СОВРЕМЕННЫЕ ТЕНДЕНЦИИ В КЛАСТЕРНОМ АНАЛИЗЕ

В.Б. Бериков, Г.С. Лбов

Институт математики им. С.Л. Соболева СО РАН
630090, г. Новосибирск, пр. Академика Коптюга, д. 4

Аннотация. В статье сделан обзор существующих подходов к решению задачи кластерного анализа. Рассматриваются новые разработки в области кластерного анализа, основанные на привлечении ансамблей алгоритмов и логических моделей. Описываются преимущества таких алгоритмов. Ансамблевые методы позволяют значительно повысить устойчивость группировочных решений. Логические модели позволяют группировать разнотипные данные, а также давать объяснение результатов анализа на языке логических высказываний. Формулируются перспективные направления развития кластерного анализа.

Annotation. In this paper, we overview the existing approaches to solving the problem of cluster analysis. We consider new developments in the field of the cluster analysis, based on ensembles of algorithms and logical models. The advantages of such algorithms are described. The ensemble methods allow to raise significantly the stability of grouping decisions. Logical models allow to group heterogeneous data, as well as give an explanation of results of the analysis on the language of logical statements. The perspective directions of cluster analysis are formulated.

1. Введение

В последние десятилетия наблюдается рост интереса к новому направлению в обработке информации – интеллектуальному анализу данных (Data Mining). В отличие от классических способов анализа, в этой области большое внимание уделяется моделированию поведения человека, решающего сложные интеллектуальные задачи обобщения, выявления закономерностей, нахождения ассоциаций и т.д. В большой степени развитию этой дисциплины способствовало проникновение в сферу анализа данных идей, возникших в теории искусственного интеллекта.

В предлагаемой работе рассматривается частная задача интеллектуального анализа данных – задача кластерного анализа, известная также как задача автоматической группировки объектов, классификации без учителя или таксономии. Основной целью в кластерном анализе является выделение сравнительно небольшого числа групп объектов, как можно более схожих между собой внутри группы, и как можно более отличающихся в разных группах. Этот вид анализа широко используется в информационных системах при решении задач классификации и обнаружения закономерностей в данных: при работе с базами данных, анализе интернет-документов, сегментации изображений и т.д. В настоящее время разработано достаточно большое число алгоритмов кластерного анализа (см. [1,2,17]). Однако в этой области существует ряд актуальных проблем, рассмотрению которых и посвящена статья.

Работа имеет следующий план. В параграфе 2 вводятся основные понятия и обозначения, используемые в работе. В следующем параграфе кратко перечисляются известные подходы, применяемые в кластерном анализе. В четвертом параграфе формулируются актуальные методологические проблемы кластерного анализа. В параграфе 5 описывается ансамблевый подход к решению задачи кластер-анализа. Шестой параграф посвящен логическим моделям кластерного анализа. В заключении делаются выводы, обрисовываются перспективные направления дальнейших исследований.

2. Основные понятия и обозначения

Пусть имеется выборка объектов исследования $s=\{o^{(1)},...,o^{(N)}\}$, которая сформирована в результате отбора некоторых представителей генеральной совокупности. Требуется сформировать $K \geq 2$ классов (групп объектов); число классов может быть как выбрано заранее, так и не задано (в последнем случае оптимальное количество кластеров должно быть определено автоматически). Каждый объект генеральной совокупности описывается с помощью набора переменных $X_1,...,X_n$. Набор $X=\{X_1,...,X_n\}$ может включать переменные разных типов (количественные и качественные, под которыми будем понимать номинальные и булевы, а также порядковые). Пусть D_j обозначает множество значений переменной X_j. Обозначим через $x=x(o)=x_1(o),...,x_n(o)$ набор наблюдений переменных для объекта o, где $x_j(o)$ есть значение переменной X_j для данного объекта. Соответствующий выборке набор наблюдений переменных будем представлять в виде таблицы данных V с N строками и n столбцами: $V=\{x_j^{(i)}\}$, $i=1,2,\ldots,N$, $j=1,2,\ldots,n$; при этом значение $x_j^{(i)}$, находящееся на пересечении i-й строки и j-го столбца соответствует наблюдению j-й переменной для i-го объекта: $x_j^{(i)}=X_j(o^{(i)})$. В некоторых задачах исходная информация представляет собой таблицу попарных расстояний между объектами.

Можно выделить следующие основные этапы кластерного анализа.

1. *Формирование системы переменных.* Часто исследователь не может с уверенностью сказать, какие именно переменные действительно важны для анализа, поэтому стремится включить как можно больше потенциально информативных факторов. Нередко требуется предварительно выбрать из исходного множества переменных наиболее эффективную подсистему (в зарубежной литературе этот процесс называется «feature selection»). Кроме того, в некоторых задачах целесообразно трансформировать исходные переменные так, чтобы образовать новые, более информативные показатели («feature extraction»). Чтобы избежать «доминирования» переменных с большим масштабом измерения, проводят предварительную нормировку исходных переменных.

2. *Определение способа вычисления расстояния между объектами или группами объектов.* Этот способ должен отражать специфику решаемой прикладной задачи. Для каждой пары объектов $o^{(i)}$ и $o^{(l)}$ обозначим расстояние между ними через $\rho(o^{(i)},o^{(l)})$, где $i \neq l$. Например, в случае непрерывных переменных может быть задано евклидово расстояние $\rho_E(o^{(i)},o^{(l)}) = \sqrt{\sum_{j=1}^{n}(x_j^{(i)} - x_j^{(l)})^2}$. Чтобы исключить влияние сильных линейных корреляций между переменными, применяют расстояние Махалонобиса $\rho_M^2(o^{(i)},o^{(l)}) = (\mathbf{x}^{(i)} - \mathbf{x}^{(l)})^T \Sigma^{-1}(\mathbf{x}^{(i)} - \mathbf{x}^{(l)})$, где $\mathbf{x}^{(i)}$ и $\mathbf{x}^{(l)}$ - вектора-столбцы значений переменных для соответствующих объектов, Σ - ковариационная матрица (оцененная по выборке, либо полагаемая известной априори). Для номинальных переменных может использоваться расстояние Хэмминга. Для групп объектов также определяется способ нахождения расстояния, например, по принципу «дальнего соседа», «ближнего соседа» и др. [2]. Принцип «дальнего соседа» оправдан в случае, когда есть априорная информация о том, что таксоны имеют компактную сферическую форму. Принцип «ближнего соседа» имеет смысл применять, если известно, что таксоны могут иметь «вытянутую» форму или концентрически расположены.

