КЛАСТЕРИЗАЦИЯ НЕПОЛНЫХ ДАННЫХ

Владимир Рязанов, Кирилл Тишин, Антон Щичко

***Abstract**: В работе рассматриваются различные подходы к решению задач кластеризации данных при наличии в них пропусков. Применяются методы коллективных решений, а также алгоритмы заполнения пропущенных значений признаков. Также предлагаются способы вычисления оценок качества кластеризации данных с прочерками.*

***Keywords**: кластеризация, неполные данные, прочерки, коллективные решения*

***ACM Classification Keywords**: I.5.3 Computing Methodologies - Pattern recognition - Clustering*

Введение

Случаи наличия пропусков в обучающих выборках являются обычным явлением в практической классификации без учителя (кластеризации). В настоящее время сформировалось два общих подхода для решения задач кластеризации данных с пропусками. Первый подход состоит в восстановлении пропущенных значений признаков и последующей кластеризации полученных полных данных. При втором подходе осуществляется прямая кластеризация неполных данных. Оба подхода имеют свои преимущества и недостатки. При первом подходе требуется создание методов восстановления значений признаков, при их восстановлении теряется некоторая информация. Однако здесь возможно последующее использование существующих алгоритмов и программ кластеризации полных данных. При прямой кластеризации неполных данных мы не теряем дополнительно информацию, но здесь необходимо создание новых алгоритмов кластеризации неполных данных или модификация известных методов на случай неполноты данных. Существуют многочисленные алгоритмы для восстановления значений признаков и ряд методов прямой кластеризации неполных данных.

Алгоритмы восстановления значений признаков можно условно разделить на два типа. Первый тип алгоритмов (marginalization) предполагает простое исключение из обучающих выборок неполных признаковых описаний. Ясно, что этот путь может быть целесообразным только при относительно малой доле неполных объектов. При втором типе алгоритмов пропущенные значения признаков заменяются их оценками (imputation). Здесь применяются обычно простейшие подходы (замена прочерков на усредненные статистические оценки значений признаков: means, random, the nearest neighbor method, и т.п.) [1,2], и регрессионные модели, когда неизвестное значение признака вычисляется с помощью найденной функции регрессии по известным признакам (линейная регрессия, SVR [3]). Широкое распространение получил EM-алгоритм, предполагающий вероятностную модель построения выборок [1].

Второй подход при кластеризации неполных данных состоит в адаптации методов к случаям неполных данных. В данном случае не требуется предварительного восстановления значений признаков. В [4] предложена модификация метода «fuzzy k-means». В работе [5] описаны два метода разбиения неполных данных на линейные нечеткие кластеры. Модификация метода «k-means», где из-за пропусков создаются системы ограничений, изложена в [6].

Результаты сравнения различных подходов показывают, что при решении практических задач априори не ясно, какой из подходов окажется наиболее приемлемым. Таким образом, создание новых подходов и алгоритмов кластеризации неполных данных является актуальной задачей. Другой актуальной проблемой в данной области является оценка качества кластеризаций (степени «определенности» кластеризаций) как результат отсутствия значений некоторых признаков.

В настоящей статье предлагается два подхода для решения задач кластеризации с прочерками на заданное число кластеров с оценкой степени определенности решений. Первый подход основан на восстановлении значений неизвестных признаков, решении задачи кластеризации данных без прочерков и вычислении степени определенности полученных кластеризаций на основе оценки их устойчивости. Второй подход основан на решении конечного множества задач кластеризации $Z_i, i = 1,..., N$, выборок допустимых полных описаний, соответствующих выборкам исходных неполных описаний, и построении коллективного решения. Разброс решений задач $Z_i, i = 1,..., N$, относительно коллективного решения и используется для оценки степени определенности кластеризации неполных данных (коллективного решения).

