Построение ансамбля деревьев решений в кластерном анализе*

В.Б. Бериков
Институт математики им. С.Л. Соболева СО РАН, Новосибирск, Россия
e-mail: berikov@math.nsc.ru

Разработан и исследован коллективный алгоритм кластерного анализа. Проведено теоретическое обоснование ансамблевых алгоритмов, основанных на попарной классификации объектов. Предложен алгоритм кластерного анализа, использующий ансамбль деревьев решений. Алгоритм позволяет проводить классификацию в разнотипном пространстве переменных. Проведено исследование алгоритма с помощью статистического моделирования и решения тестовых задач.

Ключевые слова: кластерный анализ, коллектив алгоритмов, логическая модель, дерево решений.

Введение

В задаче кластерного анализа (см., например, [1, 2]) требуется сформировать сравнительно небольшое число групп объектов, которые были бы как можно более схожими между собой внутри каждой группы и как можно более различающимися в разных группах. Известные подходы к решению этой задачи зависят от способа понимания "похожести" и "различия" объектов, разного рода дополнительных предположений и т. д. Так, в вероятностном подходе считается, что наблюдаемые в многомерном пространстве объекты принадлежат различным классам, причем каждый класс характеризуется вероятностным распределением с неизвестными параметрами. Геометрический подход использует аналогии с классификацией, которую проводит исследователь при анализе изображений на плоскости или в трехмерном пространстве. При использовании логического подхода предполагается, что каждый кластер описывается некоторой достаточно простой логической закономерностью.

Одной из актуальных задач кластерного анализа является группировка объектов, описываемых разнотипными (количественными или качественными) факторами. В случае разнотипного пространства возникает методологическая проблема определения в нем метрики.

Другая актуальная проблема — повышение устойчивости группировочных решений. В большинстве алгоритмов кластерного анализа результаты могут сильно меняться в зависимости от выбора начальных условий, порядка объектов, параметров работы алгоритма и т. п.

Наконец, определенную трудность вызывает неоднозначность нумерации кластеров, особенно в случае большого числа классов. Поскольку номера кластеров не играют

*Работа выполнена при финансовой поддержке РФФИ (гранты № 08-07-00136а и 09-07-12087-ofi_m).

роли, удобнее использовать попарную классификацию, т. е. определять, относится ли каждая пара объектов к одному и тому же классу либо к разным классам.

Одним из перспективных подходов к решению задач кластерного анализа в разнотипном пространстве является подход, основанный на логических решающих функциях (логических моделях). Логические модели широко используются для решения задач распознавания и прогнозирования [3 – 5]. Это объясняется хорошей интерпретируемостью результатов, имеющих вид логических закономерностей, высокой прогнозирующей способностью, возможностью обрабатывать разнотипные переменные, выделять наиболее важные факторы. Разработке алгоритмов построения логических моделей кластерного анализа была посвящена, например, работа [6]. Впервые алгоритм кластерного анализа с использованием логических решающих функций был предложен в [7]. В работе [8] был описан метод построения логической функции в задаче кластерного анализа, основанный на рекурсивном алгоритме построения дерева решений. Этот алгоритм позволяет путем увеличения глубины перебора находить более сложные логические закономерности, описывающие структуру кластеров.

Известно, что устойчивость решений в кластерном анализе может быть повышена путем применения ансамблей алгоритмов (см, например, [9, 10]). При этом используются результаты, полученные различными алгоритмами или одним алгоритмом, но с разными параметрами настройки, по различным подсистемам переменных и т. д. После построения ансамбля находится итоговое коллективное решение. Идея построения коллективных решений, основанных на комбинации простых алгоритмов, активно используется в современной теории и практике интеллектуального анализа данных, распознавания образов и прогнозирования (см., например, алгоритмы оценок [11], алгоритмы бэггинга [12], бустинга [13] и др.). Теоретический анализ алгоритмов коллективной классификации (см., например, [12, 14, 15]) показывает, что качество решений, как правило, улучшается при увеличении числа алгоритмов, входящих в ансамбль.

Целью настоящей работы является:

1 — теоретическое обоснование алгоритмов ансамблевого кластерного анализа, основанных на попарной классификации объектов;

2 — описание методики, использующей сочетание логических моделей классификации и ансамблевых алгоритмов;

3 — практическое подтверждение эффективности предложенной методики.

Материал статьи изложен в следующем порядке. В первом разделе даются основные определения и понятия, используемые в работе, проводится теоретическое обоснование эффективности ансамблевых алгоритмов. Во втором разделе описывается алгоритм построения логических решающих функций (деревьев группировочных решений) в кластерном анализе, предлагается методика построения коллективного группировочного решения. В третьем разделе рассматриваются примеры решения модельных и тестовых задач. В заключении приводятся основные выводы работы.