3. *Группировка объектов.* На этом шаге проводится создание групп объектов. Разбиение на группы может быть «жестким» (формируется разбиение исходного множества объектов), а может быть и «нечетким» (вычисляется степень принадлежности каждого объекта к группам). В данной работе будем рассматривать группировку первого типа. Пусть сформировано группировочное решение $G=\{C^{(1)}, C^{(2)},\ldots, C^{(K)}\}$, где $C^{(k)} = \{o^{(i_1)}, o^{(i_2)}, \ldots, o^{(i_{N_k})}\}$, N_k – число объектов, входящих в k-й кластер, $k=1,2,\ldots,K$. Под группировочной решающей функцией будем понимать отображение f: $s \rightarrow \{1,2\ldots,K\}$.

Существует большое многообразие алгоритмов группировки. Основные подходы, на которых базируются эти алгоритмы, представлены в параграфе 3.

4. *Представление результатов.* Требуется получить простое и информативное описание полученных кластеров. Часто для такого описания выбирается «типичный объект» или определяется набор усредненных по группе показателей. Используется

также описание в виде набора таксонов. Под таксоном будем понимать подобласть пространства переменных минимального объема, имеющую некоторую заданную форму и содержащую точки соответствующей группы.

5. *Определение качества полученной группировки.* Специалисту прикладной области необходимо удостовериться в том, что сформированные группы действительно отражают внутренние закономерности, характерные для решаемой задачи, способствуют достижению целей анализа, помогают открыть новые свойства изучаемых объектов. Существуют также более формальные способы проверки качества, связанные с нахождением вероятности случайного образования групп, которую можно вычислить в рамках той или иной модели распределения (с проверкой статистических гипотез об однородности наблюдений различных классов); с бутстрэп-методом; с вычислением различных показателей качества (внутригруппового разброса, индекса Гудмана-Крускаля; Ранда; С-индекса и т.д.) [25].

3. Основные подходы, используемые в кластерном анализе

В настоящее время существует несколько подходов к решению задачи кластерного анализа, которые основаны на различных представлениях о задаче, использовании специфичной для каждой предметной области дополнительной информации и т.д. Кратко перечислим наиболее часто используемые подходы. Заметим, что описанная ниже классификация не является четкой; некоторые методы могут быть разработаны на основе комбинации различных подходов.

- *Вероятностный подход.* Предполагается, что каждый объект генеральной совокупности принадлежит одному из K классов, однако номера классов непосредственно ненаблюдаемы. Объекты выбираются из генеральной совокупности случайно и независимо, поэтому переменные, описывающие объекты, случайны. Для каждого класса определено вероятностное распределение заданного семейства; параметры распределения неизвестны. Имеющаяся выборка наблюдений представляет собой реализацию смеси распределений. Необходимо определить наиболее правдоподобные значения параметров, восстановив закон распределения для каждого класса. Существуют приближенные алгоритмы расщепления смеси (EM-алгоритм) [1]. Проверка значимости разделения может быть проведена с использованием аппарата проверки статистических гипотез. Известны также алгоритмы, основанные на непараметрических оценках плотности.

- Подход, использующий *аналогию с центром тяжести*. Для каждой группы определяется вектор средних значений показателей, интерпретируемый как «центр тяжести» группы. Используется критерий внутригруппового рассеяния: $d(G)=\sum_{k=1}^{K}\sum_{l=1}^{N_k}\sum_{j=1}^{n}(x_j^{(i_l)}-g_{kj})^2$, где $g_{kj}=\frac{1}{N_k}\sum_{l=1}^{N_k}x_j^{(i_l)}$ – координата «центра тяжести» k-го кластера по переменной X_j, j=1,2,…,n, k=1,2,…,K. Оптимальная группировка, при заданном K, соответствует минимальному значению критерия.

В алгоритме K-средних [2] группировочное решение формируется динамически из некоторой исходной группировки путем поэтапного перераспределения объектов в группы с ближайшими центрами тяжести. Это перераспределение идет до получения

устойчивого разделения. Аналогичная методика используется также в алгоритме FOREL [3].

- Подход, основанный на *теории графов*. Наиболее известный алгоритм этого семейства – алгоритм кратчайшего незамкнутого пути [29]. Предварительно строится минимальное остовное дерево графа, в котором вершины соответствуют объектам, а ребра имеют длину, равную расстоянию между соответствующими объектами. Для образования кластеров из построенного дерева удаляются ребра максимальной длины.

- *Иерархический подход.* Данное направление также имеет отношение к теоретико-графовому подходу. Результаты группировки представляются в виде дерева группировки (дендрограммы). Алгоритмы, основанные на этом подходе, можно разделить на агломеративные (поэтапно объединяющие ближайшие группы или объекты) и дивизимные (в которых поэтапно осуществляется разделение исходной группы на наиболее удаленные подгруппы; те в свою очередь также разделяются на подгруппы и т.д.). Группировочные решения представляют собой вложенную иерархию подгрупп.

- Подход, основанный на понятии *ближайшего соседа*. Группировка осуществляется последовательно путем приписывания объекта кластеру, в котором находится ближайший объект, при условии, что расстояние до объекта не превышает заданный порог. Существуют различные варианты определения расстояния; при определении меры близости может учитываться и расположение других соседних точек [14].

- *Нечеткие алгоритмы* кластерного анализа. При использовании данного подхода предполагается, что каждый кластер представляет собой нечеткое множество объектов. К наиболее популярным алгоритмам этого семейства можно отнести алгоритм нечетких *C*-средних [8].

