Вісник Харківського національного університету
Серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління»
№ 703, 2005, с.144-157

УДК 004.652/.942

Об одном классе алгоритмов кластер-анализа

А. А. Карпенко
Харьковский национальный университет им. В.Н. Каразина, Украина

Classification - one of fundamental processes in a science which is basic in practical and scientific activity of the person. Mathematical theory were offered, on the basis of which algorithms were developed, which application results in reduction of capacity of set of probable classifications, due to allocation of set of regular classifications was suggested in given article. It allows to facilitate a search extremum of functional qualities of splitting, and therefore a finding of the best classification for some configuration in tasks of automatic classification (cluster-analysis).

Часто утверждается, что *классификация* – один из фундаментальных процессов в науке. Это понятие тесно связано с такими терминами, как группировка, типологизация, кластеризация, систематизация, дискриминация, и является одним из основополагающих в практической и научной деятельности человека. Факты и явления должны быть упорядочены, прежде чем мы сможем их понять и разработать общие принципы, объясняющие их появление и видимый порядок. С этой точки зрения, классификация является интеллектуальной деятельностью высокого уровня, необходимой нам для понимания природы. Но поскольку классификация – это упорядочение объектов по их схожести, а объектом можно назвать все что угодно, включая процессы и действия – все, чему можно приписать вектор дескрипторов (вектор значений признаков)**,** – то можно прийти к заключению, что классификация не ограничена рамками усилий человеческого интеллекта и в действительности является фундаментальным свойством всех живых организмов.

Определяющим моментом в выборе подходящей математической постановки для конкретной задачи классификации является ответ на вопрос, на какой исходной информации будут основаны наши выводы. Исходную информацию целесообразно подразделять на:

- априорные сведения об искомых классах;
- предварительную выборочную информацию (так называемые обучающие выборки);
- наблюдения, подлежащие классификации.

Мы рассмотрим задачу классификации объектов без использования обучающего множества. Этот вид классификации называется *автоматической классификацией* или *классификацией без учителя*. Методы самообучения получили широкое распространение в интеллектуальных системах, в частности – в экспертных системах распознавания образов и классификации и т.д. В системах распознавания образов и классификации соответствующий класс задач обучения без учителя получил название *кластер-анализа* (т.е.

самопроизвольного разбиения исходной выборки на компактные подмножества, или *кластеры*).

Задача автоматической классификации не полностью определена, если не указаны свойства, которыми должны обладать искомые классы объектов. Выбор этих свойств или, что то же, определение класса – это основной вопрос теории автоматической классификации. Если имеется адекватное определение класса, становится возможным отличать хорошие классификации от плохих.

Попытки разработать методы автоматической классификация сделали необходимым оценивать сходство количественно. Одним из подходов к количественному определению оценки схожести заключается в попытке найти основу для суждений о сходстве. Это обычно достигается с помощью детального описания свойств, на основе которых, как полагают, можно выразить сходство. Этот подход привел к детализации и дроблению дескрипторов объектов, которые необходимо классифицировать. Каждому объекту приписываются длинные списки дескрипторов, т.е. векторы значений признаков, а классификация проводится по матрице данных, скомпонованной из набора таких векторов. От природы основных признаков объекта зависят важные теоретические выводы, но они различаются в зависимости от области применения.

Несмотря на широкое применение кластер-анализа, общепринятого определения кластеров не существует. Среди большинства разработчиков существует интуитивное понимание, что элементы одного кластера ближе друг к другу, чем к другим элементам, однако особенности этих отношений явно не называются. Для определения кластеров использовались различные параметры: плотность операционных таксономических единиц в признаковом гиперпространстве, объем, занимаемый кластером, связанность элементов определенного кластера, а также промежутки между соседними кластерами в сравнении с их диаметрами.

Кластер-анализ накладывает на объекты отношения на основе численных значений парных функций на этих объектах. Эти отношения нужны для выявления собственной структуры данных, но они зачастую сами налагают структуру, соответствующую особенностям алгоритма группировки. Таким образом, кластер-анализ не только раскрывает истинный порядок, регулярность или естественные законы, но также подгоняет данные под некоторую, заранее заготовленную модель.

При обзоре литературы были использованы источники [1, 2, 3].

1. Математическая постановка задачи автоматической классификации

Для того чтобы построить автоматическую процедуру решения задачи автоматической классификации, необходимо дать более строгое определение класса. Один из возможных путей – это конструирование *критерия качества классификации* [3]. Критерий качества классификации J (*функционал качества разбиения*) ставит в соответствие каждой возможной классификации множества объектов некоторое число. Областью определения J является множество всех возможных классификаций объектов, а областью значений – множество действительных чисел. Предполагается, что классификации, хорошие в смысле

принятого определения класса, соответствуют экстремальным значениям критерия J.

Таким образом, если критерий J задан, то можно оценить любую классификацию. Как правило, однако, нереально вычислить J для каждой возможной классификации. Это связано с тем, что *числа Белла* $B_n = e^{-1} \cdot \sum_{k \geq 0} \frac{k^n}{k!}$ (число способов разбиения n-элементного множества на подмножества) асимптотически равны $m(n)^n \cdot e^{m(n)-n-1/2} / \sqrt{\ln n}$, где $m(n) \cdot \ln m(n) = n - 1/2$ [4]. Поэтому для эффективного определения наилучшей в смысле критерия J классификации необходим *алгоритм автоматической классификации*. В соответствии с принятой нами системой определения критерий качества классификации и алгоритм автоматической классификации вместе составляют *процедуру решения задачи автоматической классификации*.

