A. A. Novosyolov

PARAMETRIZATION OF MODELS OF CONTROLLED SYSTEMS

The paper describes application of orthogonal series method for construction of controlled systems models under non-parametric uncertainty. A key element of the method is draw of orthogonal expansion length based on observations, in other words, defining parametric structure of the model. The method is demonstrated for estimation of distribution density and regression function. Directions for generalizing onto multi-dimensional case are also presented.

Keywords: distribution density, regression function, orthogonal series, non-parametric estimate.

УДК 004.932.2

И. А. Пестунов, В. Б. Бериков, Ю. Н. Синявский

СЕГМЕНТАЦИЯ МНОГОСПЕКТРАЛЬНЫХ ИЗОБРАЖЕНИЙ НА ОСНОВЕ АНСАМБЛЯ НЕПАРАМЕТРИЧЕСКИХ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ*

Предложен метод сегментации многоспектральных изображений на основе ансамбля непараметрических алгоритмов; дано теоретическое обоснование метода. Результаты статистического моделирования на модельных данных и реальных изображениях подтверждают эффективность предложенного метода.

Ключевые слова: сегментация многоспектральных изображений, непараметрические алгоритмы кластеризации, ансамблевый подход.

Сегментация является одним из важнейших этапов анализа цифровых изображений [1]. Она заключается в разбиении изображения на сегменты на основе подобия спектральных, текстурных и других характеристик пикселов. Методы сегментации нашли широкое применение во многих прикладных областях, в том числе в дистанционном зондировании Земли (ДЗЗ) [2; 3], интерес к которому в последние годы непрерывно возрастает.

Один из наиболее распространенных подходов к сегментации многоспектральных изображений основан на статистических методах кластеризации [4; 5]. В этом случае задачу кластеризации, как правило, приходится решать при отсутствии каких-либо априорных сведений о числе классов и их вероятностных характеристиках. Для этого наиболее подходящими являются непараметрические алгоритмы, позволяющие получить хорошие результаты при минимальной априорной информации. Их общим недостатком является высокая чувствительность к входным параметрам, что существенно усложняет процесс настройки алгоритма для решения конкретной задачи.

Известно [6–8], 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 различными алгоритмами либо одним алгоритмом с различными значениями параметров, по разным подсистемам переменных и т. д. В настоящее время ансамблевый подход является одним из наиболее перспективных направлений в кластерном анализе [9].

В данной работе предложен алгоритм сегментации многоспектральных изображений с использованием ансамбля непараметрических алгоритмов кластеризации, основанных на оценках плотности Розенблатта–Парзена [10; 11]. Для формирования ансамбля используются результаты выполнения непараметрического алгоритма MeanSC (представляющего собой оптимизацию предложенного ранее алгоритма [12]) с различными значениями параметра сглаживания h. Итоговое коллективное решение строится на основе попарной классификации объектов. Дано теоретическое обоснование предложенного алгоритма, приведены результаты статистического моделирования на модельных данных и реальных изображениях, подтверждающие его эффективность.

Непараметрический подход к задаче кластеризации данных ДЗЗ. Предположим, что произведена k-спектральная съемка участка местности, содержащего N элементов разрешения, тогда результат съемки можно представить в виде множества $X = \{x^{(i)} = (x_1^{(i)}, ..., x_k^{(i)}) \in R^k,\ i = \overline{1, N}\}$, где $x_j^{(i)}$ – значение яркости i-го элемента разрешения в j-м диапазоне спектра $(j = \overline{1, k})$.

* Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (код проекта 09-07-12087-офи_м).

Пусть каждый вектор $x^{(i)}$ – реализация k-мерного случайного вектора x, плотность распределения которого $f(x)$, $x=(x_1,\ldots,x_k)\in R^k$ неизвестна и нет какой-либо априорной информации о ее параметрическом виде. В этих условиях для оценивания плотности $f(x)$ в точке $x\in R^k$ целесообразно воспользоваться непараметрической оценкой Розенблатта–Парзена $\hat{f}_N(x)$, определяемой выражением

$$\hat{f}_N(x)=\frac{1}{Nh^k}\sum_{i=1}^{N}\Phi\left(\frac{x-x^{(i)}}{h}\right),$$

где h – параметр сглаживания; $\Phi(x)$ – колоколообразная функция (ядро), удовлетворяющая определенным условиям сходимости [10; 13].

Среди ядер, удовлетворяющих этим условиям, наибольшей популярностью пользуются радиально-симметричные ядра, представимые в виде

$$\Phi(x)=c_k\phi\left(\|x\|^2\right),$$

где константа $c_k>0$; $\phi:[0;\infty)\to R$ – непрерывная функция, удовлетворяющая условиям:

$$\phi(t)\ge 0;\ \phi(t_1)\ge\phi(t_2),\ \text{если } t_1<t_2;\ \int_0^{\infty}\phi(t)dt<\infty.$$

К таким ядрам относятся, например, ядро Епанечникова [13]:

$$\Phi_{\mathrm{E}}=\begin{cases}\frac{1}{2}V_k^{-1}(k+2)(1-\|x\|^2), & \text{если } \|x\|\le 1,\\ 0, & \text{иначе,}\end{cases}$$

где V_k – объем k-мерного единичного шара и многомерное нормальное ядро;

$$\Phi_{\mathrm{N}}(x)=(2\pi)^{-k/2}e^{-\frac{\|x\|^2}{2}}.$$

При использовании радиально-симметричных ядер, оценки неизвестной плотности распределения $f(x)$ и ее градиента $\nabla f(x)$ могут быть записаны ([14]) в виде

$$\hat{f}_N(x)=\frac{c_k}{Nh^k}\sum_{i=1}^{N}\phi\left(\left\|\frac{x-x_i}{h}\right\|^2\right);$$

$$\hat{\nabla}f(x)=\nabla\hat{f}_N(x)=\frac{2c_k}{Nh^{k+2}}\sum_{i=1}^{N}(x-x_i)\,\phi'\left(\left\|\frac{x-x_i}{h}\right\|^2\right).$$

Обозначим $\psi(t)=-\phi'(t)=-d\phi(t)/dt$, предполагая, что функция $\phi(t)$ дифференцируема $\forall t\in[0;\infty)$ за исключением конечного множества точек.

