Кілт сөздер: бағдарламалық қамтамасыз ету, автоматты басқару, жүйесі, тәсілі.

МРНТИ 50.07

Е.Н. Амиргалиев[1], Б. Калыбек уулу[2]

[1]Университет имени Сулеймана Демиреля, Каскелен, Казахстан
[2]Институт информационных технологий и автоматики НАН Кыргызстана

ПОСТРОЕНИЕ КОМПАКТНЫХ ПОДМНОЖЕСТВ – ЦЕНТРОВ КЛАСТЕРОВ В ЗАДАЧАХ ГРУППОВОГО СИНТЕЗА АЛГОРИТМОВ КЛАССИФИКАЦИИ

Аннотация: В статье рассмотрена задача выделения компактных ядер с целью построения алгоритма группового синтеза на множестве решении алгоритмов, входящих в группу, роль и значение построения компактных подмножеств - центров кластеров в задачах группового синтеза алгоритмов классификации, представлены задачи группового синтеза, построения результирующего разбиения, рассмотрены основные подходы к формализации описаний таксономических разбиений, описаны трудности, связанные с разным числом, размерами, а также порядком формирования классов в каждом из алгоритмов комитета, что накладывает целых ряд существенных ограничений при построении искомого результирующего разбиения.

Ключевые слова: синтез, подмножество, задача, алгоритм, кластер.

Основная задача группового синтеза алгоритмов классификации состоит в построении оптимального результирующего разбиения исследуемого объектного множества во множестве разбиений, порождаемых каждым алгоритмом, составляющим базовый набор комитета (группы). Понятие оптимальности конкретизируется для реальных задач классификации.

Формулируем задачу группового синтеза.

Пусть задано множество $S = \{S_1, S_2, ..., S_m\}$ объектов классификации. Каждый объект $S_i \in S$ однозначно описывается набором из n числовых параметров $S_i = (\alpha_{i1}, ..., \alpha_{in})$, называемых признаками характеризующих объектов. Таким образом, S - подмножество n -мерного пространства.

Имеется множество $A = (A_1, A_2, ..., A_t)$ - базовый набор таксономических алгоритмов комитета.

Применительно к множеству S каждый алгоритм $A_u \in A$ строит разбиение

$$A_u(S) = R_u;\ R_u = K_1, K_2, ..., K_l;$$
$$K_i \cap K_j = 0,\quad i \neq j,\quad i, j = 1,2,...,l,\ u = 1,2,...,t. \qquad (1)$$

Ставится задача построения результирующего разбиения R^*, объединяющего результаты каждого алгоритма из A.

В качестве такого результирующего разбиения можно рассмотреть, например, пересечение R^* разбиений $R_1, R_2, ..., R_t$:

$$R^* = \cap\ R_u; \quad u = 1,...,t. \qquad (2)$$

Однако построение фактического пересечения разбиений является весьма сложной переборной задачей, требующей просмотра большего числа возможных вариантов. Поэтому актуальным является вопрос о построении содержательно оптимальных результирующих разбиений во множестве разбиений базового набора алгоритмов комитета.

Существует два основных подхода к формализации описаний таксономических разбиений. Первый подход состоит в представлении разбиения в виде информационной матрицы, характеризующей принадлежность объектов исходного множества тем или иным классам, второй – в виде матрицы смежности, отражающей вхождение каждой пары объектов (S_i, S_j) в один или разные классы.

Информационная матрица $\|a_{ij}\|_{m\,x\,l}$ разбиения R является аналогом информационной матрицы алгоритма распознавания. Единственным их различием является то, что каждый вектор матрицы $\|a_{ij}\|_{m\,x\,l}$ содержит ровно одну единичную компоненту, что единственным образом определяет разбиение исходного множества S на непересекающиеся классы. А информационная матрица в задачах распознавания допускает наличие пересекающихся классов.

