{\mathrm{err}}(\Theta) = (1 - P_h)P_U^2 + P_h(1 - P_U)^2 - (P_U + (1 - 2P_U)P_h)^2.$$

После преобразований получим

$$\mathbf{Var}_{\Theta} P_{\mathrm{err}}(\Theta) = (1 - 2P_U)^2 P_h(1 - P_h),$$

что и требовалось доказать.

Обозначим через $P_{\mathrm{err}}(\Theta_1, \ldots, \Theta_L)$ случайную функцию, значение которой при фиксированных аргументах равно вероятности ошибки, которую может совершить ансамблевый алгоритм при классификации a и b. Здесь $\Theta_1, \ldots, \Theta_L$ — независимые статистические копии случайного вектора Θ. Рассмотрим поведение вероятности ошибки для коллективного решения.

Утверждение 2. Начальный момент k-го порядка для величины вероятности ошибки коллективного решения

$$\mathbf{E}_{\Theta_1, \ldots, \Theta_L} P_{\mathrm{err}}^k(\Theta_1, \ldots, \Theta_L) = (1 - P_{H,L})P_U^k + P_{H,L}(1 - P_U)^k,$$

где

$$P_{H,L} = P\Big(\frac{1}{L}\sum_{l=1}^{L} h(\Theta_l) \geq \frac{1}{2}\Big) = \sum_{l = \lfloor L/2 \rfloor + 1}^{L} C_l^l P_h^l (1 - P_h)^{L-1},$$

$\lfloor \cdot \rfloor$ означает целую часть числа.

Доказательство данного утверждения подобно доказательству утверждения 1 (вероятность ошибки коллективного решения определяется по формуле, аналогичной формуле (1)). Кроме того, ясно, что распределение числа голосов, отданных за решение $h = 1$, является биномиальным: $\mathrm{Bin}(L, P_h)$.

Как и в утверждении 1, можно показать, что математическое ожидание и дисперсия величины вероятности ошибки для коллективного решения равны соответственно

$$\mathbf{E}_{\Theta_1, \ldots, \Theta_L} P_{\mathrm{err}}(\Theta_1, \ldots, \Theta_L) = P_U + (1 - 2P_U)P_{H,L},$$

$$\mathbf{Var}_{\Theta_1, \ldots, \Theta_L} P_{\mathrm{err}}(\Theta_1, \ldots, \Theta_L) = (1 - 2P_U)^2 P_{H,L}(1 - P_{H,L}).$$

Воспользуемся следующей априорной информацией об алгоритме кластерного анализа. Будем считать, что ожидаемая вероятность ошибочной классификации $\mathbf{E}_{\Theta} P_{\mathrm{err}}(\Theta) < 1/2$, т. е. предполагается, что алгоритм μ проводит классификацию с лучшим качеством, нежели алгоритм случайного равновероятного выбора. Из следствия 1 видно, что выполняется один из двух вариантов: а) $P_h > 1/2$ и $P_U > 1/2$; б) $P_h < 1/2$ и $P_U < 1/2$. Рассмотрим для определённости первый случай.

Утверждение 3. Если $\mathbf{E}_{\Theta} P_{\mathrm{err}}(\Theta) < 1/2$ и при этом $P_h > 1/2$ и $P_U > 1/2$, то с увеличением мощности (числа элементов) ансамбля ожидаемая вероятность ошибочной классификации уменьшается, стремясь в пределе к $1 - P_U$, а дисперсия величины вероятности ошибки стремится к нулю.

Доказательство. Из интегральной теоремы Муавра — Лапласа следует, что при увеличении L

$$P_{H,L} = 1 - P\Big(\frac{1}{L}\sum_{l=1}^{L} h(\Theta_l) < \frac{1}{2}\Big)$$

сходится к

$$1 - \Phi\Big(\frac{1/2 - P_h}{\sqrt{P_h(1-P_h)/L}}\Big),$$

где $\Phi(\cdot)$ — функция распределения стандартного нормального закона. Значит, при $L \to \infty$ величина $P_{H,L}$ монотонно увеличивается, стремясь в пределе к единице. Из

$$E_{\Theta_1,\ldots,\Theta_L} P_{\text{err}}(\Theta_1,\ldots,\Theta_L) = P_U + (1-2P_U)P_{H,L},$$

где $(1-2P_U) < 0$, и утверждения 2 следует справедливость утверждения 3.

Очевидно, что во втором случае ожидаемая вероятность ошибки также уменьшается при увеличении мощности ансамбля, стремясь в пределе к величине P_U, при этом дисперсия ошибки стремится к нулю.