Тогда, согласно [15; 16]

$$m_h(x)=\frac{\sum_{i=1}^{N}x_i\psi\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^{N}\psi\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}-x-$$

вектор среднего сдвига.

Этот вектор интересен тем, что его направление совпадает с направлением градиента оценки плотности $f(x)$ в точке x.

Итерационную процедуру, заключающуюся в переходе от $x_0\in R^k$ к $x_0^1=x_0+m_h(x_0)$, затем от x_0^1 к $x_0^2=x_0^1+m_h(x_0^1)$ и т. д. до точки x_0^*, для которой $m_h(x_0^*)=0$, называют алгоритмом среднего сдвига. Доказано [14], что эта процедура сходится к локальным максимумам (модам) плотности распределения $f(x)$. Путь, пройденный от точки x_0 до моды x_0^*, будем называть траекторией среднего сдвига и обозначать $\left[x_0,\ldots,x_0^*\right]$.

Процедура среднего сдвига порождает естественное разбиение множества X на компоненты связности: точки x_i и x_j связны, если итеративные процессы среднего сдвига, начинающиеся с этих точек, сходятся к одной и той же моде. Эта процедура достаточно трудоемка, поэтому ее непосредственное применение ограничено выборками небольшого объема. В работах [17; 18] процедура среднего сдвига применяется не ко всей исходной выборке, а к некоторому ее подмножеству значительно меньшего объема.

В следующем разделе приведено описание быстрого алгоритма кластеризации многоспектральных данных MeanSC (являющегося оптимизацией предложенного ранее алгоритма [12]), в котором стартовое множество точек для запуска процедуры среднего сдвига порождается клеточной структурой данных, формируемой в пространстве спектральных признаков. В этом алгоритме вектор $m_h(x)$ и оценка плотности $\hat{f}_N(x)$ вычисляются с использованием финитных радиально-симметричных ядер.

Описание алгоритма MeanSC. Предлагаемый алгоритм опирается на использование двух характерных особенностей многоспектральных данных. Первая из них заключается в ограниченности диапазонов изменения значений спектральных признаков (значения лежат в диапазоне целых чисел от 0 до $K-1$, где K – число уровней квантования видеосигнала, обычно не превышающее 256), а вторая – в высокой частоте повторяемости векторов спектральных яркостей. Повторяемость обусловливается ограниченностью диапазона спектральных яркостей, наличием корреляции между спектральными диапазонами, а также относительной однородностью и достаточной протяженностью природных объектов. Для описания алгоритма MeanSC с параметрами h, ε, T введем следующие определения.

Определение 1. Пусть в точке $x^*\in R^k$ достигается локальный максимум оценки плотности $\hat{f}_N(x)$. Тогда точка $x\in R^k$ *связна* с x^*, если процедура среднего сдвига, стартовавшая из x, сходится к x^*. В дальнейшем через a^* будем обозначать моду плотности, к ко-

торой сойдется процедура среднего сдвига, стартовавшая в *a*.

Определение 2. *Компонентой связности*, определяемой локальным максимумом x^* ($\hat{f}_N(x^*) \geq \varepsilon$), назовем непустое подмножество точек $Q(x^*) \subseteq X$, связных с x^*. Точку $x \in X$ будем считать *«шумом»*, если она связна с локальным максимум x_0^*: $\hat{f}_N(x_0^*) < \varepsilon$. Здесь $\varepsilon > 0$ – порог «шума».

Определение 3. *Кластером* назовем подмножество $C \subseteq X$, которое либо задано одной компонентой связности, либо задано множеством компонент связности $\mathbb{C}$ и удовлетворяет условиям:

1) $\forall\, x \in C$ выполнено $Q(x^*) \in \mathbb{C}$;

2) $\forall\, x_1^*, x_2^*$ таких, что $Q(x_1^*), Q(x_2^*) \in \mathbb{C}$ существует непрерывная траектория $P \subset R^k$, соединяющая x_1^* и x_2^*, вдоль которой $\dfrac{\hat{f}_N(x)}{\min\left(\hat{f}_N(x_1^*), \hat{f}_N(x_2^*)\right)} > T$. Здесь $T \geq 1$ – параметр, задаваемый пользователем и отвечающий за уровень детализации результата.

В соответствии с введенными определениями, алгоритм MeanSC можно записать в виде следующей последовательности шагов.

1. Формируем клеточную структуру данных в пространстве спектральных признаков. Для этого разбиваем все пространство значений спектральных признаков $[0; K-1]_1 \times \ldots \times [0; K-1]_k$ на гиперкубические клетки со стороной $2h$ (h – параметр сглаживания). Вводим общую нумерацию клеток (последовательно от одного слоя клеток к другому) и с каждой клеткой связываем набор попавших в нее спектральных векторов из X.

2. Формируем таблицу «весов» векторов множества X. Здесь под «весом» вектора x понимаем число вхождений x в множество X. При обработке спутниковых изображений таблица «весов» позволяет значительно (иногда в несколько десятков раз) сократить объем вычислений при выполнении процедуры среднего сдвига и вычислении оценок плотности распределения.

3. Формируем множество начальных (стартовых) векторов S для запуска процедуры среднего сдвига. Для каждой клетки, которая содержит векторы из X, вычисляем вектор средних значений по всем точкам, попавшим в эту клетку. Совокупность полученных таким образом векторов образуют множество S.

4. Для каждого вектора $s \in S$ находим моду s^* оценки плотности распределения $\hat{f}_N(x)$, связную с s. Из найденных мод формируется множество $Z_0 = \left\{s^* \mid s \in S, \hat{f}_N(s^*) \geq \varepsilon\right\}$. По мере нахождения мод заполняем множество $\overline{S} = \bigcup\limits_{s \in S} \{\overline{s}_0 = s, \ldots, \overline{s}_i = \overline{s}_{i-1} + m_h(\overline{s}_{i-1}), \ldots, s^*\}$, которое содержит все точки, пройденные при выполнении процедуры среднего сдвига.

