УДК 681.3: 519.2

Выбор информативных признаков и поиск наилучшей кластеризации: единый подход к классификации методов

к.ф.-м.н., доц. Сарычева Л.В. , , Национальный горный университет, г. Днепропетровск, sarycheva@prognoz.dp.ua

Рассмотрен единый подход к классификации методов кластеризации и методов выбора информативных признаков в распознавании образов, в котором и кластеризация, и выбор признаков являются задачами дискретной оптимизации на решетках разбиений. Проведено соответствие между алгоритмами кластеризации и выбора признаков. Предложен новый иерархический двусторонний алгоритм кластеризации.

Введение

Методы кластеризации и методы выбора информативных признаков из совокупности исходных широко применяются в интеллектуальном анализе данных [1, 2]. Задачи кластеризации и выбора признаков являются, в общем случае, многоэкстремальными. Большинство известных алгоритмов решения этих задач не позволяют достичь глобального экстремума.

В данной работе предлагается рассматривать алгоритмы поиска наилучшей кластеризации и поиска информативных признаков с единых позиций. Это позволяет находить соответствие между этими алгоритмами и предлагать новые пути поиска глобального экстремума, присущие только одной группе алгоритмов. Поэтому вырисовываются новые пути развития методов поиска наилучшей кластеризации (при незаданном числе кластеров) на основе соответствия с аналогичными известными методами выбора признаков.

Исходные данные представляют собой таблицы типа «объект-признак»:

$$X=\{x_{ij}\}, \quad i=1,2,\ldots,n, \qquad j=1,2,\ldots,m,$$

где i-я строка таблицы – i-й вектор-объект (численные значения всех признаков i-го объекта), а j-й столбец таблицы – j-й вектор-признак (численные значения j-го признака для всех объектов), n – число объектов, m – число признаков.

Задача выбора информативных признаков

Применение методов распознавания образов неразрывно связано с выявлением и использованием в решающих правилах наборов информативных признаков.

Пусть имеется k классов объектов: $K_1, K_2, \ldots, K_k$. Каждый объект описывается m признаками. Класс K_l, $l=1,2,..,k,$ содержит n_l объектов:

$$n, + n_2 + \ldots + n_k = n.$$

Под выбором информативных признаков понимают сужающее отображение F:

$$\{X^m\} \xrightarrow{F} \{X^{m_1}\},$$

$$m_1 < m, \qquad F=F(K_1, K_2, \ldots, K_k),$$

при котором достигается экстремум некоторого функционала качества $J_X(F)$. Каждому такому отображению можно поставить в соответствие вектор

$$V=(V_1, V_2, \ldots, V_m),$$

где $V_i = 1$, если X^i входит в состав выбираемых признаков, и $V_i = 0$ - в противном случае. Тогда функционал $J_X(F)$ можно рассматривать как функцию $g_{m1}(V)$, заданную на вершинах единичного гиперкуба $[0,1]^m$. Из всех вершин нужно найти такую, на которой достигается экстремум $g_{m1}(V)$.

Математическая постановка задачи выбора информативной подсистемы признаков из исходной может быть представлена в виде:

$$g_{m1}(V) \to \min_{V \subset D}; \qquad D=\{V=R^m;\ v_j=0\ U\ 1\}, \qquad (1)$$

если размерность m_1 искомой подсистемы не определена, или в виде

$$g_{m1}(V) \to \min_{V \subset G}; \qquad G=\{V \subset D;\ \sum_{j=1}^{m_1} v_j = m_1\}, \qquad (2)$$

если размерность искомой подсистемы задана. Элементами множества D являются вершины решетки разбиений - единичного m-мерного гиперкуба (рис.1).

Особенность задачи состоит в том, что в большинстве случаев функция многоэкстремальна и не может быть исследована аналитически. Поэтому для нахождения ее экстремума используются различные процедуры перебора вершин с текущей оценкой их ценности. Критерий качества определяется исходя из априорных данных и предварительного анализа совокупности признаков.

Для нахождения оптимального подмножества признаков в общем случае сейчас известен только один метод – полный перебор всевозможных подмножеств исходного множества признаков. Число вариантов при этом огромно – для выбора наилучшего подмножества m_1 признаков из общего числа m признаков требуется $m! / (m_1! (m-m_1)!)$ сравнений значений функции $g_{m1}(V)$. Поэтому на практике применяют субоптимальные методы поиска, сокращающие перебор вариантов.

