© 2022 г. В.Б. БЕРИКОВ, д-р техн. наук (berikov@math.nsc.ru)
(Институт математики им. С.Л. Соболева СО РАН, Новосибирск;
Новосибирский государственный университет)

МОДЕЛЬ И МЕТОД ПОСТРОЕНИЯ РАЗНОРОДНОГО КЛАСТЕРНОГО АНСАМБЛЯ[1]

Рассматривается задача кластеризации данных с помощью разнородного ансамбля с использованием матрицы коассоциации. Формулируется вероятностная модель, учитывающая коррелированность оценочных функций, с помощью которой находятся соотношения между характеристиками ансамбля и показателями качества итогового решения. Найдено выражение для оптимальных весов базовых алгоритмов, для которых минимальна верхняя граница оценки вероятности ошибки кластеризации. Проведено экспериментальное исследование предложенного метода, показавшее его преимущество по сравнению с рядом аналогичных методов.

Ключевые слова: кластерный анализ, ансамбль алгоритмов, коассоциативная матрица, модель кластерного ансамбля, оптимальные веса алгоритмов.

DOI: 10.31857/S000523102212008X, **EDN:** KSWKSX

1. Введение

В кластерном анализе требуется получить разбиение некоторого множества объектов на относительно небольшое число однородных подмножеств (групп, кластеров, классов). Число групп может быть известно заранее или должно быть определено автоматически. В данной работе рассматривается случай, когда требуемое число групп задано пользователем. Под критерием однородности разбиения понимается некоторый функционал, зависящий от описаний объектов, например показателей внутригруппового и межгруппового разброса.

Существует большое число методов кластерного анализа [1–4]. На практике чаще всего применяются приближенные итеративные алгоритмы, работа которых управляется с помощью некоторых параметров.

В настоящее время существует необходимость в создании новых методов, которые позволяли бы обнаруживать кластеры сложной формы и получать устойчивые результаты вне зависимости от начальных инициализаций алгоритма, порядка рассмотрения объектов и наличия шумовых искажений в данных.

[1] Работа проведена при финансовой поддержке РНФ, проект 22-21-00261.

В задачах классификации и прогнозирования активно развивается подход, основанный на коллективном принятии решений [5–9]. При этом итоговое решение определяется на основе нескольких вариантов разбиений, полученных различными алгоритмами либо одним алгоритмом, с разными параметрами работы. Коллективный (ансамблевый) подход позволяет повышать устойчивость результатов группировки в случае неопределенности в выборе параметров, проводить обработку больших объемов данных (анализируя по отдельности сравнительно небольшие их части), а также использовать «простые» вычислительно эффективные алгоритмы (например, направленные на поиск кластеров сферической формы) для обнаружения сложных структур данных [10].

Существует несколько основных направлений в методах построения коллективных решений кластерного анализа [11]. В данной работе рассматривается направление, основанное на использовании *коассоциативных матриц* (называемых также *матрицами попарных совпадений, матрицами смежности, co-occurrence matrix*), устанавливающих, как часто каждая пара объектов оказывается в одном и том же кластере (или в различных кластерах) по всем вариантам разбиения. Использование такого вида матриц позволяет решить проблему взаимного соответствия кластеров в вариантах группировки: поскольку нумерация кластеров внутри каждой кластеризации является субъективной, любые перестановки меток кластеров эквивалентны (проблема подробно рассматривается в [6], где предложен один из первых алгоритмов построения кластерного ансамбля).

Элементы усредненной матрицы могут рассматриваться как меры попарного расстояния (сходства) между объектами: чем чаще пара объектов была объединена алгоритмами, входящими в ансамбль, в один кластер, тем более похожими являются данные объекты. Для получения итогового консенсусного разбиения используется какой-либо из алгоритмов кластерного анализа, основанный на попарном сходстве, например, агломеративный алгоритм построения иерархической группировки.

Вероятностное обоснование данного подхода (доказательство сходимости ансамблевых решений к «истинному» разбиению) было сделано в [12]. В [13] представлена вероятностная модель коллективного кластерного анализа, позволяющая свести задачу кластеризации к задаче попарной классификации с латентными классами. При этом учитываются веса различных базовых алгоритмов ансамбля. Сформулирован критерий оптимальности весов и предложен метод их нахождения. В настоящей работе проводится дальнейшее развитие данного подхода. Проводится исследование влияния коррелированности базовых решений ансамбля на его качество. Показано, что учет коррелированности позволяет объяснить улучшение качества ансамбля при увеличении степени разнообразия вариантов разбиения, что ранее было экспериментально установлено в ряде работ (см., например, [14–16]).

Работа имеет следующую структуру. В первом разделе даны основные понятия работы. Во втором разделе вводится вероятностная модель ансамблевого кластерного анализа, в рамках которой исследуются свойства ансамбля. В третьем разделе описывается методика выбора оптимальных весов базовых алгоритмов ансамбля. Четвертый раздел посвящен вычислительному эксперименту с алгоритмом. В Заключении подводятся итоги работы и намечаются перспективы дальнейших исследований.

2. Основные понятия и обозначения

Пусть информация о множестве объектов исследования $A = \{a_1, \ldots, a_N\}$ представлена в виде набора $\mathbf{X} = \{x_1, \ldots, x_N\}$, где $x_i = X(a_i) = (X_1(a_i), \ldots, X_d(a_i))$ — вектор признаков для объекта $a_i \in A$, $i = 1, \ldots, N$, d — размерность пространства признаков, $x_i \in \mathbf{R}^d$. Требуется разбить множество A на некоторое заданное число кластеров K.

В ансамблевом кластерном анализе рассматривается набор базовых алгоритмов группировки $\mu_1, \ldots, \mu_M$, которые строят варианты разбиения множества A. Каждый вариант состоит из $K_{l,m}$ непересекающихся подмножеств (кластеров), $m = 1, \ldots, M$, $l = 1, \ldots, L_m$, где L_m – число запусков алгоритма μ_m (величина $K_{l,m}$ может не совпадать с итоговым числом кластеров K). Варианты разбиения получены при различных значениях параметров работы (или, в более общем смысле, «условий обучения», таких как подмножество отобранных переменных или набор начальных центроидов). Обозначим множество возможных значений параметров алгоритма μ_m через $\mathbf{\Omega}_m$.

Определим для каждой пары различных объектов a_i и a_j из множества A величину

$$h(\Omega_{l,m}; i, j, \mathbf{X}) = \mathbf{I}\left[\mu_m\left(a_i, \Omega_{l,m}, \mathbf{X}\right) = \mu_m\left(a_j, \Omega_{l,m}, \mathbf{X}\right)\right],$$

где $\mathbf{I}[\cdot]$ — индикаторная функция: $\mathbf{I}[true] = 1$; $\mathbf{I}[false] = 0$, $\mu_m(a, \Omega_{l,m}, \mathbf{X})$ — номер кластера, приписанного объекту $a \in A$ согласно l-му варианту разбиения, сформированного алгоритмом μ_m для набора параметров $\Omega_{l,m} \in \mathbf{\Omega}_m$ по набору данных $\mathbf{X}$. При заданных $\Omega_{l,m}$, $\mathbf{X}$ будем записывать: $h_{l,m}(i, j) = h(\Omega_{l,m}; i, j, \mathbf{X})$. Для каждого l-го варианта разбиения можно определить коассоциативную матрицу $\mathbf{H}_l = (h_l(i, j))$.