Кластеризация неполных данных на базе восстановления значений признаков

Пусть дана стандартная выборка неполных признаковых описаний объектов $X = \{\mathbf{x}_1, \mathbf{x}_2,..., \mathbf{x}_m\}$, $\mathbf{x}_i = (x_{i1}, x_{i2},..., x_{in})$. Считаем, что $x_{ij} = \Delta, \forall j \in \Omega_i, i = 1,..., m$, где символом Δ обозначаем неизвестное значение признака (пропуск, прочерк). Иногда также будем использовать обозначение $J = \{\langle i, j\rangle, i = 1,2,..., m, j \in \Omega_i\}$. Будем использовать локальный метод заполнения прочерков, суть которого состоит в следующей итерационной процедуре. Сначала все неизвестные значения заполняются случайными числами из области допустимых значений признака $x_{ij} \in M_j \subseteq R, j = 1,2,..., n$, (область M_j определяется как конечное множество значений $j-$го признака, которые принимают объекты обучающей выборки). Далее неизвестные значения последовательно модифицируются с помощью сочетания метода k-ближайших соседей и процедуры сдвига.

Пусть фиксирована метрика ρ в R и значение целочисленного параметра k, $1 \le k \le m-1$.

Локальный алгоритм восстановления значений признаков [7]:

Шаг 0. Инициализация случайных $x_{ij}^0 \in M_j, j = 1,2,..., n$, если $x_{ij} = \Delta, \forall j \in \Omega_i, i = 1,..., m$. Получаем таблицу полных описаний $\left\| x_{ij}^{(0)} \right\|_{m \times n}$.

Шаг t=1,2,.... Пусть $x_{ij}^{(t-1)}*$ - среднее значение признака № j по k ближайшим соседям объекта $\mathbf{x}_i$. Тогда определяем $x_{ij}^{(t)} = x_{ij}^{(t-1)} + \theta(x_{ij}^{(t-1)}* - x_{ij}^{(t-1)}), \forall \langle i, j\rangle \in J$. Здесь $0 < \theta \le 1$ управляющий параметр (скорость обучения). Если не выполняется условие останова, шаг повторяется.

В качестве критерия останова используем один из стандартных критериев или их комбинацию (максимальное число итераций N, $\rho(x_{ij}^{(t)}, x_{ij}^{(t-1)}) \le \varepsilon, \forall \langle i, j\rangle \in J$, и др.). После вычисления значений $x_{ij}^{(t)}, \forall \langle i, j\rangle \in J$ положим x_{ij}^* равным значению из M_j, ближайшему к соответствующему $x_{ij}^{(t)}, \forall \langle i, j\rangle \in J$. Если $x_{ij} \ne \Delta$, положим $x_{ij}^* = x_{ij}$.

Полученную в итоге выборку полных описаний обозначим как $X* = \{\mathbf{x}*_1, \mathbf{x}*_2,..., \mathbf{x}*_m\}$. Пусть получено решение задачи кластеризации данной выборки на l кластеров некоторым алгоритмом A: $\mathrm{K} = \{K_1, K_2,..., K_l\}$, $K_i \subseteq X*, i = 1,2,..., l$, $\bigcup_{i=1}^{l} K_i = X*, K_i \cap K_j = \varnothing, i \ne j$. Обозначим $D_t = \{\mathbf{x}'_t\}$ – множество всех допустимых $\mathbf{x}'_t$, соответствующих вектору $\mathbf{x}_t$ (т.е. $x'_{tj} = \begin{cases} x_{tj}, & x_{tj} \ne \Delta, \\ \in M_j, & x_{tj} = \Delta. \end{cases}$).

Пусть $\mathbf{x}*_t \in K_\lambda$. При замене $\mathbf{x}*_t$ на некоторый $\mathbf{x}'_t$, множество $K_\lambda \setminus \{\mathbf{x}*_t\} \cup \{\mathbf{x}'_t\}$ может оказаться не кластером с позиций алгоритма A. Обозначим через $f_t(\mathrm{K})$ - долю объектов $\mathbf{x}'_t$ из D_t, для которых множество $K_\lambda \setminus \{\mathbf{x}*_t\} \cup \{\mathbf{x}'_t\}$ будет кластером, а также все остальные кластеры останутся без изменений.

Определение 1. Степенью определенности $f(\mathrm{K})$ кластеризации $\mathrm{K} = \{K_1, K_2, ..., K_l\}$ назовем величину $f(\mathrm{K}) = \frac{1}{m}\sum_{t=1}^{m} f_t(\mathrm{K})$.

