СВОЙСТВА РАЗБИЕНИЯ СИСТЕМЫ ПОДМНОЖЕСТВ ПО СИСТЕМЕ N ЛОКАЛЬНЫХ МАКСИМУМОВ С ИСПОЛЬЗОВАНИЕМ ПСЕВДОМЕТРИКИ, ПОРОЖДЕННОЙ ВЕРОЯТНОСТНЫМ РАСПРЕДЕЛЕНИЕМ

Т. В. Куприянова
Красноярский государственный университет, Россия
e-mail: tanyakv@rambler.ru

The properties of the partition of system of subsets by n local maximums are proved. These properties are the connectivity of subgraphs corresponding to the classes of partition of system of subsets $2^{\mathfrak{X}}$ and the statement that each class of the partition is the set of lattices with special structure. It means that each class with enough large capacity can be introduced by the smaller part of it. The whole class is mutually restored according to this part.

Введение

Настоящая работа дополняет один из разделов теории случайных конечных абстрактных множеств (СКАМ), а именно разбиение системы подмножеств $2^{\mathfrak{X}}$, где $\mathfrak{X}$ — конечное абстрактное множество. В работе доказываются свойства разбиения подмножеств по системе n локальных максимумов с использованием псевдометрики, порожденной вероятностным распределением:

1) если всю систему подмножеств $2^{\mathfrak{X}}$ рассматривать как граф, то подграфы, соответствующие каждому классу ее разбиения, будут связны;

2) каждый класс разбиения — это множество подрешеток двух типов, где каждая подрешетка не включает в себя подрешетку из того же класса: в первом типе — наибольшее подмножество подрешетки есть объединение ее наименьшего подмножества и моды класса; во втором типе – наименьшим подмножеством является мода класса, а наибольшим подмножеством — объединение моды класса и наименьшего подмножества соответствующей подрешетки первого типа.

При решении прикладных задач часто возникает необходимость в разбиении системы подмножеств $2^{\mathfrak{X}}$. При большой мощности $\mathfrak{X}$ анализ всех исходов или всех подмножеств множества $\mathfrak{X}$ практически невозможен из-за их неполиномиального количества. Поэтому для изучения свойств распределения СКАМ K, заданного над $\mathfrak{X}$, всю систему подмножеств $2^{\mathfrak{X}}$ делят на классы по каким-либо признакам, после чего работают с классами, число

которых значительно меньше $2^{|\mathfrak{X}|}$. Понятно, что чем проще структура каждого класса разбиения $2^{\mathfrak{X}}$, тем проще с ней работать.

Примером прикладной задачи, использующей в своей формулировке разбиение по системе n локальных максимумов, является задача классификации подмножеств случайного множества без указания учителя. Необходимость в решении задачи классификации без указания учителя существует в таких областях, как медицинское страхование, задача изучения покупательского спроса по ассортименту, анализ котировок акций, анализ политико-экономических ситуаций и т. д.

Задача классификации без указания учителя в общем виде сформулирована давно, причем существует обширный ряд как формулировок задачи классификации, так и методов их решения [1]. Понятно, что конкретная формулировка задачи классификации без указания учителя в первую очередь определяется отличиями классифицируемых объектов друг от друга и спецификой их взаимодействия, задающими пространство объектов, для которых решается эта задача.

Примером наиболее широко используемого пространства при решении задачи классификации без указания учителя (например, в экономике) является евклидово пространство, обладающее линейной структурой. Объекты, обладающие числовыми признаками, "погружают" в евклидово пространство с размерностью, равной числу этих признаков, и в качестве меры близости между объектами используют евклидово расстояние. Очевидно, что существуют задачи (задача медицинского страхования, задача изучения покупательского спроса по ассортименту, анализ котировок акций, анализ политико-экономических ситуаций и т. д.), где объекты — множества — не обладают числовыми признаками, или их единственный числовой признак — это вероятность значения. В работе [1] для подобных объектов в качестве меры близости предлагается использовать коэффициент корреляции. Однако коэффициент корреляции, используемый в качестве меры, не обладает метрическими свойствами, поэтому классы, построенные на его основе, являются несвязными. Более того, если рассматривать семейство независимых объектов с разными вероятностями покрытий, то мера близости между любыми двумя объектами, определяемая коэффициентом корреляции, будет одинакова для любых двух объектов, и не учитывает отличие объектов друг от друга.

В работе [2] вводится понятие вероятностной псевдометрики, которая учитывает не только статистические (вероятностная псевдометрика зависит от ковариации между множествами) и структурные зависимости между множествами, но и их вероятности покрытия — в семействе независимых подмножеств в общем случае подмножества не находятся на одинаковом псевдорасстоянии (в смысле вероятностной псевдометрики). Вероятностная псевдометрика обладает всеми метрическими свойствами, поэтому классы, построенные на ее основе, являются связными и обладают определенной структурой, характеризуемой свойством 2.

Рассмотрим пример использования свойства 2 при решении задачи анализа покупательского спроса по ассортименту. Пусть на оптовой базе есть множество товаров $\mathfrak{X}$. Покупатели приходят и покупают подмножества данного множества товаров. Случайного покупателя можно рассматривать как случайное множество товаров K, которое он покупает. Необходимо определить, на какие классы можно разбить все подмножества товаров или на какие классы можно разбить всех покупателей. Полученную классификацию (классы подмножеств) можно использовать при выработке стратегий поведения на том или ином рынке товаров. Но, к сожалению, при большом ассортименте товара может возникнуть проблема анализа класса подмножеств товара, имеющего неполиномиальную мощность.

Если на оптовой базе, скажем, 10 наименований товаров, то мощность отдельных классов может быть порядка 2^{10} подмножеств. Понятно, что в этом случае заниматься анализом класса покупателей, соответствующего классу подмножеств товаров с неполиномиальной мощностью, невозможно.

Руководствуясь указанным выше свойством 2 разбиения по системе n локальных максимумов, доказанным в настоящей работе (см. следствие 6), можно при анализе каждого класса просматривать (или выводить на экран компьютера) не все подмножества каждого класса, а только значительно меньшую часть, по которой однозначно восстанавливается весь класс.

В настоящей работе приводятся основные понятия теории СКАМ, необходимые для изложения материала, излагаются результаты работы, доказываются указанные выше 1-е и 2-е свойства разбиения по системе n локальных максимумов (см. лемму и следствие), приводится пример, иллюстрирующий лемму, а также приложение результата данной работы для решения задачи классификации подмножеств случайного множества без указания учителя.

1. Основные понятия теории СКАМ

Определение 1. *Случайным конечным абстрактным множеством называется измеримое отображение*

$$K : (\Omega, \mathcal{F}, \mathbf{P}) \to (2^{\mathfrak{X}}, 2^{2^{\mathfrak{X}}}),$$

где $\mathfrak{X} = \{ x_1, ..., x_N\}$ *— это некоторое конечное множество, а* $(\Omega, \mathcal{F}, \mathbf{P})$ *— вероятностное пространство. Здесь и далее в работе* $2^{\mathfrak{X}}$ *— это система подмножеств множества* $\mathfrak{X}$, $2^{2^{\mathfrak{X}}}$ *— это система подмножеств системы подмножеств множества* $\mathfrak{X}$.