1. Ансамблевый кластерный анализ с использованием попарной классификации

Пусть имеется выборка объектов исследования $s = \{o^{(1)}, \ldots, o^{(N)}\}$, которая сформирована в результате отбора некоторых представителей генеральной совокупности. Требуется сформировать $K \geq 2$ классов (групп объектов); число классов может быть как

задано, так и не задано (в последнем случае оптимальное количество кластеров должно быть определено автоматически).

Каждый объект генеральной совокупности описывается с помощью набора переменных $X = X_1, \ldots, X_n$. Набор X может включать переменные разных типов (количественные и качественные, под которыми будем понимать номинальные и булевы, а также порядковые). Пусть D_j — множество значений переменной X_j. Обозначим через $x = X(o)$ набор наблюдений переменных для объекта o, где $X(o) = (x_1, \ldots, x_n)$, $x_j = X_j(o)$ — значение переменной X_j для данного объекта, $j = 1, \ldots, n$. Соответствующий выборке набор наблюдений переменных будем представлять в виде таблицы данных с N строками и n столбцами.

Предположим, что имеется некоторая скрытая (непосредственно не наблюдаемая) переменная Y, которая задает принадлежность каждого объекта к некоторому из $K \geq 2$ классов. Каждый класс характеризуется определенным законом условного распределения $p(x|Y = k) = p_k(x)$, $k = 1, \ldots, K$. Рассмотрим следующую вероятностную модель генерации данных. Пусть для каждого объекта определяется класс, к которому он относится, в соответствии с априорными вероятностями $P_k = \mathbb{P}(Y = k)$, $k = 1, \ldots, K$, где $\sum\limits_{k=1}^{K} P_k = 1$. Затем в соответствии с распределением $p_k(x)$ определяется значение x. Указанная процедура проводится независимо для каждого объекта.

Пусть с помощью некоторого алгоритма кластерного анализа μ по таблице данных строится разбиение множества объектов s на K подмножеств. Под группировочным решением будем понимать набор $G = \{C^{(1)}, \ldots, C^{(k)}, \ldots, C^{(K)}\}$, где $C^{(k)} = \{o^{(i_1)}, \ldots, o^{(i_{N_k})}\}$, N_k — число объектов, входящих в k-й кластер, $k = 1, \ldots, K$. Группировочной решающей функцией назовем отображение $f : s \to \{1, \ldots, K\}$.

Поскольку нумерация кластеров не играет роли, удобнее рассматривать отношение эквивалентности, т. е. указывать, относит ли алгоритм μ каждую пару объектов в один и тот же класс либо в разные классы. Определим для каждой пары $o^{(i)}$ и $o^{(j)}$ величину

$$h_{\mu,\, o^{(i)},\, o^{(j)}} = \begin{cases} 0, \text{ если объекты отнесены в один класс,} \\ 1, \text{ иначе,} \end{cases} \tag{1}$$

где $i, j = 1, \ldots, N$, $i \neq j$.

Рассмотрим произвольную пару a, b различных объектов выборки. Обозначим соответствующие наблюдения через x_a и x_b.

Пусть $P_Y = \mathbb{P}(Y(a) \neq Y(b))$ — вероятность отнесения объектов к различным классам. Например, при $K = 2$ указанная вероятность равна

$$P_Y = 1 - \mathbb{P}(Y(a) = 1|x_a)\mathbb{P}(Y(b) = 1|x_b) -$$

$$-\mathbb{P}(Y(a) = 2|x_a)\mathbb{P}(Y(b) = 2|x_b) = 1 - \sum_{k=1}^{2} \frac{p_k(x_a)p_k(x_b)P_k^2}{p(x_a)p(x_b)},$$

где $p(x_o) = \sum\limits_{k=1}^{2} p_k(x_o)P_k$, $o = a, b$.

Обозначим вероятность ошибки, которую может совершить алгоритм μ при классификации a и b, через $P_{er,\,\mu}$,

$$P_{er,\,\mu} = \begin{cases} P_Y, \text{ если } h_{\mu,\, a,\, b} = 0, \\ 1 - P_Y, \text{ если } h_{\mu,\, a,\, b} = 1. \end{cases}$$

Легко заметить, что

$$P_{er,\,\mu} = (1 - h_{\mu,\,a,\,b})P_Y + h_{\mu,\,a,\,b}(1 - P_Y) = P_Y + (1 - 2P_Y)h_{\mu,\,a,\,b}. \tag{2}$$

Предположим, что алгоритм μ зависит от случайного вектора параметров $\Theta \in \mathbf{\Theta}$, где $\mathbf{\Theta}$ — некоторое допустимое множество параметров: $\mu = \mu(\Theta)$. Например, в алгоритме k-средних результаты работы зависят от случайного исходного разбиения выборки на K подмножеств. Чтобы подчеркнуть зависимость результатов работы от параметра Θ, введем обозначения $h_{\mu(\Theta),\,a,\,b} = h(\Theta)$, $P_{er,\,\mu(\Theta)} = P_{er}(\Theta)$.