Выбор метрики (или меры близости) между объектами, каждый из которых представлен значениями характеризующего его многомерного признака, является узловым моментом исследования, от которого решающим образом зависит окончательный вариант разбиения объектов на классы при любом используемом для этого алгоритме разбиения. В каждой конкретной задаче этот выбор должен производиться по-своему, в зависимости от главных целей исследования, физической и статистической природы исследуемого многомерного признака, априорных сведений о его вероятностной природе и т.д.

Предположим, что мы хотим классифицировать N объектов, каждый из которых характеризуется n-мерным вектором, т.е. дано множество векторов $\{X_1, \ldots, X_N\}$. Мы не называем эти векторы случайными, поскольку в задача автоматической классификации они предполагаются фиксированными и известными. Каждый объект должен быть отнесен к одному из M классов, $\omega_1, \ldots, \omega_M$, где число классов M может быть, а может и не быть заранее известным. Класс, к которому относится i-й объект, обозначим ω_{k_i}, $i = 1, \ldots, N$. Для удобства будем предполагать, что k_i – целое число, заключенное между 1 и M. Классификацией Ω называют вектор, составленный из ω_i, а конфигурацией X^* – вектор, составленный из X_i. Критерий качества классификации J является функцией от Ω и X^*:

$$J = J(\omega_{k_1}, \omega_{k_2}, \ldots, \omega_{k_N}; X_1, X_2, \ldots, X_N) = J(\Omega, X^*).$$

По определению, наилучшая классификация Ω_0 удовлетворяет условию

$$J(\Omega_0; X^*) = \min_{\Omega} J(\Omega; X^*),$$

либо

$$J(\Omega_0; X^*) = \max_{\Omega} J(\Omega; X^*).$$

В задаче автоматической классификации конфигурация X^* фиксирована. Алгоритм автоматической классификации модифицирует только классификацию Ω. Обычные методы поиска экстремума здесь неприменимы

вследствие дискретного и неупорядоченного характера множества возможных классификаций.

В результате данной работы предполагается разработать некоторые алгоритмы, применение которых приводит к уменьшению мощности множества возможных классификаций, что позволяет облегчить поиск экстремума функционала качества разбиения $J(\Omega, X^*)$, а, следовательно, нахождение наилучшей классификации Ω_0 для конфигурации X^* в задачах автоматической классификации (кластер-анализа).

2. Необходимые определения, обозначения и утверждения

Будем рассматривать *класс* Γ *всех конечных неориентированных графов*. Пусть $G = G(V, E) \in \Gamma$ − конечный неориентированный граф, где V − непустое множество объектов некоторой природы, называемых вершинами графа, а E − множество его ребер − есть подмножество множества $V_-^2 / \sim$ классов эквивалентности, на которые множество $V_-^2 = \{(a,b), a \neq b\}$ разбивается следующим отношением эквивалентности

$$(a_1, b_1) \sim (a_2, b_2) \Leftrightarrow (a_1, b_1) = (a_2, b_2) \text{ или } (a_1, b_1) = (b_2, a_2).$$

Заметим, что данное определение не допускает наличие у графа кратных (параллельных) ребер и петель, т.е. графы из класса Γ являются простыми. Также следует отметить, что класс Γ не содержит нуль-графа, так как мы предположили, что $V \neq \varnothing$. Множество вершин и ребер графа G, обозначим $V(G)$ и $E(G)$, соответственно [5].

Пусть $v_1, v_2, ..., v_i, v_{i+1}, ..., v_k$ − последовательность вершин графа такая, что каждая пара соседних вершин v_i, v_{i+1} определяет ребро в этом графе. Тогда данная последовательность называется *маршрутом*. Маршрут называется *простым* или *цепью*, если ни одна вершина не встречается в нем дважды, и *циклом* или *замкнутой цепью*, если ни одна вершина не встречается в нем дважды и при этом $v_1 = v_k$. Вершины v_1 и v_k называются *концевыми вершинами* цепи [5].

Рассмотрим некоторое свойство P, которое присуще графам. Введем булеву функцию $P(G)$ на классе всех конечных неориентированных графов $P(G): \Gamma \to \{0,1\}$, которую определим следующим образом

$$P(G) = \begin{cases} 1, & \textit{если для графа G можно проверить свойство P и он им обладает} \\ 0, & \textit{в противном случае} \end{cases}.$$

Назовем свойство P *связным свойством графа*, если $(\forall G \in \Gamma)(P(G) \Rightarrow P_{связн.}(G))$, где $P_{связн.}$ − свойство связности графа. Из определения следует, что само $P_{связн.}$ является связным свойством.

Будем говорить, что свойство P является *монотонным свойством графа*, если выполнено следующее условие

$$(\forall G_1, G_2 \in \Gamma)(((V(G_1) = V(G_2)) \wedge (E(G_1) \subset E(G_2)) \wedge P(G_1)) \Rightarrow P(G_2)),$$

иными словами, если некоторый граф обладает свойством P, то при добавлении ребра в этом графе это свойство не нарушается.