Матрица смежности $\|C_{ij}\|_{m x m}$ представляет матрицу, каждый элемент которой определяет принадлежность каждой пары объектов (S_i, S_j), $S_i, S_j \in S$ одному или разным классам:

$$C_{ij} = \begin{cases} 1, & если\ S_i \in K_u,\ S_j \in K_u; \\ 0, & если\ S_i \in K_u,\ S_j \notin K_u. \end{cases} \qquad (3)$$

Рассмотрим каждый из предлагаемых подходов к формализованному описанию разбиений применительно к задаче комитетного синтеза.

Применение каждого алгоритма $A_i \in A$ к множеству S дает набор информационных матриц

$$A(S) = \left\|a^1{}_{ij}\right\|_{m\,x\,l}, \left\|a^2{}_{ij}\right\|_{m\,x\,l}, ..., \left\|a^t{}_{ij}\right\|_{m\,x\,l}. \tag{4}$$

Объединение результатов работы алгоритмов комитетного A можно представить в виде результирующей матрицы $\left\|a^*{}_{ij}\right\|_{m x l}$, предварительно представляясь как

$$\left\|b_{ij}\right\|_{m x l} = \sum_{u=1}^{t} \left\|a^u{}_{ij}\right\|_{m x l}. \tag{5}$$

В дальнейшем к матрице $\left\|b_{ij}\right\|_{m\,x\,l}$ применяется решающее правило, например, вида

$$a^*{}_{ij} = \begin{cases} 1, \; если \; b_{ij} > \Omega; \\ 0, \; если \; b_{ij} \le \Omega, \end{cases} \tag{6}$$

где Ω - специально подбираемый порог.

Следует отметить, что при таком представлении результирующего разбиения приходится преодолевать весьма значительные трудности, связанные с разным числом, размерами, а также порядком формирования классов в каждом из алгоритмов комитета. Это накладывает целых ряд существенных ограничений при построении искомого результирующего разбиения.

С указанных позиций более удобным представляется второй подход, связанный с представлением результатов совместной работы алгоритмов комитета в виде результирующей матрицы $\left\|D_{ij}\right\|_{m\,x\,m}$, формирующейся по правилу большинства по результатам работы каждого из алгоритмов, составляющих комитет. Суть подхода состоит в следующем.

Пусть $S = \{S_1, ..., S_q\}$ - исходное множество объектов из некоторого объектного множества D. Предполагается, что D представимо в виде объединения некоторого числа подмножеств $K_1, ..., K_l$, называемых классами, причем значение l может быть неизвестным. Элементы каждого из множеств $K_1, ..., K_l$ эквивалентны друг другу в некотором смысле. Необходимо для любой исходной выборки $S \subset D$ по информации $J(S)$ определить число классов и провести разбиение объектов S на классы эквивалентных объектов $\widetilde{K}_1, ..., \widetilde{K}_l$, т.е. $\bigcup_{j=1}^{l} \widetilde{K}_j = S$.

Если $J(S)=J(D)$, то совокупность $\widetilde{K}_1,...,\widetilde{K}_l$ характеризует $K_1,...,K_l$.

Пусть имеется некоторый набор алгоритмов классификации $A_1,...,A_t$ и считаются заданными результаты их применения к исходной информации $J\{S_1,...,S_q\}$ в виде информационных матриц

$$A_v(J(S))=\left\|a_{ij}^v\right\|_{q\,x\,l},\; v=1,2,...,t \qquad (7)$$

где $a_{ij}^v\in\{0,1\}$, l_v - число классов, полученное в результате применения алгоритма A_v. Поскольку алгоритмы классификации работают без априорной информации о классах, то классы в информационных матрицах (7) нумеруются произвольно и для того, чтобы сравнить их результаты пользуются различными приемами. Так классифицирующий алгоритм может быть определен как алгоритм, переводящий информацию $J(S)$ в класс эквивалентности $K(\left\|a_{ij}^v\right\|)$, т.е. во множество информационных матриц, равных с точностью до перестановки столбцов.