5. Связываем каждую точку $x \in X$ с ближайшей точкой из множества $\overline{S}$, используя для сокращения вычислительных расходов введенную клеточную структуру. В результате множество X разбивается на компоненты связности в соответствии с определением 2.

6. Формируем кластеры из выделенных компонент связности. Если искомая траектория P, соединяющая $Q(x_1^*)$ и $Q(x_2^*)$ (определение 3), существует, она, наиболее вероятно, проходит через общую границу $Q(x_1^*)$ и $Q(x_2^*)$. Поэтому для ее нахождения выбираем точки $x_1 \in Q(x_1^*)$ и $x_2 \in Q(x_2^*)$, расположенные ближе всего к общей границе, и проверяем выполнение условий из определения 3 для $P = \left[x_1, \ldots, x_1^*\right] \cup \left[x_1, x_2\right] \cup \left[x_2, \ldots, x_2^*\right]$. Здесь $\left[x_1, x_2\right]$ – отрезок прямой, соединяющей точки x_1 и x_2. Исходя из определения процедуры среднего сдвига, точка $x \in P$ с наименьшей плотностью находится на отрезке $\left[x_1; x_2\right]$. Точки отрезка проверяем последовательно с шагом h.

Заметим, что благодаря финитности используемого ядра, при вычислении оценки плотности и векторов среднего сдвига достаточно использовать только векторы из клеток, которые являются соседними к клетке, содержащей точку x. Это позволяет значительно уменьшить объем вычислений.

Метод построения коллективного решения. Результаты кластеризации, получаемые с помощью алгоритма MeanSC, являются неустойчивыми к изменению параметра сглаживания h. Известно [6–8], что устойчивость результатов кластеризации может быть повышена путем использования ансамблевого подхода. При этом для построения коллективного решения используются различные принципы. Так, в [7] предлагается принцип максимизации количества взаимной информации, которую разделяет итоговое группировочное решение с исходными кластеризациями. В ряде работ используется принцип, основанный на нахождении согласованной матрицы подобия (или различия) объектов. В данной работе будет использован именно этот принцип; ансамбль предлагается формировать из L частных решений, полученных в результате выполнения алгоритма MeanSC с различными значениями параметра сглаживания h.

Предлагаемый метод построения коллективного решения может быть описан следующим образом.

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

Обозначим через $H(\theta_l)$ бинарную матрицу $H(\theta_l)=\{H_{i,j}(\theta_l)\}$ размерности $N\times N$, которая водится для l-й группировки, следующим образом:

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

где $i,j=1,\ldots,N,\ i\neq j$.

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 $\mathbb{G}$. 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 говорит о том, что шанс оказаться в одной группе у объектов незначителен.

После вычисления согласованной матрицы различий, для нахождения коллективного решения будем применять стандартный агломеративный метод построения дендрограммы, который в качестве входной информации использует попарные расстояния между объектами [9]. 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. Этот метод является очень привлекательным, потому что он позволяет выделять иерархическую структуру кластеров, которая упрощает процесс интерпретации результатов.

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

Предположим, что имеется некоторая скрытая (непосредственно ненаблюдаемая) переменная U, которая задает принадлежность каждого объекта к некоторому из $M\geq 2$ классов. Каждый класс характеризуется определенным законом условного распределения $p(x\,|\,U=r)=f_r(x),\ r=1,\ldots,M$. Рассмотрим следующую вероятностную модель генерации данных. Пусть для каждого объекта определяется класс, к которому он относится, в соответствии с априорными вероятностями $P_r=\mathbb{P}(U=r),\ r=1,\ldots,M$, где $\sum_{r=1}^{M}P_r=1$. Затем в соответствии с распределением $f_r(x)$ определяется значение x. Указанная процедура проводится независимо для каждого объекта.

Пусть с помощью некоторого алгоритма кластерного анализа μ строится разбиение множества объектов X 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. Определим для каждой пары объектов a и b следующую величину:

$$\mathbf{H}_{a,b}(\mu)=\begin{cases}0, & \text{если объекты отнесены в один кластер,}\\ 1, & \text{иначе,}\end{cases}$$

где $a,b\in X,\ a\neq b$.

Выберем произвольную пару a и b различных объектов выборки.

Пусть $P_U=\mathbb{P}(U(a)\neq U(b))$ – вероятность отнесения объектов к различным классам. Например, при $M=2$ указанная вероятность равна

$$P_U=1-\mathbb{P}(U(a)=1\,|\,a)\mathbb{P}(U(b)=1\,|\,b)-\mathbb{P}(U(a)=$$
$$=2\,|\,a)\mathbb{P}(U(b)=2\,|\,b)=1-\sum_{r=1}^{2}\frac{f_r(a)f_r(b)P_r^2}{p(a)p(b)},$$

где $p(\omega)=\sum_{r=1}^{2}f_r(\omega)P_r,\ \omega=a,b$.

Обозначим вероятность ошибки, которую может совершить алгоритм μ при классификации a и b через $P_{\text{er}}(\mu)$, где

$$P_{\text{er}}(\mu)=\begin{cases}P_U, & \text{если } \mathbf{H}_{a,b}(\mu)=0,\\ 1-P_U, & \text{если } \mathbf{H}_{a,b}(\mu)=1.\end{cases}$$

Легко заметить, что

$$P_{\text{er}}(\mu)=(1-\mathbf{H}_{a,b}(\mu))P_U+\mathbf{H}_{a,b}(\mu)(1-P_U)=$$
$$=P_U+(1-2P_U)\mathbf{H}_{a,b}(\mu).$$

Алгоритм μ зависит от случайного вектора параметров $\Theta\in\boldsymbol{\Theta}$: $\mu=\mu(\Theta)$. Чтобы подчеркнуть зависимость результатов работы от параметра Θ, в дальнейшем будем обозначать $\mathbf{H}_{a,b}(\mu(\Theta))=\mathbf{H}_{a,b}(\Theta)$, $P_{\text{er}}(\mu(\Theta))=P_{\text{er}}(\Theta)$.