Коллективная кластеризация в рамках рассматриваемого подхода полагается как процесс, включающий несколько основных этапов.

На **первом этапе** формируются различные варианты разбиения выборки $P_{1,1}, \ldots, P_{l,m}, \ldots, P_{L_M,M}$, полученные на основе случайного независимого выбора параметров алгоритмов из множеств $\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_M$ в соответствии с некоторыми распределениями (например, равномерными).

На **втором этапе** проводится анализ полученных разбиений: определяются оценочные функции $\gamma_{l,m} \geq 0$, которые зависят от индексов качества полученных вариантов, мер их разнообразия и т.п. и в общем случае являются функциями от разбиений: $\gamma_{l,m} = \gamma_{l,m}(P_{1,1}, \ldots, P_{L_M,M})$. Будем считать, что

чем выше показатели качества вариантов, тем большее значение принимает величина оценочной функции.

Например, при применении процедуры селекции ансамбля [16] проводится вычисление мер разнообразия разбиений и тем вариантам, для которых эта мера не превышает определенный порог, присваивается нулевое значение оценочной функции. Оставшимся вариантам приписывается некоторое ненулевое постоянное значение.

На **третьем этапе** вычисляется усредненная с весами коассоциативная матрица $\mathbf{H} = (h(i,j))$ с элементами

$$h(i,j) = \sum_{m=1}^{M} \sum_{l=1}^{L_m} \gamma_{l,m} h_{l,m}(i,j), \tag{1}$$

$(i,j = 1,\ldots,N, i \neq j)$. Элементы $\mathbf{H}$ рассматриваются как меры близости между парами соответствующих объектов. Для формирования окончательного разбиения на заданное число кластеров можно применять любой алгоритм, который использует матрицу $\mathbf{H}$ как исходную информацию для группирования. В данной работе с этой целью используется алгоритм, основанный на спектральном кластерном анализе [17]. В итоге получим разбиение множества A на подмножества $C_1,\ldots,C_K$.

В целях управления качеством ансамбля целесообразно ввести управляемую компоненту в (1) и представить элементы усредненной коассоциативной матрицы в виде

$$h(i,j) = \sum_{m=1}^{M} \alpha_m \sum_{l=1}^{L_m} \gamma_{l,m} h_{l,m}(i,j), \tag{2}$$

где $\alpha_1,\ldots,\alpha_M$ — веса базовых алгоритмов $\mu_1,\ldots,\mu_M$, $\sum \alpha_m = 1$, $\alpha_m \geq 0$. Веса могут назначаться пользователем в соответствии с характеристиками ансамбля. Данная возможность будет рассмотрена ниже в разделе 4.

3. Вероятностная модель ансамблевой классификации

Описывается модель коллективной классификации с латентными классами, в рамках которой по наблюдаемым характеристикам ансамбля оценивается вероятность непосредственно не наблюдаемой ошибки классификации.

3.1. Модель генерации данных и ансамбля алгоритмов

При формулировании модели будем использовать основные положения работы [13]. Сформулируем следующую модель генерации данных. Пусть объекты множества A случайным и независимым образом выбираются из некоторой генеральной совокупности Γ, разделенной на K классов. Предположим, что в описании произвольного объекта $a \in \Gamma$ имеется признак Y, который указывает на его принадлежность к определенному классу: $Y(a) \in \{1,\ldots,K\}$.

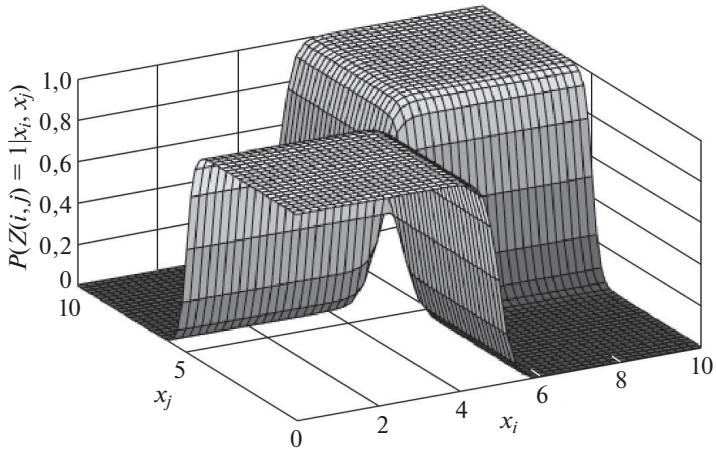

График функции $\mathrm{P}[Z(i,j)=1 \mid x_i, x_j]$ в зависимости от значений $x_i,\ x_j$.

Этот признак является скрытым (непосредственно не наблюдаемым) для алгоритма кластеризации. Каждому классу соответствует априорная вероятность $P_k = P(Y(a)=k),\ k=1,\ldots,K$, где $\sum\limits_{k=1}^{K} P_k = 1$. Классу с номером k сопоставляется закон условного распределения в пространстве признаков: $p(X(a)=x \mid Y(a)=k) = p_k(x),\ k=1,\ldots,K$.

Для произвольной пары различных объектов $a_i, a_j \in A$ определим случайную величину

$$Z(i,j) = \mathbf{I}\left[Y(a_i) = Y(a_j)\right],$$

которая показывает действительный статус пары, т.е. принадлежит ли она к одному классу или к разным классам. Пусть набор данных $\mathbf{X}$ фиксирован; при этом источником случайности в модели служит выбор параметров алгоритма кластеризации, а также истинный статус пар объектов. Распределение величины $Z(i,j)$ при условии заданных значений $X(a_i)=x_i$ и $X(a_j)=x_j$ определяется через условное распределение $P[Y \mid X(a)=x]$, $a \in A$:

$$\mathrm{P}\left[Z(i,j)=1 \mid x_i, x_j\right] = \sum_{k=1}^{K} \mathrm{P}\left[Y(a_i)=k \mid x_i\right] \mathrm{P}\left[Y(a_j)=k \mid x_j\right] =$$

$$= \sum_{k=1}^{K} \frac{p_k(x_i) p_k(x_j) P_k^2}{p(x_i) p(x_j)},$$

где $p(x) = \sum\limits_{k=1}^{K} p_k(x) P_k$. На рисунке показан пример графика функции $\mathrm{P}\left[Z(i,j)=1 \mid x_i, x_j\right]$ для случая двух классов, подчиняющихся одномерным нормальным распределениям: $x_i \sim N(2,1)$, $x_j \sim N(8,1)$ и равными априорными вероятностями.

Обозначим через Ω_m случайный набор параметров, выбираемый из $\mathbf{\Omega}_m$. Предположим, что алгоритм μ_m проработал некоторое число L_m раз при независимо выбранных, в соответствии с одним и тем же распределением, параметрах. Тогда величины $\Omega_{m,1}, \ldots, \Omega_{m,L_m}$ являются независимыми и одинаково распределены с Ω_m.