Рассмотрим вопрос вычисления степени определенности на примере алгоритма k –межгрупповых средних.

Пусть для выборки полных описаний $X* = \{\mathbf{x}*_1, \mathbf{x}*_2, ..., \mathbf{x}*_m\}$ получена кластеризация $\mathrm{K} = \{K_1, K_2, ..., K_l\}$ с помощью метода k-внутригрупповых средних [8]. Это означает, что для $\forall \mathbf{x}*_t \in K_i$ имеет место

$$\left\|\mathbf{x}*_t - \mathbf{m}*_i\right\| \le \left\|\mathbf{x}*_t - \mathbf{m}*_j\right\|, \ \forall j \ne i, \tag{1}$$

где $\mathbf{m}*_i = \frac{1}{n_i}\sum_{\mathbf{x}*_j \in K_i} \mathbf{x}*_j$.

Вычисление $f_t(\mathrm{K})$ сводится к проверке неравенств, выполненных при замене $\mathbf{x}*_t$ на различные допустимые $\mathbf{x}'_t = (x'_{t1}, x'_{t2}, ..., x'_{tn})$. Во-первых, проверяется, что объект остался ближе к своему измененному центру, чем к остальным при замене на допустимое описание:

$$\left\|\mathbf{x}'_t - \mathbf{m}'_i\right\| \le \left\|\mathbf{x}'_t - \mathbf{m}*_j\right\|, \ \forall j \ne i. \tag{2}$$

Во-вторых, проверяется, что после замены одного объекта на допустимое описание остальные объекты остались ближе к своим центрам:

$$\left\|\mathbf{x}*_p - m*_{i_p}\right\| \le \left\|\mathbf{x}*_p - \mathbf{m}'_i\right\|, \ \forall p \ne t. \tag{3}$$

Неравенства (2) можно переписать в следующем виде:

$$\left\|\mathbf{x}*_t - \Delta\mathbf{x}_t - \mathbf{m}*_i + \frac{\Delta\mathbf{x}_t}{n_i}\right\|^2 \le \left\|\mathbf{x}*_t - \Delta\mathbf{x}_t - \mathbf{m}*_j\right\|^2, \text{ где } \Delta\mathbf{x}_t = \mathbf{x}*_t - \mathbf{x}'_t,$$

После элементарных преобразований имеем неравенство

$$\left\|\mathbf{x}*_t - \mathbf{m}*_i\right\|^2 - \left\|\mathbf{x}*_t - \mathbf{m}*_j\right\|^2 + 2(\Delta\mathbf{x}_t, \frac{\mathbf{x}*_t}{n_i} - \mathbf{m}*_j + \mathbf{m}*_i \frac{(n_i - 1)}{n_i}) - \left\|\Delta\mathbf{x}_t\right\|^2 \frac{(2n_i - 1)}{n_i^2} \le 0. \tag{4}$$

Неравенства (3) можно преобразовать аналогично:

$$\left\|\mathbf{x}*_p - m*_{i_p}\right\|^2 \le \left\|\mathbf{x}*_p - \mathbf{m}*_i + \frac{\Delta\mathbf{x}_t}{n_i}\right\|^2,$$

$$\left\|\mathbf{x}*_p - m*_{i_p}\right\|^2 - \left\|\mathbf{x}*_p - m*_i\right\|^2 + 2(\Delta\mathbf{x}_t, \frac{\mathbf{m}*_i - \mathbf{x}*_p}{n_i}) - \left\|\Delta\mathbf{x}_t\right\|^2 \frac{1}{n_i^2} \le 0. \tag{5}$$

Вычисление $f_t(\mathrm{K})$ сводится к проверке выполнения неравенств вида $a+\sum_{i\in\Omega_t} y_i c_i + b\sum_{i\in\Omega_t} y_i^{\,2} \le 0$,

$p+\sum_{i\in\Omega_t} y_i r_i + q\sum_{i\in\Omega_t} y_i^{\,2} \le 0$ где $a,b,p,q,c_i,r_i,\ i=1,2,...,k$ – константы для заданного t, а $y_i \in \{x^*_{ti} - x'_{ti} : x'_{ti} \in M_i\}$. При малых k здесь возможен перебор, при больших – оценка по случайной выборке.