- Подход, использующий *искусственные нейронные сети,* основан на аналогии с процессами, происходящими в биологических нейронных системах. Известно большое число алгоритмов данного семейства [16]. Типичная архитектура представляет собой однослойную сеть, в которой каждый нейрон соответствует некоторому кластеру. В процессе обучения сети происходит итеративное изменение передаточных весов между входными и выходными узлами сети; тем самым осуществляется поиск оптимального

значения критерия группировки. Нейронные сети позволяют эффективно использовать параллельные методы вычислений. Привлекательным свойством самоорганизующихся сетей Кохонена [21] является то, что они формируют наглядное двумерное отображение множества объектов. Существует определенное подобие в процессе группировки между алгоритмами, основанными на нейронных сетях и некоторыми классическими методами кластерного анализа.

Эволюционный (генетический) подход. Алгоритмы данного семейства построены на аналогии с природной эволюцией. В них используются понятия популяции – набора различных вариантов группировки (называемых также хромосомами, по аналогии с соответствующими биологическими объектами), и эволюционных операторов – процедур, позволяющих из одной или нескольких родительских хромосом получить одну или несколько хромосом-потомков. Этими процедурами являются: селекция, рекомбинация и мутация. Генетический алгоритм [13] способен осуществлять поиск решения, доставляющего глобальный минимум критерию качества группировки. Опишем подробнее основные шаги алгоритма.

1) Формируется случайная популяция группировочных решений (рис. 1). Каждый вариант группировки представляется в виде последовательности целых чисел длины N, кодирующих номера кластеров. Для каждой последовательности («хромосомы») определяется значение критерия качества.

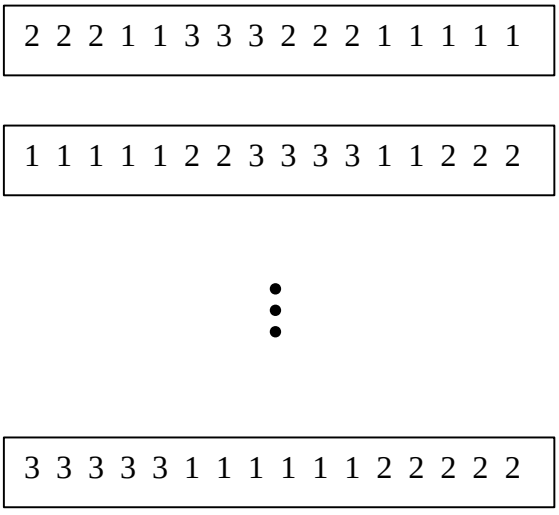

Рис. 1. Популяция хромосом; число классов K=3, число объектов N=16

2) Генерируется следующая популяция путем использования эволюционных операторов. Оператор селекции служит для случайного отбора «родительских» хромосом, наилучших с точки зрения критерия качества. Рекомбинация служит для образования из отобранных хромосом новых последовательностей. Наиболее известным является оператор рекомбинации «кроссовер». Этот оператор для каждой пары родительских хромосом образует пару хромосом-потомков с помощью перестановки одного или нескольких сегментов (рис. 2).

Оператор мутации случайным образом изменяет последовательность в соответствии с определенным правилом (например, заменяет в случайно выбранном элементе один номер группы другим). Для новой популяции вычисляются соответствующие значения критерия.

3) Шаг 2 повторяется, пока не будет выполняться заданное условие остановки.

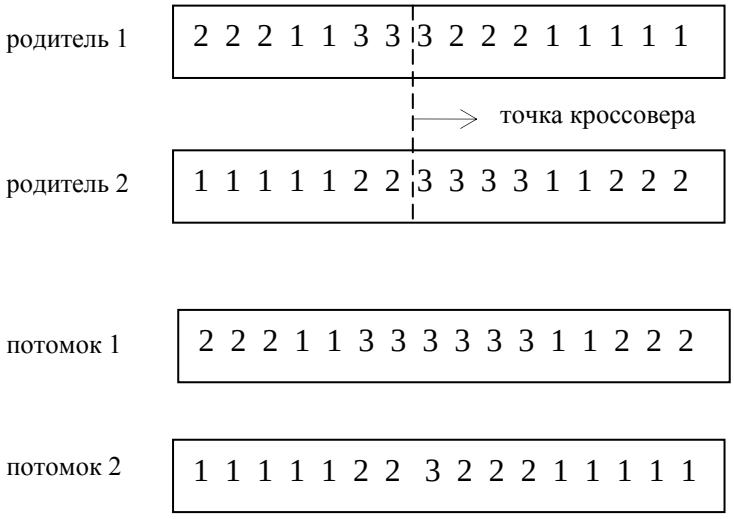

Рис. 2. Оператор кроссовера

Описанный оператор рекомбинации обладает рядом недостатков, к числу которых относят *недействительность* группировочных решений и *нечувствительность* к контексту. Недействительность решений возникает, когда образуются потомки с меньшим числом кластеров. Например, после применения

оператора кроссовера к последовательностям (2 2 1 1 3 3) и (3 3 1 2 1 2) в точке между вторым и третьим элементами, возникают две новые последовательности (3 3 1 1 3 3) и (2 2 1 2 1 2), у которых только два кластера. Нечувствительность к контексту проявляется, когда одно и то же группировочное решение кодируется разными последовательностями. Например, последовательности (1 1 1 2 2 2) и (2 2 2 1 1 1) представляют одно и то же разбиение объектов на группы. При этом потомки этих последовательностей (1 1 1 1 1 1) и (2 2 2 2 2 2) значительно отличаются от оригиналов.

Указанные недостатки приводят к значительному ухудшению качества группировки. Один из возможных способов решения проблемы, предложенный в работе [15], описан в параграфе 5.

Существуют и другие подходы к решению задачи глобальной оптимизации, которые могут применяться в кластерном анализе. Так, например, известен подход *имитации отжига* [19], также базирующийся на идее моделирования природных процессов. Отметим также работу [5], в которой предлагался метод случайного поиска с адаптацией, явившийся одним из первых в семействе эволюционных алгоритмов.