Предполагается, что все требования, необходимые для корректной работы базовых алгоритмов (например, относящиеся к свойствам меры близости между парами объектов), соблюдены.

Результаты работы алгоритма μ_m (т.е. полученные элементы матрицы коассоциации) можно рассматривать как детерминированные функции $h_{l,m}(i,j) = h(i,j,\Omega_{l,m},\mathbf{X})$ от независимых многомерных случайных величин $\Omega_{l,m} \in \mathbf{\Omega}_m$, $l = 1, \ldots, L_m$, которые подчиняются одному и тому же распределению на $\mathbf{\Omega}_m$. Обозначим: $\Psi = \{\Omega_{1,1}, \ldots, \Omega_{L_M,M}\}$.

Рассмотрим условные вероятности правильного решения для каждого алгоритма (объединения пары объектов из одного класса в один кластер или разделения пары из разных классов в различные кластеры):

$$q_m^{(1)}(i,j,\mathbf{X}) = \mathrm{P}\left[h(i,j,\Omega_m,\mathbf{X}) = 1 \mid Z(i,j) = 1\right],$$

$$q_m^{(0)}(i,j,\mathbf{X}) = \mathrm{P}\left[h(i,j,\Omega_m,\mathbf{X}) = 0 \mid Z(i,j) = 0\right].$$

В результате работы алгоритма получим набор независимых случайных величин — решений

$$h(i,j,\Omega_{1,m},\mathbf{X}), \ldots, h(i,j,\Omega_{L_m,m},\mathbf{X}), \quad m = 1, \ldots, M.$$

Будем также считать, что решения условно независимы:

$$\begin{aligned}\mathrm{P}\left[h\left(i,j,\Omega_{i_1,m_1},\mathbf{X}\right) = h_{r_1}, \ldots, h\left(i,j,\Omega_{i_j,m_j},\mathbf{X}\right) = h_{r_j} \mid Z(i,j) = z\right] = \\ = \mathrm{P}\left[h\left(i,j,\Omega_{i_1,m_1},\mathbf{X}\right) = h_{r_1} \mid Z(i,j) = z\right] \times \ldots \times \\ \times \mathrm{P}\left[h\left(i,j,\Omega_{i_j,m_j},\mathbf{X}\right) = h_{r_j} \mid Z(i,j) = z\right],\end{aligned}$$

где $\Omega_{i_1,m_1}, \ldots, \Omega_{i_j,m_j}$ — произвольные параметры; индексы $m_1, \ldots, m_j$ соответствуют разным сочетаниям алгоритмов, $h_{r_1}, \ldots, h_{r_j}$, $z \in \{0,1\}$.

Величина оценочной функции $\gamma_{l,m}$, определяемая на втором этапе построения ансамбля, зависит от полного набора результатов группировки $P_{1,1}, \ldots, P_{L_M,M}$. Необходимость учета других вариантов, помимо текущего, может быть обоснована, например, тем, что при наличии совпадающих или очень похожих вариантов их вес было бы разумно уменьшить. Часть дублирующих разбиений можно исключить, что эквивалентно обнулению их весов. Так как результаты группировки, в свою очередь, определяются выбором параметров работы алгоритмов, то можно считать, что оценочная функция зависит от всего набора параметров: $\gamma_{l,m} = \gamma_{l,m}(\Psi,\mathbf{X})$. Поскольку функционал

качества определяется по всему набору данных, можно полагать, что он практически не зависит от любых других величин, определенных для конкретной пары объектов. То есть будем считать, что величины $\gamma_{l,m} = \gamma_{l,m}(\Psi, \mathbf{X})$ и $h_{l,m}(i,j) = h\,(i,j,\Omega_{l,m},\mathbf{X})$ некоррелированы.

Можно полагать, что для каждого базового алгоритма μ_m, $m = 1,\ldots,M$, оценка различных вариантов его работы проводится по одинаковой схеме; тогда распределение величин $\gamma_{l,m} = \gamma_{l,m}(\Psi,\mathbf{X})$, $l = 1,\ldots,L_m$ будет одинаково.

3.2. Характеристики ансамбля

Для того чтобы использовать модель ансамблевой классификации для управления качеством коллективного решения, необходимо найти зависимость между наблюдаемыми характеристиками работы ансамбля и показателями качества коллективной классификации.

3.2.1. Вероятность ошибки ансамблевой классификации

Введем предварительно следующие понятия. Для фиксированной пары различных индексов $i,j \in \{1,\ldots,N\}$, ансамблевым решением, полученным для усредненной матрицы коассоциации с элементами, вычисляемыми согласно (2), назовем величину

$$D(i,j,\Psi,\mathbf{X}) = \mathbf{I}\left[\sum_{l,m:h_{l,m}(i,j)=1}\gamma_{l,m} > \sum_{l,m:h_{l,m}(i,j)=0}\gamma_{l,m}\right].$$

Величина $D(i,j,\Psi,\mathbf{X}) \in \{0,1\}$ и показывает, превышает ли суммарный вес голосов за объединение пары в один кластер (когда $D(i,j,\Psi,\mathbf{X}) = 1$) сумму голосов за ее разделение (при $D(i,j,\Psi,\mathbf{X}) = 0$). Ансамблевое решение зависит от набора случайных параметров алгоритма, а также от исходного множества данных.

Вероятность ошибки ансамблевой классификации для пары определяется как

$$P_{er}\,(i,j;\mathbf{X}) = \mathrm{P}_{\Psi,Z(i,j)}\left[D(i,j,\Psi,\mathbf{X}) \neq Z(i,j)\right]. \tag{3}$$

Здесь индексы $\Psi, Z(i,j)$ обозначают источник случайности.

3.2.2. Разложение ошибки на составляющие

В данном подразделе, для краткости, будем опускать индексы i,j и обозначение $\mathbf{X}$ в величинах, определенных выше (подразумевая, что соответствующие значения i, j, $\mathbf{X}$ фиксированы). Пусть $D = D(\Psi)$.

Выражение для вероятности ансамблевой ошибки классификации (3) можно представить в следующем виде: $P_{er} = E_{\Psi,Z}[Z-D]^2 = \mathrm{Var}[Z-D] + {} + (E[Z-D])^2$. Так как $\mathrm{Var}[Z-D] = \mathrm{Var}[Z] + \mathrm{Var}[D] - 2\,\mathrm{Cov}[Z,D]$, то

$$P_{er} = \mathrm{Var}[Z] + \mathrm{Var}[D] + (E[Z-D])^2 - 2\,\mathrm{Cov}[Z,D]. \tag{4}$$

Здесь через $\mathrm{Var}[\cdot]$ и $\mathrm{Cov}[\cdot,\cdot]$ обозначены соответственно дисперсия и ковариация случайных величин. Полученное выражение (4) можно проинтерпретировать следующим образом. Ошибка классификации для пары объектов имеет несколько источников возникновения:

— неустранимая ошибка $\mathrm{Var}[Z]$, которая связана с близостью или пересечением истинных непосредственно не наблюдаемых классов;

— разброс ансамблевых решений $\mathrm{Var}[D]$;

— несоответствие между истинным статусом пары объектов и решением ансамбля, выражаемое через квадрат смещения: $(E[Z-D])^2$;

— ковариация с обратным знаком $\mathrm{Cov}[Z,D]$ между истинным статусом и решением ансамбля.