Аналогично можно вычислить степень определенности кластеризаций, полученных другими алгоритмами A. Для этого достаточно посчитать (оценить) долю полных описаний, которые будут соответствовать критериям останова алгоритма A.

Кластеризация неполных данных на базе построения коллективных решений.

По выборке $X=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_m\}$ формируется N выборок полных описаний $X'^{(i)}=\{\mathbf{x}'^{(i)}_1,\mathbf{x}'^{(i)}_2,...,\mathbf{x}'^{(i)}_m\}, i=1,2,...,N$, где $x'^{(i)}_{tj}=\begin{cases} x_{tj}, & x_{tj}\neq\Delta, \\ \in M_j, & x_{tj}=\Delta \end{cases}$ (вероятность присвоения $x'^{(i)}_{tj}$ некоторого значения из M_j равна его частоте встречаемости на обучающей выборке). Для каждой из полученных полных выборок решается задача кластеризации на l кластеров и находятся N решений $\mathrm{K}^{(i)}=\{K^{(i)}_1,K^{(i)}_2,...,K^{(i)}_l\}, i=1,2,...,N$. Далее по данным решениям строится коллективная кластеризация $\mathrm{K}=\{K_1,K_2,...,K_l\}$, которая и принимается как решение задачи кластеризации с прочерками.

Определение 2. Степенью определенности $\Phi(\mathrm{K})$ кластеризации $\mathrm{K}=\{K_1,K_2,...,K_l\}$ назовем величину $\Phi(\mathrm{K})=\sum_{i=1}^{N}\max_{<t_1,t_2,...,t_l>}\sum_{j=1}^{l}\left|K_j\cap K^i_{t_j}\right| / mN$.

Определение 3. Степенью определенности $F(\mathrm{K})$ кластеризации $\mathrm{K}=\{K_1,K_2,...,K_l\}$ назовем величину $F(\mathrm{K})=\min_{i=1,...,N}\max_{<t_1,t_2,...,t_l>}\sum_{j=1}^{l}\left|K_j\cap K^i_{t_j}\right| / m$.

Здесь $<t_1,t_2,...,t_l>$ некоторая перестановка от $<1,2,...,l>$. Величина $\max_{<t_1,t_2,...,t_l>}\sum_{j=1}^{l}\left|K_j\cap K^i_{t_j}\right|$ характеризует близость кластеризаций K и $\mathrm{K}^{(i)}, i=1,2,...,N$ (при равных кластеризациях она равна m). Критерий $\Phi(\mathrm{K})$ характеризует нормированную среднюю близость коллективной кластеризации K относительно допустимых выборок. Критерий $F(\mathrm{K})$ характеризует наихудший случай.

Впервые задача построения коллективных классификаций и комитетный алгоритм ее решения были предложены в работах [9-10]. В настоящем подходе к кластеризации неполных данных использовался метод [11]. Ранее коллективное решение строилось некоторым алгоритмом по множеству кластеризаций, полученных для одной и той же выборки различными методами кластеризации. В нашем случае коллективное решение будет строиться также фиксированным алгоритмом, но по множеству кластеризаций, полученных одним методом кластеризации для различных выборок $X'^{(i)}=\{\mathbf{x}'^{(i)}_1,\mathbf{x}'^{(i)}_2,...,\mathbf{x}'^{(i)}_m\}, i=1,2,...,N$. Опишем алгоритм.

Результаты кластеризации выборок $X'^{(i)} = \{\mathbf{x}'^{(i)}{}_1, \mathbf{x}'^{(i)}{}_2, ..., \mathbf{x}'^{(i)}{}_m\}, i = 1,2,..., N$, некоторым методом кластеризации можно записать в виде трехмерной информационной матрицы $\left\|\alpha_{ij}^{\nu}\right\|_{m \times l \times N}$, $\alpha_{ij}^{\nu} \in \{0,1\}$, $\sum_{j=1}^{l} \alpha_{ij}^{\nu} = 1$, $i = 1,...,m$, $j = 1,...,l$, $\nu = 1,...,N$.