Здесь предлагается следующий подход для использования информации, полученной в результате работы алгоритмов $A_1,...,A_t$.

Алгоритм A_v дает разбиение объектов S на классы эквивалентных объектов в виде $\left\|a_{ij}^v\right\|_{q\,x\,lv}$, что позволяет множество, состоящее из всевозможных пар объектов из S, однозначно разбить на два подмножества $\{(S_{i,}\,S_j)\}_v^1$ и $\{(S_{i,}\,S_j)\}_v^2$ следующим образом:

$$\{(S_{i,}\,S_j)\}_v^1=\{(S_{i,}\,S_j)\mid\exists_K\in\{1,2,...,l_v\},i=1,...,m-1,\; j=i+1,...,m,\quad a_{ik}=a_{jk}=1;$$

(8)

$$\{(S_{i,}\,S_j)\}_v^2=\{(S_{i,}\,S_j)\mid\exists_K\in\{1,2,...,l_v\},i=1,...,m-1,\; j=i+1,...,m,\quad a_{ik}\neq a_{jk}. \qquad (9)$$

Содержательно множества (4.29) и (4.30) определяют разбиение всех пар на одновременно принадлежащие и не принадлежащие, соответственно, какому-нибудь из классов $K_1,...,K_l$.

Множества S, $\{(S_{i,}\,S_j)\}_v^1$, $\{(S_{i,}\,S_j)\}_v^2$ определяют два взаимно дополнительных графа

$$G_1^v=\left(S,\{(S_i,S_j)\}_v^1\right)\;,\qquad G_2^v=\left(S,\{(S_i,S_j)\}_v^2\right),$$

где граф G_1^v- неориентированный, состоящий из суммы полных подграфов.

Задача состоит в нахождении такого разбиения, которое было бы наиболее согласовано со всеми разбиениями, полученными алгоритмами $A_i \in A, \quad i = 1,2,...,t$.

Предлагаемый подход основан на построении и анализе $Q(t)$ - результирующего графа $G = (S, \{S_i, S_j\})$, каждое ребро которого «наследует» свойство не менее, чем $Q(t)$ графов $G_1,...,G_t$, $[t/2] \le Q(t) \le t$.

Пусть B - совокупность n - мерных булевых векторов $b = (b_1,...,b_t)$ с нормой $\|b\| = Q(t)$ и $J(b) = \{i / b_i = 1\}$, тогда

$$G = (S, \{(S_i, S_j)\}) = \bigcup_{b \in B} \bigcap_{v \in J(b)} G_1^v, \tag{10}$$

$$\{(S_i, S_j)\} = \bigcup_{b \in B} \bigcap_{v \in J(b)} \{(S_i, S_j)\}_v^1. \tag{11}$$

Множество ребер графа G может быть записано в более приемлемой с вычислительной точки зрения форме. Введем характеристические функции $\chi_1,...,\chi_t$ на парах (S_u, S_v):

$$\chi(S_u, S_v) = \begin{cases} 1, & \textit{если } (S_u, S_v) \in \{(S_u, S_v)\}_v^1 \\ 0, & \textit{в противном случае} \end{cases}, \tag{12}$$

тогда

$$\{(S_i, S_j)\} = \{(S_u, S_v) \mid \sum \chi_v(S_u, S_v) \ge Q(t)\} \atop u, v, i, j = 1,2,...,q \quad . \tag{13}$$

Покажем, что бинарное отношение R, определяемое $Q(t)$ - результирующим графом G, наиболее согласовано со всеми отношениями, определяемыми алгоритмами $A_1,..., A_t$.

Дальнейший анализ $Q(t)$-результирующего графа G заключается в выделении в нем наиболее представительных частей и может развиваться в нескольких направлениях. Например, в нахождении максимально полных частей графа G или в поиске ядер на структурных элементах графа G, т.е. совокупности вершин или ребер графа, удовлетворяющих некоторым соображениям содержательного характера.