Таким образом, для любого подмножества $E \subseteq \mathfrak{X}$ существует вероятность $p(E) = \mathbf{P}(K = E)$, причем $\sum\limits_{E \subseteq \mathfrak{X}} p(E) = 1$.

В работе используется обозначение

$$\boldsymbol{p} = \left\{ p(E) = \mathbf{P}(K = E),\ E \in 2^{\mathfrak{X}} \right\}$$

— для произвольного распределения над $\mathfrak{X}$.

Определение 2. *Пусть* K *— это некоторое СКАМ, заданное над* $\mathfrak{X}$ *с распределением* $\boldsymbol{p}$. *Тогда множество* E_1^* *будем называть глобальным максимумом (или первым представителем), если*

$$\mathbf{P}(K = E_1^*) \geq \mathbf{P}(K = E), \quad E, E_1^* \subseteq \mathfrak{X}, \quad E \neq E_1^*.$$

В работе [3] на множестве $2^{\mathfrak{X}}$ для СКАМ, заданного над $\mathfrak{X}$ с распределением $\boldsymbol{p}$, вводится псевдометрика

$$d(A, B) = \mathbf{P}(A \subseteq K) + \mathbf{P}(B \subseteq K) - 2\mathbf{P}(A \cup B \subseteq K), \quad A, B \in \mathfrak{X}. \tag{1}$$

Определение 3. *Пусть* K *— это некоторое СКАМ, заданное над* $\mathfrak{X}$ *распределением* $\boldsymbol{p}$, *а также на множестве* $2^{\mathfrak{X}}$ *задана псевдометрика* d, *определяемая формулой (1).*

Пусть E_1^* — это глобальный максимум распределения $\boldsymbol{p}$. Пусть также заданы $(n-1)$ неравных друг другу множества $E_2^*, ..., E_n^*$ таких, что

$$d(E_1^*, E_i^*) > 0, \quad d(E_i^*, E_j^*) > 0, \quad i \neq j, \quad 2 \leq i \leq n, \quad 2 \leq j \leq n.$$

Для каждого множества построим класс множеств

$$\mathcal{A}_i = \Big\{ E \in 2^{\mathfrak{X}} | d(E_1^*, E) > d(E_i^*, E), d(E_j^*, E) > d(E_i^*, E),$$

$$d(E_k^*, E) \geq d(E_i^*, E), i < k \leq n, 2 \leq j < i \Big\}, \quad 2 \leq i \leq n. \tag{2}$$

Соответственно формируется класс

$$\mathcal{A}_1 = \left\{ E \in 2^{\mathfrak{X}} | d(E, E_1^*) \leq d(E, E_i^*), 2 \leq i \leq n \right\}. \tag{3}$$

Если для любого i $(2 \leq i \leq n)$ множество E_i^* в классе $\mathcal{A}_i$ имеет максимальную вероятность значения, то система множеств $E_1^*, ..., E_n^*$ называется системой n локальных максимумов, а система классов $\mathcal{A}_1, ..., \mathcal{A}_n$ — это n классов системы подмножеств $2^{\mathfrak{X}}$ или разбиение $2^{\mathfrak{X}}$ по n локальным максимумам.

Очевидно, что система n локальных максимумов существует тогда и только тогда, когда существует n множеств с ненулевыми вероятностями значений.

Определение 4. *Решетке подмножеств множества $\mathfrak{X}$ взаимно однозначно соответствует граф $G = (V, U)$. Здесь V — множество вершин, отвечающее множеству $2^{\mathfrak{X}}$, системе подмножеств множества $\mathfrak{X}$,*

$$V = 2^{\mathfrak{X}}.$$

U — множество ребер; две вершины соединены ребром тогда и только тогда, когда мощность симметрической разности соответствующих вершинам подмножеств равна 1,

$$(v_1, v_2) \in U \iff |v_1 \triangle v_2| = 1, \quad v_1, v_2 \in V.$$

Граф $G = (V, U)$ называется графом, соответствующим решетке подмножеств $2^{\mathfrak{X}}$.

2. Разбиение системы подмножеств

Пусть на множестве $\mathfrak{X}$ задано некоторое СКАМ K, и пусть существует система n локальных максимумов $E_1^*, ..., E_n^*$, которая формирует n классов $\mathcal{A}_1, ..., \mathcal{A}_n$. Так как n классов $\mathcal{A}_1, ..., \mathcal{A}_n$ — это разбиение системы подмножеств $2^{\mathfrak{X}}$, а решетке системы подмножеств $2^{\mathfrak{X}}$ соответствует связный граф G, то каждому классу $\mathcal{A}_i$ $(1 \leq i \leq n)$ можно взаимнооднозначно сопоставить подграф G_i $(1 \leq i \leq n)$. Связность подграфа G_i $(1 \leq i \leq n)$ доказывает лемма 5.

Лемма 5. *Пусть K — это некоторое СКАМ, заданное над $\mathfrak{X}$ с распределением $\boldsymbol{p}$. Пусть $E_1^*, ..., E_n^*$ — это система n локальных максимумов, а $\mathcal{A}_1, ..., \mathcal{A}_n$ — соответствующие им классы. Пусть G — это граф, соответствующий решетке системы подмножеств $2^{\mathfrak{X}}$, а $G_1, ..., G_n$ — подграфы, соответствующие классам $\mathcal{A}_1, ..., \mathcal{A}_n$. Тогда каждый подграф G_i $(1 \leq i \leq n)$ является связным.*

Доказательство проводится методом от противного. Предположим, что существует i $(1 \le i \le n)$ такое, что подграф G_i несвязный. В этом случае существует такое множество $B \in \mathcal{A}_i$ (т. е. вершина B подграфа G_i), что не существует пути из вершины E_i^* в вершину B подграфа G_i.

Доказательством связности подграфа G_i является существование пути из вершины E_i^* в вершину B в подграфе G_i в случае, когда $E_i^* \subset B$ и $E_i^* \not\subset B$.