Ее подматрицу $\left\|\alpha_{ij}^{\nu}\right\|_{l \times N}$, $i = 1,2,...,m$ можно рассматривать как новое признаковое описание объекта $\mathbf{x}'_i$. В качестве коллективного решения задачи кластерного анализа принимается кластеризация выборки данных m матричных описаний.

Рассмотрим реализацию описанного подхода, на примере использования метода «k – внутригрупповых средних». На каждом итерационном шаге алгоритма нам необходимо выполнять две основные операции, а именно, нахождение центра выделенного множества объектов и нахождение расстояний между объектами и центром. Далее в качестве кластеризуемой выборки будем рассматривать выборку $\tilde{X} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, ..., \tilde{\mathbf{x}}_m\}, \tilde{\mathbf{x}}_t = \left\|\alpha_{tj}^{\nu}\right\|_{l \times N}, t = 1,2,..., m.$

В качестве расстояния между объектами используется расстояние Хемминга:

$$\rho(\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_k) = \sum_{j=1}^{l} \sum_{\nu=1}^{N} \left|\alpha_{tj}^{\nu} - \alpha_{kj}^{\nu}\right|. \tag{6}$$

Центральной матрицей подмножества матриц $\tilde{X}' \subseteq \tilde{X}$ будем называть матрицу $\tilde{\mathbf{x}}^* = \left\|y_j^{\nu}\right\|_{l \times N}$:

$$\tilde{\mathbf{x}}^* = \arg\min_{\tilde{\mathbf{x}} \in \tilde{X}'} \sum_{\tilde{\mathbf{x}}_i \in \tilde{X}'} \rho(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}).$$

(центральная матрица может быть не единственной). Нахождение центральной матрицы подмножества $\tilde{X}'$, $L = \left|\tilde{X}'\right|$ (для простоты обозначений считаем, что $\tilde{X}' = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, ..., \tilde{\mathbf{x}}_L\}$) представляется в виде задачи дискретной оптимизации с Nl переменными (5-7):

$$\sum_{i=1}^{L} \sum_{\nu=1}^{N} \sum_{j=1}^{l} \left|y_j^{\nu} - \alpha_{ij}^{\nu}\right| \longrightarrow \min, \tag{7}$$

$$\sum_{j=1}^{l} y_j^{\nu} = 1, \ \nu = 1,...N, \tag{8}$$

$$y_j^{\nu} \in \{0,1\}, \ j = 1,...,l, \ \nu = 1,...,N. \tag{9}$$

Задача (7-9) распадается на N независимых оптимизационных задач

$$\sum_{i=1}^{L} \sum_{j=1}^{l} \left|y_j^{\nu} - \alpha_{ij}^{\nu}\right| \longrightarrow \min, \ \sum_{j=1}^{l} y_j^{\nu} = 1, \ y_j^{\nu} \in \{0,1\}, \ \nu = 1,2,..., N.$$

Обозначим $\mathrm{H}(\mu) = \sum_{i=1}^{L} \left(\sum_{\substack{j=1 \\ j \neq \mu}}^{l} \alpha_{ij}^{\nu} + 1 - \alpha_{i\mu}^{\nu} \right)$ и пусть $\eta = \arg\min_{\mu} \Phi(\mu)$.

Тогда бинарный вектор $\mathbf{y}^{*\nu}$, $y_i^{*\nu} = \begin{cases} 0, i \neq \nu \\ 1, i = \nu \end{cases}$ и будет решением данной задачи.

Таким образом, используя метрику (6) для бинарных объектов размерности $l \times N$ и выше определенный алгоритм вычисления бинарных центральных бинарных матриц подмножеств матриц, в качестве алгоритма коллективной кластеризации может быть использован стандартный алгоритм

$k-$внутригрупповых средних (данный алгоритм назван в [11] алгоритмом коллективных $k-$внутригрупповых средних).