С каждым бинарным отношением $R_v \subset \{S\}x\{S\}$ и соответствующим графом G_v связана матрица смежности $\|\chi_v(S_i, S_j)\|_{qxq}, \quad v = 1,...,t$. Рассматривая эти матрицы как векторы размерности q^2, т.е.

$$\rho(R_u, R_v) = \sum_{i,j=1}^{q} |\, \chi_u(S_u, S_v) - \chi_v(S_i, S_j) \,|, \tag{14}$$

находим некоторое бинарное отношение $\widetilde{R}$, на котором достигается минимальное значение

$$\sum_{v=1}^{t} \rho(R^1, R_v)$$

для всех допустимых $R^1 \subset \{S\}x\{S\}$. В этом случае $\widetilde{R}$ - согласованное бинарное отношение для заданного набора отношений $R_1,...,R_t$.

Приведем теорему о критерии большинства.

Теорема. Граф G из (4.31) при соответствующем выборе $Q(t)$, $1 \le Q(t) \le t$ определяет согласованное бинарное отношение R, удовлетворяющее условию:

$$\sum \rho(R, R_v) = \min_{R^1 \subset \{S\}x\{S\}} \sum_{v=1}^{t} \rho(R^1, R_v). \qquad (15)$$

Смысл теоремы состоит в том, что определяется значение $Q^*(t)$, при котором бинарное отношение результирующего графа $G(Q^*(t))$ соответствует среднему значению для всех бинарных отношений, получаемых в результате применения заданного набора алгоритмов $A_1,...,A_t$ к исходному множеству объектов. Оказывается, достаточно использовать критерий большинства $Q^*(t) \ge]\frac{t}{2}[+1$ для получения оптимального графа $G(Q(t))$.

Задача поиска максимально полных частей графа состоит в нахождении полных порожденных подграфов G_i, $i = 1,2,...,l$,удовлетворяющих условиям $V(G_i) \bigcap V(G_j) = 0$ для любых $i \ne j$, $i, j = 1,2,...,l$, таких, что если $S_t \subset V(G_i)$ и $S_w \subset V(G_j)$, то $(S_t, S_w) \notin \bigcup(G)$, где $V(G_k)$ - множество вершин, а $\bigcup(G_k)$ - множество ребер произвольного графа.

Таким образом, это – задача нахождения в G связных частей, являющихся эквивалентными в смысле реберных связей. Данная задача, в силу сложности существующих алгоритмов ее решения не может быть решена в приемлемое время, если m велико. С другой стороны, не всегда гарантировано существование полных частей в графе достаточной мощности.

Поэтому выделение полных частей в графе заменяется более упрощенной задачей выделения частей в графе G, более насыщенными, чем другие, полными графами, состоящими из трех вершин.

Результирующий граф G определяет бинарное отношение $R \subset \{S\} \times \{S\}$ на множество вершин S не являющееся, вообще говоря, отношением эквивалентности.

Определение. Бинарное отношение $R_C \subset \{S\} \times \{S\}$ назовем отношением слабой эквивалентности, если для любой вершины S_ω существует хотя бы две вершины S_u и S_v такие, что из $(S_\omega, S_u) \subset R_C$ и $(S_\omega, S_v) \subset R_C$ следует $(S_u, S_v) \subset R_C$.

Очевидно, что отношение слабой эквивалентности разбивает некоторое подмножество $\widetilde{S} \subset S$ на непересекающиеся классы связанных вершин. Если класс один, то элементы множества $\widetilde{S}$ связаны отношением слабой эквивалентности.

Решение задачи о выделении наиболее насыщенных частей полными графами, состоящими из трех вершин - есть выделение из всех подмножеств слабо эквивалентных вершин в графе множества, для которого слабая эквивалентность наиболее близка к полной эквивалентности относительно реберных связей.