Рассмотрим случай, когда $E_i^* \subset B$ (рис. 1). Псевдорасстояние между множествами E_i^*, B и E_j^*, $E_i^* \subset B$. Пусть множества E_i^*, E_j^* — представители класса $\mathcal{A}_i$ и $\mathcal{A}_j$ соответственно. Если $B \in \mathcal{A}_i$, то из построения классов псевдорасстояние между B и E_i^* меньше, чем псевдорасстояние между B и E_j^*. Возьмем множество D такое, что

$$E_i^* \subset D, \quad D \subset B.$$

Пусть множество $D \notin \mathcal{A}_i$, значит, множество D принадлежит некоторому классу $\mathcal{A}_j$ $(i \ne j)$. Если $j > i$, то из определения классов (2)

$$\mathcal{A}_j = \Big\{ E \in 2^{\mathfrak{X}} | d'(E_1^*, E) > d'(E_j, E), d'(E_i, E) > d'(E_j, E), \\ d'(E_k, E) \ge d'(E_j, E), j < k \le n, 2 \le i < j \Big\},$$

и из предположения $D \in \mathcal{A}_j$ следует, что

$$d(E_i^*, D) > d(E_j^*, D). \tag{4}$$

Если же $j < i$, то из определения классов (2)

$$\mathcal{A}_j = \Big\{ E \in 2^{\mathfrak{X}} | d'(E_1^*, E) > d'(E_j, E), d'(E_k, E) > d'(E_j, E), \\ d'(E_i, E) \ge d'(E_j, E),\ j < i \le n,\ 2 \le k < j \Big\},$$

и из предположения $D \in \mathcal{A}_j$ следует, что

$$d(E_i^*, D) \ge d(E_j^*, D).$$

Для доказательства рассмотрим оба случая. Сначала рассмотрим случай, когда

$$(j > i) \text{ или } \Big((j < i) \text{ и } ((d(E_i^*, D) > d(E_j^*, D)) \Big).$$

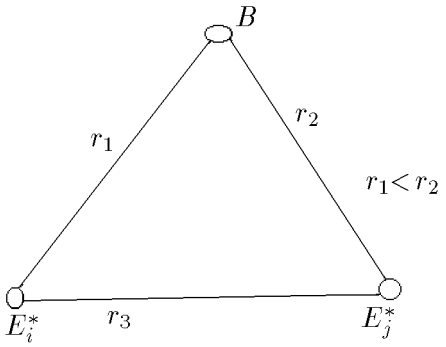


Рис. 1.

Распишем (4) в терминах вероятности:

$$\begin{gathered}\mathbf{P}(E_i^* \subseteq K) + \mathbf{P}(D \subseteq K) - 2\mathbf{P}(E_i^* \cup D \subseteq K) > \\ > \mathbf{P}(E_j^* \subseteq K) + \mathbf{P}(D \subseteq K) - 2\mathbf{P}(E_j^* \cup D \subseteq K).\end{gathered} \tag{5}$$

Из начального предположения доказываемой леммы $B \in \mathcal{A}_i$, т. е. в терминах вероятности,

$$\begin{gathered}\mathbf{P}(E_j^* \subseteq K) + \mathbf{P}(B \subseteq K) - 2\mathbf{P}(E_j^* \cup B \subseteq K) > \\ > \mathbf{P}(E_i^* \subseteq K) + \mathbf{P}(B \subseteq K) - 2\mathbf{P}(E_i^* \cup B \subseteq K).\end{gathered} \tag{6}$$

Сложим неравенства (5), (6):

$$\begin{gathered}\mathbf{P}(D \subseteq K) - 2\mathbf{P}(E_i^* \cup D \subseteq K) + \mathbf{P}(B \subseteq K) - 2\mathbf{P}(E_j^* \cup B \subseteq K) > \\ > \mathbf{P}(D \subseteq K) - 2\mathbf{P}(E_j^* \cup D \subseteq K) + \mathbf{P}(B \subseteq K) - 2\mathbf{P}(E_i^* \cup B \subseteq K).\end{gathered} \tag{7}$$

Приведем подобные в (7):

$$\mathbf{P}(E_j^* \cup D \subseteq K) + \mathbf{P}(E_i^* \cup B \subseteq K) > \mathbf{P}(E_i^* \cup D \subseteq K) + \mathbf{P}(E_j^* \cup B \subseteq K). \tag{8}$$

Так как $E_i^* \subset B$ и $E_i^* \subset D$, то (8) можно представить следующим образом:

$$\mathbf{P}(E_j^* \cup D \subseteq K) + \mathbf{P}(B \subseteq K) > \mathbf{P}(D \subseteq K) + \mathbf{P}(E_j^* \cup B \subseteq K). \tag{9}$$

Преобразуем (9):

$$\mathbf{P}(B \subseteq K) - \mathbf{P}(E_j^* \cup B \subseteq K) > \mathbf{P}(D \subseteq K) - \mathbf{P}(E_j^* \cup D \subseteq K), \tag{10}$$

откуда следует, что

$$\mathbf{P}(B \subseteq K, E_j^* \not\subseteq K) > \mathbf{P}(D \subseteq K, E_j^* \not\subseteq K). \tag{11}$$

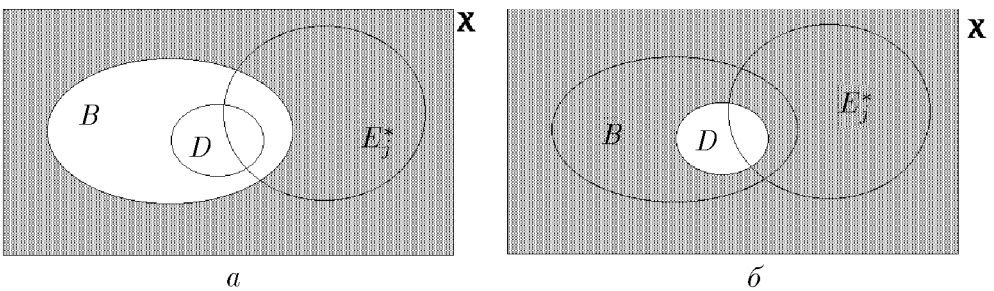
Неравенство (11) неверно, так как $B \supset D$ (рис. 2). Рассмотрим три множества $B \supset D$ и E_j^*. На рис. 2, *а* заштрихованная область схематично представляет $\mathbf{P}(B \subseteq K, E_j^* \not\subseteq K)$, а на рис. 2, *б* — $\mathbf{P}(D \subseteq K, E_j^* \not\subseteq K)$. Как видно, $\mathbf{P}(B \subseteq K, E_j^* \not\subseteq K) < \mathbf{P}(D \subseteq K, E_j^* \not\subseteq K)$.

Рассмотрим случай, когда $j < i$ и $d(E_i^*, D) = d(E_j^*, D)$. Запишем равенство значений псевдометрики в терминах вероятности:

$$\begin{gathered}\mathbf{P}(E_i^* \subseteq K) + \mathbf{P}(D \subseteq K) - 2\mathbf{P}(E_i^* \cup D \subseteq K) = \\ = \mathbf{P}(E_j^* \subseteq K) + \mathbf{P}(D \subseteq K) - 2\mathbf{P}(E_j^* \cup D \subseteq K).\end{gathered} \tag{12}$$

Приведем подобные в выражении (12), после чего получим следующее равенство:

$$\mathbf{P}(E_i^* \subseteq K) - 2\mathbf{P}(E_i^* \cup D \subseteq K) = \mathbf{P}(E_j^* \subseteq K) - 2\mathbf{P}(E_j^* \cup D \subseteq K). \tag{13}$$



а *б*

Рис. 2.

