УДК 517

КОМПАКТНАЯ КОМПОНЕНТНАЯ И СОКРАЩЕННАЯ ИНТЕРВАЛЬНАЯ СТРУКТУРЫ ПРИЗНАКОВОГО ПРОСТРАНСТВА, ПОРОЖДАЕМЫЕ ЭМПИРИЧЕСКИМИ ДАННЫМИ

Ильченко А.В.

ТАВРИЧЕСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ ИМ. В.И.ВЕРНАДСКОГО,
ФАКУЛЬТЕТ МАТЕМАТИКИ и ИНФОРМАТИКИ
ПР-Т ВЕРНАДСКОГО,4, Г.СИМФЕРОПОЛЬ, КРЫМ, УКРАИНА, 95007

Abstract.
The compact component structure of attributes space, clustering algorithm based on this concept, the brief interval structure concept of attributes space and its construction algorithm are considered in the paper.

Введение

Алгоритмы и методы кластерного анализа во многом составляют основу приложений интеллектуального анализа данных. Большие массивы многомерных данных, подлежащих обработке, предъявляют ряд требований к используемым алгоритмам кластеризации: возможность находить кластеры в пространстве большой размерности, легкость интерпретации и наглядность полученных результатов, отсутствие необходимости приведения исходных данных к какому-либо каноническому виду и так далее. Кроме этого, во многом открытыми и актуальными остаются вопросы обоснованности и объективности полученных результатов.

Целью данной работы является попытка определения некоторых формальных понятий, в терминах которых удобно определять данные, обладающие хорошей, отчетливо просматриваемой кластерной структурой; построение и обоснование алгоритмов кластеризации и описания таких данных.

Компонентная структура признакового пространства

Пусть $A = \{A_1, A_2, \ldots, A_n\}$ — семейство конечных вполне упорядоченных доменов. $S = A_1 \times A_2 \times \cdots \times A_n$ – n-мерное признаковое пространство.

Иногда на обозначение A_j будем ссылаться как на измерение (атрибут) признакового пространства.

G – множество объектов.

$H_{\bar{0}}$ – проекция множества G в признаковое пространство S.

$H_0 = S \setminus H_{\bar{0}}$ — дополнение пространства $H_{\bar{0}}$ до пространства S.

$H_{\bar{0}}$ – «занятая область».

H_0 – «свободная область».

$S = H_0 \bigcup H_{\bar{0}}$ — представление признакового пространства в виде объединения «свободной» и «замкнутой» областей.

Определение 1. Пусть $H \in \{H_0, H_{\bar{0}}\}$ — одна из областей пространства S. Компонентой множества H называется максимальное (по включению подмножеств) связное подмножество множества H.

Множество всех компонент области H называется компонентной структурой этой области.

Каждую из областей признакового пространства можно представить в виде объединения компонент ее компонентной структуры: $H_i = \bigcup\limits_{j=1}^{K_i} H_{ij}$, где

K_i — число компонент области $H_i, i \in \{0, \bar{0}\}$.

H_{ij} — компоненты области $H_i, j = 1, 2, \ldots, K_i$.

Множество всех компонент, определяемых «свободной» и «занятой» областями, называется компонентной структурой признакового пространства, порождаемой проекцией множества G в пространство S.

Пространство S можно представить в виде объединения компонент его компонентной структуры:$S = \bigcup\limits_{i=0}^{\bar{0}} \bigcup\limits_{j=1}^{K_i} H_{ij}$.

Компактная компонента структуры признакового пространства

Пусть множество объектов G представимо в виде объединения попарно непересекающихся подмножеств (классов): $G = \bigcup\limits_{i=1}^{K} G_i$.

Семейство классов $\{G_i\}$ называется классовой структурой множества G.

Каждому классу G_i соответствует подмножество точек H_i, признакового пространства S – проекция класса G_i в пространство S. Проекция H_i не обязательно односвязное подмножество. В общем случае H_i распадается на несколько компонент, то есть каждый класс объектов G_i порождает соответствующее этому классу семейство компонент $\{H_{ij}\}$ – компонентную структуру класса G_i (компонентную структуру проекции H_i).

Множество всех компонент всех классов называется компонентной структурой классов множества G в признаковом пространстве S.

$K(G)$ — обозначение компонентной структуры классов множества G.

Проекция множества G в признаковое пространство S образует «занятую» область $H_{\bar{0}}$ признакового пространства.

$K(H_{\bar{0}})$ — обозначение компонентной структуры «занятой» области $H_{\bar{0}}$.

$H_0 = S \setminus H_{\bar{0}}$ — «свободная» область признакового пространства. Ее компоненты – компонентная структура «свободной области».

Две точки признакового пространства называются соседними, если они различаются значениями только одного атрибута и эти различающиеся значения являются соседними относительно полного порядка по этому атрибуту.

Точка признакового пространства называется соседней подмножеству точек $H \subseteq S$, если эта точка является соседней хотя бы к одной из точек подмножества H.