4. Актуальные проблемы кластерного анализа

Несмотря на большое число исследований в области кластерного анализа, в этой области существует ряд актуальных проблем. Перечислим основные проблемы.

1. Проблема обоснования качества результатов анализа. Известно, что процесс группировки в значительной степени носит субъективный характер. Это выражается, в частности, в том, что один и тот же набор объектов может классифицироваться по-разному в зависимости от прикладной области, степени полноты знаний об объектах изучения и т.д. Поэтому необходимо разрабатывать методы, позволяющие максимально полно учитывать имеющиеся экспертные знания, а также разрабатывать соответствующие критерии качества группировки.

2. Многие трудноформализуемые области исследований характеризуются недостаточностью знаний об изучаемых объектах, что затрудняет формулировку их математических моделей. В этих условиях, в частности, проблематичным становится применение алгоритмов расщепления смеси распределений (например, EM-алгоритма [1]), базирующихся на представлении о том, что каждый класс описывается некоторым известным (с точностью до параметров) распределением в пространстве переменных.

3. Проблема анализа большого числа разнотипных (количественных или качественных) факторов. В случае разнотипного пространства, возникает методологическая проблема определения в нем метрики (некоторые способы введения такого рода метрик изложены в работе [10]). С другой стороны, даже в пространстве однотипных (количественных) переменных при увеличении их числа усиливается «проклятие размерности», что может привести к почти полной неразличимости точек. Так, расстояние от любой точки до ее «ближайшего соседа» для некоторых видов расстояний может практически совпадать (с учетом машинной точности) с расстоянием до ее «дальнего соседа». Зрительные аналогии, уместные в пространстве малой размерности, становятся совершенно неприемлемыми в пространстве большой размерности. Например, в 20-мерном евклидовом пространстве объем гиперкуба превышает объем вписанной в него гиперсферы более чем в 40 000 000 раз, что кажется странным с точки зрения двух- или трехмерных пространств.

4. Нелинейность взаимосвязей; наличие пропусков, погрешностей измерения переменных. Классические методы снижения размерности (метод главных компонент; метод независимых компонент), используемые в кластерном анализе, в основном ориентированы на линейные зависимости между переменными. Для выявления более сложных взаимосвязей требуются такие алгоритмы, как нелинейные (ядерные) методы главных компонент [27] и т.п.

5. Необходимость представления результатов анализа в форме, понятной специалистам прикладной области. Помимо хорошей прогнозирующей способности для любого алгоритма анализа данных важно, насколько понятными и интерпретируемыми являются его результаты. Для улучшения интерпретируемости решений можно использовать логические модели [6,24]. Такого рода модели используются для решения задач распознавания образов и прогнозирования количественных показателей, например, в методах построения решающих деревьев или логических решающих функций.

6. Проблема поиска глобального экстремума у критерия качества группировки. Критерий качества, как правило, является функцией, зависящей от большого числа факторов, нелинейным, обладающим множеством локальных экстремумов. Для нахождения кластеров необходимо решить сложную комбинаторную задачу поиска оптимального варианта классификации. Как известно, число различных вариантов разбиения N объектов на K групп равно [23]

$$M(N,K)=\frac{1}{K!}\sum_{i=1}^{K}(-1)^i\binom{K}{i}(K-i)^N.$$

Поэтому алгоритм полного перебора вариантов имеет трудоемкость, экспоненциально зависящую от размерности. Если число групп заранее неизвестно, то переборная задача становится еще сложнее. Таким образом, при увеличении размерности таблиц данных происходит «комбинаторный взрыв». Классические алгоритмы кластерного анализа осуществляют направленный поиск в сравнительно небольшом подмножестве пространства решений, используя различного рода априорные ограничения (на число кластеров или их форму, на порядок включения объектов в группы и т.д.). При этом нахождение строго-оптимального решения не гарантируется. Для поиска оптимального решения применяются более сложные методы, такие как генетические (эволюционные)

алгоритмы [13], нейронные сети [21] и т.д. Существуют экспериментальные исследования, подтверждающие преимущества таких алгоритмов перед классическими алгоритмами [20]. Однако и при использовании эволюционных методов возникают проблемы [22], связанные со спецификой решаемой задачи кластер-анализа: с трудностью интерпретации используемых операторов рекомбинации и кроссовера, о чем говорилось в параграфе 3 .

7. Проблема устойчивости группировочных решений. В классических алгоритмах решения задач кластер-анализа (например, алгоритме К-средних) результаты группировки могут сильно меняться в зависимости от выбора начальных условий, порядка объектов, параметров работы алгоритмов и т.п. В последнее время различными авторами [11,12,28] предлагаются способы повышения устойчивости группировочных решений, основанные на применении ансамблей алгоритмов. При этом используются результаты группировки, полученные различными алгоритмами, или одним алгоритмом, но с разными параметрами настройки, по различным подсистемам переменных и т.д. После построения ансамбля проводится нахождение итогового коллективного решения. Такие способы описаны в следующем разделе.

5. Группировка на основе ансамбля алгоритмов

Идея построения коллективных решений, основанных на комбинации простых алгоритмов, активно используется в современной теории и практике интеллектуального анализа данных, распознавания образов и прогнозирования. Так, широко известны алгоритмы оценок, предложенные Ю.И. Журавлевым [4], алгоритмы бэггинга [9], бустинга [26] и т.д.

Пусть построен набор группировочных решений $\mathbf{G} = \{G^{(1)}, G^{(2)}, ..., G^{(L)}\}$, где $G^{(i)}$ – *i*-й вариант группировки, содержащий $K^{(i)}$ кластеров. Соответствующий набор группировочных решающих функций обозначим через $\mathbf{f}=\{f^{(1)}, f^{(2)}, ..., f^{(L)}\}$. Согласующей функцией назовем отображение $\mathbf{f} \to g$, где g – некоторая группировочная решающая функция.