Из (13) выразим вероятность покрытия $\mathbf{P}(E_i^* \subseteq K)$:

$$\mathbf{P}(E_i^* \subseteq K) = \mathbf{P}(E_j^* \subseteq K) + 2\mathbf{P}(E_i^* \cup D \subseteq K) - 2\mathbf{P}(E_j^* \cup D \subseteq K). \tag{14}$$

Так как $B \in \mathcal{A}_i$ и $(j < i)$, то из определения классов (2)

$$\mathcal{A}_i = \Big\{ E \in 2^{\mathfrak{X}} | d'(E_1^*, E) > d'(E_i, E), d'(E_j, E) > d'(E_i, E), \\ d'(E_k, E) \geq d'(E_i, E),\ i < k \leq n,\ 2 \leq j < i \Big\},$$

следует, что

$$d(E_i^*, B) < d(E_j^*, B). \tag{15}$$

Запишем выражение (15) в терминах вероятности:

$$\begin{gathered} \mathbf{P}(E_i^* \subseteq K) + \mathbf{P}(B \subseteq K) - 2\mathbf{P}(E_i^* \cup B \subseteq K) < \\ < \mathbf{P}(E_j^* \subseteq K) + \mathbf{P}(B \subseteq K) - 2\mathbf{P}(E_j^* \cup B \subseteq K). \end{gathered} \tag{16}$$

В неравенстве (16) приведем подобные

$$\mathbf{P}(E_i^* \subseteq K) - 2\mathbf{P}(E_i^* \cup B \subseteq K) < \mathbf{P}(E_j^* \subseteq K) - 2\mathbf{P}(E_j^* \cup B \subseteq K). \tag{17}$$

Подставим вместо $\mathbf{P}(E_i^* \subseteq K)$ в выражение (17) правую часть равенства выражения (14):

$$\begin{gathered} \mathbf{P}(E_j^* \subseteq K) + 2\mathbf{P}(E_i^* \cup D \subseteq K) - 2\mathbf{P}(E_j^* \cup D \subseteq K) - 2\mathbf{P}(E_i^* \cup B \subseteq K) < \\ < \mathbf{P}(E_j^* \subseteq K) - 2\mathbf{P}(E_j^* \cup B \subseteq K). \end{gathered} \tag{18}$$

В неравенстве (18) приведем подобные и поделим все неравенство на 2:

$$-\mathbf{P}(E_j^* \cup D \subseteq K) + \mathbf{P}(E_i^* \cup D) - \mathbf{P}(E_i^* \cup B) < \mathbf{P}(E_j^* \cup B). \tag{19}$$

Так как $E_i^* \subset B$ и $E_i^* \subset D$, то

$$\mathbf{P}(E_i^* \cup B \subseteq K) = \mathbf{P}(B \subseteq K), \quad \mathbf{P}(E_i^* \cup D \subseteq K) = \mathbf{P}(D \subseteq K).$$

Следовательно, неравенство (19) эквивалентно выражению

$$-\mathbf{P}(E_j^* \cup D) + \mathbf{P}(D \subseteq K) < \mathbf{P}(B \subseteq K) - \mathbf{P}(E_j^* \cup B \subseteq K). \tag{20}$$

Из (20) следует, что

$$\mathbf{P}(D \subseteq K, E_j^* \not\subseteq K) < \mathbf{P}(B \subseteq K, E_j^* \not\subseteq K). \tag{21}$$

Неравенство (21) неверно, так как $B \supset D$ (см. рис. 2).

Таким образом, доказано, что множество D также принадлежит классу $\mathcal{A}_i$, следовательно, вершина D графа G принадлежит подграфу G_i, и доказано, что для любого множества B из класса $\mathcal{A}_i$ такого, что $E_i^* \subset B$, существует путь из вершины B в вершину E_i^* подграфа G_i.

Пусть заданы четыре множества E_i^*, E_j^*, B, D. Множества E_i^*, E_j^* — представители классов $\mathcal{A}_i$ и $\mathcal{A}_j$ соответственно, при этом $E_i^* \subset D \subset B$. Известно, что $B \in \mathcal{A}_i$. В таком случае псевдорасстояние от множества D до множества E_j^* будет больше, чем до множества E_i^*, т. е. множество D также принадлежит классу $\mathcal{A}_i$ (рис. 3).

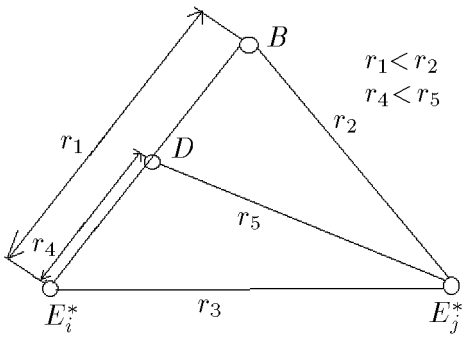


Рис. 3.

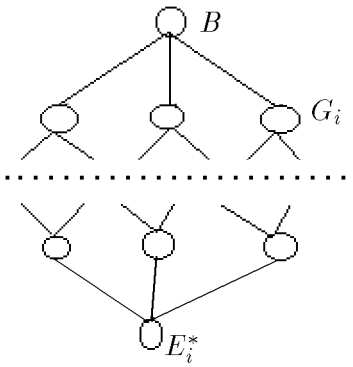


Рис. 4.

Пусть G — это граф, соответствующий решетке $2^{\mathfrak{X}}$, и пусть $\mathcal{A}_i$ — это i-й класс решения задачи классификации подмножеств случайного множества с представителем E_i^*. Классу $\mathcal{A}_i$ соответствует подграф G_i, и существует подмножество $B \supset E_i^*$ такое, что $B \in \mathcal{A}_i$. Тогда все подмножества из сет-отрезка $[E_i^*, B]$ принадлежат классу $\mathcal{A}_i$, следовательно, соответствующие им вершины в графе G также принадлежат подграфу G_i (рис. 4).

Необходимо отметить, что из вышеизложенной части доказательства следует

$$\Big(E_i^* \subset D \subset B, \quad B \in \mathcal{A}_i\Big) \Longrightarrow D \in \mathcal{A}_i. \tag{22}$$

Так как $B \in \mathcal{A}_i$ и $E_i^* \in \mathcal{A}_i$, то выражение (22) эквивалентно выражению

$$\Big(E_i^* \subseteq D \subseteq B, \quad B \in \mathcal{A}_i\Big) \Longrightarrow D \in \mathcal{A}_i. \tag{23}$$

Рассмотрим случай, когда $E_i^* \not\subseteq B$. В этом случае доказательство существования пути из вершины B в вершину E_i^* в подграфе G_i сводится к доказательству двух фактов:

— для любого множества A такого, что $B \subset A \subseteq E_i^* \cup B$, соответствующая вершина A графа G также принадлежит подграфу G_i;

— для любого подмножества D такого, что $E_i^* \subset D \subset E_i^* \cup B$, соответствующая вершина D графа G также принадлежит подграфу G_i.

Докажем, что для любого множества A такого, что $B \subset A \subseteq E_i^* \cup B$, соответствующая вершина A графа G также принадлежит подграфу G_i.