S – окрестностью подмножества H называется подмножество точек признакового пространства, полученное в результате объединения подмножества H и всех его соседних точек.

H_s – обозначение S - окрестности подмножества H.

Определение 2. Пусть H – компонента одного из классов, H_s – ее S - окрестность. Компонента H называется компактно расположенной в признаковом пространстве, если ее S - окрестность не пересекается ни с одной из других компонент классов, то есть $\forall H' \in (K(G) \setminus H) : H_s \bigcap H' = \varnothing$.

Компактная расположенность означает, что такая компонента класса отделена компонентами «свободной» области от других компонент классов.

Определение 3. Компонентная структура классов называется компактной, если каждая ее компонента является компактно расположенной.

В общем случае компонентная структура классов множества G не совпадает с компонентной структурой «занятой» области $H_{\bar{0}}$, то есть $K(G) \neq K(H_{\bar{0}})$.

Теорема 1. *$K(G) = K(H_{\bar{0}})$ тогда и только тогда, когда $K(G)$ – компактная компонентная структура.*

Доказательство. Докажем справедливость отношения «$\subseteq$» : Пусть $H \in K(G)$ – произвольная компонента класса.

Допустим, что $H \notin K(H_{\bar{0}})$. Тогда существует $H' \in K(H_{\bar{0}} : H \subset H' \subseteq H_{\bar{0}})$.

Следовательно, $H' \setminus H \subseteq H_{\bar{0}}$, то есть все точки, дополняющие H до H', являются «занятыми».

Среди точек подмножества $H' \setminus H$ есть хотя бы одна, которая является соседней к H, так как иначе было бы нарушено требование связности H'. Как и все точки подмножества $H' \setminus H$, эта точка является «занятой».

Таким образом существует точка соседняя к H, но не принадлежащая H, которая является «занятой». Это противоречит тому, что H – компактно расположенная.

Следовательно, $H \in K(H_{\bar{0}})$ и, в силу произвольности выбора компоненты H, $K(G) \subseteq K(H_{\bar{0}})$.

Докажем справедливость отношения « $\supseteq$ ».

Пусть $H \in K(H_{\bar{0}})$ – произвольная компонента «занятой» области.

Допустим, что $H \in K(G)$. Тогда существует $H' \in K(G) : H \subset H'$. Следовательно, $H' \setminus H \subseteq H_{\bar{0}}$, то есть подмножество точек, дополняющих H до H', состоит из «занятых» точек.

Среди точек подмножества $H' \setminus H \subseteq H_{\bar{0}}$ существует хотя бы одна точка, которая является соседней к H, так как иначе нарушалось бы требование связности подмножества H'.

Пусть $h \in H' \setminus H$ – соседняя к H точка. Как и все точки подмножества $H' \setminus H$, эта точка является «занятой».

Тогда $H \subset H \bigcup h \subseteq H_{\bar{0}}$и, следовательно, H не является максимальным связным подмножеством области $H_{\bar{0}}$. Значит $H \notin K(H_{\bar{0}})$. Противоречие.

Поэтому, $H \in K(G)$ и, в силу произвольности выбора компоненты H, $K(H_{\bar{0}}) \subseteq K(G)$.

Определение 4. Компонентная структура признакового пространства называется компактной, если все компоненты классов являются компактно расположенными в признаковом пространстве.

Именно случай компактной компонентной структуры признакового пространства и рассматривается далее. Для компактной компонентной структуры поиск компонент компонентной структуры классов множества G, в силу теоремы 1, заменяется поиском компонент «занятой» области $H_{\bar{0}}$ признакового пространства.

Задача кластеризации и компактная компонентная структура признакового пространства

Задача кластеризации может быть сформулирована следующим образом [3].

Задано:

1. Признаковое пространство S;
2. $V \subseteq G$ - выборка объектов из множества G.

Требуется разбить объекты выборки V на группы так, чтобы в каждой группе оказались объекты, «схожие» по значениям атрибутов признакового пространства S.

Будем считать объекты выборки V «схожими», если их проекции в признаковое пространство содержатся в одной компоненте компактной компонентной структуры признакового пространства.

Если компонентная структура признакового пространства известна, то объекты выборки V , проекции которых попали в одну компоненту, рассматриваются как кластер [2].

Таким образом, задача кластеризации сводится к следующей задаче определения компактной компонентной структуры: по заданной выборке $V \subseteq G$ определить (приближенно восстановить) компактную компонентную структуру признакового пространства S.

Алгоритм кластеризации выборки, использующий понятие компонентной структуры

Компоненты «занятой» области представляют собой дополнение «свободной» области до признакового пространства. Если «свободная» область определена, то можно считать, что определена и «занятая» область.