Выделим из всевозможных свойств неориентированных конечных графов классы связных и монотонных свойств:

1. C – класс связных свойств графов, т.е. $(\forall P \in \mathsf{C})(P - \textit{связное свойство})$. Очевидно, что $P_{связн.} \in \mathrm{X}$.
2. M – класс монотонных свойств графов, т.е. $(\forall P \in \mathsf{M})(P - \textit{монотонное свойство})$.

В дальнейшем нас будут особенно интересовать такие свойства, которые являются одновременно и монотонными, и связными. Обозначим класс таких свойств $\mathsf{MC} = \mathsf{M} \bigcap \mathsf{C}$, т.е. $(\forall P \in \mathsf{MC})(P \in \mathsf{M} \wedge P \in \mathsf{C})$.

Очевидно, что $(\forall G \in \Gamma)(\forall P \in \mathsf{MC})(P_{связн.}(G) \Leftarrow P(G) \Leftarrow P_{полн.}(G))$, где $P_{связн.}$ – свойство связности графа, а $P_{полн.}$ – свойство полноты. Ниже будет показано, что $P_{связн.} \in \mathsf{MC}$ и $P_{полн.} \in \mathsf{MC}$. Таким образом, $P_{связн.}$ можно назвать минимальным монотонным связным свойством графа, а $P_{полн.}$ – максимальным.

Приведем еще некоторые примеры монотонных связных свойства графов: n-связность ($P_{n-связн.}$), неразделимость ($P_{неразд.}$), гамильтоновость (P_{Γ}) и полугамильтоновость ($P_{\Pi\Gamma}$). В принципе, если свойство P некоторого графа G монотонное, но не связное, то, если это позволяет свойство P, вполне можно требовать от графа G связности, что позволит рассматривать свойство P как элемент класса MC (таким свойством, например, является наличие цикла у графа).

Введем неотрицательную функцию $\Delta : V \times V \to \mathbb{R}_{+}$ такую, что $\forall a,b,c,d \in V$ выполняются следующие свойства:

1. $\Delta(a,b) = 0 \Leftrightarrow a = b$ – различаемость;
2. $\Delta(a,b) = \Delta(b,a)$ – симметричность;
3. $(\forall \varepsilon > 0)(\exists \delta > 0)((\Delta(a,b) < \delta \wedge \Delta(c,d) < \delta) \Rightarrow |\Delta(a,c) - \Delta(b,d)| < \varepsilon)$ – устойчивость.

Заметим, что свойство 3 является более слабым свойством, чем неравенство треугольника. Действительно, если выполняется неравенство треугольника, т.е. $(\forall a,b,c \in V)(\Delta(a,b) + \Delta(b,c) \geq \Delta(a,c))$, то $\forall \varepsilon > 0$ возьмем $\delta = \varepsilon / 2$ и рассмотрим

$$|\Delta(a,c) - \Delta(b,d)| \leq |\Delta(a,b) + \Delta(b,c) - \Delta(b,d)| = |\Delta(a,b) - (\Delta(b,d) - \Delta(b,c))| \leq$$

$$\leq \Delta(a,b) + |\Delta(b,d) - \Delta(b,c)| \leq \Delta(a,b) + \Delta(c,d) \leq \delta + \delta = \varepsilon / 2 + \varepsilon / 2 = \varepsilon . \blacksquare$$

Следовательно, в качестве функции Δ можно взять обычное расстояние (метрику) между вершинами графа.

Рассмотрим граф $G_{\varepsilon}(V) \in \Gamma$ такой, что $(\forall a,b \in V)(\Delta(a,b) \leq \varepsilon \Leftrightarrow (a,b) \in E(G_{\varepsilon}))$, и зафиксируем некоторое свойство графа P из класса MC.

Для непустого множества $A \subset V$ введем функцию, характеризующую *рассеянность множества* A

$$D_P(A)=\begin{cases}\inf\{\varepsilon>0\mid P(G_\varepsilon(A))\},\ |A|>1\\ 0,\ |A|=1\end{cases},$$

а для пары непустых множеств $A,B\subset V$, симметричную функцию, характеризующую *удаленность множеств A и B* (расстояние между множествами)

$$S(A,B)=\min\{\Delta(a,b)\mid a\in A,b\in B\}.$$

Величина, обратная к рассеянности множества, характеризует его зацепление, а обратная к удаленности двух множеств − их связность.

Следует отметить, что из определения функции $D_P(A)$ следует, что если $(\forall\varepsilon>0)(\neg P(G_\varepsilon(A)))$, то $D_P(A)=+\infty$. Очевидно также, что функции $D_P(A)$ и $S(A,B)$ являются неотрицательными. Причем, из определения функции $D_P(A)$ следует, что

$$D_P(A)=0\Leftrightarrow|A|=1,$$

а из свойства различаемости функции Δ следует, что

$$S(A,B)=0\Leftrightarrow A\bigcap B\neq\varnothing.$$

Также нетрудно показать, что если $A\bigcap B=\varnothing$, то для любого непустого подмножества C множества A выполняется следующее неравенство: $S(C,B)\geq S(A,B)$. Этот означает, что функция $S(A,B)$ монотонна по любому аргументу, так как она симметрична.