Если множество B принадлежит i-му классу $\mathcal{A}_i$, то выполняется неравенство

$$d(E_j^*, B) \geq d(E_i^*, B), \quad i \neq j,$$

или

$$\begin{aligned} &\mathbf{P}(B \subseteq K) + \mathbf{P}(E_j^* \subseteq K) - 2\mathbf{P}(B \cup E_j^* \subseteq K) \geq \\ &\geq \mathbf{P}(B \subseteq K) + \mathbf{P}(E_i^* \subseteq K) - 2\mathbf{P}(B \cup E_i^* \subseteq K). \end{aligned} \tag{24}$$

Неравенство (24) можно записать следующим образом:

$$\mathbf{P}(E_i^* \subseteq K) - 2\mathbf{P}(B \cup E_i^* \subseteq K) \leq \mathbf{P}(E_j^* \subseteq K) - 2\mathbf{P}(B \cup E_j^* \subseteq K). \tag{25}$$

Если множество $A = B + D$ ($D \subseteq E_i^* \setminus B$) принадлежит i-му классу $\mathcal{A}_i$, то после замены в неравенстве (25) множества B на множество A неравенство (25) сохраняется. Очевидно, что $B \cup E_i^* = A \cup E_i^*$, поэтому левая часть неравенства (25) после замены множества B на

множество A сохраняет то же значение. Так как множество $A \supset B$, то $E_j^* \cup B \subseteq E_j^* \cup A$, поэтому

$$\mathbf{P}(E_j^* \cup A \subseteq K) \le \mathbf{P}(E_j^* \cup B \subseteq K),$$

из чего следует, что правая часть неравенства (25) при замене B на A увеличится. Получившееся в результате неравенство

$$\mathbf{P}(E_i^* \subseteq K) - 2\mathbf{P}(A \cup E_i^* \subseteq K) \le \mathbf{P}(E_j^* \subseteq K) - 2\mathbf{P}(A \cup E_j^* \subseteq K)$$

эквивалентно неравенству

$$d(E_i^*, A) \le d(E_j^*, A).$$

Следовательно, множество A принадлежит i-му классу $\mathcal{A}_i$.

Таким образом, доказано, что если $B \not\supset E_i^*$ и $B \in \mathcal{A}_i$, то любое множество

$$A = B + D, \quad D \subseteq E_i^* \setminus B,$$

принадлежит классу $\mathcal{A}_i$ ($A \in \mathcal{A}_i$), т. е. вершина A принадлежит подграфу G_i (рис. 5). Пусть граф G соответствует решетке $2^{\mathfrak{X}}$, и пусть $\mathcal{A}_i$ — это i-й класс решения задачи классификации подмножеств случайного множества, а E_i^* — его представитель. Классу $\mathcal{A}_i$ соответствует граф G_i, подмножество $B \in \mathcal{A}_i$, причем $E_i^* \not\subset B$. Тогда все подмножества из сет-отрезка $[B, E_i^* \cup B]$ принадлежат классу $\mathcal{A}_i$, следовательно, соответствующие им вершины графа G принадлежат графу G_i. Из этого вытекает важный частный случай, что вершина $E_i^* \cup B$ принадлежит подграфу G_i.

Докажем, что для любого подмножества D такого, что $E_i^* \subset D \subset E_i^* \cup B$, соответствующая вершина D графа G также принадлежит подграфу G_i.

Так как $E_i^* \subseteq E_i^* \cup B$ и $E_i^* \cup B \in \mathcal{A}_i$ и выше доказано, что для любого множества B из класса $\mathcal{A}_i$ такого, что $E_i^* \subset B$, существует путь из вершины B в вершину E_i^* подграфа G_i, то любое подмножество

$$D \supset E_i^*, \quad D \subset E_i^* \cup B$$

также принадлежит классу $\mathcal{A}_i$, следовательно, вершина D принадлежит подграфу G_i (рис. 6).

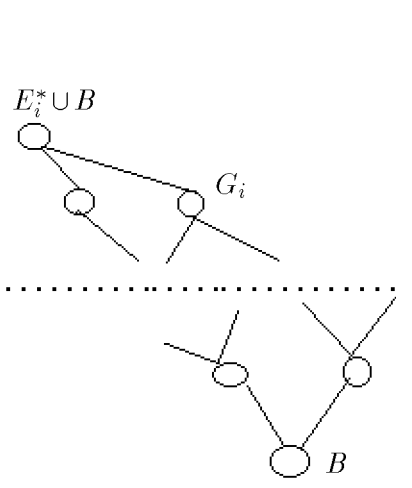


Рис. 5.

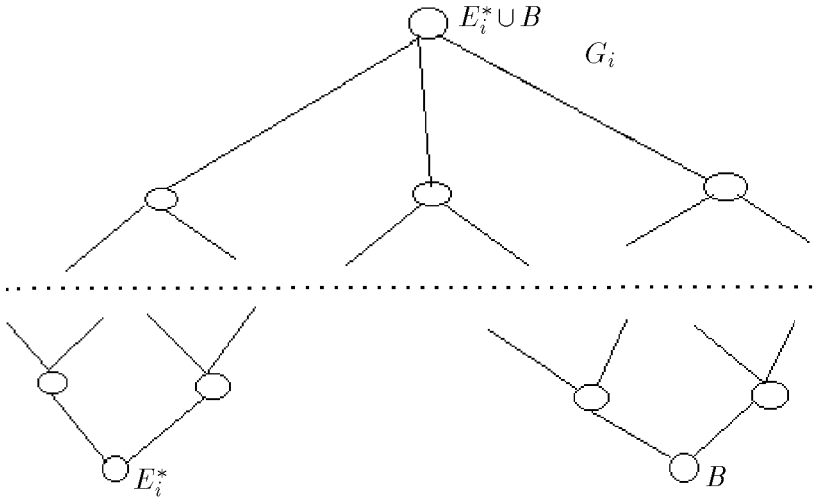


Рис. 6.

Пусть граф G соответствует решетке $2^{\mathfrak{X}}$, $\mathcal{A}_i$ — это i-й класс решения задачи классификации подмножеств случайного множества, а E_i^* — его представитель. Классу $\mathcal{A}_i$ соответствует граф G_i, и подмножество $B \in \mathcal{A}_i$, причем $E_i^* \not\subset B$. Тогда все подмножества из сет-отрезка $[B, E_i^* \cup B]$ принадлежат классу $\mathcal{A}_i$ и все подмножества из сет-отрезка $[E_i^*, E_i^* \cup B]$ принадлежат классу $\mathcal{A}_i$. Следовательно, вершины, соответствующие подмножествам из сет-отрезка $[B, E_i^* \cup B]$, и вершины, соответствующие подмножествам из сет-отрезка $[E_i^*, E_i^* \cup B]$, графа G принадлежат графу G_i.

Таким образом, доказано, что для любого множества B из класса $\mathcal{A}_i$ такого, что $E_i^* \not\subset B$, существует путь из вершины B в вершину E_i^* подграфа G_i.

Необходимо отметить, что из вышеизложенной части доказательства следует

$$\Big(B \subset D \subset E_i^* \cup B, \quad E_i^* \not\subseteq B, \quad B \in \mathcal{A}_i\Big) \Longrightarrow D \in \mathcal{A}_i. \tag{26}$$

Так как $B \in \mathcal{A}_i$, то выражение (26) эквивалентно выражению

$$\Big(B \subseteq D \subseteq E_i^* \cup B, \quad E_i^* \not\subseteq B, \quad B \in \mathcal{A}_i\Big) \Longrightarrow D \in \mathcal{A}_i. \tag{27}$$

Поскольку доказано, что для подграфа G_i существует путь из вершины E_i^* в вершину B в подграфе G_i в случаях, когда $E_i^* \subset B$ и $E_i^* \not\subset B$, подграф G_i является связным.