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 определенности, будем считать, что L – нечетно. Коллективным (ансамблевым) решением по большинству голосов будем называть функцию

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

В рамках описанной модели для предложенного коллективного решения справедливы следующие утверждения [6].

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

$$\mathbb{E}_{\Theta}P_{\text{er}}(\Theta)=P_U+(1-2P_U)P_H,$$
$$\text{Var}_{\Theta}P_{\text{er}}(\Theta)=(1-2P_U)^2P_H(1-P_H),$$

где $P_H=\mathbb{P}(H(\Theta)=1)$.

Обозначим через $P_{\text{er}}(\Theta_1, \ldots, \Theta_L)$ случайную функцию, принимающую при фиксированных аргументах значение, равное вероятности ошибки, которую может совершить ансамблевый алгоритм при классификации a и b. Здесь через $\Theta_1, \ldots, \Theta_L$ обозначены статистические копии случайного вектора Θ. Рассмотрим поведение вероятности ошибки для коллективного решения.

Утверждение 2. Математическое ожидание и дисперсия величины вероятности ошибки для коллективного решения равны соответственно:

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

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

где $P_{\mathbf{H},L} = \mathbb{P}\left(\frac{1}{L}\sum_{l=1}^{L} H(\theta_l) \geq \frac{1}{2}\right) = \sum_{l=\left[\frac{L}{2}\right]+1}^{L} C_L^l P_H^l (1 - P_H)^{L-l}$,

$[\cdot]$ означает целую часть числа.

Воспользуемся следующей априорной информацией об алгоритме кластерного анализа. Будем считать, что ожидаемая вероятность ошибочной классификации $\mathbb{E}_\Theta P_{\text{er}}(\Theta) < 1/2$, т. е. ожидается, что алгоритм μ проводит классификацию с лучшим качеством, нежели алгоритм случайного равновероятного выбора. Из утверждения 1 следует, что выполняется один из двух вариантов: $P_H > 1/2$ и $P_U > 1/2$; $P_H < 1/2$ и $P_U < 1/2$. Рассмотрим, для определенности, первый случай.

Утверждение 3. Если $\mathbb{E}_\Theta P_{\text{er}}(\Theta) < 1/2$ и при этом $P_H > 1/2$ и $P_U > 1/2$, то при увеличении мощности ансамбля ожидаемая вероятность ошибочной классификации уменьшается, стремясь в пределе к $1 - P_U$, а дисперсия величины вероятности ошибки стремится к нулю.

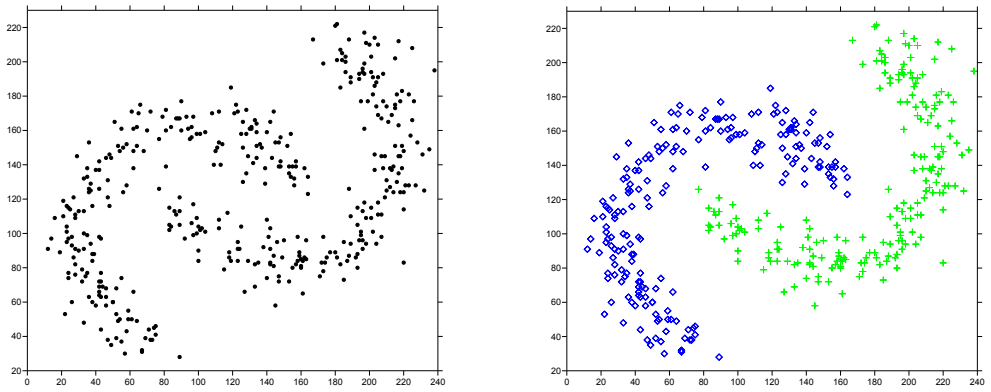
Последнее утверждение позволяет сделать вывод о том, что при выполнении вполне естественных условий использование ансамблевого подхода позволяет улучшить качество кластеризации.

Результаты экспериментальных исследований. В соответствии с описанной выше схемой, разработан и программно реализован на языке программирования C++ ансамблевый алгоритм $\text{EMeanSC}\left(\bar{h} = \{h_1, \ldots, h_L\}, \varepsilon, T, T_d\right)$. Здесь $h_1, \ldots, h_L$ – значения параметра сглаживания; ε – порог «шума»; T – порог объединения компонент связности, а T_d – параметр дендрограммы.

Ниже приведены результаты нескольких экспериментов на модельных данных и реальных изображениях. Эксперименты 1 и 2 подтверждают эффективность алгоритма для выделения классов сложной формы. Эксперимент 3 демонстрирует возможность разделения пересекающихся классов разной плотности. Эксперимент 4 демонстрирует применение алгоритма для обработки спутниковых изображений. В эксперименте 5 показано влияние параметра T_d на уровень детализации картосхемы.

Эксперимент 1. Использовались двумерные данные, состоящие из 400 точек, сгруппированных в два равновеликих линейно неразделимых класса (рис. 1), имеющих форму бананов (рис. 1, *а*). Модель построена с помощью инструментария PRTools (PRTools: the Matlab Toolbox for Pattern Recognition – http://www.prtools.org). Результаты кластеризации, построенные по ансамблю из шести элементов, представлены на рис. 1, *б*. Выделено 2 кластера, содержащих по 200 точек.

Эксперимент 2. Использовались двумерные данные (рис. 2), состоящие из 200 двумерных точек, сгруппированных в два спиралевидных класса по 100 точек (рис. 2, *а*). Сложность этой модели заключается в том, что плотность точек по мере удаления от центра спирали уменьшается. На рис. 2, *б* представлены результаты выполнения алгоритма MeanSC с двумя значениями параметра сглаживания. Несложно заметить, что искомые классы оказались раздробленными. Ансамбль, сформированный на основе указанных результатов, успешно выделил кластеры, совпадающие с искомыми классами (рис. 2, *в*).