Результаты экспериментов на модельных и практических данных

Предложенные алгоритмы кластеризации неполных данных были апробированы на модельных примерах и двух практических задачах. В качестве модельных задач использовались выборки смеси нормальных распределений с независимыми признаками. Математические ожидания и дисперсии классов выбирались такими, чтобы результат их кластеризации совпадал с их априорной классификацией. Визуализация одного модельного примера для четырех классов (проекция многомерных данных на плоскость обобщенных признаков, см. [8,12]) приведена на рис. 1. Рассматриваемые обучающие выборки полных описаний преобразовывалась в выборки с прочерками при различных уровнях неполноты данных: задается процент w неизвестных значений признаков и в каждой строке таблицы обучения по равномерному закону распределения $w\%$ признаков считались неизвестными. Далее, полученные выборки частичных описаний восстанавливались в выборки полных описаний с помощью локального алгоритма и решалась задача их кластеризации на 4 кластера. Отдельно решалась задача кластеризации неполных выборок с помощью коллективных кластеризаций. Были построены зависимости критериев $f(\mathrm{K}),\quad \Phi(\mathrm{K}),\quad F(\mathrm{K})$, а также зависимости показателей $\varphi(\mathrm{K}) = \max\limits_{<t_1,t_2,\ldots,t_l>} \sum\limits_{j=1}^{l} \left|K_j \cap K^*_{t_j}\right| / m$ (где $\mathrm{K}^* = \{K^*_1, K^*_2, \ldots, K^*_l\}$ - априорная классификация исходной модельной выборки, а K – коллективная кластеризация данных с прочерками) и $\varphi_{средн.}(\mathrm{K}) = \max\limits_{<t_1,t_2,\ldots,t_l>} \sum\limits_{j=1}^{l} \left|K_j \cap K^*_{t_j}\right| / m$ (где K – кластеризация выборки при замене прочерков на средние по выборке значения) от параметра w. Критерий $f(\mathrm{K})$ применен к двум видам кластеризаций: K^1 – кластеризации, полученные после заполнения прочерков локальным алгоритмом; K^2 – кластеризации при замене прочерков на средние по выборке значения. Параметры локального алгоритма: $\Theta = 0.4, \quad N = 50, \quad \varepsilon = 0.1, \quad k = 5.$

На рисунках 1-3, и 4-6, показаны, соответственно, визуализации и графики модельной и двух практических задач. Задача **«breast»** [13] представляет выборку из 344 описаний пациентов, имеющих доброкачественное (класс №1) или злокачественное (класс №2) новообразование, для описания пациентов использовалось девять $k-$значных признаков. Задача **«ion»** [14] (распознавание сигналов в атмосфере) имела следующие характеристики: 34 числовых признака, два класса, 351 объект. Обе задачи имеют кластерную структуру, удовлетворительно совпадающую с априорной классификацией. Вид полученных зависимостей соответствует априорным ожиданиям. Критерий $F(\mathrm{K})$ соответствует наихудшему возможному результату и его значение быстро падает с ростом w. Графики критериев $f(\mathrm{K}^1)$ и $f(\mathrm{K}^2)$ близки друг к другу, что говорит о том, что метод восстановления признаков менее важен, чем выбор вида самого критерия. Тем не менее, по результатам экспериментов видно, что значения критерия $f(\mathrm{K}^1)$ обычно выше $f(\mathrm{K}^2)$. Это косвенно показывает, что кластеризации, полученные после заполнения прочерков локальным алгоритмом, «лучше» кластеризаций данных при замене прочерков на средние по выборке значения. На всех задачах оказываются близки показатели $\varphi(\mathrm{K})$ и $\varphi_{средн.}(\mathrm{K})$, что, по-видимому, свидетельствует о надежности «усреднений» по множеству допустимых кластеризаций, т.е. использовании коллективных решений. Как видно из графиков, значение показателя $\varphi(\mathrm{K})$ в целом выше значения $\varphi_{средн.}(\mathrm{K})$. Это показывает, что коллективное решение позволяет получить более точную кластеризацию, чем полученную при замене прочерков на средние по

выборке значения. Критерий $\Phi(\mathrm{K})$ представляется наиболее объективным и обоснованным, что выражается и в стабильности его зависимости от w на всех рассмотренных задачах.