Пусть в результате L-кратного применения алгоритма μ со случайно и независимо отобранными параметрами $\theta_1, \ldots, \theta_L$ получен набор решений $h(\theta_1), \ldots, h(\theta_L)$. Для определенности будем считать, что L — нечетно. Коллективным (ансамблевым) решением по большинству голосов будем называть функцию

$$H(h(\theta_1), \ldots, h(\theta_L)) = \begin{cases} 0, \text{ если } \dfrac{1}{L}\displaystyle\sum_{l=1}^{L} h(\theta_l) < \dfrac{1}{2}, \\ 1, \text{ иначе.} \end{cases}$$

Интересно исследовать поведение коллективного решения в зависимости от мощности ансамбля L. Заметим, что одиночный алгоритм $\mu(\Theta)$ также можно рассматривать как вырожденный случай ансамбля с $L = 1$.

Утверждение 1. Математическое ожидание и дисперсия величины вероятности ошибки для алгоритма $\mu(\Theta)$ равны соответственно

$$\mathbb{E}_\Theta P_{er}(\Theta) = P_Y + (1 - 2P_Y)P_h,$$

$$\mathbf{Var}_\Theta P_{er}(\Theta) = (1 - 2P_Y)^2 P_h(1 - P_h),$$

где $P_h = \mathbb{P}(h(\Theta) = 1)$.

Доказательство. Справедливость выражения для математического ожидания следует из (2) и из того, что $\mathbb{E}_\Theta h(\Theta) = P_h$. Рассмотрим выражение для дисперсии. По определению, $\mathbf{Var}_\Theta P_{er}(\Theta) = \mathbb{E}_\Theta P_{er}^2(\Theta) - (\mathbb{E}_\Theta P_{er}(\Theta))^2$. Далее,

$$\mathbb{E}_\Theta P_{er}^2(\Theta) = \mathbb{E}_\Theta (P_Y + (1 - 2P_Y)h(\Theta))^2 =$$

$$= \mathbb{E}_\Theta (P_Y^2 + 2P_Y(1 - 2P_Y)h(\Theta) + (1 - 2P_Y)^2 h(\Theta)^2).$$

Так как $\mathbf{E}_\Theta h^2(\Theta) = \mathbf{E}_\Theta h(\Theta) = P_h$, то получим

$$\mathbb{E}_\Theta P_{er}^2(\Theta) = P_Y^2 + 2(1 - 2P_Y)P_Y P_h + P_h(1 - 2P_Y)^2 =$$

$$= P_Y^2 + (1 - 2P_Y)P_h(2P_Y + 1 - 2P_Y) = P_Y^2 + (1 - 2P_Y)P_h.$$

Отсюда

$$\mathbf{Var}_\Theta P_{er}(\Theta) = P_Y^2 + (1 - 2P_Y)P_h - (P_Y + (1 - 2P_Y)P_h)^2.$$

После преобразований имеем

$$\mathbf{Var}_\Theta P_{er}(\Theta) = (1 - 2P_Y)^2 P_h(1 - P_h),$$

что и требовалось доказать.

Обозначим через $P_{er}(\Theta_1, \ldots, \Theta_L)$ случайную функцию, значение которой при фиксированных аргументах равно вероятности ошибки, которую может совершать ансамблевый алгоритм при классификации a и b. Здесь $\Theta_1, \ldots, \Theta_L$ — статистические копии случайного вектора Θ. Рассмотрим поведение вероятности ошибки для коллективного решения.

Утверждение 2. Математическое ожидание и дисперсия величины вероятности ошибки для коллективного решения равны соответственно

$$\mathbb{E}_{\Theta_1,\ldots,\Theta_L} P_{er}(\Theta_1, \ldots, \Theta_L) = P_Y + (1 - 2P_Y)P_{H,L},$$

$$\mathbf{Var}_{\Theta_1,\ldots,\Theta_L} P_{er}(\Theta_1, \ldots, \Theta_L) = (1 - 2P_Y)^2 P_{H,L}(1 - P_{H,L}),$$

где $P_{H,L} = \mathbb{P}\left(\frac{1}{L}\sum_{l=1}^{L} h(\theta_l) \geq \frac{1}{2}\right) = \sum_{l=\lceil \frac{L}{2} \rceil + 1}^{L} C_L^l P_h^l (1 - P_h)^{L-l}$, $\lceil \cdot \rceil$ означает целую часть числа.