а *б*

Рис. 1. Двумерные данные, состоящие из 400 точек:
а – исходные данные; *б* – результат выполнения EMeanSC с параметрами
$\bar{h} = \{10; 10{,}5; 11; 11{,}5; 12; 12{,}5\}$, $\varepsilon = 0$, $T = 4$, $T_d = 0{,}99$ (2 кластера)

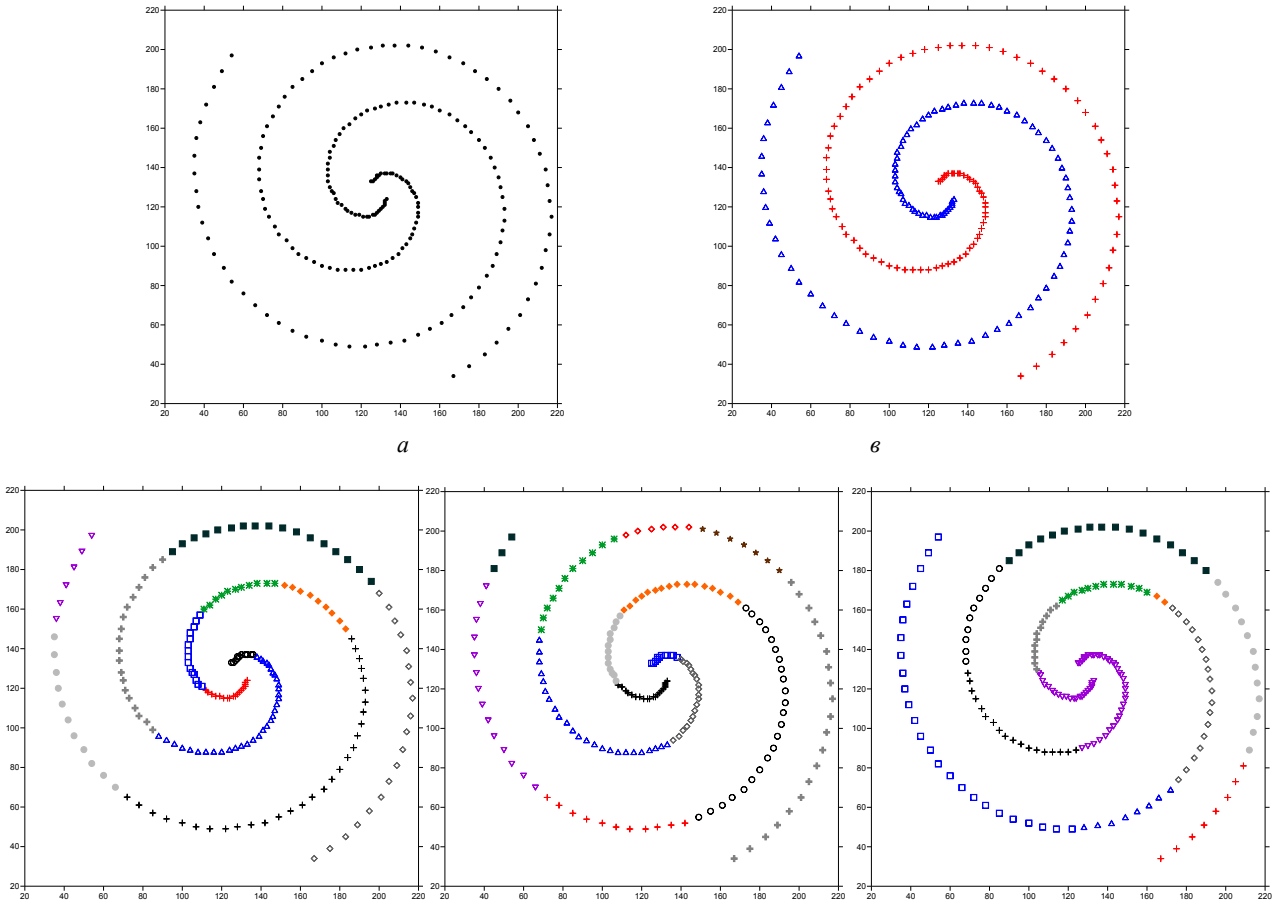


Рис. 2. Двумерные данные, состоящие из 200 точек:
а – исходные данные; *б* – результаты выполнения алгоритма MeanSC (12, 14 и 12 кластеров соответственно)
с параметрами $h \in \{8; 8,5; 9\}$, $\varepsilon = 0$, $T = 5,5$; *в* – результат выполнения алгоритма EMeanSC
с параметрами $\overline{h} = \{8; 8,5; 9\}$, $\varepsilon = 0$, $T = 5,5$, $T_d = 0,999$ (2 кластера)