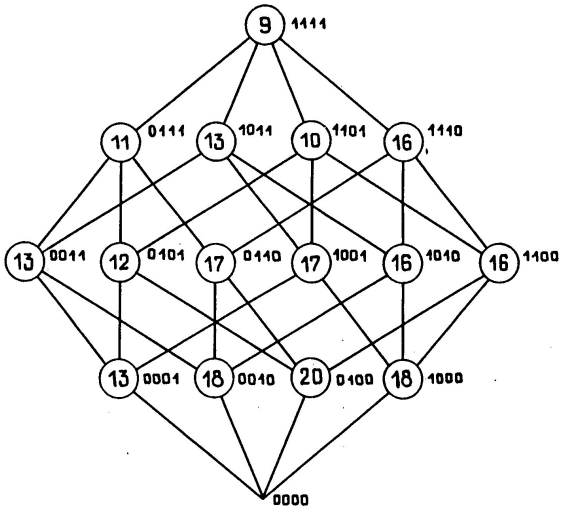

Рис.1. Пример решетки выбора признаков
(m=4, вершины – элементы множества D)

В зависимости от способа построения алгоритма, задающего последовательность прохождения вершин в гиперкубе (рис.1), методы выбора признаков можно разделить на следующие группы.

1. Последовательный перебор вариантов [3]:
а) полный перебор,
б) ранжировка,
в) последовательное присоединение,
г) последовательное отбрасывание,
д) различные комбинации последовательного присоединения и последовательного отбрасывания,
е) двусторонний поиск и др.
2. Случайный поиск и его модификации [3, 4, 11]:
а) обычный случайный поиск (метод Монте-Карло),
б) случайный поиск с адаптацией,
в) случайный поиск с возвратом,
г) модифицированный случайный поиск с адаптацией и др.
3. Комбинированные методы [4, 5, 10]:
а) метод ветвей и границ,
б) минимаксный метод,
в) метод синтеза,
г) *beam*-поиск,
д) *(r, s)* –поиск и др.
4. Генетические методы (комбинаторика + случайный поиск) [6]:
а) при доминирующем влиянии оператора скрещивания,
б) при доминирующем влиянии оператора мутации.

Сравнение некоторых методов выбора признаков проведено в [3].

Задача поиска наилучшей кластеризации

Под кластеризацией, или автоматической классификацией, понимают группировку заданного множества объектов в подмножества в соответствии со свойствами самих объектов. При этом требуется, чтобы подмножества включали в себя объекты, в некотором смысле более похожие на объекты из того же подмножества, чем на объекты из других подмножеств.

Пусть задано конечное множество объектов $X = \{X_1, X_2, \ldots, X_n\}$.
Кластеризацией

$$K(X) = \{K_1(X), K_2(X), \ldots, K_k(X)\},\ 1 \le k \le n,$$

множества X называется семейство непустых, попарно непересекающихся подмножеств (кластеров) $K_l(X)$,

$l=1,2,..,k$, множества X, объединение которых совпадает с X:

$$K_1 \cup K_2 \cup \ldots \cup K_k = X, \quad K_i \cap K_j = \varnothing \text{ при } i \neq j, \quad K_l \neq \varnothing, \; l=1,2,..,k.$$

Число кластеров k может быть заранее неизвестным.

Для решения задачи кластеризации необходимо:

а) дать определение кластера – указать свойства, общие для всех объектов отдельного кластера (меру сходства между объектами);

б) задать способ образования кластеров (*сортировка, перегруппировка, объединение, разбиение, добавление, поиск*);

в) указать критерий J качества кластеризации;

г) организовать движение к максимуму (минимуму) критерия J (при этом определяется и число реально существующих кластеров).

Алгоритмы кластеризации отличаются большим разнообразием (например, алгоритмы, реализующие полный перебор сочетаний объектов или осуществляющие случайные разбиения множества объектов). В то же время большинство алгоритмов состоит из двух этапов. На первом этапе задается начальное (возможно, произвольное) разбиение множества объектов на классы и определяется некоторый математический критерий качества кластеризации. На втором этапе объекты переносятся из класса в класс до тех пор, пока значение критерия не перестанет улучшаться.