Доказательство данного утверждения аналогично доказательству утверждения 1 (вероятность ошибки коллективного решения определяется по формуле, аналогичной формуле (2)). Кроме того, ясно, что распределение числа голосов, отданных за решение $h = 1$, является биномиальным $\mathrm{Bin}(L, P_h)$.

Воспользуемся следующей априорной информацией об алгоритме кластерного анализа. Будем считать, что ожидаемая вероятность ошибочной классификации $\mathbb{E}_\Theta P_{er}(\Theta) < \frac{1}{2}$, т. е. ожидается, что алгоритм μ проводит классификацию с лучшим качеством, чем алгоритм случайного равновероятного выбора. Из утверждения 1 следует, что выполняется один из двух вариантов: а) $P_h > \frac{1}{2}$ и $P_Y > \frac{1}{2}$; б) $P_h < \frac{1}{2}$ и $P_Y < \frac{1}{2}$. Рассмотрим, для определенности, первый случай.

Утверждение 3. Если $\mathbb{E}_\Theta P_{er}(\Theta) < \frac{1}{2}$ и при этом $P_h > \frac{1}{2}$ и $P_Y > \frac{1}{2}$, то с увеличением мощности ансамбля ожидаемая вероятность ошибочной классификации уменьшается, стремясь в пределе к величине $1 - P_Y$, а дисперсия величины вероятности ошибки стремится к нулю.

Доказательство. Из интегральной теоремы Муавра—Лапласа следует, что при увеличении L величина

$$P_{H,L} = 1 - \mathbb{P}\left(\frac{1}{L}\sum_{l=1}^{L} h(\Theta_l) < \frac{1}{2}\right)$$

стремится к

$$1 - \Phi\left(\frac{1/2 - P_h}{\sqrt{P_h(1 - P_h)/L}}\right),$$

где $\Phi(\cdot)$ — функция распределения стандартного нормального закона. Значит, при $L \to \infty$ $P_{H,L}$ монотонно увеличивается, стремясь в пределе к 1. Из того что

$$\mathbb{E}_{\Theta_1,\ldots,\Theta_L} P_{er}(\Theta_1, \ldots, \Theta_L) = P_Y + (1 - 2P_Y)P_{H,L},$$

где $(1 - 2P_Y) < 0$, и из утверждения 2 следует справедливость утверждения 3.

Очевидно, что в случае б) ожидаемая вероятность ошибки при увеличении мощности ансамбля также уменьшается, стремясь в пределе к величине P_Y; при этом дисперсия ошибки стремится к нулю.

Доказанное утверждение позволяет сделать вывод о том, что при выполнении определенных вполне естественных условий применение ансамбля позволяет улучшить качество кластеризации.

2. Логические модели в кластерном анализе

Рассмотрим более подробно логический подход к задаче кластерного анализа. Под логической моделью группировки данных будем понимать дерево, в котором 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, $i = 1, \ldots, v$, $v \geq 2$ — число ветвей, выходящих из вершины, причем набор $E_j^{(1)}, \ldots, E_j^{(v)}$ есть разбиение множества значений D_j. Каждому m-му листу (концевой вершине) дерева соответствует группа объектов выборки, удовлетворяющих цепочке высказываний, проверяемых по пути из корневой вершины в этот лист. Данной цепочке можно сопоставить логическое утверждение вида

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

$$\text{то объект } o \text{ относится к } m\text{-й группе,}$$

где q_m — длина данной цепочки. Описанное дерево будем называть группировочным деревом решений.

После группировки объектов некоторым алгоритмом можно строить логическую модель, т. е. решать задачу распознавания образов в классе логических решающих функций, где под образами понимаются номера кластеров, приписанные объектам. Однако алгоритм, в котором группировка осуществляется непосредственно при построении логической модели, позволяет в наилучшей степени отразить логическую структуру данных.

2.1. Построение группировочного дерева решений

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

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

где $T_j^{(m)}$ — отрезок $\left[\min\limits_{o \in C^{(m)}} X_j(o); \max\limits_{o \in C^{(m)}} X_j(o)\right]$ 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 $T^{(m)} = T_1^{(m)} \times \cdots \times T_n^{(m)}$, соответствующую описанию группы, назовем m-м таксоном.