Рассмотрим множество $A^k=\{A_i\}_{i=1}^k$ непустых попарно непересекающихся подмножеств множества V. Для такого множества определим функции $D_P(A^k)$ и $S(A^k)$ следующим образом

$$D_P(A^k)=\max\{D_P(A_i)\mid i=\overline{1,k}\},$$

$$S(A^k)=\begin{cases}\min\{S(A_i,A_j)\mid i\neq j\wedge i=\overline{1,k}\wedge j=\overline{1,k}\},\ k>1\\ +\infty,\ k=1\end{cases}.$$

Так как при добавлении подмножества минимум в определении $S(A^k)$ может только уменьшиться, а максимум в определении $D_P(A^k)$ увеличится, то очевидно, что $(\forall A^l\subseteq A^k\mid 1<l\leq k)((D_P(A^l)\leq D_P(A^k))\wedge(S(A^l)\geq S(A^k)))$. Также из определения функции $S(A^k)$ следует, что

$$S(A^k)=+\infty\Leftrightarrow k=1.$$

Пусть π − некоторое *разбиение множества V*. Это означает, что $\pi=A^k=\{A_i\}_{i=1}^k$, такое что $\bigcup_{i=1}^k A_i=V$. Множества A_i из разбиения π назовем *атомами разбиения*. Далее будем использовать обозначение $\pi(a)$ для атома разбиения π, содержащего элемент $a\in V$.

На множестве всех разбиений множества V можно ввести отношение порядка следующим образом

$$\pi'\leq\pi''\Leftrightarrow(\forall A'\in\pi')(\exists A''\in\pi'')(A'\subseteq A'').$$

Очевидно, что

$$\pi' < \pi'' \Rightarrow (\exists A \in \pi'')(\exists! A_1, A_2, \ldots, A_l \in \pi',\ l > 1)\left(A = \bigcup_{i=1}^{l} A_i\right)$$

и

$$\pi' = \pi'' \Leftrightarrow (\pi' \le \pi'') \wedge (\pi'' \le \pi').$$

Так же очевидно, что для множества $V = \{a_1, \ldots, a_n\}$ разбиение $\pi_{\min} = \{\{a_1\}, \ldots, \{a_n\}\}$ является наименьшим, а разбиение $\pi_{\max} = \{\{a_1, \ldots, a_n\}\} = \{V\}$ − наибольшим. Эти два разбиения назовем *тривиальными разбиениями множества* V. Отметим, что для любого разбиения π множества V выполняется неравенство $1 = |\pi_{\max}| \le |\pi| \le |\pi_{\min}| = |V|$, а также $|V| = 1 \Leftrightarrow \pi_{\min} = \pi_{\max}$.

С учетом обозначения для разбиения π можно определить пару функций

$$D_P(\pi) = D_P(A^k) \text{ и } S(\pi) = S(A^k),$$

или, что то же самое,

$$D_P(\pi) = \max\{D_P(\pi(a)) \mid a \in V\}$$

и

$$S(\pi) = \begin{cases} \min\{S(\pi(a), \pi(b)) \mid a, b \in V, \pi(a) \ne \pi(b)\}, & |\pi| > 1 \\ +\infty, & |\pi| = 1 \end{cases}.$$

Рассмотрим отдельно случаи, когда разбиение π множества V является тривиальным разбиением:

1. $\pi = \pi_{\min}$ *и* $|V| > 1 \Leftrightarrow D_P(\pi) = 0$ *и* $0 < S(\pi) < +\infty$.

$$D_P(\pi_{\min}) = \max\{D_P(A) \mid A \in \pi_{\min}\} = \max\{D_P(\{a\}) \mid a \in V\} = 0,$$

$$S(\pi_{\min}) = \min\{S(A, B) \mid A, B \in \pi_{\min}, A \ne B\} = \min\{S_P(\{a\}, \{b\}) \mid a, b \in V, a \ne b\} > 0$$

и $S(\pi_{\min}) < +\infty$, так как $|\pi_{\min}| = |V| > 1$;

2. $\pi = \pi_{\max}$ *и* $|V| > 1 \Leftrightarrow D_P(\pi) > 0$ *и* $S(\pi) = +\infty$.

$$D_P(\pi_{\max}) = \max\{D_P(A) \mid A \in \pi_{\max}\} = D_P(V) > 0,$$

$S(\pi_{\max}) = +\infty$, так как $|\pi_{\max}| = 1$;

3. $\pi = \pi_{\max} = \pi_{\min} \Leftrightarrow |V| = 1 \Leftrightarrow D_P(\pi) = 0$ *и* $S(\pi) = +\infty$.

$D_P(\pi) = \max\{D_P(A) \mid A \in \pi\} = D_P(V) = 0$, так как $|V| = 1$,

$S(\pi) = +\infty$, так как $|\pi| = 1$.

3. Определение класса регулярных разбиений и исследование его свойств

Впервые понятие регулярного разбиения было введено и исследовано для свойства полноты графов в [6, 7].

Далее, если это не оговорено, считается, что рассматриваемое нами свойство графов $P \in \mathsf{MC}$. Так же будем считать, что $+\infty < +\infty$.

Назовем *разбиение* π *множества* V *регулярным*, если для него выполнено следующее условие

$$D_P(\pi) < S(\pi).$$

При этом если для разбиения π выполняется $D_P(\pi) = S_P(\pi) = +\infty$, то в силу нашего предположения, что $+\infty < +\infty$, данное разбиение будет являться регулярным.