Классификационные процедуры иерархического типа основаны на последовательном *объединении* кластеров (агломеративные процедуры) и на последовательном *разбиении* (дивизимные процедуры). Наибольшее распространение получили агломеративные процедуры. На первом шаге в таких процедурах все объекты считаются отдельными кластерами. Затем на каждом последующем шаге два ближайших кластера объединяются в один. Каждое объединение уменьшает число кластеров на один так, что в конце концов все объекты объединяются в один кластер. В отличие от оптимизационных кластерных алгоритмов, предоставляющих исследователю конечный результат группирования объектов, иерархические процедуры позволяют проследить процесс выделения группировок и иллюстрируют соподчиненность кластеров, образующихся на разных шагах какого-либо агломеративного или дивизимного алгоритма.

Функционалы качества и конкретные алгоритмы кластеризации достаточно полно и подробно рассмотрены в литературе [7, 8, 9]. Они характеризуются различной трудоемкостью и подчас требуют ресурсов высокопроизводительных компьютеров.

Поиск наилучшей кластеризации как задача дискретной оптимизации на решетке разбиений исходного множества объектов

В задаче выбора признаков и в задаче поиска наилучшей кластеризации в качестве исходных данных используется одна и та же таблица «объект-признак». Если под поиском (выбором) наилучшей кластеризации понимать сужающее отображение Φ множества исходных объектов X на множество кластеров K:

$$\{X_n\} \xrightarrow{\Phi} \{K_k\},$$

$$k \leq n, \qquad \Phi = \Phi(X^1, X^2, \ldots, X^m),$$

при котором достигается экстремум некоторого функционала качества $I_X(\Phi)$, то каждому такому отображению можно поставить в соответствие разбиение W множества исходных объектов. Тогда $I_X(\Phi)$ можно рассматривать как функцию $q_k(W)$, заданную на вершинах решетки разбиений исходного множества объектов. Из всех вершин нужно найти ту, на которой достигается экстремум $q_k(W)$.

Математическая постановка задачи поиска кластеризации может быть представлена в виде, аналогичном (1), если число кластеров не задано, или в виде, аналогичном (2), если число k кластеров задано.

Пример решетки разбиений для случая n=4 представлен на рис.2. Поэтому можно найти соответствие между методами выбора информативных признаков и методами поиска наилучшей кластеризации. Например, методу последовательного присоединения можно поставить в соответствие построение дивизимного иерархического дерева, а методу последовательного отбрасывания – агломеративного.

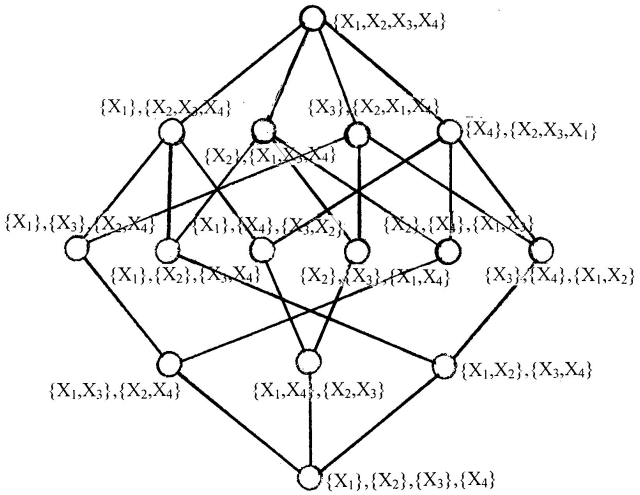

Рис.2. Пример решетки поиска наилучшей кластеризации (n=4)

Типичные пути поиска в исследуемой решетке (рис.3): а) полный перебор; б) ранжировка; в) последовательное присоединение; г) последовательное отбрасывание; д) двусторонний поиск; е) последовательное присоединение и отбрасывание; ж) метод ветвей и границ; з) *beam*- поиск; и) *(r, s)*- поиск.

Единый системный подход к классификации методов кластеризации и методов выбора признаков в распознавании образов, в котором и кластеризация, и выбор признаков являются задачами дискретной оптимизации на вершинах решетки разбиений, позволяет увидеть новые направления развития этих методов.

Например, при построении иерархических деревьев кластеризации, можно навстречу друг другу строить два дерева: одно дерево строить при помощи постепенного увеличения числа кластеров, а другое при помощи уменьшения числа кластеров; там, где они встретятся, и будет оптимальная кластеризация. Такое построение соответствует двустороннему поиску (рис.3д).