Относительной мощностью (объемом) j-й проекции таксона T назовем величину

$$\delta_j = \frac{|T_j|}{|D_j|},$$

где через $|T_j|$ обозначена длина интервала (в случае количественной или порядковой переменной X_j) либо мощность (число значений) соответствующего подмножества в случае качественной переменной X_j, $j = 1, \ldots, n$. Под объемом таксона будем понимать величину

$$V_T = \prod_{j=1}^{n} \delta_j.$$

Под критерием качества группировки при заданном числе кластеров будем понимать суммарный объем таксонов

$$\Delta = \sum_{m=1}^{M} V_{T^{(m)}}.$$

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 будем понимать величину [5]

$$F = \Delta + \gamma \frac{M}{N},$$

где $\gamma > 0$ — некоторый заданный параметр, подбираемый экспериментально. При минимизации этого критерия, с одной стороны, получаем таксоны минимального объема, с другой — стремимся уменьшить число этих таксонов.

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 данной работе 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.

Для построения дерева могут использоваться описанный в работе [3] метод последовательного ветвления LRP или рекурсивный R-метод [16]. 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 данного метода используется второй вариант критерия качества группировки F, для которого число групп заранее не задано. Суть предлагаемого метода состоит 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;

— 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 R можно увеличивать глубину перебора вариантов, что позволяет учитывать более сложные зависимости между переменными (при этом возрастают время работы и требуемый объем памяти). Показано, что алгоритм обладает полиномиальной трудоемкостью. Отличительная черта алгоритма состоит в том, что заранее число ветвей, выходящих из каждой вершины, не фиксируется, 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 в несколько более объемных таксонов так, чтобы улучшить критерий качества группировки.

2.2. Построение коллективного группировочного решения

Пусть получен набор группировочных решений $\mathbb{G} = \{G^{(1)}, \ldots, G^{(l)}, \ldots, G^{(L)}\}$, где $G^{(l)}$ — l-й вариант группировки, содержащий $K^{(l)}$ кластеров. Каждый l-й вариант формируется в результате применения рекурсивного алгоритма построения группировочного дерева решений в случайном подпространстве переменных (обозначим соответствующий алгоритм через μ_l).

Полученный набор группировочных решающих функций обозначим через $\mathbf{f} = \{f^{(1)}, \ldots, f^{(L)}\}$. 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 [9] предлагается принцип максимизации количества взаимной информации, которая относится к итоговой группировке с учетом исходных группировочных решений. Используем известный принцип, основанный на нахождении согласованной матрицы подобия (или различия) объектов.

Обозначим через $h^{(l)}$ бинарную матрицу $h^{(l)} = \{h^{(l)}(i,j)\}$ размерности $N \times N$, которая вводится для l-й группировки, следующим образом:

$$h^{(l)}(i,j) = h_{\mu_l, o^{(i)}, o^{(j)}},$$

где величина $h_{\mu_l, o^{(i)}, o^{(j)}}$ введена в разделе 1 (см. формулу (1)), $i,j = 1, \ldots, N$ $(i \neq j)$, $l = 1, \ldots, L$. После построения L группировочных решений можно сформировать согласованную матрицу различий $H = \{H(i,j)\}$,

$$H(i,j) = \frac{1}{L} \sum_{l=1}^{L} h^{(l)}(i,j),$$

Величина $H(i,j)$ равна частоте классификации объектов $o^{(i)}$ и $o^{(j)}$ в разные группы в наборе группировок $\mathbb{G}$. Близкое к нулю значение этой величины означает, что данные объекты имеют большой шанс попадания в одну и ту же группу, близкое к единице — указывает на то, что шанс оказаться в одной группе у объектов незначителен.

После вычисления согласованной матрицы различий для нахождения итогового варианта группировки будем применять стандартный агломеративный метод построения дендрограммы, который в качестве входной информации использует попарные расстояния между объектами [2]. При этом расстояния между группами будем определять по принципу "средней связи", т. е. как среднее арифметическое попарных расстояний между объектами, входящими в группы.

3. Экспериментальное исследование ансамблевого алгоритма

Для определения качества алгоритма была разработана процедура статистического моделирования. Процедура состоит в следующем:

— многократное генерирование случайных выборок в соответствии с заданным распределением для каждого класса;

— построение с помощью алгоритма согласованного группировочного решения для каждой выборки;

— определение качества группировки;

— нахождение усредненного по всем выборкам показателя качества.

Для построения деревьев использовался рекурсивный алгоритм с параметрами $R = 1$, $\gamma = 1$. Усреднение проводилось по 100 случайным выборкам, являющимися реализациями смеси указанных распределений. Качество группировки P_{cor} определяется как частота правильной классификации. Оценивался 95%-й доверительный интервал для вероятности правильной классификации. Ниже даны результаты моделирования для трех тестовых примеров.