Обозначим через Π_P^V *множество всех регулярных разбиений* множества V на основе свойства P. Из определения следует, что тривиальные разбиения $\pi_{\min} \in \Pi_P^V$ и $\pi_{\max} \in \Pi_P^V$ для любого непустого множества V и любого свойства $P \in \mathsf{MC}$, т.е. $\Pi_P^V \neq \varnothing$.

Содержательно, к классу регулярных разбиений относятся те разбиения, для которых атомы внутри себя зацеплены более сильно, чем связаны между собой.

Основным результатом исследований класса регулярных разбиений является следующая теорема:

Теорема. Пусть $\pi', \pi'' \in \Pi_P^V$, тогда имеют место следующие эквивалентности:

1. $\pi' \le \pi'' \Leftrightarrow D_P(\pi') \le D_P(\pi'') \Leftrightarrow S(\pi') \le S(\pi'')$;
2. $\pi' < \pi'' \Leftrightarrow S(\pi') \le D_P(\pi'')$.

Перед тем как доказывать данную теорему сформулируем и докажем несколько вспомогательных утверждений:

Утверждение 1. Если $\pi \in \Pi_P^V \setminus \{\pi_{\max}\}$, то $\forall A \in \pi$ разбиение $\pi' = \pi \setminus \{A\} \in \Pi_P^{V \setminus A}$. Причем

1. $D_P(\pi) \ge D_P(\pi')$;
2. $S(\pi) \le S(\pi')$.

Доказательство.

Пусть $\pi = \{A_i\}_{i=1}^k$, где $k > 1$ (так как $\pi \neq \pi_{\max}$), и $\pi' = \pi \setminus \{A_l\}$ для любого фиксированного $l = \overline{1,k}$.

Тогда рассмотрим:

1. $D_P(\pi') = \max\{D_P(A_i) \mid i = \overline{1,k} \wedge i \neq l\} \le \max\{D_P(A_i) \mid i = \overline{1,k}\} = D_P(\pi)$, т.е. $D_P(\pi) \ge D_P(\pi')$;
2. $S(\pi') = \begin{cases} \min\{S(A_i, A_j) \mid i, j = \overline{1,k} \wedge i \neq j \wedge i \neq l \wedge j \neq l\}, & |\pi'| > 1 \\ +\infty, & |\pi'| = 1 \end{cases} \ge$

$\ge \min\{S(A_i, A_j) \mid i, j = \overline{1,k} \wedge i \neq j\} = S(\pi)$, т.е. $S(\pi) \le S(\pi')$.

Так как π – регулярное разбиение множества V, то получаем $D_P(\pi') \le D_P(\pi) < S(\pi) \le S(\pi')$, откуда немедленно следует, что π' является регулярным разбиением множества $V \setminus A_l$. ■

Следствие. Если $\pi = \{A_i\}_{i=1}^k \in \Pi_P^V \setminus \{\pi_{\max}\}$ ($k > 1$, так как $\pi \neq \pi_{\max}$), то $\forall 0 < l < k$ и $\forall \{A_j\}_{j=1}^l \subset \{A_i\}_{i=1}^k$, разбиение $\pi' = \{A_j\}_{j=1}^l \in \Pi_P^{V'}$, где $V' = \bigcup_{j=1}^{l} A_j$.

Причем $D_P(\pi) \ge D_P(\pi')$ и $S(\pi) \le S(\pi')$. Данное следствие легко доказывается, последовательно применяя доказанное утверждение.

Утверждение 2. Для любого непустого множества V и любого свойства $P \in \mathsf{MC}$ выполняется следующее:

1. $(\forall A \subseteq V \mid A \neq \varnothing)(D_{P_{связн.}}(A) \le D_P(A) \le D_{P_{полн.}}(A))$;
2. $(\forall \pi - \textit{разбиение множества } V)(D_{P_{связн.}}(\pi) \le D_P(\pi) \le D_{P_{полн.}}(\pi))$;
3. $\left|\Pi^V_{P_{полн.}}\right| \le \left|\Pi^V_P\right| \le \left|\Pi^V_{P_{связн.}}\right|$.

Доказательство.

Из определения функции $D_P(A)$ и свойства монотонных связных свойств графов $(\forall G \in \Gamma)(\forall P \in \mathsf{MC})(P_{связн.}(G) \Leftarrow P(G) \Leftarrow P_{полн.}(G))$, вытекает справедливость пунктов 1 и 3 нашего утверждения.