С учетом этого замечания, алгоритм кластеризации выборки V может быть определен следующим образом:

1. По проекции выборки V оценить размер интервалов признакового пространства S, которые могут достоверно считаться «свободными» [1], [2];
2. Определить все максимальные «свободные» интервалы, размеры которых удовлетворяют полученным оценкам. Объединение этих интервалов образует «свободную» область H_0 признакового пространства S;
3. Точки признакового пространства, не попавшие ни в один из «свободных» интервалов, считаются «занятыми». Они образуют «занятую» область признакового пространства S: $H_{\bar{0}} = S \setminus H_0$;
4. Определить все компоненты $H_{\bar{0}j} \mid j = 1, 2, \ldots K_{\bar{0}}$ области $H_{\bar{0}}$.
   K_0 – число компонент области $H_{\bar{0}}$;
5. Объекты выборки V, проекции которых попали в одну компоненту, объявляются кластером. Таким образом, выборка Vраспадается на кластеры $V_j \subseteq V \mid j = 1, 2, \ldots, K_{\bar{0}}$;
6. Каждый из полученных кластеров V_j, рассматривается как отдельная выборка. Минимальный интервал S_j признакового пространства S, содержащий элементы проекции кластера V_j, рассматривается как отдельное признаковое пространство;
7. Для каждой пары (V_j, S_j) повторяются пункты 1-6 алгоритма. Процесс завершается тогда, когда ни один из кластеров V_j не распадается на более мелкие группы.

Примечание Для реализации шага 2 алгоритма, необходимо уметь находить максимальные «свободные» интервалы, удовлетворяющие оценкам шага 1. Компоненты «занятой» области (шаг 4) удобно представлять с использованием максимальных «занятых» интервалов признакового пространства. Описание кластеров, полученных в результате работы алгоритма, также удобно проводить в терминах максимальных интервалов признакового пространства. Отмеченные факты служат основанием необходимости разработки алгоритмов, позволяющих строить различные максимальные интервалы признакового пространства, порождаемые имеющимися данными.

Сокращенная интервальная структура признакового пространства

Определение 5. Пусть H – одна из компонент компонентной структуры пространства S,$H \subseteq H$ – интервал, содержащийся в этой компоненте. Интервал I, содержащийся в компоненте H, называется максимальным интервалом компоненты H, если не существует его собственного над интервала, содержащегося в компоненте H.

Множество всех максимальных интервалов компоненты H называется сокращенной интервальной структурой этой компоненты.

$I(H)$ – обозначение множества всех максимальных интервалов компоненты.

Каждую компоненту H признакового пространства S можно представить в виде объединения интервалов ее сокращенной интервальной структуры: $H = \bigcup\limits_{I \in I(H)} I$.

Множество всех максимальных интервалов, определяемых всеми компонентами компонентной структуры пространства S, называется сокращенной интервальной структурой признакового пространства, порождаемой проекцией множества G в пространство S.

Пространство S можно представить в виде объединения интервалов его сокращенной интервальной структуры: $S = \bigcup\limits_{i=0}^{\bar{0}} \bigcup\limits_{j=1}^{K_i} \bigcup\limits_{I \in I(H_{ij})} I$.

Задача определения сокращенной интервальной структуры признакового пространства

Задано:

1. Признаковое пространство S;
2. Проекция $H_{\bar{0}}$ множества объектов G в признаковое пространство S.

Требуется определить сокращенную интервальную структуру признакового пространства S, порождаемую проекцией $H_{\bar{0}}$ множества G.

Использование сокращенной интервальной структуры признакового пространства

Если задача определения сокращенной интервальной структуры признакового пространства решена, то ее решение может быть использовано для:

1. Описания в виде сокращенной интервальной структуры компонент признакового пространства;
2. Описания в виде сокращенной интервальной структуры областей признакового пространства, соответствующих классам множества G;
3. Формирования решающего правила о принадлежности классу предъявленного объекта из множества G;

4. Формирования решающего правила о классификации точек признакового пространства S.

Описание интервалов признакового пространства

$I(S)$ – обозначение множества всех интервалов признакового пространства S. Отношение включения « $\subseteq$ » определяет частичный порядок на множестве $I(S)$.

Фиксируется некоторый способ описания интервалов признакового пространства.

$D(S)$ – обозначение семейства описаний всех интервалов признакового пространства S. Способ описания интервалов выбирается так, что каждому интервалу соответствует единственное его описание и каждое описание определяет единственный интервал, то есть соответствие между $D(S)$ и $I(S)$ является взаимно однозначным.

Пусть $I, J \in I(S)$ интервалы признакового пространства, $d(I), d(J) \in D(S)$ - описания интервалов I и J соответственно. На $D(S)$ определим отношение предшествования «$\leqslant$» , порождаемое отношением порядка « $\subseteq$ », определенным на множестве $I(S) : d(I) \leqslant d(J)$, если $J \subseteq I$.