Рис. 1. Смесь нормальных распределений

Рис. 2. Задача «breast»

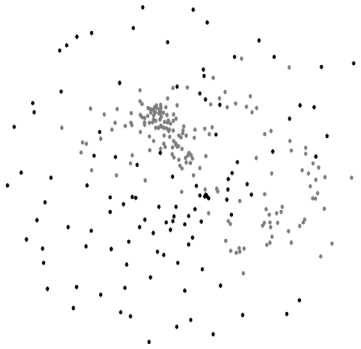


Рис. 3. Задача «Ion»

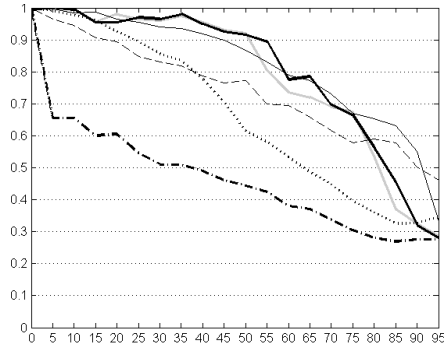


Рис. 4. Зависимости критериев и показателей от неполноты данных в модельной задаче

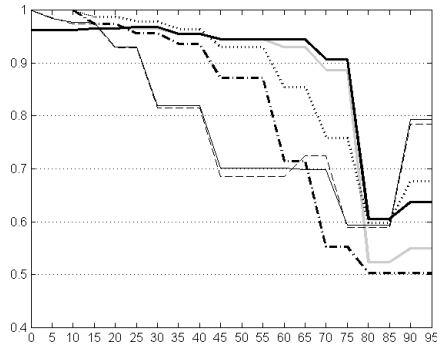


Рис. 5. Зависимости критериев и показателей от неполноты данных в задаче «breast»

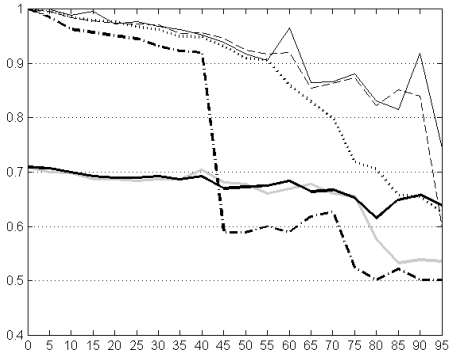


Рис. 6. Зависимости критериев и показателей от неполноты данных в задаче «Ion»

Обозначения графиков критериев и показателей.

- показатель $\varphi_{средн.}(\mathrm{K})$
- показатель $\varphi(\mathrm{K})$
- критерий $f(\mathrm{K}^1)$
- критерий $f(\mathrm{K}^2)$
- критерий $F(\mathrm{K})$
- критерий $\Phi(\mathrm{K})$

Заключение

Предложенные в настоящей работе критерии степени определенности кластеризаций неполных данных основаны на оценке устойчивости полученных кластеризаций относительно возможных вариаций неполных признаковых описаний. Рассмотрены различные подходы. По результатам настоящих предварительных исследований критерий $\Phi(K)$ представляется наиболее объективным. На основе предложенных критериев очевидным образом могут вычисляться и оценки степени определенности кластеризации отдельных объектов. Ранее, в работе [15] был предложен общий подход к оценке качества кластеризаций, основанный на оценке устойчивости кластеризаций относительно вариаций обучающей выборки. Представляет интерес исследование взаимосвязи рассмотренных в настоящей работе критериев степени определенности кластеризаций и критериев качества кластеризаций, а также создание критериев качества кластеризаций неполных данных. Данные вопросы будут рассмотрены в дальнейших исследованиях авторов.

В настоящее время не существует общих, универсальных алгоритмов кластеризации, тем более в случае неполноты данных. Создание новых подходов и алгоритмов расширяет возможности анализа данных, нахождения скрытых закономерностей и свойств по выборкам прецедентов.

Acknowledgements

Настоящая работа выполнена при поддержке Программ Президиума РАН №14 и «Фундаментальные науки - медицине», Программы №2 Отделения математических наук РАН, проектов РФФИ 09-01-00409, 10-01-90015 Бел_а, 10-01-90419 Укр_а.