Так как, если $(\forall A \in \pi)(D_{P_{связн.}}(A) \le D_P(A) \le D_{P_{полн.}}(A))$, то очевидно, что и $\max\{D_{P_{связн.}}(A) \mid A \in \pi\} \le \max\{D_P(A) \mid A \in \pi\} \le \max\{D_{P_{полн.}}(A) \mid A \in \pi\}$, а, следовательно, $D_{P_{связн.}}(\pi) \le D_P(\pi) \le D_{P_{полн.}}(\pi)$. Пункт 2 доказан. ■

Утверждение 3. Для любого множества V такого, что $|V| > 1$, имеют место следующие утверждения:

1. $(\forall A^l = \{A_i\}_{i=1}^l \mid 1 < l \le |V|)(D_P(\bigcup_{i=1}^{l} A_i) \ge S(A^l))$;
2. Если $\pi = A^k = \{A\}_{i=1}^k$ – разбиение множества V и $\pi \neq \pi_{\max}$, то
$$(\forall A^l \subseteq A^k \mid 1 < l \le k)(D_P(\bigcup_{i=1}^{l} A_i) \ge S(\pi));$$
3. Если $\pi = A^k = \{A\}_{i=1}^k \in \Pi^V_P \setminus \{\pi_{\max}\}$, то
$$(\forall A^l \subseteq A^k \mid 1 < l \le k)(\forall 1 < j \le k)(D_P(\bigcup_{i=1}^{l} A_i) > D_P(A_j)).$$

Доказательство.

Используя утверждение 1, получаем доказательство пункта 1:

$$D_P(\bigcup_{i=1}^{l} A_i) \ge D_{P_{связн.}}(\bigcup_{i=1}^{l} A_i) \ge \min\{S(A_i, A_j) \mid i \neq j \wedge i = \overline{1,l} \wedge j = \overline{1,l}\} = S(A^l),$$

так как $l > 1$. Используя свойства функции $S(A^k)$, отсюда, очевидно, следует доказательство пункта 2, так как $k \ge l > 1$ и $S(A^l) \ge S(A^k) = S(\pi)$. Если к тому же разбиение π является регулярным, то, продолжая цепочку неравенств

$$S(\pi) > D_P(\pi) \ge D_P(A_j) \text{ для любого } A_j \in \pi,$$

получаем доказательство пункта 3 нашего утверждения. ■

Теперь перейдем к *доказательству теоремы*:

Будем доказывать все три эквивалентности параллельно.

Необходимость.

Если $\pi' = \pi''$ то очевидно, что $D_P(\pi') = D_P(\pi'')$ и $S(\pi') = S(\pi'')$.

Если же $\pi' < \pi''$, то это означает, что $(\exists A \in \pi'')(\exists B_1,\ldots,B_l \in \pi' \mid l>1)(A = \bigcup_{i=1}^{l} B_i)$.

Используя регулярность разбиений π' и π'', свойства функций $D_P(\pi)$ и $S(\pi)$, а также утверждение 3, получаем следующую цепочку неравенств

$$D_P(\pi') < S(\pi') \le D_P(\bigcup_{i=1}^{l} B_i) = D_P(A) \le D_P(\pi'') < S(\pi'').$$

Откуда немедленно следует, что $D_P(\pi') < D_P(\pi'')$, $S(\pi') < S(\pi'')$, $S(\pi') \le D_P(\pi'')$.

Достаточность.

Рассмотрим сначала случай, когда $\pi'' = \pi_{\max}$. Теорема будет выполнена, так как очевидно, что любое другое регулярное разбиение будет меньше максимального.

Итак, пусть $\pi'' \neq \pi_{\max}$. Рассмотрим каждое из трех неравенств в отдельности:

1. $D_P(\pi') \le D_P(\pi'') \Rightarrow$

$$D_P(B) \le D_P(\pi') \le D_P(\pi'') < S(\pi'') \le S(A_i, A_j);$$

2. $S(\pi') \le S(\pi'') \Rightarrow$

$$D_P(B) \le D_P(\pi') < S(\pi') \le S(\pi'') \le S(A_i, A_j);$$

3. $S(\pi') \le D_P(\pi'') \Rightarrow$

$$D_P(B) \le D_P(\pi') < S(\pi') \le D_P(\pi'') < S(\pi'') \le S(A_i, A_j).$$

Таким образом, из любого из этих трех неравенств, следует, что $(\forall B \in \pi')(\forall A_i, A_j \in \pi'' \mid i \neq j)(D_P(B) < S(A_i, A_j))$.

Будем вести доказательство от противного. Пусть $\neg(\pi' \le \pi'')$ $(\neg(\pi' < \pi''))$, тогда

$$(\exists B \in \pi')(\exists A_1,\ldots,A_l \in \pi'' \mid l>1)((A_i \cap B \neq \varnothing \mid i = \overline{1,l}) \wedge (A_j \cap B = \varnothing \mid j > l)),$$

причем элементы из множества $\{A_i \cap B\}_{i=1}^{l}$ попарно не пересекаются, так как множества A_i являются атомами разбиения π''. Используя утверждение 3, рассмотрим величину $D_P(B)$:

$$D_P(B) = D_P(\bigcup_{i=1}^{l}(A_i \cap B)) \ge S(\{A_i \cap B\}_{i=1}^{l}).$$

Без ограничения общности можно считать, что минимум в определении функции $S(\{A_i \cap B\}_{i=1}^{l})$ достигается на элементах $A_1 \cap B$ и $A_2 \cap B$. Таким образом, мы имеем следующее:

$$D_P(B) = D_P(\bigcup_{i=1}^{l}(A_i \cap B)) \ge S(\{A_i \cap B\}_{i=1}^{l}) = S(A_1 \cap B, A_2 \cap B) \ge S(A_1, A_2).$$

Последнее неравенство следует из свойств симметричности и монотонности функции $S(A,B)$.