Для выбора наилучшей согласующей функции могут быть использованы различные принципы. Так, в работе [28] предлагается принцип максимизации количества взаимной информации, которую разделяет итоговая группировка с исходными группировочными решениями. При этом каждой паре группировочных решающих функций $f^{(a)}$ и $f^{(b)}$ сопоставляются случайные переменные *X* и *Y*, которые принимают значения из множеств $\{1,\ldots,K^{(a)}\}$ и $\{1,\ldots,K^{(b)}\}$ соответственно. Для этих переменных определяется нормированное количество взаимной информации

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}},$$

где *I(X,Y)* – количество взаимной информации между *X* и *Y*, *H*(*X*) и *H*(*Y*) – энтропия *X* и *Y* соответственно. Выборочной оценкой значения *NMI* служит величина

$$\phi^{(NMI)}(f^{(a)}, f^{(b)}) = \frac{\sum_{h,l} n_{h,l} \log\left(\frac{N \cdot n_{h,l}}{n_h^{(a)} n_l^{(b)}}\right)}{\sqrt{\left(\sum_h n_h^{(a)} \log\frac{n_h^{(a)}}{N}\right)\left(\left(\sum_l n_l^{(b)} \log\frac{n_l^{(b)}}{N}\right)\right)}},$$

где $n_h^{(a)}$ обозначает число объектов в *h*-м кластере *a*-го варианта группировки, $n_l^{(b)}$ – число объектов в *l*-м кластере *b*-го варианта группировки, $n_{h,l}$ есть число объектов, которые одновременно принадлежат *h*-му кластеру в *a*-й группировке и *l*-му кластеру в *b*-й группировке.

Для набора группировочных решающих функций **f** и некоторой группировочной решающей функции *g* можно определить среднее количество взаимной информации

$$\phi^{(ANMI)}(g,\mathbf{f}) = \frac{1}{L}\sum_i \phi^{(NMI)}(g, f^{(i)}). \qquad (1)$$

Для поиска оптимального варианта согласования в [28] предлагается направленная процедура, которая, начиная с некоторого исходного варианта группировки, поочередно для каждого объекта вычисляет номер группы, при отнесении к которой значение критерия (1) увеличивается в наибольшей степени.

Разработаны и другие принципы согласования. В ряде работ используется принцип, основанный на нахождении согласованной матрицы подобия объектов. Введем для *i*-й группировки бинарную матрицу подобия $S^{(i)} = \left\{S^{(i)}(j,m)\right\}$ размерности $N \times N$ следующим образом: $S^{(i)}(j,m) = 1$ если $o^{(j)}$ и $o^{(m)}$ принадлежат одному кластеру; $S^{(i)}(j,m) = 0$ иначе, где $j,m=1,2,\ldots,N$; $i=1,2,\ldots,L$. Сформируем согласованную матрицу подобия $S = \{S(j,m)\}$,

$$S(j,m) = \frac{1}{L}\sum_{i=1}^{L} S^{(i)}(j,m),$$

где $j,m=1,2,\ldots,N$. Величина $S(j,m)$ равна частоте классификации объектов $o^{(j)}$ и $o^{(m)}$ в одну и ту же группу в наборе группировок **G**. Близкое к единице значение величины $S(j,m)$ означает, что данные объекты имеют большой шанс попадания в одну и ту же группу. Близкое к нулю значение этой величины говорит о том, что шанс оказаться в одной группе у этих объектов незначителен.

После вычисления согласованной матрицы подобия, для нахождения итогового варианта группировки можно применять алгоритмы, которые в качестве входной

информации используют расстояния между объектами (например, иерархические методы построения дендрограммы, или теоретико-графовые методы).

В работе [15] предлагается оператор рекомбинации, основанный на полученной матрице подобия. При этом эволюционный алгоритм группировки состоит из следующих основных шагов.

1. Генерируются M группировочных решений; при этом каждый вариант группировки получается в результате случайного отбора переменных;

2. Выбирается популяция из L наилучших по критерию качества группировки решений; для них вычисляется матрица подобия S;

3. Генерируется случайная матрица подобия $S^{(new)}$ с элементами $S^{(new)}(j,m)=1$ если $rand(1)<S(j,m)$; $S^{(new)}(j,m)=0$ иначе, где через $rand(1)$ обозначено случайное число в интервале [0;1], $j,m=1,2,\ldots,N$. С помощью агломеративного алгоритма группировки (с использованием расстояния, вычисляемого по принципу «средней связи») формируется новое группировочное решение $G^{(new)}$.

4. Выбирается элемент популяции $G^{(worst)}$ с наихудшим значением критерия качества. Если критерий для $G^{(new)}$ окажется лучше, чем для $G^{(worst)}$, то $G^{(worst)}$ заменяется на $G^{(new)}$.

5. Элементы популяции подвергаются изменениям с помощью оператора мутации.

6. Процесс повторяется, начиная с шага 2, до тех пор, пока не будет выполнено заданное условие остановки.

После завершения эволюционных изменений, из популяции выбирается итоговый наилучший вариант группировки. К достоинствам предложенного алгоритма авторы относят то, что он дает возможность обрабатывать массивы данных в пространстве переменных большой размерности (каждое группировочное решение может строиться в пространстве переменных, число которых относительно невелико); позволяет избежать возникновения "недействительных" группировочных решений, "нечувствительных" к контексту, что является проблемой для традиционных операторов рекомбинации. В качестве иллюстрации преимущества ансамблевого метода, в работе [Hong] приводится экспериментально полученный график (рис. 3) зависимости качества группировки (оцениваемый с помощью индекса Ранда [25]) от числа итераций для трех различных генетических алгоритмов: использующего

оператор одноточечного кроссовера (нижний график); использующего только оператор мутации, т.е. без рекомбинации (средний график); и основанного на ансамблевом методе (верхний график). Исходные данные представляют собой реализацию смеси, состоящей из пяти нормальных распределений в восьмимерном пространстве. Объем выборки равен 100.