Таким образом, нами рассмотрена задача выделения компактных ядер с целью построения алгоритма группового синтеза на множестве решении алгоритмов, входящих в группу.

Список использованной литературы:

1 Амиргалиев Е.Н., Мухамедгалиев А.Ф. Оптимизационная модель алгоритмов классификации (таксономии) // Жур. вычисл. мат. и матем. физ. – 1985. – № 11. – С. 1733 – 1737

2 Айдарханов М.Б. Принципы синтеза алгоритмов и решений в задачах групповых классификации // Доклады АН РК. – № 3. – 1992. – С. 21 - 25

3 Айдарханов М. Б., Амиргалиев Е. Н., Мухамедгалиев А. Ф. О синтезе результатов алгоритмов классификации // Известия АН Каз ССР. – Серия физ.– мат. – № 1. – 1998. – С. 11-17

4 Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания и классификации. // Проблемы кибернетики. – М.: Наука. – Вып. 33. – 1978. – С. 93-103

E.N. Amirgaliev[1], B. Kalybek uulu[2]

[1]Suleyman Demirel University, Kaskelen, Kazakhstan

[2]Institute of Information Technologies and Automation National Academy of Sciences of Kyrgyzstan

BUILDING A COMPACT SUBSET - CENTER OF THE CLUSTER IN THE PROBLEM OF GROUP CLASSIFICATION ALGORITHM SYNTHESIS

Abstract: The article considers the problem of extracting the compact cores in order to build the algorithm group synthesis on a set of decision algorithms within the group, the role and importance of building compact subsets - clusters centers in problems of group synthesis classification algorithms, presented the group synthesis problem, the construction of the resulting partition considered the main approaches to the formalization of the taxonomic descriptions of partitions, described the difficulties associated with a different number, size and order of the formation of classes in each of the Committee of the algorithms that puts whole number of significant limitations in constructing the desired result of the partition.

Keywords: synthesis, a subset of the task, the algorithm cluster.

Е.Н. Әмірғалиев[1], Б. Қалыбек уулу[2]

[1]Сулейман Демирел атындағы университеті, Қаскелең, Қазақстан
[2]Қырғызстан ҰҒА Ақпараттық технологиялар мен автоматтандыру институты

АЛГОРИТМДЕР ЖИНАҚТАУДЫҢ ТОПТЫҚ СИНТЕЗДЕУДЕ КЛАСТЕРЛЕР ОРТАЛАРЫНЫҢ КОМПАКТЫ ІШКІ ЖИЫНЫН ҚҰРУ

Аңдатпа: Мақалада негізгі тәсілдер, соның нәтижесінде бөлімнің құрылыс, топтық синтез мәселелері ұсынылады, топтық синтез жіктеу алгоритмдерін мәселелері кластерлер орталықтары - бап, топ ішіндегі ықшам жиындарын құрылыс рөлі мен маңыздылығын шешім алгоритмдерін жиынтығы бойынша алгоритм тобы синтезін құру мақсатында жинақы өзегін алу мәселесін қарайды ресімдеу сипаттамалар таксономикалық қалқалар, қалқаның қажетті нәтижеге салу елеулі шектеулер бүкіл санын қояды алгоритмдерін комитетінің әрбір сынып қалыптастыру түрлі саны, көлемі мен тәртібіне байланысты қиындықтарды сипатталған.

Кілт сөздер: синтез, ішкі жиын, міндет, алгоритм, кластер.

ҒТАХР 50.13

М.Н. Калимолдаев[1], А.А. Абдилдаева[1], Ф.М. Галиева[1]
[1]Ақпараттық және есептеуіш технологиялар институты

СЫЗЫҚТЫ ЕМЕС РЕГУЛЯТОРЛЫ ФАЗАЛЫҚ ЖҮЙЕМЕН БАЙЛАНЫСТЫ ГЛОБАЛЬДЫ АСИМТОТИКАЛЫҚ ОРНЫҚТЫЛЫҚ