Можно заметить, что в «простых» задачах кластеризации, т.е. при высокой степени разделимости классов (например, когда дисперсия $\mathrm{Var}[Z]$ мала для всех пар объектов) и достаточно точном соответствии решений ансамбля истинному статусу, малое значение вероятности ошибки в (4) достигается при небольшом разнообразии ансамбля (малом разбросе решений, т.е. величины D для пары).

Однако в случае трудноразделимых классов, когда вариабельность Z велика, для уменьшения смещения и увеличения ковариации требуется увеличивать разнообразие ансамбля. Например, в идеальном случае, когда $\mathrm{P}[Z=D]=1$ (когда алгоритм построения ансамбля точно определяет истинный статус пар объектов), смещение будет нулевым, ковариация максимальной, а дисперсия решений (совпадающая с дисперсией Z) также большой.

Соотношение (4) позволяет на качественном уровне судить об источниках возникновения ошибки ансамбля. С использованием рассматриваемого ниже понятия маржинальной функции можно получить более детальные выводы относительно поведения ансамбля.

3.2.3. Основные характеристики маржинальной функции

Рассмотрим усредненную коассоциативную матрицу в виде (2). Оценкой решения (маржинальной функцией кластерного ансамбля [13]) назовем величину

$$(5)\quad mg(i,j,\Psi,Z(i,j),\mathbf{X}) = \sum_{l,m:h_{l,m}(i,j)=Z(i,j)} \alpha_m\gamma_{l,m} - \sum_{l,m:h_{l,m}(i,j)\neq Z(i,j)} \alpha_m\gamma_{l,m},$$

где $Z(i,j)\in\{0,1\}$ — истинный статус пары. Эта функция показывает, насколько взвешенное число голосов за правильное решение превосходит взвешенное число голосов за неправильное решение, и зависит от параметров алгоритмов, истинного статуса пары и от выборки.

С использованием маржинальной функции можно представить условную вероятность ошибки предсказания величины Z для пары объектов в следую-

щем виде:

$$P_{err}(i,j,z) = \mathrm{P}\left[mg(i,j,\Psi,Z(i,j),\mathbf{X}) < 0 \mid Z(i,j) = z\right].$$

Переход к условной вероятности обусловлен более простым видом получаемых выражений. Чтобы определить свойства данной величины, необходимо знать характеристики маржинальной функции. Найдем ее первые условные моменты.

Утверждение 1. Условное математическое ожидание маржинальной функции равно

$$E_{\Psi|z}\left[mg\left(i,j,\Psi,Z(i,j),\mathbf{X}\right)\right] = \sum_m \alpha_m L_m \Gamma_m \left(2q_m^{(z)}(i,j,\mathbf{X}) - 1\right), \tag{6}$$

а ее условная дисперсия —

$$\begin{aligned}\mathrm{Var}_{\Psi|z}[mg(\Psi,Z)] = &\sum_m \alpha_m^2 L_m \left[V_m + 4(\Gamma_m)^2 q_m^{(z)}(i,j,\mathbf{X})(1 - q_m^{(z)}(i,j,\mathbf{X}))\right] + \\ &+ \sum_m \alpha_m^2 L_m (L_m - 1)\left(2q_m^{(z)}(i,j,\mathbf{X}) - 1\right)^2 C_m + \\ &+ \sum_{\substack{m,m':\\ m\neq m'}} \alpha_m \alpha_{m'} L_m L_{m'} \left(2q_m^{(z)}(i,j,\mathbf{X}) - 1\right)\left(2q_{m'}^{(z)}(i,j,\mathbf{X}) - 1\right) C_{m,m'},\end{aligned} \tag{7}$$

где оператор $E_{X|z}[\cdot] = E_X[\cdot \mid Z = z]$ обозначает условное математическое ожидание, $\Gamma_m = E_\Psi[\gamma_{l,m}]$ есть математическое ожидание оценочного функционала на множестве Ω_m, $V_m = \mathrm{Var}_\Psi[\gamma_{l,m}]$ обозначает дисперсию оценочной функции, $C_m = \mathrm{cov}[\gamma_{l,m}, \gamma_{l',m}]$ — ковариация между оценочными функциями для алгоритма μ_m, $C_{m,m'} = \mathrm{cov}[\gamma_{l,m}, \gamma_{l',m'}]$ — ковариация между оценочными функциями для различных алгоритмов $\mu_m, \mu_{m'}$ $(m \neq m', m, m' = 1, \ldots, M)$.

Доказательство утверждения 1 приведено в Приложении.

3.2.4. Влияние характеристик ансамбля на вероятность ошибки

Известно, что для любой случайной величины U с конечным математическим ожиданием $E[U] > 0$ и дисперсией $\mathrm{Var}\,[U]$ справедливо

$$P[U < 0] \leq P[\{U < E[U] - \varepsilon\} \bigcup \{U > E[U] + \varepsilon\} = P[\mid U - E[U] \mid > \varepsilon],$$

где $\varepsilon = E[U]$. Из неравенства Чебышева вытекает, что

$$P[U < 0] < \frac{\mathrm{Var}[U]}{(E[U])^2}.$$

Следовательно, для условной вероятности ошибки

$$P_{err}(i,j,z) = \mathrm{P}\left[mg(i,j,\Psi,Z(i,j),\mathbf{X}) < 0 \mid Z(i,j) = z\right]$$

выполняется

$$P_{err}(i,j,z) < \frac{\mathrm{Var}_{\Psi|z}\left[mg(i,j,\Psi,Z(i,j),\mathbf{X})\right]}{\left(E_{\Psi|z}\left[mg(i,j,\Psi,Z(i,j),\mathbf{X})\right]\right)^2} \tag{8}$$

в случае, когда $E_{\Psi|z}[mg(i,j,\Psi,Z(i,j),\mathbf{X})] > 0$. Из утверждения 1 следует, что данное условие будет выполнено, в частности, если для всех $m = 1,\ldots,M$ справедливо

$$0{,}5 + \varepsilon < q_m^{(z)}(i,j,\mathbf{X}) \leq 1, \tag{9}$$

т.е. алгоритм кластеризации принимает решение не наугад (здесь ε — некоторая малая положительная величина).

Для уменьшения вероятности ошибки можно дополнительно к (8) и (9) потребовать, чтобы дисперсия маржинальной функции (7) была бы минимальной. На основании утверждения 1 можно сделать несколько качественных выводов о поведении ансамблевых решений.

1) При прочих равных условиях верхняя граница вероятности ошибки снижается при уменьшении ковариации оценочных функций C_m и $C_{m,m'}$ для алгоритмов μ_m, $\mu_{m'}$, $m, m' = 1,\ldots,M$. Заметим, что уменьшение ковариации соответствует повышению степени рассогласованности оценочных функций (высокой степени разнообразия результатов группировки).