Лемма доказана.

Следствие 6. (О структуре класса разбиения). *Пусть K — это некоторое СКАМ, заданное над $\mathfrak{X}$ с распределением $\boldsymbol{p}$. Пусть $E_1^*, ..., E_n^*$ — это система n локальных максимумов, а $\mathcal{A}_1, ..., \mathcal{A}_n$ — соответствующие им классы. Тогда для каждого i ($i : 1, ..., n$) существует набор множеств $\widehat{B'}_i$ из класса $\mathcal{A}_i$ ($\widehat{B'}_i \subseteq \mathcal{A}_i$), такой что каждый класс $\mathcal{A}_i$ есть объединение подрешеток*

$$\bigcup_{B' \in \widehat{B'}_i} \left\{\mathcal{B}^i\right\}_{B'} \cup \left\{\mathcal{E}^i\right\}_{B'} \tag{28}$$

таких, что

$$\begin{gathered}\left\{\mathcal{B}^i\right\}_{B'} \not\subseteq \left\{\mathcal{E}^i\right\}_{B''}, \quad \left\{\mathcal{E}^i\right\}_{B'} \not\subseteq \left\{\mathcal{B}^i\right\}_{B''}, \quad B', B'' \in \widehat{B'}_i, \\ \left\{\mathcal{B}^i\right\}_{B'} \not\subseteq \left\{\mathcal{B}^i\right\}_{B''}, \quad \left\{\mathcal{E}^i\right\}_{B'} \not\subseteq \left\{\mathcal{E}^i\right\}_{B''}, \quad B', B'' \in \widehat{B'}_i.\end{gathered} \tag{29}$$

Здесь подрешетка $\{\mathcal{B}^i\}_{B'}$ определяется наименьшим множеством $B' \in \mathcal{A}_i$ и наибольшим множеством $E_i^* \cup B'$, а подрешетка $\{\mathcal{E}^i\}_{B'}$ определяется наименьшим множеством E_i^* и наибольшим множеством $E_i^* \cup B'$. Аналогично подрешетка $\{\mathcal{B}^i\}_{B''}$ определяется наименьшим множеством $B'' \in \mathcal{A}_i$ и наибольшим множеством $E_i^* \cup B''$, а подрешетка $\{\mathcal{E}^i\}_{B''}$ определяется наименьшим множеством E_i^* и наибольшим множеством $E_i^* \cup B''$.

Доказательство. Используем соотношения (23), (27), полученные в процессе доказательства леммы 5, а именно: если подмножество $B \subseteq \mathfrak{X}$ принадлежит классу $\mathcal{A}_i$ ($B \in \mathcal{A}_i$), то любое подмножество $D \subseteq \mathfrak{X}$, удовлетворяющее хотя бы одному из соотношений

$$B \subseteq D \subseteq E_i^* \cup B, \tag{30}$$

$$E_i^* \subseteq D \subseteq E_i^* \cup B, \tag{31}$$

также принадлежит классу $\mathcal{A}_i$ ($D \in \mathcal{A}_i$).

Очевидно, что соотношение (30) задает подрешетку $\{\mathcal{B}^i\}_B$ с минимальным элементом B и максимальным элементом $E_i^* \cup B$, а соотношение (31) — подрешетку $\{\mathcal{E}^i\}_B$ с минимальным элементом E_i^* и максимальным элементом $E_i^* \cup B$.

Так как каждый элемент подрешеток $\{\mathcal{B}^i\}_B$ и $\{\mathcal{E}^i\}_B$ принадлежит классу $\mathcal{A}_i$, то для любого множества $B \subseteq \mathfrak{X}$, принадлежащего классу $\mathcal{A}_i$, объединение двух подрешеток $\{\mathcal{B}^i\}_B \cup \{\mathcal{E}^i\}_B$ также принадлежит классу $\mathcal{A}_i$:

$$\left\{\mathcal{B}^i\right\}_B \cup \left\{\mathcal{E}^i\right\}_B \subseteq \mathcal{A}_i, \quad i : 1, ..., n.$$

Следовательно, класс $\mathcal{A}_i$ — это объединение подрешеток, "построенных на основе" всех подмножеств B, принадлежащих классу $\mathcal{A}_i$:

$$\mathcal{A}_i = \bigcup_{B \in \mathcal{A}_i} \left\{\mathcal{B}^i\right\}_B \cup \left\{\mathcal{E}^i\right\}_B, \quad i : 1, ..., n. \tag{32}$$

Из доказательства леммы 5 (см. выражение (27)), следует: если подмножество $B \subseteq \mathfrak{X}$ принадлежит классу $\mathcal{A}_i$, то для любого подмножества $D \subseteq \mathfrak{X}$ такого, что

$$B \subset D \subseteq E_i^* \cup B,$$

подмножество D также принадлежит классу $\mathcal{A}_i$. Отсюда подрешетка $\{\mathcal{B}^i\}_B$ содержит подрешетку $\{\mathcal{B}^i\}_D$ и подрешетка $\{\mathcal{E}^i\}_B$ — подрешетку $\{\mathcal{E}^i\}_D$:

$$\left\{\mathcal{B}^i\right\}_B \supset \left\{\mathcal{B}^i\right\}_D, \quad \left\{\mathcal{E}^i\right\}_B \supset \left\{\mathcal{E}^i\right\}_D, \quad B \subset D \subset E_i^* \cup B. \tag{33}$$

Из доказательства леммы 5 (см. выражение (23)), следует: если подмножество $B \subseteq \mathfrak{X}$ принадлежит классу $\mathcal{A}_i$, то для любого подмножества $D \subseteq \mathfrak{X}$ такого, что

$$E_i^* \subset D \subseteq E_i^* \cup B,$$

подмножество D также принадлежит классу $\mathcal{A}_i$. Отсюда подрешетка $\{\mathcal{E}^i\}_B$ содержит подрешетку $\{\mathcal{E}^i\}_D$:

$$\left\{\mathcal{E}^i\right\}_B \supset \left\{\mathcal{E}^i\right\}_D, \quad E_i^* \subset D \subset E_i^* \cup B. \tag{34}$$

Таким образом, из (33) и (34) следует, что в объединении выражения (32) в общем случае присутствуют подрешетки, содержащие друг друга.