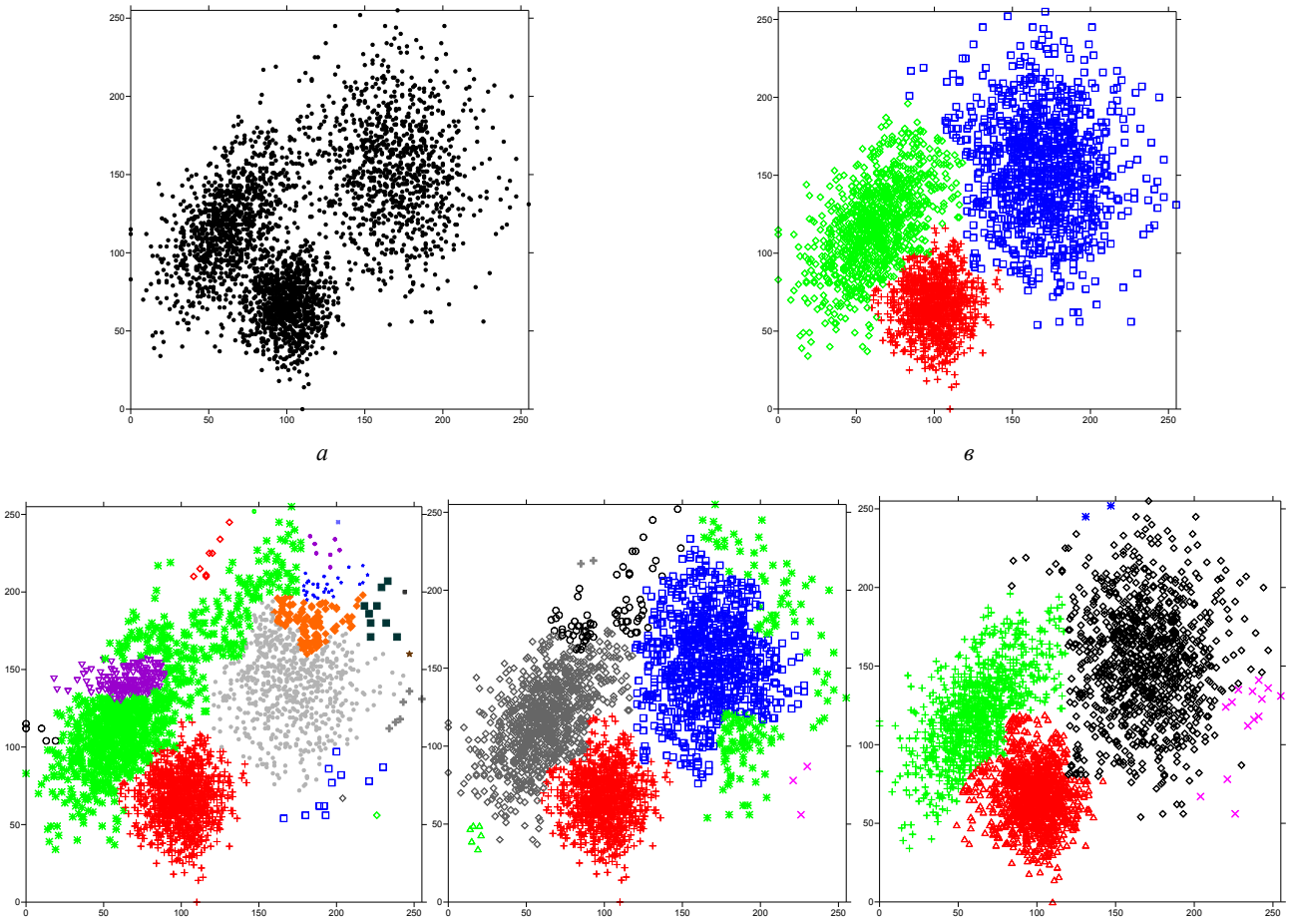
Эксперимент 3. Использовались двумерные данные, состоящие из 3 000 точек (рис. 3), сгруппированных в 3 нормально распределенных класса (рис. 3, *а*). Большинство кластеризаций, на основе которых формировался ансамбль, содержат достаточно грубые ошибки (дробление и пересечение классов) (рис. 3, *б*). При этом результаты выполнения ансамблевого алгоритма (рис. 3, *в*) не содержат грубых ошибок.

Эксперимент 4. Использовался фрагмент снимка Болотнинского района Новосибирской области (размером 500 × 450), полученного со спутника ALOS/ANVIR-2 17 июня 2007 г. (рис. 4). Исследуемый участок ограничен 55°52'14.8" и 55°55'2.13" северной широты и 83°50'45.41" и 83°54'51.6" восточной долготы. Исходный фрагмент представлен на рис. 4, *а*. Обработка выполнялась по трем спектральным каналам. Обработка производилась на ПК с тактовой частотой 2,4 ГГц (объем оперативной памяти 2 Гб). Время обработки с двумя значениями параметра сглаживания составляет 18 с. Алгоритм выделил пять классов (рис. 4, *б*).

Эксперимент 5. Использовалось цветное изображение (рис. 5, *а*) размером 510×604 пикселов. Кластеризация выполнялась в цветовом пространстве $R \times G \times B$. Каждый кластер соответствовал однородной области на изображении. Срезам дендрограммы, построенной в ходе выполнения алгоритма EMeanSC, на уровнях, соответствующих различным значениям параметра T_d, соответствуют рис. 5, *б–г*. Параметр дендрограммы управляет степенью раздробленности получаемых кластеров, что позволяет получить картосхему с необходимым пользователю уровнем детализации.

В работе представлен метод комбинирования ансамблевого и непараметрического подходов для кластеризации изображений, позволяющий повысить качество и устойчивость получаемых результатов. Дано его теоретическое обоснование. В соответствии с этим методом создан алгоритм EMeanSC, основанный на непараметрических оценках плотности Розенблатта–Парзена. Показано, что алгоритм способен разделять кластеры сложной структуры и может быть использован для сегментации многоспектральных изображений.

Заметим, что предложенный метод построения ансамблевых алгоритмов допускает распараллеливание наиболее трудоемких этапов обработки, позволяющее повысить быстродействие при реализации его на многопроцессорных вычислительных системах.