Пример 1. Распределение для каждого из $K = 2$ классов является многомерным нормальным с одной и той же ковариационной матрицей Σ. Вектор математических ожиданий для каждого класса выбирается случайно из множества вершин единичного гиперкуба; ковариационная матрица является диагональной: $\Sigma = \sigma I$, где σ принимает значения из множества $\{0.1; 0.2; 0.3; 0.4\}$. Из общего числа переменных 50 являются количественными (их номера выбираются случайно), а 50 — булевыми. Для булевых переменных исходные значения, полученные с помощью датчика случайных чисел, округляются до ближайшего целого из множества $\{0; 1\}$. Объем выборки для первого и второго классов равен 25. Число деревьев в ансамбле задано $L = 10$; каждое дерево строится в подпространстве размерности 2. На рис. 1 приведены значения полученных показателей качества. Для сравнения указаны аналогичные усредненные показатели для одиночных деревьев. На графиках также отмечены соответствующие доверительные интервалы. Как видно из рисунка, применение ансамбля позволяет существенно улучшить качество группировки при условии, что классы не очень сильно пересекаются (при $\sigma \leq 0.3$).

Пример 2. В отличие от предыдущего примера, число классов $K = 3$; для количественных переменных векторы математических ожиданий для каждого класса выбираются случайно из множества $\{1, 2, \ldots, 10\}$. Некоторые переменные (их номера определяются случайно) являются шумовыми; для остальных переменных дисперсия $\sigma = 3$. Каждое дерево строится в случайно выбранном подпространстве переменных размерности 3. На рис. 2 представлены полученные усредненные значения частоты правильной классификации в зависимости от числа деревьев, входящих в ансамбль; при различном числе шумовых переменных.

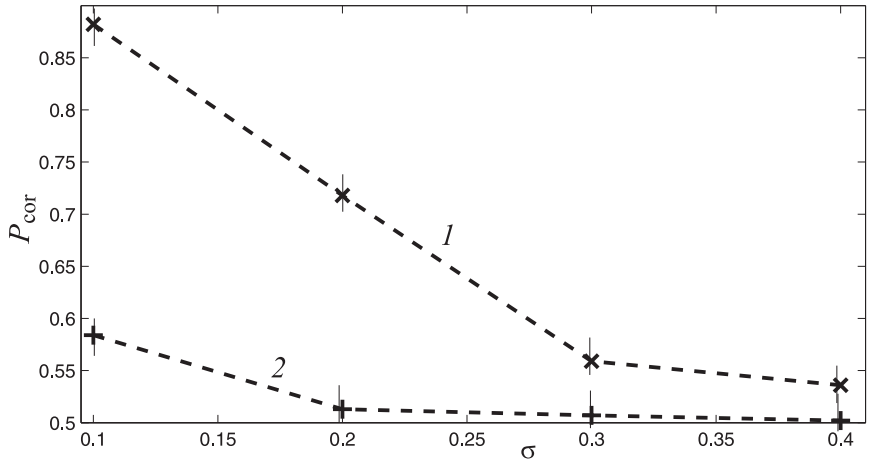

Рис. 1. Частота правильных решений P_{cor} ансамблевого алгоритма (*1*) и алгоритма построения одиночного дерева (*2*) в зависимости от дисперсии σ

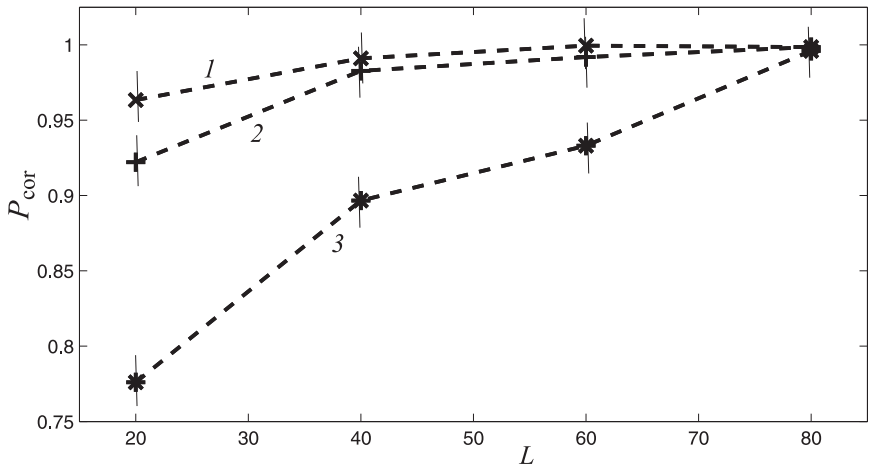

Рис. 2. Частота правильных решений ансамблевого алгоритма в зависимости от мощности ансамбля при различном числе шумовых переменных: *1* — 20, *2* — 40, *3* — 80; L — число деревьев в ансамбле

Отметим, что при достаточно большой мощности ансамбля частота правильных решений приближается к 1, т. е. практически не зависит от числа шумовых переменных.