Очевидно, что класс $\mathcal{A}_i$ можно задать как объединение подрешеток, не содержащих друг друга, для чего в равенстве (32) необходимо объединять подрешетки не по всем множествам $B \in \mathcal{A}_i$, а только по принадлежащим семейству подмножеств $\widehat{B'_i} \subseteq \mathcal{A}_i$ такому, что выполняется равенство

$$\mathcal{A}_i = \bigcup_{B' \in \widehat{B'}_i} \left\{\mathcal{B}^i\right\}_{B'} \cup \left\{\mathcal{E}^i\right\}_{B'}, \tag{35}$$

при этом подрешетки из выражения (35) не содержат друг друга, что означает

$$\begin{gathered}\left\{\mathcal{B}^i\right\}_{B'} \not\subseteq \left\{\mathcal{E}^i\right\}_{B''}, \quad \left\{\mathcal{E}^i\right\}_{B'} \not\subseteq \left\{\mathcal{B}^i\right\}_{B''}, \quad B', B'' \in \widehat{B'}_i, \\ \left\{\mathcal{B}^i\right\}_{B'} \not\subseteq \left\{\mathcal{B}^i\right\}_{B''}, \quad \left\{\mathcal{E}^i\right\}_{B'} \not\subseteq \left\{\mathcal{E}^i\right\}_{B''}, \quad B', B'' \in \widehat{B'}_i.\end{gathered}$$

Следствие доказано.

Пример 7. Пусть задано некоторое СКАМ K над множеством $\mathfrak{X} = \{x, y, z, w\}$ со следующим распределением:

$$\begin{array}{ll}
\mathbf{P}(K = \{x\}) = 0,11, & \mathbf{P}(K = \{y, w\}) = 0,1, \\
\mathbf{P}(K = \{x, y, z, w\}) = 0,09, & \mathbf{P}(K = \{y\}) = 0,09, \\
\mathbf{P}(K = \{z, w\}) = 0,09, & \mathbf{P}(K = \{x, z, w\}) = 0,08, \\
\mathbf{P}(K = \{x, w\}) = 0,08, & \mathbf{P}(K = \emptyset) = 0,06, \\
\mathbf{P}(K = \{x, y\}) = 0,06, & \mathbf{P}(K = \{x, z\}) = 0,06, \\
\mathbf{P}(K = \{y, z\}) = 0,05, & \mathbf{P}(K = \{w\}) = 0,04, \\
\mathbf{P}(K = \{x, y, w\}) = 0,04, & \mathbf{P}(K = \{z\}) = 0,02, \\
\mathbf{P}(K = \{y, z, w\}) = 0,03, & \mathbf{P}(K = \{x, y, z\}) = 0.
\end{array}$$

В этом случае существуют две системы двух локальных максимумов

$$\Big(\{x\}, \{y, w\}\Big), \quad \Big(\{x\}, \emptyset\Big).$$

Каждая система образует свои классы. Для первой системы разбиение системы множеств $2^{\mathfrak{X}}$ представлено на рис. 7. Над множеством $\mathfrak{X} = \{x, y, z, w\}$ задано СКАМ. Решетка $2^{\mathfrak{X}}$. Вероятность значения подмножества соответствует радиусу круга на рисунке. Чем больше вероятность, тем больше круг. Разные цвета кругов демонстрируют разбиение для первой системы двух локальных максимумов: $(\{x,\}, \{y, w\})$.Черные круги представляют подмножества из первого класса $\Big(\{x\}, \{x, w\}, \{x, z\}, \emptyset\Big)$, белые круги и светло-серый — подмножества из второго класса. Множество $\{y, w\}$ — мода второго класса.

Аналогично на рис. 8 представлено разбиение для второй системы двух локальных максимумов: черные круги представляют подмножества из первого класса, белые круги и светло-серый круг — подмножества из второго класса $\Big(\emptyset, \{y\}, \{w\}\Big)$. Пустое множество также есть мода второго класса.

Как видно из рис. 7, 8, в рассматриваемом примере подграфы, соответствующие классам $\mathcal{A}_1, \mathcal{A}_2$ как первой так и второй систем двух локальных максимумов, являются связными.

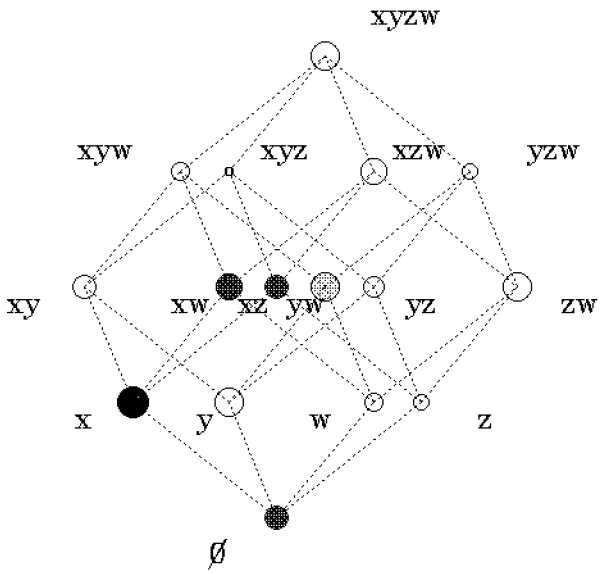


Рис. 7.

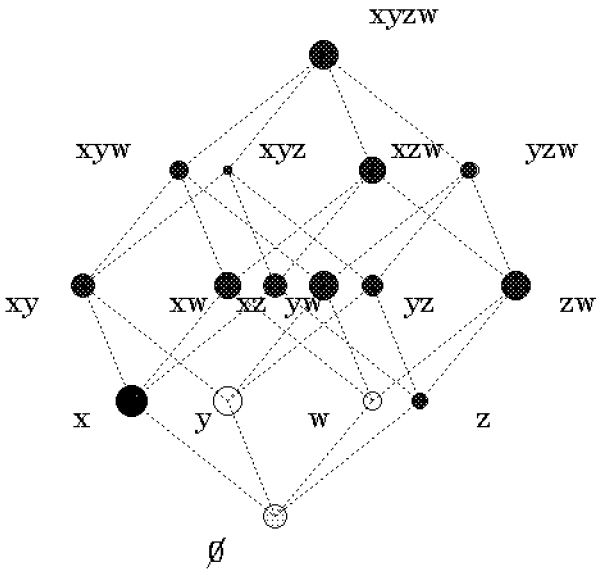


Рис. 8.

В первой системе двух локальных максимумов класс $\mathcal{A}_1$ представлен объединением трех подрешеток

$$\mathcal{A}_1 = \bigcup\Big\{ \{x\}, \emptyset \Big\} \bigcup\Big\{ \{x\}, \{x,z\} \Big\} \bigcup\Big\{ \{x\}, \{x,w\} \Big\}. \tag{36}$$

Как видно из выражения (36), множество

$$\widehat{B}'_1 = \Big\{ \emptyset, \{x,z\}, \{x,w\} \Big\},$$

причем для каждого множества $B' \in \widehat{B}'_1$ одна из двух подрешеток $\{\mathcal{B}^1\}_{B'}$, $\{\mathcal{E}^1\}_{B'}$ содержит другую подрешетку. Например, если $B' = \emptyset$, то

$$\Big\{\mathcal{B}^1\Big\}_{B'} = \Big\{\mathcal{B}^1\Big\}_{\emptyset} = \Big\{ \emptyset, \{x\} \Big\},$$

а

$$\Big\{\mathcal{E}^1\Big\}_{B'} = \Big\{\mathcal{E}^1\Big\}_{\emptyset} = \{x\}.$$