а *в* *б*

Рис. 3. Двумерные данные, состоящие из 3 000 точек:
а – исходные данные; *б* – 3 из 5 элементов ансамбля (19, 8 и 5 кластеров, соответствующих значениям $h \in \{10; 15; 20\}$);
в – результат выполнения EMeanSC с параметрами $\overline{h} = \{5; 10; 15; 20; 25\}$, $\varepsilon = 0$, $T = 2,5$, $T_d = 0,75$ (3 кластера)

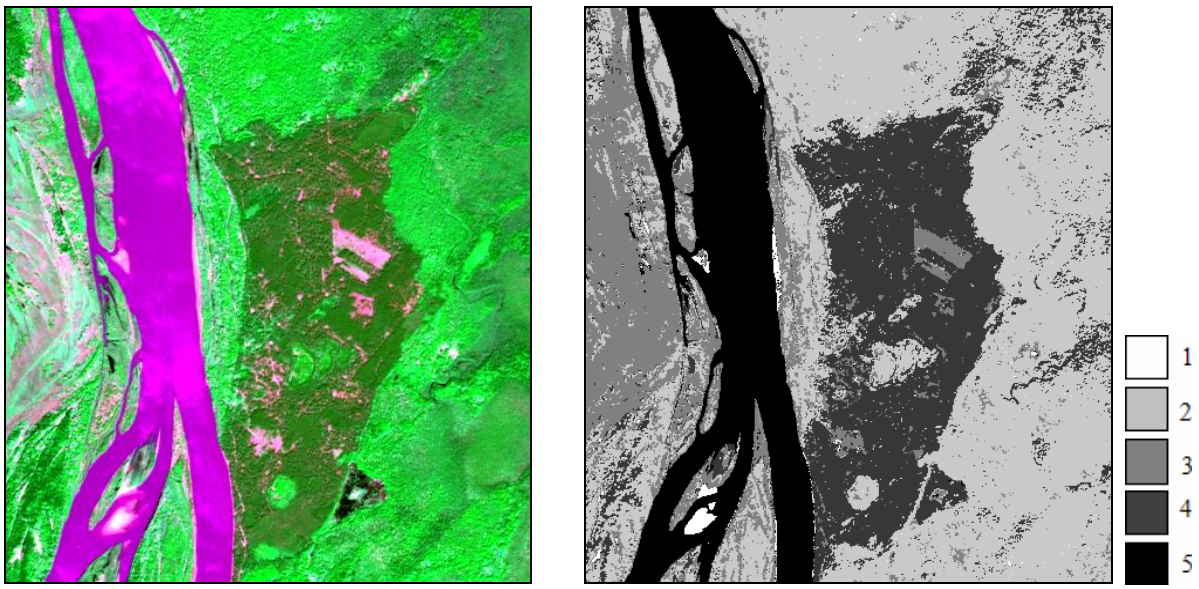
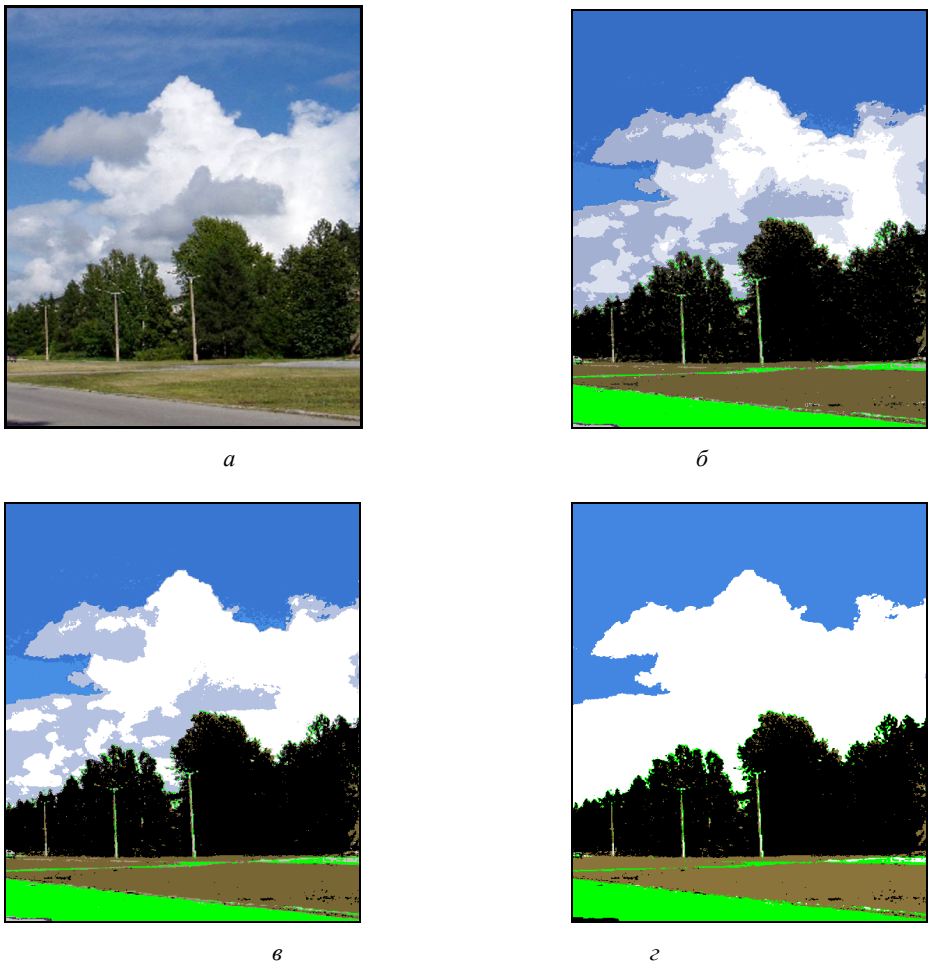


Рис. 4. Фрагмент снимка Болотнинского района Новосибирской области:
а – снимок ALOS; *б* – результаты выполнения EMeanSC с параметрами $\overline{h} = \{5; 10; 15\}$, $\varepsilon = 0$, $T = 1,3$, $T_d = 0,5$.
Выделено 6 кластеров: *1* – песчаные отложения; *2* – березовые и березово-осиновые травяные леса; *3* – луговая и кустарниковая растительность в пойме реки; *4* – сосновые травяные и травяно-кустарничковые леса; *5* – река Обь



а *б*

в *г*

Рис. 5. Изображение:
а – исходное изображение; *б*, *в*, *г* – результаты выполнения EMeanSC с параметрами $\overline{h} = \{5; 7; 9\}$, $\varepsilon = 0$, $T = 1,5$, $T_d = 0,4$ (18 кластеров), 0,6 (12 кластеров) и 0,95 (6 кластеров) соответственно

Библиографические ссылки

1. Гонсалес Р., Вудс Р. Цифровая обработка изображений. М. : Техносфера, 2006. С. 812.