Библиография

[1] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data. Wiley, New York, 1987.

[2] Zloba, E. Statistical methods of reproducing of missing data / E.Zloba, I.Yatskiv // Computer Modelling & New Technologies. — 2002. — Vol.6, No.1 — P. 51-61.

[3] F. Honghai, C. Guoshun, Y. Cheng, Y. Bingru and C. Yumei, A SVM Regression Based Approach to Filling in Missing Values, LNCS - Knowledge-Based Intelligent Information and Engineering Systems, Springer Berlin - Heidelberg, vol. 3683, 2005, 581-587.

[4] Manish Sarkar and Tze-Yun Leong, Fuzzy K-means Clustering with Missing Values. Proc AMIA Symp. 2001: 588–592.

[5] Katsuhiro Honda, and Hidetomo Ichihashi, Linear Fuzzy Clustering Techniques With Missing Values and Their Application to Local Principal Component Analysis, IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 12, NO. 2, APRIL 2004, pp. 183-193.

[6] Kiri Wagstaff, Clustering with Missing Values: No Imputation Required. In "Classification, Clustering, and Data Mining Applications". Studies in Classification, Data Analysis, and Knowledge Organization, 2004, Volume 0, Part VII, 649-658.

[7] Михайлова Е.И., Рязанов В.В., Штаюра В.А. //Распознавание по прецедентам при наличии пропусков значений признаков. Математические методы распознавания образов: 14-я Всероссийская конференция. Владимирская обл., г.Суздаль, 21-26 сентября 2009 г.: Сборник докладов.-М.: МАКС Пресс, 2009. Стр. 163-164.

[8] Дуда Р., Харт П., Распознавание образов и анализ сцен. Издательство "Мир", Москва, 1976, 511 с.

[9] Рязанов В.В. Комитетный синтез алгоритмов распознавания и классификации // ЖВМ и МФ. 1981. Том 21, №6. С.1533-1543.

[10] Рязанов В.В. О синтезе классифицирующих алгоритмов на конечных множествах алгоритмов классификации (таксономии) // ЖВМ и МФ, 1982. Том 22, №2. С.429-440.

[11] Бирюков А.С., Рязанов В.В., Шмаков А.С. Решение задач кластерного анализа коллективами алгоритмов . Журнал вычислительной математики и математической физики, М.: Наука. Т.48, 2008, N 1, стр. 176-192

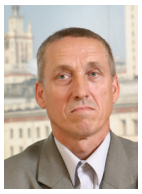
[12] Ю.И.Журавлев, В.В.Рязанов, О.В.Сенько, РАСПОЗНАВАНИЕ. Математические методы. Программная система. Практические применения. Изд.во «ФАЗИС», Москва, 2006, 178 стр.

[13] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[14] Sigillito, V. G., Wing, S. P., Hutton, L. V., \& Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest, 10, 262-266

[15] Арсеев А.С., Коточигов К.Л., Рязанов В.В. : Универсальные критерии кластеризации и вопросы устойчивости. Труды 13-й Всероссийской конференции «Математические методы распознавания образов», Санкт-Петербург, 2007, с. 63-64.

Информация об авторах



Vladimir Ryazanov *– Heard of Department; Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Russia,* 119991 Moscow, Vavilov's street, 40; *e-mail:* rvv@ccas.ru

Major Fields of Scientific Research: Pattern recognition, Data mining, Artificial Intelligence

Kirill Tishin *– The last year student; Lomonosov Moscow State University; Faculty of Computational Mathematics and Cybernetics; Department of Mathematical Forecasting Methods, Russia, 119991, Moscow, GSP-1, Leninskie Gory, 1, p. 52; e-mail: kirill.tishin@gmail.com*

Major Fields of Scientific Research: Pattern recognition, Data mining, Artificial Intelligence

Anton Schichko *– The last year student; Lomonosov Moscow State University; Faculty of Computational Mathematics and Cybernetics; Department of Mathematical Forecasting Methods, Russia, 119991, Moscow, GSP-1, Leninskie Gory, 1, p. 52; e-mail: anton.schichko@gmail.com*

Major Fields of Scientific Research: Pattern recognition, Data mining, Artificial Intelligence