2) Предположим, что $L_1 = \ldots = L_M = L$, все оценочные функции совпадают и равны $\gamma_{l,m} \equiv 1/L$, $\alpha_m \equiv 1/M$. Значит, $\forall m$, $\Gamma_m = 1/L$, $V_m = C_m = 0$ и $\forall m, m'$ $(m \neq m')$, $C_{m,m'} = 0$. Из (6) получим:

$$E_{\Psi|z}\left[mg(i,j,\Psi,Z(i,j),\mathbf{X})\right] = \frac{1}{M}\sum_m \left(2q_m^{(z)}(i,j,\mathbf{X}) - 1\right),$$

а из (7) —

$$\mathrm{Var}_{\Psi|z}\left[mg(i,j,\Psi,Z(i,j),\mathbf{X})\right] = \frac{4}{M^2 L}\sum_m q_m^{(z)}(i,j,\mathbf{X})\left(1 - q_m^{(z)}(i,j,\mathbf{X})\right).$$

На основании (8), (9) можно сделать вывод о том, что $P_{err}(i,j,z) \to 0$ при $L \to \infty$ и $M = \mathrm{const}$. То есть при выполнении условий модели и увеличении количества элементов ансамбля вероятность ошибки стремится к нулю.

3) Предположим, что ожидаемая оценка качества Γ_m для некоторого значения m увеличивается при условии максимальной степени устойчивости $q_m^{(z)} = 1$ и при прочих неизменных характеристиках ансамбля. При этом верхняя граница вероятности ошибки будет уменьшаться. Однако при низких значениях вероятности правильного решения $q_m^{(z)}$ дисперсия маржинальной функции увеличится за счет первой группы слагаемых в (7). Таким образом, «переоценка» ансамбля может привести к ухудшению его качества.

4. Выбор оптимальных весов

4.1. Решение оптимизационной задачи

Рассмотрим задачу выбора весов $\alpha_1, \ldots, \alpha_M$ базовых алгоритмов так, чтобы минимизировать верхнюю границу вероятности ошибки. Пусть выбрана произвольная пара различных объектов $a_i, a_j \in A$, таких что выполняется условие (9), тогда из (8) получим следующую задачу минимизации (индексы i, j и обозначение $\mathbf{X}$ для краткости опущены):

$$\frac{\mathrm{Var}_{\Psi|z}[mg(\Psi, Z)]}{\left(E_{\Psi|z}[mg(\Psi, Z)]\right)^2} \to \min_{\alpha_1,\ldots,\alpha_M}$$

при ограничениях $\sum \alpha_m = 1$, $\alpha_m \geq 0$, $m = 1, \ldots, M$.

Приближенное численное решение можно получить, например, с помощью градиентного метода. В данной работе рассматривается упрощенная задача, при решении которой удается вывести ответ в аналитической форме:

$$\mathrm{Var}_{\Psi|z}[mg(\Psi, Z)] \to \min_{\alpha_1,\ldots,\alpha_M}$$

при тех же ограничениях. Для ее решения используем метод множителей Лагранжа. Пусть функция Лагранжа следующая:

$$\begin{aligned}\mathbf{L} = \sum_m \alpha_m^2 L_m \left[V_m + 4(\Gamma_m)^2 q_m^{(z)} \left(1 - q_m^{(z)}\right)\right] + \\ + \sum_m \alpha_m^2 L_m (L_m - 1) \left(2q_m^{(z)} - 1\right)^2 C_m + \\ + \sum_{\substack{m,m':\\ m \neq m'}} \alpha_m \alpha_{m'} L_m L_{m'} \left(2q_m^{(z)} - 1\right)\left(2q_{m'}^{(z)} - 1\right) C_{m,m'} - \lambda\left(\sum \alpha_m - 1\right),\end{aligned}$$

где λ — множитель Лагранжа. Тогда необходимое условие экстремума будет

$$\begin{aligned}\frac{\partial \mathbf{L}}{\partial \alpha_m} = 2\alpha_m L_m \left[V_m + 4(\Gamma_m)^2 q_m^{(z)} \left(1 - q_m^{(z)}\right) + (L_m - 1)\left(2q_m^{(z)} - 1\right)^2 C_m\right] + \\ + \sum_{\substack{m':\\ m \neq m'}} \alpha_{m'} L_m L_{m'} \left(2q_m^{(z)} - 1\right)\left(2q_{m'}^{(z)} - 1\right) C_{m,m'} - \lambda = 0,\end{aligned}$$

$m = 1, \ldots, M$. Для удобства введем матричные обозначения: пусть $\mathbf{A}_{M\times M} = (A(m', m''))$ — матрица с элементами

$$A(m', m'') = \tag{10}$$

$$= \begin{cases} 2L_m \left[V_m + 4(\Gamma_m)^2 q_m^{(z)} \left(1 - q_m^{(z)}\right) + (L_m - 1)\left(2q_m^{(z)} - 1\right)^2 C_m\right], & m' = m'', \\ L_m L_{m'} \left(2q_m^{(z)} - 1\right)\left(2q_{m'}^{(z)} - 1\right) C_{m,m'}, & m' \neq m''. \end{cases}$$

Получим систему линейных уравнений в виде

$$A \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_M \end{pmatrix} = \begin{pmatrix} \lambda \\ \vdots \\ \lambda \end{pmatrix}.$$

Добавив условие $\sum \alpha_m = 1$, запишем систему в матричной форме:

$$\mathbf{C}a = b,$$

где

$$a = (\alpha_1, \ldots, \alpha_M, \lambda)^{\mathrm{T}}, \quad \mathbf{C} = \begin{pmatrix} & & & -1 \\ & \mathbf{A} & & \vdots \\ & & & -1 \\ 1 & \ldots & 1 & 0 \end{pmatrix}, \quad b = (0, \ldots, 0, 1)^{\mathrm{T}}.$$

Тогда решение определяется как

$$a = \mathbf{C}^{-1}b, \tag{11}$$

если обратная матрица существует (в случае каких-либо особенностей матрицы условимся назначать алгоритмам одинаковые веса).

Для оценки оптимальных коэффициентов заменим характеристики оценочных функций, входящие в (10), их выборочными аналогами: выборочным средним, дисперсией и ковариацией. Эти величины определяются по полученным в ходе экспериментов значениям оценочных функций (для оценивания ковариации, рассматриваются различные пары объектов).

Элементы матрицы $\mathbf{A}$ зависят от величин $q_m^{(z)}$, которые являются непосредственно не наблюдаемыми, а также могут меняться для разных пар объектов. Однако по условию задачи оптимальные коэффициенты должны быть постоянными для всех пар. Будем считать величину $q_m^{(z)}$ параметром, принимающим значение q, близкое к 1, так как полученные выражения справедливы для любых пар, в том числе правильно кластеризуемых с большой вероятностью (предполагается, что такие пары объектов существуют).

4.2. Основные шаги алгоритма

Опишем основные шаги предложенного алгоритма построения взвешенного кластерного ансамбля (назовем его CEOW — Cluster Ensemble with Optimized Weights).