Пример 3. С помощью статистического моделирования проводилось сравнение разработанного алгоритма с алгоритмом k-средних и алгоритмом построения дерева решений (последние два работали в пространстве всех переменных). При этом из 100 количественных переменных 90 являлись шумовыми (их номера выбирались случайно); для остальных переменных величина $\sigma = 0.25$, число классов $K = 2$, объем выборки для каждого класса равен 25. Полученный график зависимости частоты правильной классификации от мощности ансамбля представлен на рис. 3. Видно, что при увеличении мощности ансамбля качество коллективного алгоритма становится лучше, чем двух других рассматриваемых алгоритмов.

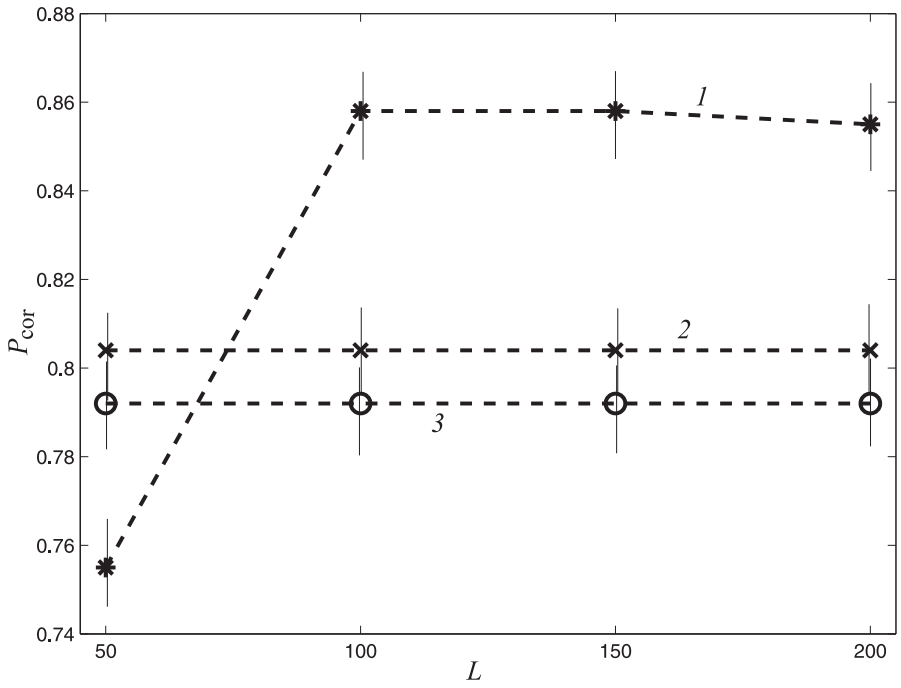

Рис. 3. Результаты сравнения коллективного алгоритма (*1*) с алгоритмами k-средних (*2*) и построения одиночного дерева (*3*)

Пример 4. Разработанный алгоритм тестировался на трех таблицах реальных данных, полученных непосредственно от специалистов прикладных областей либо из сети Интернет (репозитарий UCI [17]). Во всех анализируемых таблицах известна принадлежность объектов к характерным классам, что позволяет определить ошибку классификации, возникающую при использовании предлагаемого алгоритма кластерного анализа (естественно, переменная, задающая номера классов, при построении деревьев не используется). Заметим, что такое априорное разделение объектов на группы не всегда полностью совпадает с "объективной" классификацией, однако может служить для получения приближенной оценки качества тестируемого алгоритма.

1. Таблица данных "антропология" включает описания антропологических находок эпохи неолита на территории Сибири [18]. Объекты исследования описываются множеством из 23 переменных, представляющих собой измерения линейных и угловых размеров костей скелета. Была проанализирована информация о 252 антропологических объектах, которые принадлежали к двум антропологическим типам монголоидной и европеоидной расовых ветвей.

2. В таблице "наконечники" собраны археологические данные о 102 наконечниках стрел, обнаруженных в древних памятниках культуры на территории Новосибирской области [19]. Каждый наконечник описывается восемью числовыми и четырьмя номинальными переменными (число имен варьируется от 2 до 10). Указанные памятники относятся к двум основным типам культур.