В результате мы получили, что $(\exists B \in \pi')(\exists A_1, A_2 \in \pi'')(D_P(B) \ge S(A_1, A_2))$, но это приводит нас к противоречию. Следовательно, предположение о том, что $\neg(\pi' \le \pi'')$ $(\neg(\pi' < \pi''))$, неверно.

Таким образом, мы имеем $\pi' \le \pi''$ ($\pi' < \pi''$), что и требовалось доказать.

Очевидно, что из равенства $S(\pi') = D_P(\pi'')$ не может следовать равенство разбиений π' и π''. Это следует из регулярности разбиений, так как по определению у регулярного разбиения $D_P(\pi) < S(\pi)$. ∎

Следствие. Если $\pi', \pi'' \in \Pi_P^V$, то

$$\pi' = \pi'' \Leftrightarrow D_P(\pi') = D_P(\pi'') \Leftrightarrow S(\pi') = S(\pi'').$$

Доказательство следует из эквивалентности $\pi' = \pi'' \Leftrightarrow (\pi' \le \pi'') \wedge (\pi'' \le \pi')$.

Замечание. Следует отметить, что эквивалентность

$$\pi' < \pi'' \Leftrightarrow S(\pi') \le D_P(\pi'')$$

в формулировке теоремы является более сильной, чем другие эквивалентности, так как из неравенства $S(\pi') \le D_P(\pi'')$ и регулярности разбиений следуют неравенства $D_P(\pi') < D_P(\pi'')$ и $S(\pi') < S(\pi'')$:

$$D_P(\pi') < S(\pi') \le D_P(\pi'') < S(\pi'').$$ ∎

Доказанная теорема дает нам право утверждать, что множество регулярных разбиений $\Pi_P^V = \{\pi_1, \ldots, \pi_N\}$ некоторого непустого множества V на основе монотонного связного свойства P является линейно упорядоченным множеством относительно порядка на разбиениях, т.е. из всех элементов этого множества можно выстроить следующую цепочку неравенств:

$$\pi_{\min} = \pi_1 < \pi_2 < \ldots < \pi_{N-1} < \pi_N = \pi_{\max}.$$

Откуда немедленно следует, что $\left|\Pi_P^V\right| \le |V|$.

4. Описание алгоритмов построения всех регулярных разбиений некоторого множества

Здесь мы рассмотрим два алгоритма построения множества Π_P^V для произвольного непустого множества V и произвольного монотонного связного свойства графов P – алгоритмы построения объединением и дроблением.

Алгоритм построения множества Π_P^V объединением:

1. $R = \varnothing$ – здесь мы будем накапливать полученные регулярные разбиения;
2. Строим множество $S = \{\Delta(a,b) \mid a,b \in V, a \ne b\} \bigcup \{0, +\infty\}$;
3. Сортируем множество S по возрастанию;
4. Для каждого $\alpha \in S$:
   a. Строим граф $G_\alpha(V)$;
   b. Выделяем компоненты связности $\{G_i\}_{i=1}^k$ графа $G_\alpha(V)$;
   c. Если выполнено одно из следующих трех условий, то добавляем разбиение $\pi_\alpha = \{V(G_i)\}_{i=1}^k$ к множеству R (конечно, если еще $\pi_\alpha \notin R$):
      1) $\alpha = 0$ – $\pi_\alpha = \pi_{\min}$;
      2) $\alpha = +\infty$ – $\pi_\alpha = \pi_{\max}$;

3) $(\forall i = \overline{1,k})(P(G_i))$.

5. $\Pi_P^V = R$. Конец.

Алгоритм построения множества Π_P^V *дроблением* идентичен предыдущему алгоритму, за исключением того, что множество S нужно сортировать не по возрастанию, а по убыванию.

Докажем корректность приведенных алгоритмов:

Пусть множество R − это множество, полученное в результате работы любого из приведенных выше алгоритмов. Очевидно, что элементами данного множества являются различные разбиения множества V, так как компоненты связности графа разбивают его на попарно непересекающиеся по вершинам части. Покажем, что любое разбиение π_α из множества R является регулярным. Действительно, если π_{α_0} − разбиение, добавленное к множеству R на некотором шаге работы алгоритма, то по построению π_{α_0} является либо тривиальным разбиением, либо $(\forall A \in \pi_{\alpha_0})(P(G_{\alpha_0}(A)))$. Если π_{α_0} − тривиальное разбиение, то оно является регулярным. Теперь рассмотрим случай, когда π_{α_0} не является тривиальным разбиением. В этом случае по построению мы имеем, что

$$(\forall A \in \pi_{\alpha_0})(D_P(A) \le \alpha_0)$$

и

$$(\forall A, B \in \pi_{\alpha_0} \mid A \ne B)((A \bigcap B = \varnothing) \wedge ((\forall a \in A)(\forall b \in B)(\Delta(a,b) > \alpha_0))).$$

А это означает, что $(\forall A, B, C \in \pi_{\alpha_0} \mid A \ne B)(D_P(C) \le \alpha_0 < S(A,B))$. Следовательно, $D_P(\pi_{\alpha_0}) < S(\pi_{\alpha_0})$, что и требовалось доказать. Также из доказанной нами теоремы, следует, что множество R состоит из всевозможных регулярных разбиений множества V, так как используемый алгоритм построения множества R перебирает все возможные различные $D_P(\pi_\alpha)$. Таким образом мы получаем, что $\Pi_P^V = R$. ■

Следует отметить, что полученное, в результате работы рассмотренных нами алгоритмов, множество Π_P^V уже является линейно упорядоченным. Соответственно, построение объединением упорядочивает множество Π_P^V по возрастанию, а построение дроблением − по убыванию.