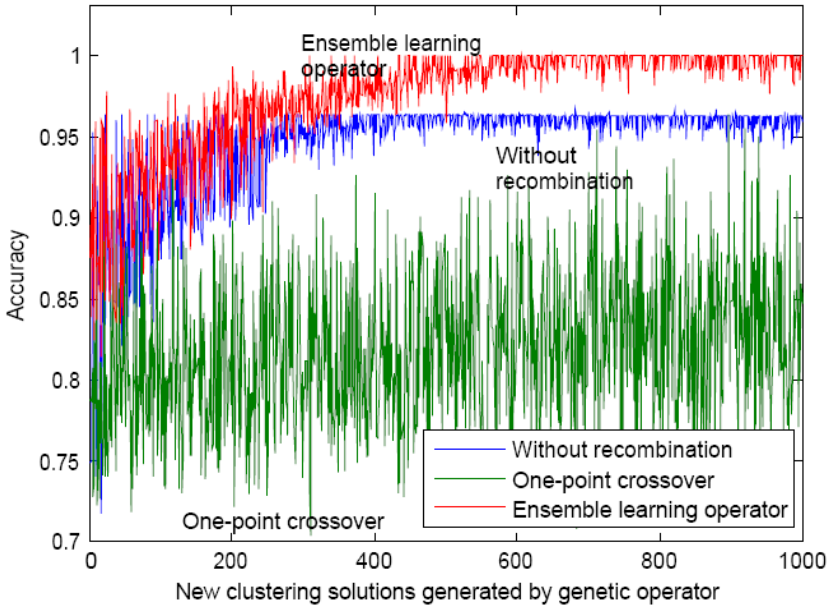

Рис.3. Зависимость качества группировки от числа итераций для трех различных генетических алгоритмов кластерного анализа

Помимо принципа согласования, использующего матрицу подобия, существуют и другие направления. В работе [28] описывается подход, основанный на представлении набора группировочных решений в виде гиперграфа. При этом вершинам гиперграфа соответствуют исходные объекты. В отличие от обычного графа, в котором ребру соответствует пара вершин, гиперребру может соответствовать любое подмножество вершин. Таким образом, каждому кластеру в каждом варианте группировки соответствует некоторое гиперребро. Задача заключается в том, чтобы найти разбиение гиперграфа по минимальному числу гиперребер; при этом полученные части должны сбалансированы, т.е. иметь примерно одинаковое число вершин. В

теории графов существуют специальные методы разбиения гиперграфов, например, с помощью алгоритма HMETIS [18].

В работе [28] приведен еще один подход, также использующий теорию графов. Подход основан на нахождении групп кластеров (метакластеров) с последующей их декомпозицией. Для нахождения расстояния между кластерами предлагается использовать индекс Джаккарда, представляющий собой отношение меры пересечения между двумя группами объектов к мере их объединения. Строится соответствующий метаграф, в котором каждой вершине соответствует определенный кластер из набора вариантов группировки. Для разбиения метаграфа применяется алгоритм HMETIS. Декомпозиция найденных метакластеров осуществляется путем определения для каждого объекта степени вхождения в метакластеры и последующем приписывании объекту номера метакластера с наибольшей степенью вхождения. При этом не гарантируется, что каждый метакластер будет содержать хотя бы один объект; поэтому в итоговой группировке может оказаться меньше K кластеров.

В работе [11] развивается идея о целесообразности включения в ансамбль, по которому определяется коллективное решение, не всех имеющихся вариантов группировки, а только тех из них, которые обладают достаточно хорошим "качеством" и "разнообразием". Для вычисления этих показателей предлагается использовать меру количества взаимной информации.

6. Логические модели в кластерном анализе

Логические модели широко используются для решения задач распознавания и прогнозирования. Это объясняется хорошей интерпретируемостью моделей, имеющих вид логических закономерностей, высокой прогнозирующей способностью, возможностью обрабатывать разнотипные переменные, выделять наиболее важные факторы. Логическую модель можно строить после группировки объектов некоторым алгоритмом, то есть решать задачу распознавания образов в классе логических решающих функций, где под образами понимаются номера кластеров, приписанные объектам. Существуют и алгоритмы, в которых группировка осуществляется непосредственно в ходе построения логической модели.

Рассмотрим алгоритм кластерного анализа, основанный на построении дерева решений [6,7]. Деревом, как известно, называется связный неориентированный граф, не содержащий циклов. Дерево называется корневым, если в нем выделена произвольная вершина, называемая корнем. В дальнейшем, говоря "дерево" будем подразумевать корневое дерево. В дереве решений каждой внутренней вершине (узлу) соответствует некоторая переменная X_j, а ветвям, выходящим из данной вершины соответствует истинность определенного высказывания вида $X_j(o) \in E_j^{(i)}$ где o – некоторый объект, $i=1,\ldots,l$, $l \geq 2$ – число ветвей, выходящих из данной вершины, причем набор $E_j^{(1)},\ldots,E_j^{(l)}$ есть разбиение множества значений D_j. Каждому m-му листу (концевой вершине) дерева приписывается решение (номер соответствующего кластера) $m=1,\ldots,M$, где M – число листьев дерева, $M=K$.

Цепочке ветвей дерева, ведущих из корневой вершины в m-й лист, можно сопоставить логическое утверждение вида

$$J^{(m)}=\text{"Если } X_{j_1}(o) \in E_{j_1}^{(i_1)} \text{ И } X_{j_2}(o) \in E_{j_2}^{(i_2)} \text{ И } \ldots \text{ И } X_{j_q}(o) \in E_{j_{q_m}}^{(i_{q_m})},$$

$$\text{то объект } o \text{ относится к } m\text{-му кластеру"},$$

где q_m – длина данной цепочки. В узлах дерева использован самый простой вид предиката. При увеличении сложности предиката (например, при проверке условия относительно линейной комбинации переменных) увеличивается сложность класса

разбиений пространства переменных. Однако в дальнейшем такая возможность не используется, так как лишь в случае, когда решающая функция задана в виде набора конъюнкций простых предикатов, результаты анализа представляются на языке, близком к естественному языку логических суждений.