Алгоритм CEOW.

Входные данные:

— $\mathbf{X}$ — таблица данных «объект-свойство»;

— $\mu_1, \ldots, \mu_M$ — набор базовых алгоритмов кластерного анализа;

— $L_1, \ldots, L_M$ — число запусков каждого базового алгоритма.

Шаги:

1. Запустить каждый базовый алгоритм заданное число раз со случайными значениями параметров, которые выбираются случайным образом из заданного множества возможных значений (распределение, в соответствии с которым инициализируются параметры, фиксировано).
2. Для каждого полученного разбиения определить оценку его качества.
3. Найти оценки величин Γ_m, V_m, C_m, для каждого алгоритма μ_m, а также оценки ковариаций $C_{m,m'}$ для комбинаций различных алгоритмов μ_m, $\mu_{m'}$.
4. Вычислить оптимальные коэффициенты $\alpha_1, \ldots, \alpha_M$ по формуле (11).
5. Вычислить взвешенную коассоциативную матрицу $\mathbf{H}$ по формуле (2).
6. Рассматривая матрицу $\mathbf{H}$ как матрицу попарного сходства объектов, найти итоговое разбиение с помощью алгоритма кластерного анализа, на вход которого подается полученная матрица сходства (для этого можно использовать агломеративный алгоритм построения иерархического разбиения либо спектральный алгоритм).

Конец работы алгоритма.

5. Экспериментальное исследование

Разработанный алгоритм реализован программно на языке Python. Для исследования алгоритма был проведен численный эксперимент. Выборка для тестирования была взята из репозитория задач машинного обучения [18]. Она представляет собой уличные изображения, сегментированные на семь классов.

Эксперимент проводился следующим образом: в качестве базовых алгоритмов использовался алгоритм K-means с разным фиксированным числом кластеров (от семи до максимально допустимого значения, определяемого числом алгоритмов в ансамбле, с шагом один). То, что в качестве базовых алгоритмов используются варианты алгоритма K-means, никак не влияет на применимость полученных теоретических результатов. Каждый алгоритм запускался 20 раз с различной случайной инициализацией центроидов. Оценка качества разбиения находилась с помощью нормированного по всем разбиениям Силуэт-индекса качества [19]. В (10) было взято значение $q = q_m^{(z)} = 1$. Для построения итогового разбиения использовался спектральный алгоритм кластерного анализа [17]. Полученное разбиение сравнивалось с истинным разбиением с помощью исправленного индекса Ранда (ARI) [20]. Более высокое значение индекса свидетельствует о лучшем соответствии разбиений.

В табл. 1 приведены результаты работы алгоритма CEOW, отдельного алгоритма K-means, а также кластерного ансамбля, основанного на вычислении коассоциативной матрицы с равными весами базовых алгоритмов (CEEW — Cluster Ensemble with Equal Weights).

На экспериментальных данных предложенный алгоритм показал более точные результаты, чем алгоритм K-means и аналогичный ансамблевый алго-

Таблица 1. Результаты эксперимента: индекс ARI

Число алгоритмов в ансамбле	CEOW	K-means	CEEW
4	0,410	0,351	0,402
10	0,443	0,351	0,443
16	0,467	0,351	0,453

ритм с равными весами. Также можно заметить, что результаты улучшаются с увеличением числа базовых алгоритмов в ансамбле.

6. Заключение

В работе проведено теоретическое и экспериментальное исследование разнородного кластерного ансамбля, основанного на наборе различных алгоритмов кластерного анализа. Коллективное решение строится путем анализа усредненной коассоциативной матрицы, при нахождении которой учитываются оценки качества полученных вариантов группировки. Для обоснования разработанного метода предложена вероятностная модель ансамблевой классификации, учитывающая коррелированность оценочных функций. В модели делается предположение о существовании «истинных» непосредственно не наблюдаемых классов, которое позволяет вывести оценки качества работы ансамбля. С помощью модели получены аналитические зависимости между оценками качества решения и характеристиками ансамбля (числом его элементов, ожидаемым значением и дисперсией индекса качества, показателями коррелированности алгоритмов). В рамках модели найдено выражение для оптимальных весов, для которых минимальна верхняя граница оценки вероятности ошибки классификации.

Разработан алгоритм, в котором реализован метод построения ансамбля и вычисления оптимальных весов. Экспериментальное исследование подтвердило эффективность предложенного метода: предложенный алгоритм дал более высокую точность, чем отдельный алгоритм K-means и аналогичный ансамблевый алгоритм, не использующий оптимизацию весов.

Проведенное исследование имеет ряд ограничений. Работа направлена в основном на получение теоретических зависимостей (в рамках предложенной модели) между характеристиками ансамбля и показателями его качества. При выборе оптимальных весов базовых алгоритмов решалась лишь упрощенная задача оптимизации, поскольку в общем случае удается получить только приближенное численное решение.

В перспективе планируется провести дополнительные экспериментальные исследования модели и метода на различных прикладных задачах. Было бы интересно сравнить решения общей и упрощенной задач оптимизации. Кроме того, планируется рассмотреть более широкий круг базовых алгоритмов кластеризации, входящих в ансамбль (DBSCAN, спектральный алгоритм и

т.д), а также применить разработанный метод в задаче слабоконтролируемого обучения с использованием коассоциативной матрицы ансамбля как матрицы сходства.

Автор благодарит магистранта НГУ В.В. Баранова за помощь в проведении вычислительных экспериментов.

ПРИЛОЖЕНИЕ

Доказательство утверждения 1. При доказательстве для краткости будем опускать обозначения i, j, $\mathbf{X}$. Маржинальная функция (5) может быть переписана в виде

$$\begin{aligned} mg(\Psi, Z) &= \sum_{l,m} \alpha_m \gamma_{l,m} \left\{ \mathbf{I}\left[h(\Omega_{l,m}) = Z\right] - \mathbf{I}\left[h\left(\Omega_{l,m}\right) \neq Z\right] \right\} = \\ &= \sum_{l,m} \alpha_m \gamma_{l,m} (2Z-1)(2h(\Omega_{l,m})-1), \end{aligned} \tag{П.1}$$

так как для булевых u, v выполняется $\mathbf{I}[u = v] - \mathbf{I}[u \neq v] = (2u-1)(2v-1)$. Отсюда получим

$$E_{\Psi|z}[mg(\Psi, Z)] = \sum_{l,m} E_{\Psi|z}\left[\alpha_m \gamma_{l,m}(2z-1)(2h(\Omega_{l,m})-1)\right] =$$

$$= (2z-1) \sum_{l,m} \alpha_m E_{\Psi}[\gamma_{l,m}] \left(2E_{\Omega_{l,m}|z}\left[h(\Omega_{l,m})\right] - 1 \right)$$