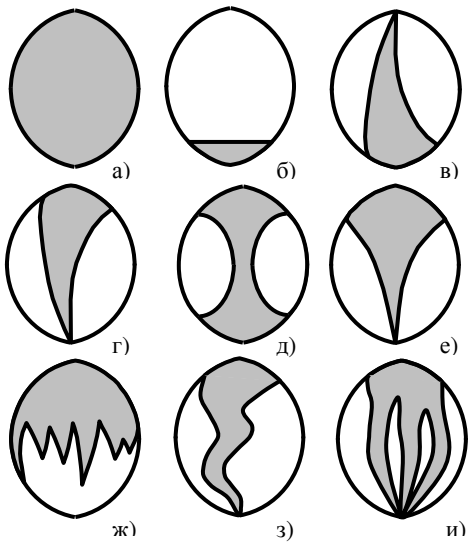

Рис.3. Типичные пути поиска в исследуемом гиперкубе:
а) полный перебор; б) ранжировка; в) последовательное присоединение;
г) последовательное отбрасывание; д) двусторонний поиск;
е) последовательное присоединение и отбрасывание;
ж) метод ветвей и границ; з) *beam*- поиск; и) *(r, s)*- поиск

Иерархический алгоритм двустороннего поиска

Новый иерархический двусторонний алгоритм кластеризации реализует способ образования кластеров, основанный на *объединении и разбиении*, используя дивизимный и агломеративный методы (рис.4).

Выбор числа кластеров в большинстве алгоритмов кластеризации не автоматизирован и является предопределенным. Поэтому возникает потребность принятия решения о реально существующем числе кластеров, что приводит к различного рода погрешностям при анализе структуры данных.

Характерной особенностью иерархического двухстороннего алгоритма является автоматическое определение числа кластеров k при достижении оптимального показателя сходства между кластеризациями K и Q, полученными соответственно дивизимным и агломеративным алгоритмами:

$$k^* = \arg\min_k \; d(K(k), Q(k)), \quad k \in \{2,3,...,n-1\}. \qquad (3)$$

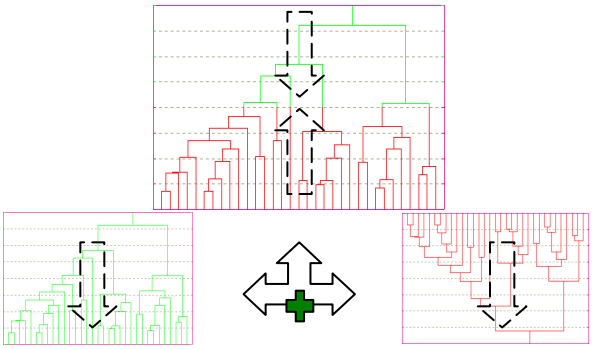

Рис.4. Построение иерархического дерева в двустороннем поиске

Для оценки сходства между двумя различными кластеризациями K и Q конечного множества объектов используется мера близости:

$$d(K,Q) = \frac{\frac{1}{2}\left(\sum_{i=1}^{k_1} |K_i|^2 + \sum_{i=1}^{k_2} |Q_i|^2\right) - \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} |K_i \cap Q_j|^2}{\frac{1}{2}\left(\sum_{i=1}^{k_1} |K_i|^2 + \sum_{i=1}^{k_2} |Q_i|^2\right)}, \qquad (4)$$

где k_1, k_2 – число кластеров (подмножеств исходного множества) в кластеризациях K и Q соответственно; $|K_i|, |Q_j|, i = 1,2,...,k_1;\ j = 1,2,...,k_2$,– мощности соответствующих подмножеств, т.е. число элементов в кластерах.

Величина $d(K,Q)$ принимает значения от *0* до *1*:

0 – при полностью совпадающих разбиениях в кластеризациях K и Q,

1 – при полностью несовпадающих, когда $\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} |K_i \cap Q_j|^2 = 0$.