Рассмотрим дерево решений с M листьями. Этому дереву соответствует разбиение пространства переменных на M попарно непересекающихся подобластей $E^{(1)},\ldots,E^{(M)}$, так что каждому m-му листу сопоставляется подобласть $E^{(m)}$. Разбиению пространства переменных, в свою очередь, соответствует разбиение выборки на подмножества $C^{(1)},C^{(2)},\ldots,C^{(M)}$. Рассмотрим произвольную группу объектов $C^{(m)}$. Описанием этой группы назовем следующую конъюнкцию высказываний:

$$H(C^{(m)})=\text{"}\, X_1 \in T_1^{(m)} \text{ И } \ldots \text{ И } X_j \in T_j^{(m)} \text{ И } \ldots \; X_n \in T_n^{(m)}\text{"},$$

где $T_j^{(m)}$ – отрезок $[\min\limits_{o\in C^{(m)}} X_j(o);\ \max\limits_{o\in C^{(m)}} X_j(o)]$ в случае количественной или порядковой переменной X_j, либо множество принимаемых значений $\{X_j(o)\,|\,o\in C^{(m)}\}$ в случае качественной переменной. Подобласть пространства переменных $T^{(m)}=T_1^{(m)}\times\ldots\times T_n^{(m)}$, соответствующую описанию группы, назовем таксоном.

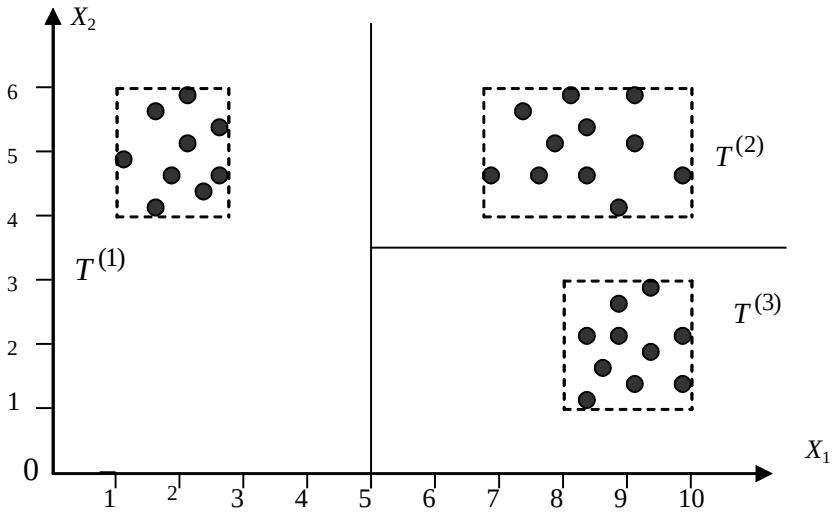

Рис. 4

В примере на рис. 4 плоскость разбита с помощью дерева на три подобласти. Описанием соответствующих групп объектов является набор высказываний: $H(C^{(1)})$="$X_1 \in [1,3]$ И $X_2 \in [4,6]$"; $H(C^{(2)})$="$X_1 \in [7,10]$ И $X_2 \in [4,6]$"; $H(C^{(3)})$="$X_1 \in [8,10]$ И $X_2 \in [1,3]$".

Важно отметить, что хотя и в дереве решений может использоваться лишь часть переменных, в описании каждой группы участвуют обязательно все имеющиеся переменные.

Относительной мощностью (объемом) таксона назовем величину $\delta^{(m)} = \prod_{j=1}^{n} \frac{|T_j^{(m)}|}{|D_j|}$, где через $|T_j^{(m)}|$ обозначена длина интервала (в случае количественной или порядковой переменной) либо мощность (число значений) соответствующего подмножества в случае качественной переменной; $|D_j|$ - длина интервала между минимальным и максимальным значением переменной X_j для исходной выборки объектов (для количественной или порядковой переменной) либо общее число значений этой переменной (для качественной переменной).

Под критерием качества группировки, при заданном числе кластеров, будем понимать величину суммарного относительного объема таксонов: $q = \sum_{m=1}^{M} \delta^{(m)}$. Оптимальной группировкой будем считать группировку, для которой значение данного критерия минимально. Заметим, что в случае, когда все переменные количественные, минимизация критерия означает минимизацию суммарного объема многомерных параллелепипедов, «охватывающих» группы.

Если же число кластеров заранее не задано, под критерием качества можно понимать величину $Q = q + \alpha M$, где α>0 – некоторый заданный параметр, подбираемый экспериментально. При минимизации этого критерия, с одной стороны, мы получаем таксоны минимального объема, а с другой стороны, стремимся уменьшить число этих таксонов.

Для построения дерева могут использоваться описанные в работе [6] метод последовательного ветвления LRP или рекурсивный R-метод [7]. На каждом шаге

алгоритма LRP некоторая группа объектов, соответствующая висячей вершине дерева, разделяется на две новых подгруппы. Разделение происходит с учетом критерия качества группировки, т.е. минимизируется суммарный объем полученных таксонов. Перспективной для дальнейшего ветвления считается вершина, для которой относительный объем соответствующего таксона больше, чем заданный параметр. Разделение продолжается до тех пор, пока не останется более перспективных вершин, либо не будет получено заданное число групп.

В случае сложной зависимости между переменными, метод последовательного ветвления, как правило, не позволяет достичь удовлетворительного решения задачи. Можно привести примеры, из которых видно, что для выявления структуры разбиения при построении дерева решений необходимо учитывать одновременно несколько переменных, что невозможно при последовательном ветвлении. В этом случае целесообразно применять рекурсивный метод. Для этого метода используется второй вариант критерия качества группировки Q, для которого число групп заранее не задано. Суть предлагаемого метода заключается в следующем. Строится "начальное" дерево с корневой вершиной B и максимально возможным числом дочерних вершин, для которого затем рекурсивным образом строятся (локально) оптимальные по заданному критерию поддеревья. Затем происходит последовательное объединение тех дочерних для B вершин, которые при объединении и рекурсивном построении соответствующего (локально) оптимального поддерева дают наилучшее значение критерия. Максимальная глубина рекурсивной вложенности задается параметром R. Путем увеличения параметра R можно увеличивать глубину перебора вариантов, что позволяет учитывать более сложные зависимости между переменными (при этом увеличивается время работы и требуемый объем памяти). Показано, что алгоритм обладает полиномиальной трудоемкостью. Отличительной чертой алгоритма является то, что заранее не фиксируется число ветвей, выходящих из каждой вершины, а ищется их оптимальное число. Кроме того, для алгоритма характерно, что при построении «начального» дерева образуются таксоны небольшого объема, которые затем «сливаются» в один или несколько более объемных таксонов так, чтобы улучшить критерий качества группировки.