Класс $\mathcal{A}_2$ в первой системе двух локальных максимумов представлен следующим образом:

$$\mathcal{A}_2 = \bigcup\Big\{\mathcal{B}^2\Big\}_{\{y\}} \bigcup\Big\{\mathcal{B}^2\Big\}_{\{w\}} \bigcup\Big\{\mathcal{B}^2\Big\}_{\{z\}} \cup \Big\{\mathcal{E}^2\Big\}_{\{z\}} \bigcup$$
$$\bigcup\Big\{\mathcal{B}^2\Big\}_{\{x,y\}} \cup \Big\{\mathcal{E}^2\Big\}_{\{x,y\}} \bigcup\Big\{\mathcal{B}^2\Big\}_{\{x,y,z\}} \cup \Big\{\mathcal{E}^2\Big\}_{\{x,y,z\}} \bigcup\Big\{\mathcal{B}^2\Big\}_{\{x,z,w\}} \cup \Big\{\mathcal{E}^2\Big\}_{\{x,z,w\}}. \tag{37}$$

Таким образом, из выражения (37) следует, что множество $\widehat{B}'_2$, определенное в следствии 6, в рассматриваемом примере равно

$$\widehat{B}'_2 = \Big\{ \{y\}, \{w\}, \{z\}, \{x,y\}, \{x,y,z\}, \{x,z,w\} \Big\},$$

а класс $\mathcal{A}_2$ равен

$$\mathcal{A}_2 = \bigcup_{B' \in \widehat{B}'_2} \Big\{\mathcal{B}^2\Big\}_{B'} \cup \Big\{\mathcal{E}^2\Big\}_{B'}.$$

Заключение

В работе исследуется один из видов разбиения системы подмножеств $2^{\mathfrak{X}}$, а именно разбиение по n локальным максимумам с псевдометрикой (1), навязываемой вероятностным распределением, и доказаны два свойства этого разбиения:

1. Если каждому классу $\mathcal{A}_i$ поставить в соответствие подграф G_i графа решетки подмножеств множества $\mathfrak{X}$, то все подграфы G_i будут связны, что свидетельствует о том, что данное разбиение — это разбиение, где множества в каждом классе связаны друг с другом, т. е. каждый класс $\mathcal{A}_i$ по праву "несет свое название "класс"".

2. Каждый класс разбиения — это множество подрешеток двух типов, где каждая подрешетка не включает в себя подрешетку из того же класса: в первом типе — наибольшее подмножество подрешетки есть объединение ее наименьшего подмножества и моды класса; во втором типе наименьшим подмножеством является мода класса, а наибольшим подмножеством — объединение моды класса и наименьшего подмножества соответствующей подрешетки первого типа.

Предлагается использовать свойство 2 в задаче классификации множеств без указания учителя. Если $\mathfrak{X}$ имеет большую мощность n (например, $n \gg 10$), то мощность отдельных

классов, полученных при решении задачи классификации, может быть порядка 2^n. В этом случае решение задачи классификации (классы подмножеств) не то, что анализировать, просматривать трудно. Поэтому на основании свойства 2 предлагается просматривать не все подмножества каждого класса, а только наименьшее и наибольшее подмножество подрешеток каждого класса, не включающих в себя подрешетки того же класса.

Математическая формулировка свойства 2 приведена в следствии 6 настоящей работы. В качестве примера использования следствия 6 приведем решение задачи классификации подмножеств случайного множества K, заданное над множеством $\mathfrak{X} = \{x, y, z, w\}$ с распределением из примера 7, без указания учителя. Здесь результатом решения задачи классификации подмножеств случайного множества без указания учителя будут два решения — две системы двух локальных максимов:

$$\Big(\{x\}, \{y, w\}\Big), \Big(\{x\}, \emptyset\Big).$$

В таком случае говорят, что множество $\{x\}$ — представитель первого класса для первого решения задачи классификации, множество $\{y, w\}$ — представитель второго класса для первого решения задачи классификации; множество $\{x\}$ — представитель первого класса для второго решения задачи классификации, $\emptyset$ — представитель второго класса для второго решения задачи классификации.

В табл. 1, 2 приведены наименьшие подмножества подрешеток каждого класса каждой системы двух локальных максимумов (т. е. для двух решений).

Из рис. 8 следует, что во втором решении задачи классификации первый класс содержит 13 подмножеств. Как видно из таблиц, число наименьших подмножеств, по которым из следствия 6 однозначно восстанавливается весь класс, не превышает 6.

Таблица 1

Наименьшие подмножества подрешеток каждого класса первой системы локальных максимумов ($\{x\}, \{y, w\}$)

Представители классов	$\{x\}$	$\{y, w\}$
Наименьшее подмножество 1-й подрешетки	$\{x, w\}$	$\{z\}$
Наименьшее подмножество 2-й подрешетки	$\emptyset$	$\{y\}$
Наименьшее подмножество 3-й подрешетки	$\{x, z\}$	$\{x, z, w\}$
Наименьшее подмножество 4-й подрешетки	—	$\{x, y\}$
Наименьшее подмножество 5-й подрешетки	—	$\{w\}$
Наименьшее подмножество 6-й подрешетки	—	$\{x, y, z\}$

Таблица 2

Наименьшие подмножества подрешеток каждого класса второй системы локальных максимумов ($\{x\}, \emptyset$)

Представители классов	$\{x\}$	$\emptyset$
Наименьшее подмножество 1-й подрешетки	$\{y, w\}$	$\{y\}$
Наименьшее подмножество 2-й подрешетки	$\{z, w\}$	$\{w\}$
Наименьшее подмножество 3-й подрешетки	$\{y, z, w\}$	—
Наименьшее подмножество 4-й подрешетки	$\{y, z\}$	—
Наименьшее подмножество 5-й подрешетки	$\{z\}$	—

Таким образом, следствие 6 позволило значительно сократить представление каждого класса, полученного при решении задачи классификации подмножеств случайного множества без указания учителя. Для того чтобы получить все 13 подмножеств первого класса для

второй системы двух локальных максимумов, необходимо рассмотреть семейство подрешеток, у которых наименьшее подмножество будет из первого столбца табл. 2, а наибольшее подмножество — это объединение наименьшего подмножества и представителя соответствующего класса, в данном случае представителем является множество $\{x\}$. Аналогичным образом можно получить все подмножества каждого класса для каждой системы двух локальных максимумов.

Список литературы

[1] Вэн Дж. Классификация и кластер. М.: Мир, 1980. 389 с.

[2] Куприянова Т. В. Задача классификации подмножеств случайного множества и ее применение: Автореф. дис. канд. ф.-м. н. Красноярск: КГТУ, 2002. 20 с.

[3] Розанов Ю. А. Теория вероятностей, случайные процессы и математическая статистика. М.: Наука, Гл. редакция физ.-мат. лит-ры, 1985. 320 с.

[4] Емеличев В. А. Лекции по теории графов. М.: Наука, 1990. 384 с.

Поступила в редакцию 11 октября 2000 г.
в переработанном виде — 18 марта 2002 г.