3. Анализировалась таблица "zoo" из репозитария UCI. В таблице, содержащей 101 наблюдение, указаны значения двух числовых и 15 булевых переменных, описывающих признаки различных животных. Каждое животное относится к одному из семи классов. Для определения качества алгоритма классификации в данном случае удобнее использовать индекс Ранда IR, представляющий собой отношение числа пар объек-

Результаты работы алгоритмов кластерного анализа: P_{ans} (IR_{ans}) — частота правильных классификаций (индекс Ранда) для ансамбля, $\overline{P}_{\text{tree}}$ (IR_{tree}) — средняя частота правильных классификаций (индекс Ранда) для одиночного дерева. Мощность ансамбля $L = 100$

Название таблицы	Качество ансамбля	Качество одиночного дерева
Антропология	$P_{\text{ans}} = 1$	$\overline{P}_{\text{tree}} = 0.85$
Наконечники	$P_{\text{ans}} = 0.83$	$\overline{P}_{\text{tree}} = 0.61$
zoo	$IR_{\text{ans}} = 0.89$	$IR_{\text{tree}} = 0.76$

тов, у которых либо одинаковые, либо разные номера классов в полученной и истинной группировках, к общему числу пар различных объектов (значение индекса, близкое к 1, говорит о хорошей согласованности группировок).

Результаты тестирования приведены в таблице. Во всех случаях размерность подпространства переменных выбиралась случайно. Данные таблицы позволяют сделать вывод о том, что во всех проведенных экспериментах использование ансамбля деревьев решений позволяет заметно улучшить качество кластеризации.

Таким образом, в работе проведено теоретическое обоснование ансамблевых алгоритмов кластерного анализа, основанных на попарной классификации. Предложен алгоритм кластерного анализа, использующий ансамбль деревьев решений. При построении коллективного решения используется согласованная матрица различий между объектами. Исследование с помощью статистического моделирования показало, что применение предложенного метода построения ансамбля деревьев решений позволяет значительно улучшить качество классификации по сравнению с качеством алгоритмов несогласованных деревьев решений и k-средних, в том числе в задачах, характеризующихся наличием шумовых переменных и их разнотипностью.

Список литературы

[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] Лбов Г.С. Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука, 1981.

[4] Лбов Г.С., Старцева Н.Г. Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск: Изд. Ин-та математики СО РАН, 1999. 212 с.

[5] Лбов Г.С., Бериков В.Б. Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. Новосибирск: Изд. Ин-та математики СО РАН, 2005. 218 с.

[6] Michalski R., Stepp R., Diday E. Automated construction of classifications: conceptual clustering versus numerical taxonomy // IEEE Trans. Pattern Anal. Machine Intell. 1983. Vol. 5. P. 396–409.

[7] Лбов Г.С., Пестунова Т.М. Группировка объектов в пространстве разнотипных переменных // Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985. С. 141–149.

[8] Бериков В.Б., Лбов Г.С., Вишневская Е.А. Статистическое моделирование для исследования одного метода автоматической группировки // Сб. науч. статей V Междунар. конф. "Компьютерный анализ данных и моделирование". Минск, Белорусский гос. ун-т, 1998. Часть 3:А-К. С. 54–59.

[9] Strehl A., Ghosh J. Clustering ensembles — a knowledge reuse framework for combining multiple partitions // J. Machine Learning Res. 2002. Vol. 3. P. 583–617.

[10] Бирюков А.С., Рязанов В.В., Шмаков А.С. Решение задач кластерного анализа коллективами алгоритмов // Журн. вычисл. математики и мат. физики. 2008. Т. 48, № 1. С. 176–192.

[11] Журавлёв Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006.

[12] Breiman L. Bagging predictors // Machine Learning. 1996. Vol. 24. P. 123–140.

[13] Schapire R. The boosting approach to machine learning: An overview // Nonlinear Estimation and Classification. Lecture Notes in Statistics / Eds. D.D. Denison, M.H. Hansen, C.C. Holmes, B. Mallick, B. Yu. 2003. Vol. 171. P. 149–172.

[14] Topchy A., Law M., Jain A., Fred A. Analysis of consensus partition in cluster ensemble // Fourth IEEE Intern. Conf. on Data Mining (ICDM'04). 2004. P. 225–232.

[15] Kuncheva L. Combining Pattern Classifiers. Methods and Algorithms. Hoboken, N.J.: John Wiley & Sons, 2004.

[16] Lbov G.S., Berikov V.B. Recursive method of formation of the recognition decision rule in the class of logical functions // Pattern Recognit. and Image Analysis. 1993. Vol. 3, No. 4. P. 428–431.

[17] http://archive.ics.uci.edu/ml/

[18] Деревянко Е.И., Лбов Г.С., Бериков В.Б. и др. Компьютерная система анализа погребальных памятников эпохи неолита и ранней бронзы // Интеграционные программы фундаментальных исследований СО РАН. Новосибирск: Изд-во СО РАН, 1998.

[19] Сальникова И.В. Костяные наконечники стрел из комплексов Западной Сибири. Проблемы классификации и моделирования: Автореф. дис. ... канд. ист. наук. Новосибирский гос. ун-т, 2002.

Поступила в редакцию 20 июля 2009 г.