в силу предполагаемой некоррелированности оценочного функционала $\gamma_{l,m}$ и величины $h(\Omega_{l,m})$, характеризующей данную пару объектов. Так как для каждого m-го множества параметров величины $\Omega_{m,1}, \ldots, \Omega_{m,L_m}$ одинаково распределены с Ω_m, то имеем

$$E_{\Psi|z}[mg(\Psi, Z)] = (2z-1) \sum_{m} \alpha_m \Gamma_m \sum_{l=1}^{L_m} \left(2E_{\Omega_{l,m}|z}\left[h(\Omega_{l,m})\right] - 1 \right) =$$

$$= (2z-1) \sum_{m} \alpha_m \Gamma_m L_m \left(2E_{\Omega_m|z}\left[h(\Omega_m)\right] - 1 \right).$$

При $z = 0$ выполняется

$$2E_{\Omega_m|z}[h(\Omega_m)] - 1 = 2P\left[h_m(\Omega_m) = 1 \mid Z = 0\right] - 1 = 2\left(1 - q_m^{(0)}\right) - 1 = 1 - 2q_m^{(0)},$$

а при $z = 1$ —

$$2E_{\Omega_m|z}[h(\Omega_m)] - 1 = 2P\left[h_m(\Omega_m) = 1 \mid Z = 1\right] - 1 = 2q_m^{(1)} - 1.$$

Значит, для произвольного $z \in \{0, 1\}$ можно записать:

$$2E_{\Omega_m|z}[h(\Omega_m)] - 1 = (2z-1)\left(2q_m^{(z)} - 1\right). \tag{П.2}$$

Отсюда

$$E_{\Psi|z}\left[mg(\Psi,Z)\right]=(2z-1)^2\sum_m\alpha_m\Gamma_m L_m\left(2q_m^{(z)}-1\right)=$$

$$=\sum_m\alpha_m L_m\Gamma_m\left(2q_m^{(z)}-1\right).$$

Рассмотрим теперь условную дисперсию маржинальной функции. Из выражения (П.1), свойств дисперсии, предположения об условной независимости решений и о некоррелируемости решений для пары объектов и оценочным функционалом группировки получим:

$$\mathrm{Var}_{\Psi|z}[mg(\Psi,Z)]=(2z-1)^2\,\mathrm{Var}_{\Psi|z}\left[\sum_{l,m:}\alpha_m\gamma_{l,m}(2h(\Omega_{l,m})-1)\right]=$$

$$=\mathrm{Var}_{\Psi|z}\left[\sum_{l,m:}\alpha_m\gamma_{l,m}(2h(\Omega_{l,m})-1)\right]=$$

(П.3)

$$=\sum_{l,m}\mathrm{Var}_{\Psi|z}\left[\alpha_m\gamma_{l,m}(2h(\Omega_{l,m})-1)\right]+$$

$$+\sum_{\substack{l,m,l',m':\\(l,m)\neq(l',m')}}\mathrm{cov}\left[\alpha_m\gamma_{l,m}(2h(\Omega_{l,m})-1),\alpha_{m'}\gamma_{l',m'}(2h(\Omega_{l',m'})-1)\right].$$

Рассмотрим группы слагаемых в (П.3) по очереди начиная с первой (обозначив ее через A). В силу предположения о некоррелируемости решений для пары объектов и оценочным функционалом группировки, получим:

(П.4)
$$A=\sum_{l,m}\mathrm{Var}_{\Psi|z}\left[\alpha_m\gamma_{l,m}(2h(\Omega_{l,m})-1)\right]=$$

$$=\sum_{l,m}\alpha_m^2\Big\{\mathrm{Var}_\Psi[\gamma_{l,m}]\mathrm{Var}[2h(\Omega_{l,m})-1]+(E_\Psi[\gamma_{l,m}])^2\,\mathrm{Var}[2h(\Omega_{l,m})-1]+$$

$$+(E[2h(\Omega_{l,m})-1])^2\,\mathrm{Var}_\Psi[\gamma_{l,m}]\Big\}=$$

$$=\sum_{l,m}\alpha_m^2\Big\{4V_m\,\mathrm{Var}[h(\Omega_{l,m})]+4(\Gamma_m)^2\,\mathrm{Var}[h(\Omega_{l,m})]+(2E[h(\Omega_{l,m})]-1)^2V_m\Big\}=$$

$$=\sum_m\alpha_m^2L_m\Big\{4V_m\,\mathrm{Var}\left[h(\Omega_m)\right]+4(\Gamma_m)^2\,\mathrm{Var}\left[h(\Omega_m)\right]+$$

$$+(2z-1)^2\left(2q_m^{(z)}-1\right)^2V_m\Big\}$$

(в последнем выражении было использовано свойство (П.2)). Далее,

$$\mathrm{Var}_{\Omega_m|z}\left[h(\Omega_m)\right]=E_{\Omega_m|z}\left[h(\Omega_m)^2\right]-\left(E_{\Omega_m|z}\left[h\left(\Omega_m\right)\right]\right)^2=$$

$$=E_{\Omega_m|z}\left[h\left(\Omega_m\right)\right]\left(1-E_{\Omega_m|z}\left[h\left(\Omega_m\right)\right]\right)$$

(так как $h(\Omega_m)^2 = h(\Omega_m)$). Отсюда для $z = 0$ получим:

$$\mathrm{Var}_{\Omega_m|z=0}\left[h\left(\Omega_m\right)\right] = \left(1 - q_m^{(0)}\right) q_m^{(0)};$$

а для $z = 1$:

$$\mathrm{Var}_{\Omega_m|z=1}[h(\Omega_m)] = q_m^{(1)}\left(1 - q_m^{(1)}\right).$$

Можно записать:

$$\mathrm{Var}_{\Omega_m|z}\left[h(\Omega_m)\right] = q_m^{(z)}\left(1 - q_m^{(z)}\right).$$

Воспользовавшись тождеством

$$\left(2q_m^{(z)} - 1\right)^2 = 1 - 4q_m^{(z)}\left(1 - q_m^{(z)}\right),$$

перепишем (П.4):

$$A = \sum_{l,m} \mathrm{Var}_{\Psi|z}\left[\alpha_m \gamma_{l,m}(2h(\Omega_{l,m}) - 1)\right] =$$

$$= \sum_m \alpha_m^2 L_m \Big\{ 4V_m q_m^{(z)}\left(1 - q_m^{(z)}\right) + 4(\Gamma_m)^2 q_m^{(z)}\left(1 - q_m^{(z)}\right) + \tag{П.5}$$

$$+ \left(1 - 4q_m^{(z)}\left(1 - q_m^{(z)}\right)\right) V_m \Big\} = \sum_m \alpha_m^2 L_m \left\{ V_m + 4(\Gamma_m)^2 q_m^{(z)}\left(1 - q_m^{(z)}\right)\right\}.$$

Преобразуем теперь вторую группу слагаемых в (П.3), обозначив ее через B:

$$B = \sum_{\substack{l,m,l',m':\\ (l,m)\neq(l',m')}} \mathrm{cov}\left[\alpha_m \gamma_{l,m}\left(2h\left(\Omega_{l,m}\right) - 1\right), \alpha_{m'}\gamma_{l',m'}\left(2h\left(\Omega_{l',m'}\right) - 1\right)\right] =$$

$$= \sum_{\substack{l,m,l',m':\\ (l,m)\neq(l',m')}} \alpha_m \alpha_{m'} \Big\{ E\left[\gamma_{l,m}\gamma_{l',m'}\left(2h\left(\Omega_{l,m}\right) - 1\right)\left(2h\left(\Omega_{l',m'}\right) - 1\right)\right] -$$

$$- E\left[\gamma_{l,m}\left(2h\left(\Omega_{l,m}\right) - 1\right)\right] E\left[\gamma_{l',m'}\left(2h\left(\Omega_{l',m'}\right) - 1\right)\right] \Big\}.$$