По результатам кластеризаций K и Q, для которых выполняется условие (3), строится матрица $\boldsymbol{A}=(a_{ij})$, $i, j=1,2,\ldots, k$, где элемент a_{ij} определяет число объектов, одновременно входящих в кластер K_i и кластер Q_j:

K	Q					
	Q_1	Q_2	…	Q_j	…	Q_k
K_1	a_{11}	a_{12}	…	a_{1j}	…	a_{1q}
K_2	a_{21}	a_{22}	…	a_{2j}	…	a_{2k}
…	…	…	…	…	…	…
K_i	a_{i1}	a_{i2}	…	a_{ij}	…	a_{ik}
…	…	…	…	…	…	…
K_k	a_{k1}	a_{k2}	…	a_{k3}	…	a_{kk}

Элементы a_{ij} матрицы $\boldsymbol{A}$ подсчитываются следующим образом:

$$a_{ij} = \sum_{\substack{e:(X_e \in K_i)\wedge(X_e \in Q_j)=1 \\ e \in \{1,2,...,n\}}} 1, \qquad i, j=1,2,\ldots,k. \qquad (5)$$

Сумма элементов матрицы $\boldsymbol{A}$ равна числу кластеризуемых объектов:

$$\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij} = n.$$

Номер кластера является не более чем меткой, то есть можно поменять нумерацию кластеров, сохранив при этом состав входящих в них элементов. Поэтому для определения «преемственности» кластеризации Q по отношению к K в матрице $\boldsymbol{A}$ производится перестановка столбцов таким образом, чтобы максимальный элемент i-й строки попал на главную диагональ:

$$Q_i \leftrightarrow \max_i a_{ii}.$$

Таким образом находятся ядра кластеров K_{1o}, K_{2o},…,K_{ko}. Ядро i-го кластера K_{io} образуют объекты (их число $|K_{io}| = a_{ii}$), входящие в один и тот же кластер в ближайших ($min\ d(K,Q)$) кластеризациях, полученных агломеративной и дивизимной процедурами. Выделенные ядра кластеров расширяют путем доклассификации

$$n - \sum_{i=1}^{k} |K_{io}| = \sum_{i=1}^{k} \sum_{\substack{j=1 \\ (j \neq i)}}^{k} a_{ij}$$

оставшихся объектов по методу ближайшего соседа или используя в качестве меры близости расстояние между объектом и "центром тяжести" ядра кластера.

Заключение

Рассмотрение задач выбора признаков и поиска наилучшей кластеризации как задач дискретной оптимизации на решетке разбиений указывает соответствующие пары методов: последовательное присоединение – объединение (агломеративный), последовательное отбрасывание – разбиение (дивизимный), а также пути создания новых методов поиска наилучшей кластеризации – аналогичных методам выбора признаков.

Предложен новый алгоритм кластеризации – двусторонний поиск, преимущества которого: 1) автоматическое определение числа реально существующих кластеров; 2) учет результатов агломеративного и дивизимного разбиений для выделения начальных ядер кластеров.

Литература

Дюк В., Самойленко А. Data Mining. Учебный курс. –С.П.: Питер, 2001. – 368с.

Загоруйко Н.Г. Прикладные методы анализа данных и знаний. - Новосибирск: Изд-во ИМ СО РАН, 1999. - 270 с.

Мирошниченко Л.В. Сравнение алгоритмов выбора признаков в распознавании образов / Статистические проблемы управления.- Вильнюс: ИМК АН Литвы, 1990. Вып.93. С.78-91.

Методы, критерии и алгоритмы, используемые при преобразовании, выделении и выборе признаков в анализе данных. Чепонис К.А., Жвиренайте Д.А., Мирошниченко Л.В., Бусыгин Б.С.- Вильнюс: Изд-во ИМК АН ЛитССР, 1988.- 149 с.

Бусыгин Б.С., Мирошниченко Л.В. Распознавание образов при геолого-геофизическом прогнозировании. Днепропетровск: ДГУ, 1991. - 168 с.

Курейчик В.М. Генетические алгоритмы. Монография. Таганрог: ТРТУ, 1998.-242с.

Жамбю М. Иерархический кластер-анализ и соответствия. - М: Финансы и статистика, 1988. - 342 с.

Классификация и кластер / Под ред. Дж.Вэн Райзина.- М: Мир, 1980.- 389 с.

Ту Д., Гонсалес Р. Принципы распознавания образов: Пер. с англ.- М.: Мир, 1978-411с.

Siedelcki W., Sklansky J. On automatic feature selection // IEEE Pattern Recognition and Art. Int., 1988. Vol. 2, Nr.2.- P.197-220.

Раудис Ш. Определение числа полезных признаков в задаче классификации / Статистические проблемы управления. – Вильнюс: ИМК АН ЛитССР, 1976. Вып. 14. С. 137-150.