Отношение « $\leqslant$ » определяет частичный порядок на элементах семейства $D(S)$. В силу определения семейства $D(S)$ и способа определения отношения порядка « $\leqslant$» на $D(S)$, соответствие между множествами $D(S)$ и $I(S)$ является взаимно однозначным и антитонным, то есть $D(S)$ и $I(S)$ - двойственные множества.

Если описание $d(I)$ предшествует описанию $d(J)$, то $d(I)$ называется подописанием описания $d(I)$,$d(J)$ – надописанием описания $d(I)$.

Пусть $d_1, d_2 \in D(S)$.

Отношение строго предшествования « $<$ » на элементах семейства $D(S)$: $d_1 < d_2 \Leftrightarrow d_1 \leqslant d_2$ и $d_1 \neq d_2$.

Отношение непосредственного предшествования « $\prec$ » на элементах семейства $D(S) : d_1 \prec d_2 \Leftrightarrow d_1 < d_2$ и $\bar{\exists} d \in D(S) : d_1 < d < d_2$ Отметим, что требование непосредственного предшествования более жесткое, чем требование строгого предшествования, то есть, справедливо

Утверждение 1. $d_1 \prec d_2 \Rightarrow d_1 < d_2$.

Часто бывает удобно точки «занятой» области признакового пространства называть единичными точками (1-точки), точки «свободной» области -нулевыми точками (0-точки) так, как если бы была задана индикаторная функция «занятой» области.

Пусть $I \in I(S)$ - некоторый интервал признакового пространства S. Если все точки интервала I - 1-точки, то I называется 1-интервалом, а соответствующее этому интервалу описание $d(I)$ называется 1-описанием. Если все точки интервала I - 0-точки, то I - О-интервал, $d(I)$ - 0-описание. Если в интервале I содержатся и нулевые и единичные точки, то I называется 0-1-интервалом, $d(I)$ - 0-1 – описанием.

Для интервалов множества $I(S)$ справедливы следующие утверждения:

Утверждение 2. *Всякий подинтервал 1-интервала является 1-интервалом.*

Утверждение 3. *Всякий подинтервал 0-интервала является 0-интервалом.*

Утверждение 4. *Всякий надинтервал 0-1 – интервала является 0-1 - интервалом.*

В силу двойственности $D(S)$ и $I(S)$, для описаний интервалов справедливы утверждения, двойственные утверждениям 1– 3:

Утверждение (2'). *Всякое надописание 1-описания является 1-описанием.*

Утверждение (3'). *Всякое надописание 0-описания является 0-описанием.*

Утверждение (4'). *Всякое подописание 0-1 – описания, является 0-1 – описанием.*

Также, в силу двойственности $D(S)$ и $I(S)$, максимальным единичным и нулевым интервалам множества $I(S)$ соответствуют минимальные (в смысле порядка, определяемого отношением «$\leqslant$») единичные и нулевые описания семейства $D(S)$.

Из утверждения 4' следует, что подсемейство всех 0-1 – описания является порядковым идеалом на частично упорядоченном множестве $D(S)$. Этот факт влечет все нижеследующие определения и утверждения текущего раздела.

Определение 6. Описание $D(S)$называется кандидатом (кандидатом в 0-1 – описание), если все строго предшествующие ему описания являются 0-1 – описаниями.

Утверждение 5. *Описание $D(S)$ является кандидатом тогда и только тогда, когда каждое непосредственно предшествующее его подописание является 0-1 – описанием.*

Утверждение 6. *Всякий кандидат является либо минимальным 1-описанием, либо минимальным 0-описанием, либо 0-1 – описанием.*

Доказательство. Пусть $d \in D(S)$ - кандидат. Если d является 1- описанием, то d - минимальное 1-описание, так как все строго предшествующие ему подописания являются 0-1- описаниями.

Если d, является 0-описанием, то d - минимальное 0-описание, так как все строго предшествующие ему подописания являются 0-1 – описаниями.

Если d не является единичным или нулевым описанием, то d – 0-1 – описание.

Из утверждения 6 следует, что одним из возможных подходов к построению всех минимальных единичных и нулевых описаний является построение всех кандидатов, а, затем, извлечение из семейства кандидатов 1-описаний и 0-описаний.

Алгоритмы именно этого направления рассматриваются далее.

Определение 7. Описание $d \in D(S)$ называется потенциальным кандидатом, если хотя бы одно из непосредственно предшествующих ему описаний является 0-1 – описанием.

Утверждение 7. *Каждый кандидат является потенциальным кандидатом.*

Утверждение 8. *Потенциальный кандидат, не являющийся кандидатом, либо не минимальное 1-описание, либо не минимальное 0-описание.*

Доказательство. Пусть описание $d \in D(S)$ потенциальный кандидат, но не кандидат. Так как d – не кандидат, то среди непосредственно предшествующих d описаний хотя бы одно является единичным или нулевым описанием. Следовательно, в силу утверждений 2'-3', d является единичным или нулевым надописанием единичного или нулевого описания. Поэтому d, не является минимальным единичным или нулевым описанием.