Представляется перспективной задачей разработка алгоритмов кластерного анализа, основанных на ансамбле логических моделей. Использование таких моделей позволит не только группировать разнотипные данные, но и давать объяснение результатов анализа на языке логических высказываний. Кроме того, это даст возможность ранжировать переменные по их значимости для группировки (переменные, относительно которых формируются логические высказывания, получают более высокую оценку значимости).

Заключение

Итак, несмотря на достаточно большое число разработанных моделей и алгоритмов кластерного анализа, при решении прикладных задач исследователи часто сталкиваются с рядом проблем, к числу которых относятся:

- трудность в обосновании качества результатов анализа, учитывающего специфику конкретной задачи;
- формулировка вероятностных моделей исследуемых объектов, особенно в случае малого объема выборки;
- необходимость обработки большого числа разнотипных (количественных или качественных) факторов;
- нелинейность взаимосвязей; наличие пропусков, погрешностей измерения переменных;
- необходимость представления результатов анализа в форме, удобной и понятной специалистам прикладной области;
- проблема поиска глобального экстремума у критерия качества группировки;
- неустойчивость группировочных решений при небольших изменениях выборки или параметров работы алгоритма.

Отсюда можно сделать вывод о целесообразности дальнейшего развития методов кластерного анализа, позволяющих решать указанные проблемы. К числу перспективных тенденций относятся разработка ансамблевых методов группировки, в том числе для логических моделей.

Данная работа проведена при поддержке Российского фонда фундаментальных исследований (грант № 08-07-00136а).

Литература

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности - М.: Финансы и статистика, 1989. - 450 с.
2. Дуда Р., Харт П. Распознавание образов и анализ сцен - М.: Мир, 1976. - 559 с.
3. Загоруйко Н.Г. Прикладные методы анализа данных и знаний - Новосибирск: Изд. Института математики, 1999. - 270 с.
4. Журавлёв Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения - М.: ФАЗИС, 2006.
5. Лбов Г. С. Выбор эффективной системы зависимых признаков // Вычислительные системы, 1965. Вып. 19. С 21-34.
6. Лбов Г.С., Старцева Н.Г. Логические решающие функции и вопросы статистической устойчивости решений - Новосибирск: Изд-во Ин-та математики, 1999. - 212 с.
7. Лбов Г.С., Бериков В.Б. Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации - Новосибирск: Изд-во Ин-та математики, 2005. - 218 с.
8. Bezdek J. C. Pattern Recognition With Fuzzy Objective Function Algorithms – NY: Plenum Press, 1981.
9. Breiman L. Bagging predictors // Machine Learning, 1996. V. 24. P. 123-140.
10. Diday, E., Simon, J. C. Clustering analysis // In: Digital Pattern Recognition, K. S. Fu, Ed. Springer-Verlag, Secaucus, NJ. P. 47–94.
11. Fern, X.Z., Brodley, C.E. Clustering ensembles for high dimensional data clustering // In Proc. International Conference on Machine Learning, 2003. P.186-193.
12. Fred, A., Jain, A.K. Combining multiple clusterings using evidence accumulation // IEEE Tran. on Pattern Analysis and Machine Intelligence, 2005. V. 27. P. 835-850.
13. Goldberg D. E. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.
14. Gowda, K. C., Krishna, G. Agglomerative clustering using the concept of mutual nearest neighborhood // Pattern Recognition, 1977. V. 10. P. 105–112.

15. Hong Y., Kwong S. To combine steady-state genetic algorithm and ensemble learning for data clustering // Pattern Recognition Letters, 2008. V. 29 (9). P. 1416-1423.
16. Jain, A. K., Mao Artificial neural networks: A tutorial // IEEE Computer, 1996. V. 29. P. 31–44.
17. Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review // ACM Computing Surveys, 1999. V.31, N 3. P. 264-323.
18. Karypis G., Aggarwal R., Kumar V., Shekhar S. Multilevel hypergraph partitioning: Applications in VLSI domain // Proceedings of the Design and Automation Conference, 1997.
19. Kirkpatrick S., Gelatt C. D., Vecchi M. P. Optimization by simulated annealing // Science, 1983. V. 220(4598). P. 671–680.
20. K. Krishna, M. Murty. Genetic K-means algorithm // IEEE Transaction on System, Man and Cybernetics- Part B, 1999. V.29. P. 433-439.
21. Kohonen T. Self-Organization and Associative Memory - 3rd ed. Springer information sciences series. Springer-Verlag, New York, NY. 1989.
22. Lu Y., Li S., Fotouhi F., Deng Y., Brown S. Incremental genetic k-means algorithm and its application in gene expression data analysis // BMC Bioinformatics, 2004.
23. Lui G. L. Introduction to the combinatorial mathematics - McGraw Hill, 1968.
24. Michalski R., Stepp R., Diday E. Automated construction of classifications: conceptual clustering versus numerical taxonomy // IEEE Trans. Pattern Anal. Mach. Intell. PAMI-5, 1983. V.5. P. 396–409.
25. Rand W. Objective criteria for the evaluation of clustering methods // Journal of American Statistical Association, 1971. V.66. P.846-850.
26. Schapire R. The boosting approach to machine learning: An overview // In MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA, Mar. 2001.
27. Scholkopf B., Smola A., Muller K. Kernel Principal Component Analysis, Advances in Kernel Methods-Support Vector Learning, 1999.
28. Strehl A., Ghosh J. Clustering ensembles - a knowledge reuse framework for combining multiple partitions // The Journal of Machine Learning Research, 2002. V.3. P.583-617.
29. Zahn C. T. Graph-theoretical methods for detecting and describing gestalt clusters // IEEE Trans. Comput., 1971. C-20. P. 68–86.