В силу предположения о некоррелируемости оценочных функций и решений для пары объектов и так как $\Omega_{l,m}$, $\Omega_{l',m'}$ независимы для различных сочетаний $(l,m) \neq (l',m')$, выполняется:

$$B = \sum_{\substack{l,m,l',m':\\ (l,m)\neq(l',m')}} \alpha_m \alpha_{m'} E\left[2h\left(\Omega_{l,m}\right) - 1\right] E\left[2h\left(\Omega_{l',m'}\right) - 1\right] \times$$

$$\times \left\{ E\left[\gamma_{l,m}\gamma_{l',m'}\right] - E\left[\gamma_{l,m}\right] E\left[\gamma_{l',m'}\right]\right\}.$$

Так как $\Omega_{l,m}$, $\Omega_{l',m'}$ одинаково распределены с Ω_m, $\Omega_{m'}$ соответственно, то, воспользовавшись (П.2), получим:

$$B = \sum_{\substack{l,m,l',m':\\(l,m)\neq(l',m')}} \alpha_m\alpha_{m'}(2z-1)^2\left(2q_m^{(z)}-1\right)\left(2q_{m'}^{(z)}-1\right)\operatorname{cov}[\gamma_{l,m},\gamma_{l',m}] =$$

$$= \sum_{\substack{l,m,l',m':\\(l,m)\neq(l',m')}} \alpha_m\alpha_{m'}\left(2q_m^{(z)}-1\right)\left(2q_{m'}^{(z)}-1\right)\operatorname{cov}[\gamma_{l,m},\gamma_{l',m}].$$

Суммирование в последнем выражении ведется как внутри каждого m-го множества параметров, так и по сочетаниям различных множеств с индексами $m,\ m'$, поэтому

$$B = \sum_m \alpha_m\left(2q_m^{(z)}-1\right)^2 \sum_{\substack{l,l'=1\\l\neq l'}}^{L_m} \operatorname{cov}\left[\gamma_{l,m},\gamma_{l',m}\right] +$$

$$+ \sum_{\substack{m,m':\\m\neq m'}} \alpha_m\alpha_{m'}\left(2q_m^{(z)}-1\right)\left(2q_{m'}^{(z)}-1\right)\sum_{l,l'}\operatorname{cov}\left[\gamma_{l,m},\gamma_{l',m'}\right].$$

Так как $\operatorname{cov}[\gamma_{l,m},\gamma_{l',m}] = C_m$, а $\operatorname{cov}[\gamma_{l,m},\gamma_{l',m'}] = C_{m,m'}$, то

$$B = \sum_m \alpha_m^2 L_m\left(L_m-1\right)\left(2q_m^{(z)}-1\right)^2 C_m +$$

$$+ \sum_{\substack{m,m':\\m\neq m'}} \alpha_m\alpha_{m'}L_mL_{m'}\left(2q_m^{(z)}-1\right)\left(2q_{m'}^{(z)}-1\right)C_{m,m'}. \qquad \text{(П.6)}$$

Таким образом, подставив (П.5) и (П.6) в (П.3), получим искомое выражение для дисперсии (7).

Утверждение 1 доказано.

СПИСОК ЛИТЕРАТУРЫ

1. *Миркин Б.Г.* Методы кластер-анализа для поддержки принятия решений: обзор. М.: Изд. дом НИУ ВШЭ, 2011.
2. *Duda R.O., Hart P.E., Stork D.G.* Pattern Classification. Second Edition. Wiley, NY. 2000.
3. *Jain A.K., Dubes R.C.* Algorithms for clustering data. Prentice Hall, NJ, 1988.
4. *Jain A.K.* Data clustering: 50 years beyond k-means. Pattern Recognition Letters. 2010. Vol. 31, No. 8. P. 651–666.
5. *Zhuravlev Yu.I., Nikiforov V.V.* Algorithms for recognition based on calculation of evaluations // Kibernetika. 1971. Vol. 3. P. 1–11.

6. *Ryazanov V.V.* On the synthesis of classifying algorithms in finite sets of classification algorithms (taxonomy). USSR Computational Mathematics and Mathematical Physics. 1982. Vol. 22, Iss. 2. P. 186–198.
7. *Breiman L.* Random Forests // Machine Learning. 2001. Vol. 45(1). P. 5–32.
8. *Kuncheva L.* Combining Pattern Classifiers. Methods and Algorithms. Wiley, NJ. 2004.
9. *Schapire R., Freund Y., Bartlett P., Lee W.* Boosting the margin: a new explanation for the effectiveness of voting methods. Annals of Statistics. 1998. Vol. 26(5). P. 1651–1686.
10. *Ghosh J., Acharya A.* Cluster ensembles. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011. Vol. 1(4). P. 305–315.
11. *Vega-Pons S., Ruiz-Shulcloper J.A.* Survey of Clustering Ensemble Algorithms. 2011. IJPRAI 25(3), 337–372.
12. *Topchy A., Law M., Jain A., Fred A.* Analysis of Consensus Partition in Cluster Ensemble // Fourth IEEE International Conference on Data Mining (ICDM'04). 2004. P. 225–232.
13. *Berikov V., Pestunov I.* Ensemble clustering based on weighted co-association matrices: Error bound and convergence properties // Pattern Recognition. 2017. Vol. 63. P. 427–436.
14. *Wu Y., Liu L., Xie Z., Chow K.H., Wei W.* Boosting Ensemble Accuracy by Revisiting Ensemble Diversity Metrics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. P. 16469–16477.
15. *Rashidi F., Nejatian S., Parvin H., Rezaie V.* Diversity based cluster weighting in cluster ensemble: an information theory approach. Artificial Intelligence Review. 2019. Vol. 52(2). P. 1341–1368.
16. *Wang Z., Parvin H., Qasem S.N., Tuan B.A., Pho K.H.* Cluster ensemble selection using balanced normalized mutual information. Journal of Intelligent & Fuzzy Systems, (Preprint). 2020. P. 1–23.
17. *Liu J., Han J.* Spectral clustering. Data Clustering. Chapman and Hall/CRC. 2018. P. 177–200.
18. http://archive.ics.uci.edu/ml/datasets/image+segmentation.
19. *Rousseeuw P.J.* Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics. 1987. Vol. 20. P. 3–65.
20. *Hubert L., Arabie P.* Comparing partitions. Journal of Classification. 1985. Vol. 2(1). P. 193–218.

Статья представлена к публикации членом редколлегии А.А. Лазаревым.

Поступила в редакцию 01.02.2022

После доработки 25.06.2022

Принята к публикации 29.06.2022