СОКРАЩЕННАЯ ИНТЕРВАЛЬНАЯ СТРУКТУРА БУЛЕВОГО ПРИЗНАКОВОГО ПРОСТРАНСТВА

Пусть $\mathrm{A}_j = \{0, 1\}$, $j = 1, 2, \ldots, n$.

$S = \{0, 1\}^n$ – признаковое пространство.

$X = \{x_1, x_2, \ldots, x_n\}, x_j \in A_j, j = 1, 2, \ldots, n$ – множество переменных, которые принимают значения из соответствующих доменов. На множестве переменных определим полный порядок в соответствии с номерами переменных: $X = \{x_1 < x_2, \ldots, x_n\}$.

$L(x) = \{\overline{x_1} < x_1 < \cdots < \overline{x_n} < x_n\}$ – упорядоченное множество литералов, соответствующих переменным множества X.

Каждому интервалу булевого признакового пространства S соответствует элементарная конъюнкция. Набор литералов такой конъюнкции рассматривается как описание соответствующего интервала. $D(S)$ – обозначение семейства всех наборов литералов, являющихся описаниями интервалов пространства S.

Отношение включения «$\subseteq$» определяет естественный порядок на элементах семейства $D(S)$, равносильный порядку «$\leqslant$», который был определен ранее. Поэтому для набора литералов $D(S)$ справедливы все вышеприведенные утверждения об элементах семейства $D(S)$.

Описание $d \in D(S)$ называется k-элементным описанием, если набор d состоит из k литералов.

Утверждение 9. *Для каждого k-элементного описания ($k > 0$) существует k непосредственно предшествующих ему $(k-1)$-элементных подописания.*

Каждое из таких подописаний получается в результате удаления из исходного описания одного из элементов (одного из литералов), входящих в состав этого k-элементного описания.

Утверждение 10. *С учетом порядка на множестве литералов $L(x)$, каждый k-элементный кандидат может быть единственным образом представлено в виде*

объединения двух его $(k-1)$*-элементных 0-1 – подписаний, у которых общие первые* $(k-2)$ *элемента:*

$$d = \{l_{i_1} < \cdots < l_{i_{k-2}} < l_{i_{k-1}} < l_{i_k}\} = \\ = \{l_{i_1} < \cdots < l_{i_{k-2}} < l_{i_{k-1}}\} \cup \{l_{i_1} < \cdots < l_{i_{k-2}} < l_{i_k}\}, \quad (1)$$

где $l_{i_j} \in L(X), j = 1, 2, \ldots, k.$

Доказательство. Пусть $d = \{l_{i_1}, \ldots, l_{i_k}\}$ – кандидат и $l_{i_1} < \cdots < l_{i_k}$. Так как d – кандидат, то все непосредственно предшествующие ему подописания являются 0-1 – описаниями.

В частности, $\{l_{i_1} < \cdots < l_{i_{k-2}} < l_{i_{k-1}}\}$ и $\{l_{i_1} < \cdots < l_{i_{k-2}} < l_{i_k}\}$ – непосредственно предшествующие подописания описания d и, следовательно, являются 0-1 – описаниями. Их объединение порождает представление (1).

Представление (1) позволяет строить описания интервалов по уровням, используя уже построенные $(k-1)$-элементные 0-1 – описания для построения k-элементных описаний.

Утверждение 11. *Некоторые из потенциальных кандидатов могут быть представлены в виде (1).*

Это те потенциальные кандидаты, у которых $(k-1)$-элементные подописания являются 0-1 – описаниями.

Утверждение 12. *Каждый кандидат, является потенциальными кандидатом, преставимым в виде (1).*

Алгоритм построения сокращенной интервальной структуры булевого признакового пространства

Таблица 1. Обозначения алгоритма

Обозначение	Пояснение
k	Номер уровня, номер итерации, количество элементов в описании интервала
$\widetilde{C_k}$	Семейство k-элементных потенциальных канидатов, полученных с использованием представления (1)
C_k	Семейство всех k-элементных кандидатов
L_{k0}	Семейство всех минимальных k-элементных 0-описаний
L_{k1}	Семейство всех минимальных k-элементных 1-описаний
L_k	Семейство всех k-элементных 0-1-описаний

Псевдокод алгоритма

Вход. S - булево признаковое пространство;
$H_{\bar{0}} \subseteq S$ - проекция множества объектов G в пространство S.

Выход. (D_0, D_1) - сокращенная интервальная структура признакового пространства.
D_0 - семейство 0-описаний, соответствующих максимальным 0-интервалам.
D_1 - семейство 1-описаний, соответствующих максимальным 1-интервалам.

Алгоритм.