Доказанное утверждение позволяет сделать вывод о том, что при выполнении вышеуказанных вполне естественных условий применение ансамбля даёт возможность улучшить качество кластеризации.

4. Результаты экспериментальных исследований. В соответствии с предложенным в разд. 2 методом разработан и программно реализован на языке Java ансамблевый алгоритм ECCA$(m_{\min}, L, T, \tau, T_d)$. Для работы алгоритма необходимо задать значения пяти параметров: $m_{\min}, L, T, \tau, T_d$. Многочисленные экспериментальные исследования, проведённые на модельных и реальных данных, показали, что при использовании десяти элементов ансамбля получаемые результаты являются устойчивыми к выбору параметра сетки $m_{\min}$. Параметр T оказывает слабое влияние на результат кластеризации. При обработке изображений этот параметр выбирался равным 0,8, а порог шума $\tau \in \{0;1\}$. Алгоритм ECCA позволяет получать иерархическую структуру данных. Проведённые исследования показывают, что параметр T_d, задающий глубину иерархии, достаточно выбирать из множества $\{0,0{,}1,\ldots,0{,}9\}$. Далее приведены результаты экспериментов на модельных и реальных данных, подтверждающие эффективность предложенного алгоритма. Обработка проводилась на ПЭВМ с тактовой частотой 3 ГГц.

Эксперимент 1. Использовалась известная таблица данных по ирисам [15]. Множество состояло из 150 точек четырёхмерного пространства признаков, сгруппированных в три класса по 50 точек. Обозначим через $|C_i^O|$ фактическое число точек i-го класса, а через $|C_i^S|$ число точек класса C_i^O, содержащихся в соответствующем кластере, выделенном алгоритмом ECCA. Аналогично [4] точность кластеризации и меру покрытия кластерами C_i^S классов C_i^O определим по формулам $|C_i^O \cap C_i^S|/|C_i^S|$ и $|C_i^O \cap C_i^S|/|C_i^O|$ соответственно, где $|\cdot|$ — мощность множества. В таблице приводятся значения вычисленных критериев после применения алгоритма ECCA с параметрами $m_{\min} = 25$, $L = 10$, $T = 0{,}9$, $\tau =$

Результаты работы алгоритма ECCA на данных по ирисам

Параметры	Классы		
	$i=1$	$i=2$	$i=3$
$\lvert C_i^O\rvert$	50	50	50
$\lvert C_i^S\rvert$	50	52	48
$\lvert C_i^O \cap C_i^S\rvert$	50	48	46
Точность, %	100	96	92
Мера покрытия, %	100	92,31	95,83

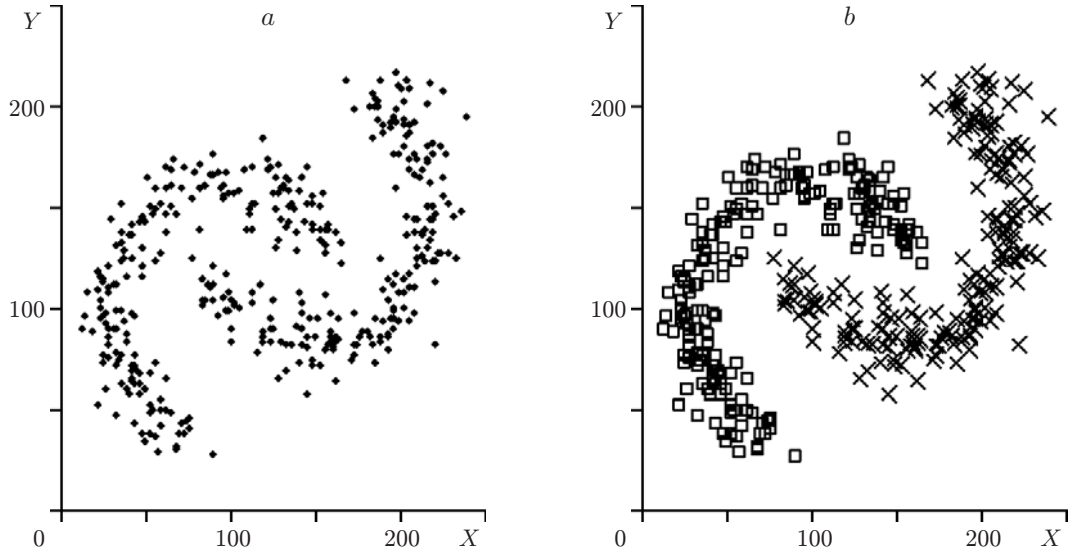

Рис. 1

$= 0$, $T_d = 0{,}7$. По этим критериям результаты алгоритма ECCA значительно превосходят результаты алгоритма GCOD [4].