Проведем оценку данных алгоритмов:

Оценку сложности алгоритмов будем производить в зависимости от количества элементов множества V. Пусть $|V| = n$.

Оценим отдельно некоторые шаги алгоритмов:

1. Построение множества S: $O(n^2)$;
2. Сортировка множества S независимо от направления: $O(n^2 \cdot \log_2 n)$;
3. Построение графа $G_\alpha(V)$: $O(n^2)$;
4. Выделение компонент связности графа $G_\alpha(V)$: $O(n^2)$;

5. Вычисление функции $P(G_i)$: $O(p(|V(G_i)|))$, где $p(n)$ – оценка сложности алгоритма проверки выполнимости свойства P для графа с n вершинами.

Теперь легко видеть, что сложность как алгоритма построения объединением, так и алгоритма построения дроблением, не превосходит $O(n^3 \cdot \max\{n, p(n)\})$.

5. Некоторые примеры регулярных разбиений

Ниже приведены все регулярные разбиения множества из 10 точек. В качестве Δ-функции взята евклидова метрика, а монотонного связного свойства графов – полнота.

Рис. 1.1. Первое регулярное разбиение.

Рис. 1.2. Второе регулярное разбиение.

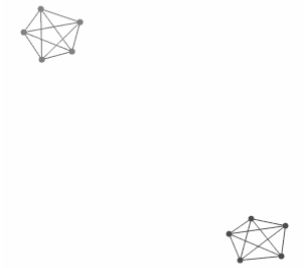

Рис. 1.3. Третье регулярное разбиение.

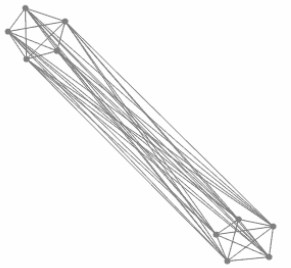

Рис. 1.4. Четвертое регулярное разбиение.

На рисунках 1.1 и 1.3 находятся тривиальные разбиения.

Вот еще несколько примеров регулярных разбиений, различных множеств, при различных параметрах:

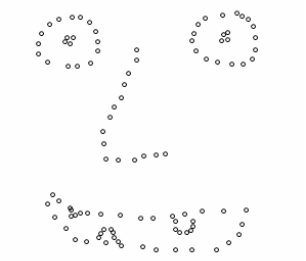

Рис. 2.1. Множество точек.

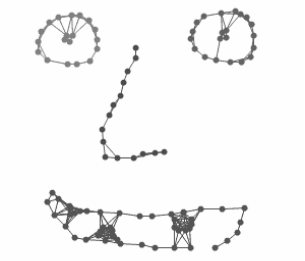

Рис. 2.2. Регулярное разбиение. Евклидова метрика. Связность.

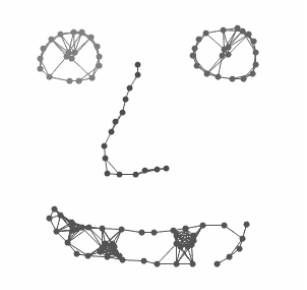

Рис. 2.3. Регулярное разбиение. Чебышевская метрика. Связность.

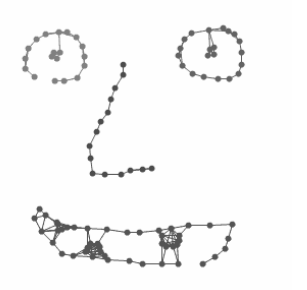

Рис. 2.4. Регулярное разбиение. Уличная метрика. Связность.

Рис. 3. Регулярное разбиение. Уличная метрика. Связность.

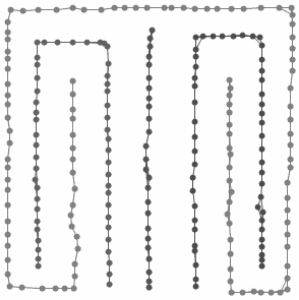

Рис. 4. Регулярное разбиение. Уличная метрика. Связность.

ЛИТЕРАТУРА

1. Патрик Э.А. Основы теории распознавания образов. – М.: Советское радио, 1980. – 408 с.
2. Сокал Р.Р. Классификация и кластер. – М.: Мир, 1980.
3. Фукунага К., Введение в статистическую теорию распознавания образов. – М.: Наука, 1979. – 368 с.
4. Грэхем Р., Кнут Д., Паташник О. Конкретная математика. Основание информатики. – М.: Мир, 1998. – 703 с.
5. Касьянов В.Н., Евстигнеев В.А. Графы в программировании: обработка, визуализация и применение. – СПб.: БХВ-Петербург, 2003. – 1104 с.
6. Жолткевич Г.Н. Автоматизация проектирования технологической оснастки: теория и практика. – К.: Техніка, 1988. – 263 с.
7. Жолткевич Г.Н., Сергеев Л.Е. Многоуровневая структура компонентных архитектур программного обеспечения. – Інформаційно-керуючі системи в залізничному транспорті. – №6, 2003. – с. 8-10.