1. $if\ \varnothing \in D(S)$ - *0-описание then return* $(\{\varnothing\}, \varnothing)$;
2. $if\ \varnothing \in D(S)$ - *1-описание then return* $(\varnothing, \{\varnothing\})$;
3. $C_1 := \{\{l\} | l \in L(X)\}$;
4. $L_{10} := \{d \in C_1 | d$ - *0-описание*$\}$;
5. $L_{11} := \{d \in C_1 | d$ - *1-описание*$\}$;
6. $k := 1$;
7. $while\ L_k \neq \varnothing\ do\ begin$
8. $k++$;
9. $\widetilde{C_k} := \{l_{i_1} < l_{i_2} < \cdots < l_{i_k}\} | \{l_{i_1} < \cdots < l_{i_{k-2}} < l_{i_{k-1}}\}, \{l_{i_1} < \cdots < l_{i_{k-2}} < l_{i_k}\} \in L_{k-1}$;
10. $C_k := \{d \in \widetilde{C_k} | \forall l \in d : d \setminus \{l\} \in L_{k-1}\}$;
11. $L_{k0} := \{d \in C_k | d$ - *0-описание*$\}$;
12. $L_{k1} := \{d \in C_k | d$ - *1-описание*$\}$;
13. $L_k := C_k \setminus (L_{k0} \cup L_{k1})$
14. end;
15. $return\ \left(\bigcup_{i=1}^{k} L_{i0}, \bigcup_{i=1}^{k} L_{i1}\right)$.

Пояснения к псевдокоду алгоритма

Алгоритм начинает работу с того, что проверяет, не является ли 0-элементное описание интервала (все признаковое пространство, рассматриваемое как интервал) 0-описанием (шаг 1) или 1-описанием (шаг 2).

Затем все 1-элементные описания рассматриваются как кандидаты (шаг 3) и из них отбираются те, которые являются 0-описаниями (шаг 4, утв. 6) или 1-описаниями (шаг 5, утв. 6). Описания, оставшиеся после такого отбора, являются 0-1-описаниями (шаг 6, утв. 6). Они используются при построении кандидатов на следующий уровень.

Условие $L_k \neq \varnothing$ (шаг 8, утв. 10) определяет условие выполнения очередной итерации алгоритма, реализующей постороение необходимых конструкций следующего уровня.

Шаг 8 определяет номер следующей итерации.

Затем (шаг 90), с использованием представления (1), строится $\widetilde{C_k}$ – семейство k-элементных потенциальных кандидатов (утв. 11), из которых (шаг 10) отбираются кандидаты (утв. 12), формирующие семейство C_k.

Шаг 11 формирует L_{k0} – семейство минимальных k-элементных 0-описаний (утв. 6).

Шаг 12 формирует L_{k1} – семейство минимальных k-элементных 1-описаний (утв. 6).

Описания, оставшиеся после такого отбора, являются 0-1-описаниями L_k (шаг 13, утв. 6).

Шаг 15 объединяет результаты, полученные на разных уровнях, и возвращает полученную пару как результат работы алгоритма.

Насколько известно автору, процедура генерации семейства $\widetilde{C_k}$ потенциальных кандидатов порядкового идеала, использующая тот или иной линейный порядок на множестве значений атрибутов (представление (1)), впервые была использована в алгоритме «Априори» [4] для частного случая поиска частотных подмножеств однозначных атрибутов. Аналогичный подход используется в некоторых алгоритмах анализа формальных понятий (например, алгоритм *Titanic* [5]). Все эти процедуры отличаются друг от друга принципом отбора элементов семейства L_k, используемых для построения потенциальных кандидатов следующего уровня. В нашем случае - это 0-1 – описания.

Кроме этого, в отличии от алгоритмов [4], [5], в которых целью является построение элементов семейства L_k, нас интересуют некоторые элементы, не входящие в L_k. Это является причиной шагов 4-5, 12-13.

Сокращенная интервальная структура признакового пространства общего вида

Пусть $A = \{A_1, A_2, \ldots, A_n\}$ – семейство конечных вполне упорядоченных доменов.

$S = A_1 \times A_2 \times \cdots \times A_n$ – n-мерное признаковое пространство.

$I = I_1 \times I_2 \times \cdots \times I_n \subseteq S$ – интервал пространства S, I_j – проекция интервала I на измерение A_j.

Задать интервал I можно, задав его дополнение $\overline{I}$ до признакового пространства S, так как $I = S \setminus \overline{I}$.

Множество значений атрибута A_j, не попавших в I_j, состоит из двух подмножеств:

- значения, предшествующие нижней границе I_j;
- значения, следующие за верхней границей I_j.

Обозначим $d = (d_1, d_2)$ – двумерный вектор, в котором:

- d_1 количество значений, предшествующих нижней границе I_j;
- d_2 количество значений, следующих за верхней границей I_j.

Вектор d можно рассматривать как описание подмножества значений атрибута A_j, дополняющих I_j до домена A_j, а значит, и как описание интервала I_j.