Эксперимент 2. Использовались двумерные данные, состоящие из 400 точек, сгруппированных в два равномощных линейно неразделимых класса, имеющих форму бананов (рис. 1, *a*, исходное множество). Модель построена с помощью инструментария PRTools (The Matlab Toolbox for Pattern Recognition: http://www.prtools.org) с параметром 0,7. На рис. 1, *b* представлены результаты работы алгоритма ECCA(15, 10, 0,3, 0,8). Для сравнения исходные данные обрабатывались известным алгоритмом DBSCAN [16]. После длительной настройки его параметров удалось добиться результата, как на рис. 1, *b*. Однако обработка выполнялась более чем в 100 раз дольше, чем с помощью ECCA.

На рис. 2 приведён график зависимости числа кластеров, получаемых при работе алгоритма CCA$(m, 0{,}8, 0)$, от параметра m, определяющего размер элементов клеточной структуры. На рис. 3 показана зависимость ошибки кластеризации от значений параметров m при фиксированных параметрах T, τ для алгоритма CCA (пунктирная линия) и $m_{\min}$ при фиксированных параметрах T, L для ECCA (сплошная линия). Ошибка класте-

Рис. 2

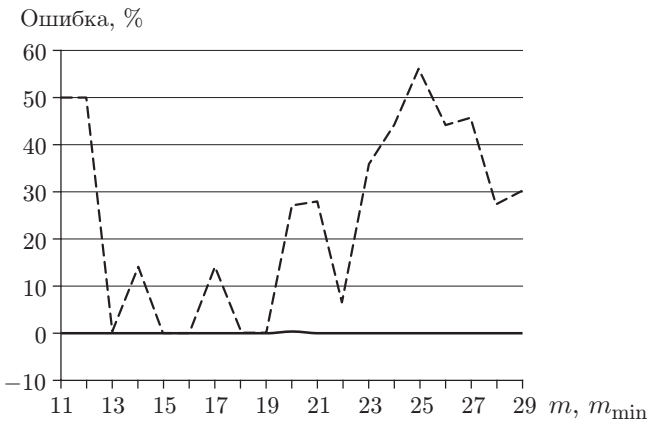

Рис. 3

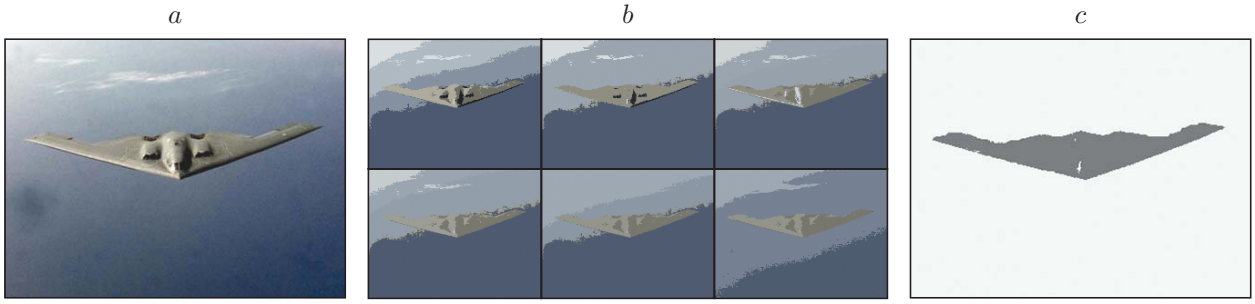

Рис. 4

ризации здесь определяется по формуле $\sum_{i=1}^{2} |C_i^O \backslash C_i^S| \Big/ \sum_{i=1}^{2} |C_i^O|$. Графики демонстрируют существенную зависимость результатов работы алгоритма CCA от настраиваемого параметра m и устойчивость получаемых решений для ансамблевого алгоритма ECCA при изменении параметра $m_{\min}$. Эта устойчивость значительно упрощает настройку параметров алгоритма ECCA.

Эксперимент 3. Обрабатывалось цветное изображение (http://commons.wikimedia.org/wiki/File:B-2_Spirit_4.jpg) (рис. 4, *a*) размером 640×480 пикселов. Кластеризация проводилась в цветовом пространстве RGB. Каждый кластер соответствовал однородной области на изображении. Использовался ансамбль из десяти элементов. Ни один из них не позволяет выделить интересующий объект на исходном изображении (рис. 4, *b*, представлены шесть элементов из десяти). На рис. 4, *c* показан результат применения ансамблевого алгоритма ECCA с параметрами $m_{\min} = 30$, $L = 10$, $T = 0{,}8$, $\tau = 0$, $T_d = 0{,}9$. Время обработки 0,88 с.