2. Dey V., Zhang Y., Zhong M. A review on image segmentation techniques with remote sensing perspective // ISPRS TC VII Symposium – 100 Years ISPRS, Vienna. Austria. July 5–7 2010. IAPRS. Vol. XXXVIII. Part 7A. P. 31–42.

3. Rekik A., Zribi M., Hamida A., Benjelloun1 M. Review of satellite image segmentation for an optimal fusion system based on the edge and region approaches // IJCSNS International Journal of Computer Science and Network 242 Security. 2007. Vol. 7. № 10. P. 242–250.

4. Jain A. K., Duin R. P. W., Mao J. Statistical Pattern Recognition: A Review // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000. Vol. 22. № 1. P. 4–37.

5. Clausi D. A. K-means Iterative Fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation // Pattern Recognition. 2002. Vol. 35. № 9. P. 1959–1972.

6. Бериков В. Б. Построение ансамбля деревьев решений в кластерном анализе // Вычислительные технологии. 2010. Т. 15. № 1. С. 40–52.

7. Strehl A., Ghosh J. Clustering ensembles – a knowledge reuse framework for combining multiple partitions // The Journal of Machine Learning Research. 2002. Vol. 38. P. 583–617.

8. Hong Y., Kwong S. To combine steady-state genetic algorithm and ensemble learning for data clustering // Pattern Recognition Letters. 2008. Vol. 29(9). P. 1416–1423.

9. Jain A. K. Data clustering: 50 years beyond K-means // Pattern Recognition Letters. 2010. Vol. 31, Is. 8. P. 651–666.

10. Parzen E. On the estimation of a probability density function and the mode // The Annals of Mathematical Statistics. 1962. Vol. 33. P. 1065–1076.

11. Rosenblatt M. Remarks on some nonparametric estimates of a density function // The Annals of Mathematical Statistics. 1956. Vol. 27. P. 832–837.

12. Пестунов И. А., Синявский Ю. Н. Непараметрический алгоритм кластеризации данных дистанци-

онного зондирования на основе grid-подхода // Автометрия. 2006. Т. 42. № 2. С. 90–99.

13. Епанечников В. А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятностей и ее применение. 1969. Т. 14. № 1. С. 156–160.

14. Comaniciu D., Meer P. Mean shift: A Robust Approach toward Feature Space Analysis // IEEE Transactions on Pattern Analysis Machine Intelligence. 2002. Vol. 24. № 5. P. 603–619.

15. Fukunaga K., Hosteeler L. D. The estimation of the gradient of a density function, with applications in patter recognition // IEEE Trasactions on Informational Theory. 1975. Vol. 21. P. 32–40.

16. Cheng Y. Mean shift, mode seeking, and clustering // IEEE Tans. Pattern Analysis and Machine Intelligence. 1995. Vol. 17. P. 790–799.

17. Comaniciu D., Meer P. Distribution Free Decomposition of Multivariate Data // Pattern Analysis and Applications. 1999. Vol. 2. P. 22–30.

18. Freedman D., Kisilev P. Fast Mean Shift by Compact Density Representation // IEEE Conference on Computer Vision and Pattern Recognition. 2009. P. 1818–1825.

I. A. Pestunov, V. B. Berikov, Yu. N. Sinyavskiy

ALGORITHM FOR MULTISPECTRAL IMAGE SEGMENTATION BASED ON ENSEMBLE OF NONPARAMETRIC CLUSTERING ALGORITHMS

The method for constructing an ensemble of nonparametric clustering algorithms is proposed. Its theoretical substantiation is resulted. Results of the model data and real images confirm the efficiency of the proposed method.

Keywords: multispectral image segmentation, nonparametric clustering algorithms, ensemble approach.

УДК 519.24

А. Н. Сергеев

О НЕПАРАМЕТРИЧЕСКИХ АЛГОРИТМАХ ПРИНЯТИЯ РЕШЕНИЙ

Рассматриваются особенности и параметры, влияющие на деятельность организаций, осуществляется постановка задачи моделирования и управления организационной системой. Приводятся математические непараметрические модели организационных систем.

Ключевые слова: организационные системы, управление, измерения, неопределенность, случайные факторы, непараметрическое моделирование, алгоритмы принятия решений.

В понятии «организационная система» используются одновременно два нетривиальных термина: «организация» и «система».

Организация может рассматриваться как процесс либо как сущность [1]. Как процесс организация – это совокупность действий, ведущих к образованию и совершенствованию взаимосвязей между частями целого. Как сущность организация – это целевое объединение под единым началом ресурсов для реализации определенной программы на основании определенных правил и процедур.

Здесь надо отметить, что некоторые организации могут сами являться ресурсами для более крупных структур, в которые они входят. Одним из наиболее важных ресурсов, несомненно, является человеческий.

Термин «система» имеет множество вариантов определений в разной литературе. Рассел Л. Акофф [1] попытался сформулировать «ядро» определения: «Система есть целое, состоящее из двух или более частей, которое удовлетворяет следующим 6 условиям:

– целое обладает одним или более определяющими свойствами или функциями;

– каждая часть в этом множестве может влиять на поведение или свойства целого;

– существует подмножество частей, которое достаточно в одном или нескольких внешних условиях для выполнения определяющей функции целого;

– способ, которым любая существенная часть воздействует на поведение или свойства системы, зависит от поведения или свойств по крайнем мере одной другой существенной части системы;

– воздействие любого подмножества существенных частей на систему в целом зависит от поведения по крайней мере еще одного другого такого подмножества;

– система есть целое, которое не может быть разделено на независимые части без потери ее существенных свойств или функций».

Акофф вводит для ресурса два различаемых свойства-термина: «целенаправленный» и «целеориентированный».

«Некая сущность является целеустремленной, если она может выбирать цели и средства в двух или более разных обстоятельствах» [1]. Если же сущность, имея