Действительно, пусть вектор $d = (d_1, d_2)$ задан. Взяв первые d_1 элементов домена A_j, получим подмножество значений, предшествующих нижней границе интервала I_j. Взяв последние d_2 элементов домена A_j, получим подмножество значений, следующих за верхней границей интервала I_j. Объединив эти подмножества, получим подмножества значений, дополняющих интервал I_j до множества A_j.

Таким образом, значение координаты d_i вектора d однозначно определяет некоторое подмножество D_i значений домена A_j и $|D_i| = d_i$ - мощность подмножества D_i.

Пусть A – некоторый домен, $I, J \subseteq A$ – интервалы этого домена.

$d(I) = (m, n)$, $d(J) = (k, l)$ – вектор-описания интервалов I и J соответственно.

$d = d(I) \cup d(J) = (\max(m, k), \max(n, l))$ – покоординатное объединение векторов.

Такая операция позволяет получать вектора-описания, соответствующие покоординатному объединению подмножеств атрибутных значений.

$n = \dim S$ – размерность пространства S, $2n$ – число сторон признакового пространства.

Перенумеруем стороны пространства $S : 1, 2, \ldots, 2n$.

Определим $2n$-мерный вектор $d = (d_1, d_2, \ldots, d_{2n})$, в котором координата d_j содержит количество значений, дополняющих интервал I до соответствующей стороны признакового пространства S.

Вектор d определяет подмножество значений всех атрибутов, значений дополняющих интервал I до пространства S. Этот вектор и будет рассматриваться в качестве описания интервала I. $D(S)$ – обозначение множества всех векторов, являющихся описаниями интервалов пространства S.

Поскольку вектор d определяет некоторое подмножество значений всех атрибутов, удобно использовать его и для обозначения этого подмножества. Поэтому, в зависимости от контекста, d будет использоваться и для обозначения $2n$-мерного вектора, и для обозначения множества значений, дополняющих соответствующий интервал до признакового пространства.

Аналогично случаю двумерного вектора, для $2n$-мерных векторов определяется операция покоординатного объединения векторов.

На векторах множества $D(S)$ существует естественный порядок (отношение доминирования) «$\leqslant$», определяемый отношением включения «$\subseteq$» на подмножестве дополняющих значений, соответствующих этим векторам.

Пусть $I, J \in I(S)$ интервалы признакового пространства, $d(I), d(J) \in D(S)$ – вектора-описания интервалов I и J соответственно. Вектор $d(I)$ предшествует вектору (доминирует вектор) $d(J)$, если подмножества дополняющих значений $d(I)$ содержатся в подмножестве дополняющих значений $d(J)$, то есть $d(I) \leqslant d(J)$, если $d(I) \subseteq d(J)$.

Так как $d(I) \subseteq d(J) \Leftrightarrow J \subseteq I$, то отношение доминирования «$\leqslant$» равносильно порядку «$\leqslant$», который был определен ранее. Поэтому для векторов множества $D(S)$ справедливы все вышеприведенные утверждения об элементах семейства $D(S)$.

Описание $d \in D(S)$ называется k-элементным описанием, если сумма значений координат вектора d равна k, то есть $\sum\limits_{j=1}^{2n} d_j = k$.

Утверждение 13. *Каждому k-элементному описанию ($k > 0$) непосредственно предшествует столько его $(k-1)$-элементных подописаний, сколько координат этого k-элементного вектора имеют положительные значения.*

Каждое из этих подописаний получается в результате уменьшения на единицу какой-либо положительной координаты этого вектора-описания.

Следствие 1. *Каждый k-элементный кандидат, имеющий положительное значение только по одной из своих координат, получается из единственного своего $(k-1)$-элементного 0-1-подописания в результате увеличения значения этой координаты на единицу.*

Утверждение 14. *С учетом порядковых номеров координат вектора-описания, каждый k-элементный кандидат, имеющий положительное значение больше чем по одной координате, может быть единственным образом представлен в виде объединения двух его $(k-1)$-элементных 0-1-подописания, у которых $k_i \neq k_j$, где k_i, k_j - первые координаты, имеющие положительные значения, при просмотре координат описания в обратном порядке:*

$$d = (k_1, \ldots, k_{i-1}, k_i + 1, 0, \ldots, k_j + 1, 0, \ldots, 0) =$$
$$(k_1, \ldots, k_{i-1}, k_i + 1, 0, \ldots, k_j, 0, \ldots, 0) \cup$$
$$\cup (k_1, \ldots, k_{i-1}, k_i, 0, \ldots, k_j + 1, 0, \ldots, 0). \qquad (2)$$

Доказательство. Пусть $d = (k_1, \ldots, k_{i-1}, k_i+1, 0, \ldots, k_j+1, 0, \ldots, 0)$ – кандидат. Тогда все непосредственно предшествующие ему подописания являются 0-1 – описаниями.