Заключение. В представленной работе предложен метод кластеризации больших массивов данных на основе ансамбля сеточных алгоритмов. Дано его теоретическое обоснование.

К основным характеристикам рассмотренного алгоритма, выделяющим его среди других алгоритмов кластерного анализа, относятся: 1) универсальность (алгоритм позволяет при наличии шума выделять кластеры, различающиеся размером, формой, плотностью); 2) высокое быстродействие при кластеризации большого числа объектов ($\sim 10^6$) (при условии, что число признаков невелико (≤ 6), это условие выполняется, в частности, в задачах анализа изображений); 3) простота настройки параметров.

Результаты приведённых экспериментов на модельных и реальных данных подтверждают высокое качество получаемых решений и их устойчивость к изменению настраиваемых параметров. Возможность получения иерархической системы вложенных кластеров значительно облегчает процесс интерпретации результатов. Высокое быстродействие алгоритма ECCA позволяет проводить обработку больших массивов данных в диалоговом режиме. Алгоритм ECCA допускает распараллеливание, позволяющее повысить быстродействие при реализации его на многопроцессорных вычислительных системах.

СПИСОК ЛИТЕРАТУРЫ

1. **Jain A. K.** Data clustering: 50 years beyond K-means // Pattern Recogn. Lett. 2010. **31**, Is. 8. P. 651–666.
2. **Mercer D. P.** Clustering large datasets. Linacre College, 2003. URL: http://www.stats.ox.ac.uk/~mercer/documents/Transfer.pdf (дата обращения: 21.03.2011).

3. **Ilango M. R., Mohan V.** A survey of grid based clustering algorithms // Intern. Journ. Eng. Sci. and Technol. 2010. **2**, N 8. P. 3441–3446.
4. **Qiu B.-Z., Li X.-L., Shen J.-Y.** Grid-based clustering algorithm based on intersecting partition and density estimation // Lect. Notes Artif. Intel. 2007. **4819**. P. 368–377.
5. **Akodjènou-Jeannin M.-I., Salamatian K., Gallinari P.** Flexible grid-based clustering // Lect. Notes Artif. Intel. 2007. **4702**. P. 350–357.
6. **Ma Eden W. M., Chow Tommy W. S.** A new shifting grid clustering algorithm // Pattern Recogn. 2004. **37**, N 3. P. 503–514.
7. **Lin N. P., Chang C.-I., Chueh H.-E. et al.** A Deflected grid-based algorithm for clustering analysis // WSEAS Transactions on Computers. 2008. **7**, N 4. P. 125–132.
8. **Shi Y., Song Y., Zhang A.** A shrinking-based approach for multi-dimensional data analysis // Proc. of the 29th VLDB Conference. Berlin, Germany, 2003. P. 440–451.
9. **Куликова Е. А., Пестунов И. А., Синявский Ю. Н.** Непараметрический алгоритм кластеризации для обработки больших массивов данных // Тр. 14 Всеросс. конф. «Математические методы распознавания образов». М.: Изд-во MAKS Press, 2009. С. 149–152.
10. **Strehl A., Ghosh J.** Clustering ensembles — a knowledge reuse framework for combining multiple partitions // Journ. Machine Learning Research. 2002. **3**. P. 583–617.
11. **Бирюков А. С., Рязанов В. В., Шмаков А. С.** Решение задач кластерного анализа коллективами алгоритмов // Журн. вычисл. матем. и матем. физики. 2008. **48**, № 1. С. 176–192.
12. **Hong Y., Kwong S.** To combine steady-state genetic algorithm and ensemble learning for data clustering // Pattern Recogn. Lett. 2008. **29**, N 9. P. 1416–1423.
13. **Berikov V. B.** Construction of the ensemble of logical models in cluster analysis // Lect. Notes Artif. Intel. 2009. **5755**. P. 581–590.
14. **Дуда Р., Харт П.** Распознавание образов и анализ сцен. М.: Мир, 1976. 559 с.
15. **Кендал М., Стьюарт А.** Многомерный статистический анализ и временные ряды. М.: Наука, 1976. С. 441–443.
16. **Ester M., Kriegel H.-P., Sander J., Xu X.** A density-based algorithm for discovering clusters in large spatial database // Proc. of the Intern. Conf. Knowledge Discovery and Data Mining (KDD'96). 1996. P. 226–231.

Поступила в редакцию 11 апреля 2011 г.