В частности $d_1 = (k_1, \ldots, k_{i-1}, k_i + 1, 0, \ldots, k_j, 0, \ldots, 0)$ и $d_2 = (k_1, \ldots, k_{i-1}, k_i, 0, \ldots, k_j + 1, 0, \ldots, 0)$ – непосредственно предшествующие подописания и, следовательно, являются 0-1 – описаниями. Их объединение порождает (2). □

Следствие 1 и представление (2) позволяет строить описания интервалов по уровням, используя уже построенные $(k-1)$-элементные 0-1 – подописания для построения k-элементных описаний.

Утверждение 15. *Некоторые из потенциальных кандидатов могут быть представлены в виде (2).*

Это те потенциальные кандидаты, у которых $(k-1)$-элементные подописания, участвующие в (2), являются 0-1 – описаниями.

Утверждение 16. *Каждый k-элементный кандидат, является потенциальным кандидатом представимым или как расширение своего единственного $(k-1)$-элементного подописания или представим в виде (2).*

Алгоритм построения сокращенной интервальной структуры признакового пространства общего вида

Для описаний интервалов признакового пространства общего вида специально подобран такой вид, чтобы алгоритм построения сокращенной интервальной структуры как можно меньше отличался от алгоритма для случая булевого признакового пространства. Поэтому отмечены только изменения, которые необходимо внести в псевдокод алгоритма для булевого признакового пространства.

Шаг 1, 2. Вместо пустого описания $\varnothing$ используется вектор-описание $d = (0, 0, \ldots, 0)$.

Шаг 3. $C_1 := \{d \in D(S) : |d| = 1\}$ – все одноэлементные описания являются кандидатами первого уровня.

Шаг 10. $\widetilde{C_k} := \{(0, \ldots, 0, k_i + 1, 0, \ldots, 0) \mid (0, \ldots, 0, k_i, 0, \ldots, 0) \in L_{k-1}\} \cup$

$$\{(k_1, \ldots, k_{i-1}, k_i + 1, 0, \ldots, 0, k_j + 1, 0, \ldots, 0) \mid$$

$$(k_1, \ldots, k_{i-1}, k_i + 1, 0, \ldots, k_j, 0, \ldots, 0),$$

$$(k_1, \ldots, k_{i-1}, k_i, 0, \ldots, k_j + 1, 0, \ldots, 0) \in L_{k-1}\}.$$

Здесь первое подмножество, участвующее в объединении, получено с использованием следствия 1. Второе подмножество получено с использованием представления (2).

Шаг 11. $C_k := \{d \in \widetilde{C_k} \mid \forall \widetilde{d} \in D(S) : \widetilde{d} \prec d \Rightarrow \widetilde{d} \in L_{k-1}\}$.

Заключение

В статье рассмотрены понятия компонентной структуры $K(H_{\overline{0}})$ признакового пространства, компонентной структуры $K(G)$ классов множества объектов G, компактно расположенной компоненты класса. Показано, в каком случае выполняется равенство $K(G) = K(H_{\overline{0}})$ и определено понятие компактной компонентной структуры признакового пространства. Предложен алгоритм кластеризации, использующий компактность компонентной структуры и оценки размеров «свободных» интервалов [1], [2].

Введено понятие сокращенной интервальной структуры признакового пространства, порождаемой проекцией множества объектов G в признаковое пространство. Предложены алгоритмы построения такой структуры для некоторых частных видов признакового пространства.

Рассмотренные алгоритмы могут применяться к пространствам большой размерности, не требуют предварительного преобразования данных к какому-либо каноническому виду, результат их работы не зависит от порядка расположения входных данных.

Для случая булевого признакового пространства результатом работы алгоритма построения сокращенной интервальной структуры являются наборы литералов, рассматриваемые как описания соответствующих максимальных интервалов. Каждый такой набор литералов определяет элементарную конъюнкцию. Это позволяет строить представления различных конструкций признакового пространства в форме сокращенной ДНФ.

Для признакового пространства общего вида результатом работы алгоритма построения сокращенной интервальной структуры являются вектора-описания, представляющие соответствующий максимальный интервал через дополняющие значения всех его атрибутов. Такие (дополняющие) описания можно преобразовать в прямые описания интервалов, которые могут быть использованы для представления различных конструкций признакового пространства в форме сокращенной ДНФ.

Такое ДНФ представление обеспечивает легкость интерпретации и наглядности получаемых результатов.

Список литературы

1. *Закревский А. Д.* Логика распознавания. — Мн.: Наука и техника, 1988. — 118 с.
2. *Ильченко А. В.* Классификация на основе компонентных структур данных в признаковом пространстве. // Таврический вестник информатики и математики, — 2004. — №2. — с. 363 — 171.
3. *Лбов Г. С., Старцева Н. Г.* Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: Изд-во Ин-та математики, 1999. — 212 с.
4. *Agrawal R., Srikant R.* Fast Algorithms for Mining Association Rules. In proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, September 1994.
5. *Ganter B., Wille R.* Formal Concept Analysis-Mathematical Foundations. Springer-Verlag, Berlin. — 1999.